SINOPSIS

Buku ini membahas secara komprehensif tentang konsep data mining dan penerapannya dalam sistem informasi. Disusun sebagai bahan ajar yang sistematis dan aplikatif, buku ini menguraikan mulai dari dasar-dasar data mining, jenis-jenis metode, hingga penerapan dalam berbagai kasus nyata di lingkungan sistem informasi.

Pembahasan dalam buku ini mencakup konsep dasar data mining, proses penambangan data, teknik-teknik analisis seperti klasifikasi, klasterisasi, asosiasi, dan prediksi, serta evaluasi hasil. Selain itu, buku ini juga mengulas bagaimana data mining diintegrasikan dengan sistem informasi untuk mendukung pengambilan keputusan yang lebih efektif dan efisien.

Dengan pendekatan yang jelas, contoh kasus, dan penjelasan yang mudah dipahami, buku ini diharapkan dapat menjadi referensi yang bermanfaat bagi mahasiswa, dosen, praktisi, dan siapa saja yang ingin memperdalam pemahaman tentang data mining dalam konteks sistem informasi.

PT Andalas
Global Research Center

Jl. Mulia 2, No. 52A, Sariharjo, Ngaglik,
Sleman, DI Yogyakarta 55581
+62 811 7488 066
globalpublisherofficial@gmail.com
www.andalaspress.com

KONSEP DATA MINING
SISTEM INFORMASI

PT Andalas
Global Research Center

PT Andalas
Global Research Center

KONSEP DATA MINING
SISTEM INFORMASI

Ir. Andrian Sah, S.Kom., M.Kom
Prof. Dr. Techn. Ahmad Ashari, M.I.Kom.
Siti Nurhayati, S.Kom., M.Kom
Dr. Rasna, ST., M.Cs



KONSEP DATA MINING SISTEM INFORMASI

Ir. Andrian Sah, S.Kom., M.Kom
Prof. Dr. Techn. Ahmad Ashari, M.I.Kom.
Siti Nurhayati, S.Kom., M.Kom
Dr. Rasna, ST., M.Cs

PT Andalas Global Research Center

Penerbit PT Andalas Global Research Center

Judul	: **Konsep Data Mining Sistem Informasi**
Penulis	: Andrian S., Ahmad A., Siti N., Rasna.
Editor	: Fitri Arianti Saputri, S.H.
Ilustator	: Rahmat Purnomo, S.Kom.
Penerbit	: **PT Andalas Global Research Center** Jl. Mulia 2, No. 52A, Sariharjo, Ngaglik, Sleman, Dareah Istimewa Yogyakarta 55581
Telp	: +62 811 7488 066
Email	: globalpublisherofficial@gmail.com
Website	: www.andalaspress.com
ISBN	: xxxx

Anggota Ikatan Penerbit Indonesia (IKAPI)
ix + 450 halaman; 14,8 cm x 21 cm

Hak Cipta Dilindungi Undang-Undang

Dilarang keras menerjemahkan, memfotokopi, atau memperbanyak sebagian atau seluruh isi buku ini dalam bentuk apapun tanpa izin tertulis dari penerbit.

Diterbitkan oleh PT Andalas Global Research Center, Yogyakarta, 2026

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas terbitnya buku ajar Data Mining ini. Perkembangan teknologi digital telah membuat organisasi menghasilkan dan menyimpan data dalam jumlah yang semakin besar, tetapi data baru bernilai ketika dapat diolah menjadi informasi dan pengetahuan yang mendukung keputusan. Karena itu, pemahaman tentang cara menemukan pola, menilai temuan, dan menggunakannya secara bertanggung jawab menjadi bekal akademik yang relevan bagi mahasiswa Sistem Informasi. Buku ini hadir pada saat kebutuhan tersebut semakin nyata, baik dalam kegiatan bisnis maupun pelayanan publik, ketika keputusan yang tepat menuntut pengelolaan data yang cermat.

Kehadiran buku ini memperkaya khazanah literatur Sistem Informasi, khususnya bahan ajar yang menjembatani konsep dan penerapan data mining. Pembahasan disusun secara bertahap, mulai dari landasan dan proses penambangan data hingga metode analisis serta pemanfaatan hasilnya. Susunan demikian membantu pembaca melihat bahwa data mining bukan sekadar kumpulan algoritma, melainkan rangkaian kerja yang

mencakup pemahaman masalah, persiapan data, pemodelan, dan evaluasi. Keunggulan pendekatan ini terletak pada upayanya menempatkan aspek teknis dalam konteks kebutuhan informasi dan pengambilan keputusan, suatu sudut pandang yang dekat dengan karakter keilmuan Sistem Informasi.

Penulis menunjukkan upaya yang terarah untuk menguraikan materi dengan bahasa yang dapat diikuti tanpa mengabaikan ketelitian konsep. Pembahasan yang bergerak dari prinsip menuju penerapan memberi ruang bagi mahasiswa untuk membangun pemahaman secara bertahap, sementara contoh dan konteks penggunaannya membantu menghubungkan materi dengan persoalan nyata. Cara penyajian seperti ini juga dapat mendukung proses belajar yang aktif: mahasiswa tidak hanya menghafal istilah atau prosedur, tetapi didorong untuk mempertanyakan kualitas data, kecocokan metode, serta arti hasil analisis. Meski demikian, pembaca tetap perlu melengkapi pemahaman melalui latihan dan eksplorasi perangkat analisis agar kompetensi praktis berkembang seiring penguasaan teori.

Buku ini layak dimanfaatkan oleh mahasiswa sebagai pegangan dalam mata kuliah Data Mining

dan sebagai pintu masuk untuk mempelajari analisis data lebih lanjut. Bagi dosen, susunan materi dan cakupan bahasannya dapat menjadi landasan dalam merancang perkuliahan, diskusi, maupun tugas berbasis kasus. Praktisi Sistem Informasi juga dapat menggunakannya untuk meninjau proses analisis data dan mempertimbangkan penerapan metode yang sesuai dengan kebutuhan organisasi. Sementara itu, peneliti dapat menjadikannya rujukan awal untuk memahami konsep dan pendekatan yang relevan sebelum menelusuri literatur yang lebih khusus.

Pada akhirnya, mutu buku ajar tidak hanya ditentukan oleh kelengkapan materi, tetapi juga oleh kemampuannya menumbuhkan cara berpikir yang kritis, teliti, dan bertanggung jawab. Buku ini diharapkan membantu pembaca memahami potensi data mining sekaligus mengenali batasan dan konsekuensi penggunaannya. Semoga kehadirannya memberi manfaat berkelanjutan bagi pembelajaran di Program Studi Sistem Informasi, memperkaya praktik akademik, serta mendorong pemanfaatan data yang lebih bermakna bagi masyarakat. Dengan harapan tersebut, buku ini patut disambut sebagai sumbangan yang berguna bagi pengembangan literasi dan keilmuan data.

DAFTAR ISI

KATA PENGANTAR..... iii
DAFTAR ISI..... vi
BAB I KONSEP DAN PERKEMBANGAN DATA MINING
.....1
BAB II DATA DAN EKSPLORASI DATASET..... 29
BAB 3 METODOLOGI DAN SIKLUS DATA MINING..... 54
BAB 4 DATA PREPROCESSING DAN FEATURE
ENGINEERING 85
BAB 5 MACHINE LEARNING UNTUK DATA MINING
.....112
BAB 6 KLASIFIKASI DATA.....141
BAB 7 REGRESI DAN PREDICTIVE ANALYTICS
.....167
BAB 8 CLUSTERING DAN SEGMENTASI DATA
.....194
BAB 9 ASSOCIATION RULE DAN PATTERN MINING
.....219
BAB 10 DATA MINING UNTUK INTERNET OF THINGS
(IOT) 245
BAB 11 DATA MINING UNTUK KEAMANAN SISTEM
INFORMASI 273
BAB 12 ANOMALY DETECTION PADA IOT DAN
KEAMANAN SISTEM.....299
BAB 13 ADVANCED MACHINE LEARNING DAN
INTELLIGENT DATA MINING.....327
BAB 14 IMPLEMENTASI DAN PENGEMBANGAN
PROYEK DATA MINING TERINTEGRASI.....355
DAFTAR PUSTAKA.....384
LAMPIRAN: KUNCI JAWABAN 393

BAB I
KONSEP DAN PERKEMBANGAN DATA MINING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep pengertian, tujuan, dan ruang lingkup data mining untuk mengidentifikasi kebutuhan analisis pada suatu permasalahan Sistem Informasi.
2. Menggunakan kerangka Knowledge Discovery in Databases (KDD) untuk menentukan posisi data mining dalam tahapan penemuan pengetahuan dari data.
3. Memilih pendekatan data mining yang sesuai dengan keterkaitan antara data mining, machine learning, dan artificial intelligence pada suatu kasus.
4. Menentukan penerapan data mining yang relevan untuk menganalisis data Internet of Things (IoT) dan mendukung keamanan Sistem Informasi.
5. Mengimplementasikan pertimbangan tren, peluang, dan tantangan data mining modern dalam

merumuskan kebutuhan analisis data secara bertanggung jawab.

B. Pendahuluan

Organisasi menghasilkan data dari beragam aktivitas, mulai dari transaksi penjualan dan interaksi pelanggan hingga pencatatan proses operasional serta komunikasi antarmesin. Data yang terus bertambah tidak serta-merta menghasilkan pengetahuan yang berguna: data dapat tersebar di berbagai sistem, memiliki format yang beragam, mengandung ketidaklengkapan, atau menyimpan pola yang sulit dikenali melalui pemeriksaan manual. Data mining menawarkan pendekatan untuk mengekstraksi pola dan hubungan yang relevan sehingga data dapat mendukung pengambilan keputusan. Dalam kerangka Knowledge Discovery in Databases (KDD), data mining merupakan salah satu tahap inti yang menerapkan metode komputasional untuk menemukan pola, sedangkan keseluruhan proses juga mencakup pemilihan, pembersihan, transformasi, dan interpretasi data (Fayyad, Piatetsky-Shapiro, dan Smyth, 1996). Bagi mahasiswa Sistem Informasi, pemahaman ini penting karena pengelolaan data tidak berhenti pada penyimpanan dan pelaporan; hasil analisis perlu dikaitkan dengan proses bisnis, kebutuhan pemangku kepentingan, serta keputusan yang hendak diperbaiki.

Penerapan data mining memerlukan pemahaman tentang hubungan sekaligus perbedaan antara data

mining, machine learning, dan artificial intelligence. Machine learning menyediakan metode yang memungkinkan sistem membangun model berdasarkan data, sedangkan artificial intelligence mencakup bidang yang lebih luas dalam perancangan sistem yang menunjukkan perilaku cerdas (Mitchell, 1997; Russell dan Norvig, 2021). Data mining memanfaatkan metode tersebut untuk menemukan pola yang dapat ditafsirkan dan digunakan dalam konteks tertentu, misalnya untuk mengelompokkan pelanggan, memperkirakan permintaan, atau mengidentifikasi transaksi yang tidak lazim. Pemilihan metode tidak dapat dilepaskan dari tujuan analisis, karakteristik data, dan konsekuensi penggunaan hasilnya. Dalam lingkungan Sistem Informasi, keputusan untuk menerapkan klasifikasi, pengelompokan, atau deteksi anomali perlu didasarkan pada kesesuaian metode dengan persoalan, bukan semata-mata pada ketersediaan perangkat lunak. Dengan demikian, kemampuan menerapkan konsep data mining juga mencakup penilaian atas konteks organisasi dan keterbatasan analisis.

Perkembangan Internet of Things memperluas sumber data melalui perangkat dan sensor yang menghasilkan aliran data secara berkelanjutan. Data tersebut dapat dimanfaatkan untuk memantau kondisi aset, memperkirakan kebutuhan pemeliharaan, atau mengenali perilaku jaringan yang menyimpang. Namun,

skala dan kecepatan data IoT sekaligus menambah tantangan dalam pengelolaan kualitas, privasi, dan keamanan; NIST menekankan bahwa karakteristik perangkat IoT perlu dipertimbangkan dalam pengelolaan risiko keamanan siber (NIST, 2019). Data mining modern juga menghadapi persoalan seperti bias data, keterjelasan model, perubahan pola dari waktu ke waktu, dan kebutuhan komputasi yang meningkat. Oleh sebab itu, penerapan analisis perlu disertai pertimbangan terhadap dampak keputusan dan tata kelola data. Bab ini membahas pengertian, tujuan, dan ruang lingkup data mining; hubungan KDD dengan data mining; keterkaitan data mining, machine learning, dan artificial intelligence; peran data mining dalam IoT dan keamanan Sistem Informasi; serta tren, peluang, dan tantangan data mining modern.

C. Pengertian, Tujuan, dan Ruang Lingkup Data Mining

Penambangan data adalah proses menggunakan metode komputasional untuk menemukan pola, keteraturan, atau hubungan yang berguna dalam kumpulan data. Proses ini tidak sekadar menjalankan algoritma: hasil analisis perlu dikaitkan dengan persoalan yang hendak diselesaikan dan ditafsirkan dalam konteks penggunaannya. Karena itu, penambangan data berada di persimpangan pengelolaan data, analisis, dan pengambilan keputusan organisasi (Unknown, 2011; Kononenko, 2007).

Tujuan analisis menentukan bentuk pengetahuan yang dicari. Organisasi dapat menggunakan penambangan data untuk mendeskripsikan kelompok yang terbentuk, mengklasifikasikan objek ke kategori tertentu, memperkirakan nilai, menemukan pola keterkaitan, atau menandai perilaku yang menyimpang. Setiap tujuan memerlukan keluaran dan cara pemanfaatan yang berbeda. Misalnya, prediksi permintaan membantu menyusun persediaan, sedangkan segmentasi pelanggan dapat mendukung penyesuaian layanan. Pemilihan metode harus mengikuti pertanyaan analisis, bukan sekadar popularitas algoritma yang tersedia (Unknown, 2019; Kononenko, 2007).

Ruang lingkup penambangan data mencakup penentuan kebutuhan analisis, penyiapan data, pemilihan pendekatan, pembentukan model, penilaian hasil, dan penafsiran pola. Tahap teknis tersebut perlu memperhatikan sumber dan kualitas data, tujuan organisasi, serta batas pemanfaatan hasil. Pembahasan terperinci mengenai karakteristik kumpulan data dan eksplorasinya terdapat pada Bab 2, sementara tahapan proses dan prapemrosesan data diuraikan pada Bab 3 dan Bab 4. Dalam bab ini, perhatian diarahkan pada posisi penambangan data sebagai kegiatan analitis yang harus menghasilkan informasi yang dapat ditindaklanjuti.

Contoh penerapan di Indonesia dapat ditemukan pada analisis penjualan kerupuk rambak. Data transaksi dapat digunakan untuk memperkirakan kategori sasaran pemasaran dengan klasifikasi naïve Bayes; nilai analisisnya terletak pada dukungannya terhadap keputusan pemasaran, bukan pada hasil algoritma semata (Wulandari, 2021). Jika tujuan pelaku usaha adalah memahami kelompok pembeli, pendekatan klasterisasi mungkin lebih sesuai. Dengan demikian, persoalan yang sama dapat menuntut pendekatan berbeda ketika kebutuhan keputusan berubah.

Tabel berikut memperlihatkan bagaimana tujuan organisasi dapat diterjemahkan menjadi bentuk analisis dan keluaran yang dapat digunakan. Contoh-contoh bersifat ilustratif dan menggambarkan pilihan awal, bukan resep algoritma yang berlaku untuk semua organisasi.

Tabel 1.1 Tujuan Analisis dan Keluaran Penambangan Data

Kebutuhan	Pertanyaan analisis	Pendekatan	Keluaran	Pemanfaatan
Segmentasi pelanggan	Kelompok apa yang	Klasterisasi	Label kelompok	Penyesuaian layanan

	terbentuk?			
Perkiraan permintaan	Berapa kebutuhan periode depan?	Regresi	Nilai prediksi	Perencanaan persediaan
Penyaringan transaksi	Transaksi mana berisiko?	Klasifikasi	Kategori risiko	Pemeriksaan lanjutan
Pola pembelian	Produk apa kerap dibeli bersama?	Pola asosiasi	Aturan keterkaitan	Penataan promosi
Pemantauan sensor	Pembacaan mana tidak lazim?	Deteksi anomali	Penanda penyimpangan	Pemeriksaan perangkat

Tabel tersebut menegaskan bahwa keluaran teknis baru bernilai ketika menjawab kebutuhan operasional tertentu. Rumusan masalah yang jelas membantu organisasi menghubungkan pola yang ditemukan dengan tindakan yang layak dilakukan.

D. Knowledge Discovery in Databases dan Data Mining

Knowledge Discovery in Databases (KDD) merujuk pada keseluruhan proses untuk memperoleh pengetahuan yang berguna dari data. Dalam kerangka ini, penambangan data merupakan tahap analitis yang menerapkan algoritma untuk menemukan pola, sedangkan KDD mencakup kegiatan yang lebih luas, mulai dari penentuan data yang relevan hingga penafsiran hasil. Pembedaan ini mencegah anggapan bahwa menjalankan algoritma saja sudah cukup untuk menghasilkan pengetahuan (Unknown, 2011; Unknown, 2019).

Secara konseptual, KDD dapat dipahami sebagai rangkaian kegiatan yang saling berhubungan: memilih data sesuai masalah, menyiapkan data agar dapat dianalisis, menerapkan metode penambangan data, lalu menilai dan menafsirkan pola yang ditemukan. Hasil penafsiran dapat menunjukkan bahwa data atau pertanyaan awal perlu ditinjau kembali. Dengan demikian, proses ini tidak selalu berjalan lurus satu arah; keputusan pada satu tahap dapat mendorong pengulangan tahap sebelumnya. Rincian tahapan dan penerapan metodologi proses dibahas lebih lanjut pada Bab 3.

Penambangan data di dalam KDD menjalankan peran inti, tetapi keberhasilannya bergantung pada mutu

keputusan di luar pemodelan. Data yang dipilih secara keliru dapat menghasilkan pola yang tidak menjawab kebutuhan. Transformasi yang tidak sesuai dapat mengaburkan informasi, sedangkan pola yang benar secara komputasional tetap dapat disalahartikan jika konteksnya tidak dipahami. Prinsip ini selaras dengan pembahasan metode penambangan data yang menempatkan persiapan, penerapan algoritma, dan interpretasi sebagai bagian yang terkait (Kononenko, 2007; Unknown, 2011).

Penerapan di Indonesia dapat dibayangkan pada puskesmas yang ingin menggunakan catatan pasien untuk merencanakan layanan. Tim analisis terlebih dahulu merumuskan kebutuhan, misalnya mengenali kelompok risiko untuk mendukung perencanaan program. Data yang sesuai kemudian dipilih dan disiapkan; tahap penambangan data dapat mengelompokkan catatan menurut kemiripan karakteristik. Hasilnya perlu diperiksa bersama pemahaman tenaga kesehatan sebelum dipakai dalam perencanaan. Penerapan klasterisasi pada data kesehatan masyarakat desa menggambarkan bagaimana hasil pengelompokan dapat diarahkan untuk mendukung keputusan program, bukan diperlakukan sebagai keputusan klinis otomatis (Ayang Prabu, 2026).

Dalam praktik, KDD juga mengharuskan tim membedakan pola yang menarik secara statistik dari

pola yang bermanfaat bagi pemangku kepentingan. Sebuah kelompok yang terbentuk belum tentu dapat ditindaklanjuti apabila tidak berkaitan dengan pilihan layanan atau alokasi sumber daya. Pengetahuan yang dihasilkan perlu disajikan secara dapat dipahami, disertai batas interpretasi, dan dipertimbangkan dampaknya. Penekanan pada tujuan penggunaan ini menjadikan KDD kerangka kerja yang menghubungkan pemrosesan data dengan kebutuhan sistem informasi dan keputusan organisasi.

Memahami hubungan KDD dan penambangan data membantu mahasiswa menempatkan algoritma secara tepat dalam suatu proyek. Ketika hasil analisis tidak dapat digunakan, penyebabnya mungkin bukan kelemahan algoritma, melainkan pemilihan data, persiapan, atau interpretasi yang kurang sesuai. Karena itu, kerangka KDD menjadi penghubung antara masalah organisasi dan proses teknis yang akan dipelajari pada bab-bab berikutnya.

E. Hubungan Data Mining, Machine Learning, dan Artificial Intelligence

Penambangan data, pembelajaran mesin, dan kecerdasan buatan memiliki wilayah yang beririsan, tetapi tidak menunjuk hal yang sama. Penambangan data berorientasi pada pencarian pola atau pengetahuan dari data untuk menjawab kebutuhan tertentu. Pembelajaran mesin menyediakan pendekatan yang memungkinkan

sistem membentuk model dari data. Kecerdasan buatan merupakan bidang yang lebih luas, yang mencakup perancangan sistem untuk melaksanakan tugas yang memerlukan perilaku cerdas. Hubungan tersebut membantu menjelaskan mengapa satu proyek dapat menggunakan metode pembelajaran mesin sebagai bagian dari kegiatan penambangan data (Kononenko, 2007; Unknown, 2011).

Dalam penerapan, pemilihan pendekatan dimulai dari keluaran yang dibutuhkan. Jika organisasi memiliki catatan transaksi berlabel dan ingin memperkirakan transaksi baru termasuk kategori apa, klasifikasi dapat menjadi pilihan. Jika label tidak tersedia dan tujuan analisisnya mencari kelompok alami, klasterisasi lebih sesuai. Kedua pendekatan tersebut termasuk tugas yang dapat dilakukan dengan metode pembelajaran mesin, tetapi pertanyaan dan informasi masukannya berbeda. Pembahasan paradigma pembelajaran serta algoritma secara lebih mendalam terdapat pada Bab 5 sampai Bab 9 (Unknown, 2019).

Kecerdasan buatan dapat mencakup sistem yang menggabungkan pembelajaran mesin dengan komponen lain, seperti aturan dan mekanisme pengambilan tindakan. Model yang belajar dari data tidak otomatis menjadi keseluruhan sistem cerdas. Model prediksi, misalnya, dapat memberi skor risiko, tetapi keputusan untuk menindaklanjuti skor tersebut memerlukan

aturan organisasi dan pertimbangan manusia. Pembedaan ini mencegah penyamaan antara model statistik, aplikasi penambangan data, dan sistem kecerdasan buatan yang lebih lengkap (Kononenko, 2007).

Di Indonesia, sebuah sistem informasi akademik dapat memanfaatkan data nilai dan kehadiran untuk memperkirakan mahasiswa yang memerlukan dukungan belajar. Model pembelajaran mesin menghasilkan prediksi berdasarkan pola data, sedangkan kegiatan penambangan data mencakup penentuan tujuan analisis serta penafsiran pola bagi pengelola program. Jika hasilnya diintegrasikan dengan prosedur layanan akademik, aplikasi tersebut menjadi bagian dari sistem pendukung keputusan yang lebih luas. Penelitian tentang penggunaan jaringan saraf untuk prediksi kasus gizi buruk di Nusa Tenggara Barat menunjukkan penerapan pendekatan pembelajaran mesin pada kebutuhan informasi kesehatan, meskipun penggunaan hasilnya tetap perlu mempertimbangkan konteks layanan dan kualitas data (Kurniati, 2021).

Penerapan deep learning memperlihatkan irisan tersebut dengan jelas. Jaringan saraf berlapis dapat digunakan untuk mempelajari representasi data, tetapi manfaatnya bergantung pada kesesuaian data, tujuan prediksi atau klasifikasi, dan kemampuan menafsirkan hasil. Pengembangan kerangka berbasis deep neural

network dan transfer learning dalam konteks prediksi serta klasifikasi menggambarkan pemanfaatan metode pembelajaran mesin sebagai sarana analisis data, bukan tujuan yang berdiri sendiri (Nurlaela, 2025). Karena itu, keputusan teknis perlu mempertimbangkan kebutuhan komputasi dan kemampuan organisasi mengelola model.

Untuk menentukan pendekatan pada suatu kasus, mahasiswa dapat menghubungkan tiga pertanyaan: pengetahuan apa yang hendak ditemukan, apakah model perlu belajar dari contoh, dan apakah kebutuhan organisasi memerlukan sistem yang lebih luas daripada model analitis. Pertanyaan pertama menempatkan penambangan data pada tujuan penemuan pola; pertanyaan kedua mengarahkan pilihan metode pembelajaran mesin; pertanyaan ketiga membuka pertimbangan integrasi dengan sistem kecerdasan buatan. Kerangka ini membantu mencegah penggunaan istilah secara saling menggantikan.

F. Peran Data Mining dalam IoT dan Keamanan Sistem Informasi

Internet untuk Segala (IoT) menghubungkan perangkat dan sensor yang mengumpulkan atau mengirimkan data tentang kondisi lingkungan maupun operasi suatu aset. Penambangan data dapat membantu mengubah catatan tersebut menjadi informasi untuk pemantauan, perkiraan, dan pengenalan pola yang tidak lazim. Namun, data sensor yang terkumpul belum tentu

langsung dapat dianalisis: konteks pengukuran, kualitas pembacaan, dan kebutuhan waktu respons perlu dipertimbangkan sebelum memilih pendekatan. Pengelolaan karakteristik data IoT akan dibahas lebih lanjut pada Bab 10.

Di Indonesia, sensor kualitas udara dapat merekam suhu, kelembapan, dan konsentrasi gas untuk mendukung pemantauan lingkungan secara berkala. Sistem IoT yang mengukur parameter tersebut menggambarkan bagaimana perangkat sensor memasok data bagi analisis kondisi udara (Alfirdaus, 2024). Penambangan data dapat digunakan untuk merangkum pola pengukuran atau membantu mengenali perubahan yang memerlukan pemeriksaan. Pada perangkat pemantau kadar CO_2 berbasis ESP32, misalnya, data sensor dan variabel lingkungan menjadi masukan bagi model prediksi; penerapannya menunjukkan keterkaitan antara pengukuran IoT dan pembelajaran mesin (Agung Suburdjati, 2026).

Gambaran perangkat sensor di lapangan membantu menghubungkan sumber data dengan kebutuhan analisis. Visual berikut menampilkan konteks pemantauan kualitas udara yang dapat menghasilkan catatan sensor untuk digunakan dalam pengambilan keputusan.



Gambar 1.1 Perangkat Pemantau Kualitas Udara Berbasis IoT. Sumber: Ilustrasi, diolah penulis

Perangkat tersebut memperlihatkan titik awal aliran data: pengukuran lingkungan yang dikumpulkan berulang oleh sensor. Agar informasi itu berguna, hasil analisis perlu ditafsirkan bersama konteks lokasi, waktu, dan kondisi operasional alat.

Pada keamanan sistem informasi, penambangan data dapat mendukung analisis log dan lalu lintas jaringan untuk mengenali aktivitas yang berbeda dari pola lazim. Sebuah model dapat membantu menyaring kejadian yang perlu diperiksa, tetapi hasilnya tidak dengan sendirinya memastikan adanya serangan. Salah klasifikasi dapat membebani tim keamanan dengan peringatan keliru atau, sebaliknya, membuat aktivitas berbahaya luput dari perhatian. Buku tentang

penambangan data dan pembelajaran mesin dalam keamanan siber membahas relevansi metode analitis dalam konteks perlindungan sistem (Dua, 2016).

Ketika volume data keamanan meningkat, penggunaan sistem deteksi intrusi berbantuan kecerdasan buatan dapat membantu menangani analisis ancaman yang lebih kompleks. Namun, keluaran sistem tetap perlu dinilai berdasarkan tujuan keamanan dan konteks operasionalnya; pemeriksaan lanjutan oleh analis tetap menjadi bagian penting dari respons insiden (Amanda, 2025). Pada jaringan perangkat IoT, pengamatan atas pola komunikasi dan kondisi perangkat dapat membantu menentukan kejadian yang perlu ditinjau, sementara pengelolaan akses dan perlindungan data tetap diperlukan untuk mengurangi risiko pada keseluruhan sistem.

G. Tren, Peluang, dan Tantangan Data Mining Modern

Penambangan data modern berkembang seiring meningkatnya skala, kecepatan, dan keragaman data. Organisasi mengolah data transaksi, sensor, media sosial, serta catatan operasional untuk mendukung kebutuhan yang beragam. Sumber yang bertambah membuka peluang menemukan pola yang sebelumnya sulit dikenali, tetapi juga menambah beban pengelolaan dan kebutuhan komputasi. Karena itu, penerapan metode perlu disesuaikan dengan karakter data serta tujuan

analisis, bukan diasumsikan cocok hanya karena mampu memproses volume besar (Unknown, 2019; KRYVENCHUK, 2022).

Salah satu arah perkembangan adalah analisis yang dapat memperbarui pola secara bertahap ketika data baru masuk. Dalam analisis media sosial, pendekatan penambangan pola inkremental ditujukan untuk mengurangi kebutuhan mengolah ulang seluruh kumpulan data setiap kali terdapat kiriman baru (Al Riyami, 2026). Peluang serupa muncul pada pemantauan IoT, ketika organisasi perlu mengenali perubahan kondisi tanpa menunggu pengumpulan data dalam jangka panjang. Penerapannya tetap perlu mempertimbangkan apakah pola yang diperbarui cukup stabil untuk mendukung keputusan.

Pembelajaran mendalam dan transfer learning juga memperluas pilihan analisis, khususnya saat data memiliki struktur kompleks. Pengembangan kerangka jaringan saraf mendalam untuk prediksi dan klasifikasi menunjukkan penggunaan pendekatan tersebut pada berbagai konteks analisis (Nurlaela, 2025). Namun, model yang lebih kompleks tidak otomatis lebih tepat. Kebutuhan akan data yang sesuai, sumber daya komputasi, dan kemampuan menjelaskan hasil perlu ditimbang terhadap manfaat yang diharapkan. Pemilihan teknik lanjutan dan evaluasi model dibahas lebih jauh pada Bab 13.

Peluang tersebut disertai persoalan privasi dan tata kelola data. Jika data sensitif dipusatkan untuk analisis, organisasi perlu mengelola risiko akses dan pengungkapan informasi. Kerangka analitik federasi yang menjaga privasi mengangkat kemungkinan menganalisis data lintas lingkungan tanpa bergantung sepenuhnya pada pemusatan data (B. & C., 2026). Pendekatan semacam ini tidak menghapus kebutuhan pengamanan: desain sistem, akses pengguna, dan risiko inferensi tetap harus diperhatikan. Pada konteks Indonesia, lembaga layanan kesehatan yang menganalisis data pasien perlu mengaitkan manfaat analitik dengan pembatasan akses dan perlindungan kerahasiaan.

Tantangan lain ialah bias pada data, perubahan pola dari waktu ke waktu, dan ketergantungan berlebihan pada hasil model. Jika data pelatihan tidak mencerminkan kelompok yang akan terdampak, prediksi dapat menghasilkan perlakuan yang tidak seimbang. Jika hubungan dalam data berubah, kinerja model yang dahulu memadai dapat menurun. Pemantauan dan peninjauan manusia membantu organisasi mengenali keterbatasan tersebut sebelum keluaran model digunakan dalam keputusan penting. Kajian tentang keamanan data sensitif pada lingkungan komputasi juga memperlihatkan bahwa perlindungan perlu mencakup

data yang sedang diproses, bukan hanya data tersimpan (Arman, 2026).

Untuk menerapkan perkembangan tersebut secara bertanggung jawab, organisasi dapat memulai dari kebutuhan keputusan, menilai kesiapan data dan infrastruktur, lalu memilih metode yang sepadan dengan risiko penggunaan. Misalnya, perusahaan transportasi di Indonesia dapat menggunakan segmentasi pelanggan untuk merancang layanan, tetapi hasil pengelompokan perlu ditafsirkan bersama tujuan bisnis dan batas penggunaan data. Penerapan klasterisasi berbasis RFM pada pengelolaan pelanggan menunjukkan bagaimana analisis dapat dihubungkan dengan strategi layanan (Priandini, 2025). Pertimbangan manfaat, keterbatasan, privasi, dan dampak keputusan menjadi bagian dari kebutuhan analisis, bukan tambahan setelah model selesai.

H. Rangkuman

Penambangan data menggunakan metode komputasional untuk menemukan pola, keteraturan, atau hubungan yang berguna dalam kumpulan data. Nilai kegiatan ini tidak terletak pada algoritma semata, tetapi pada kemampuannya menjawab persoalan dan mendukung keputusan organisasi. Karena kebutuhan keputusan berbeda, tujuan analisis perlu dirumuskan sebelum metode dipilih. Segmentasi pelanggan dapat menghasilkan label kelompok untuk penyesuaian

layanan, sedangkan perkiraan permintaan menghasilkan nilai prediksi untuk perencanaan persediaan. Klasifikasi, regresi, pola asosiasi, dan deteksi anomali pun menyediakan keluaran yang berbeda sesuai pertanyaan analisis. Ruang lingkup penambangan data meliputi penentuan kebutuhan, penyiapan data, pemilihan pendekatan, pembentukan model, penilaian hasil, serta penafsiran pola. Seluruh tahap perlu mempertimbangkan sumber dan mutu data, konteks organisasi, serta batas pemanfaatan hasil. Dalam kerangka Knowledge Discovery in Databases (KDD), penambangan data merupakan tahap analitis di dalam proses yang lebih luas: memilih data, menyiapkannya, menerapkan metode, kemudian menilai dan menafsirkan hasil. Proses ini dapat berulang; temuan dapat mendorong tim meninjau kembali data atau pertanyaan awal. Oleh sebab itu, pola yang benar secara komputasional belum tentu relevan atau bermanfaat bagi pemangku kepentingan. Penambangan data, pembelajaran mesin, dan kecerdasan buatan saling beririsan, tetapi memiliki cakupan dan penekanan berbeda. Penambangan data berfokus pada pencarian pengetahuan untuk kebutuhan tertentu, pembelajaran mesin menyediakan pendekatan untuk membentuk model dari data, sedangkan kecerdasan buatan mencakup bidang yang lebih luas dalam perancangan sistem cerdas. Dalam praktik Sistem Informasi, pemilihan metode harus mengikuti keluaran yang

dibutuhkan: klasifikasi sesuai untuk memperkirakan kategori ketika label tersedia, sementara klasterisasi dapat digunakan untuk menemukan kelompok ketika label tidak tersedia. Hasil analisis perlu ditafsirkan bersama pengetahuan konteks, disajikan agar dapat dipahami, dan dipertimbangkan dampaknya. Dengan demikian, keberhasilan proyek tidak hanya ditentukan ketepatan teknis, tetapi juga kesesuaian antara masalah, data, metode, interpretasi, dan tindakan organisasi.

I. Latihan Mahasiswa

a. Essay

1. Analisis hubungan antara rumusan kebutuhan organisasi dan pemilihan metode penambangan data. Gunakan contoh perkiraan permintaan dan segmentasi pelanggan untuk menjelaskan mengapa satu jenis algoritma tidak selalu tepat bagi semua persoalan, serta bagaimana keluaran analisis seharusnya mendukung tindakan organisasi.
2. Bandingkan peran penambangan data dengan keseluruhan proses KDD. Telaah bagaimana kesalahan dalam pemilihan atau penyiapan data dapat memengaruhi hasil algoritma, lalu rumuskan langkah yang perlu dilakukan agar pola yang ditemukan dapat ditafsirkan dan digunakan secara bertanggung jawab.
3. Evaluasi pernyataan bahwa pola yang benar secara komputasional otomatis menjadi pengetahuan yang bermanfaat. Gunakan contoh pengelompokan catatan pasien untuk membahas peran konteks,

keterlibatan pemangku kepentingan, batas pemanfaatan hasil, serta risiko jika hasil analisis diperlakukan sebagai keputusan otomatis.

4. Bandingkan penambangan data, pembelajaran mesin, dan kecerdasan buatan berdasarkan fokus dan cakupannya. Selanjutnya, analisis bagaimana ketiganya dapat berhubungan dalam proyek klasifikasi transaksi, dengan menjelaskan kontribusi masing-masing tanpa menyamakan istilah-istilah tersebut.
5. Rumuskan pendekatan analisis untuk pelaku usaha yang ingin memahami pembelian kerupuk rambak. Bandingkan pilihan klasifikasi dan klasterisasi berdasarkan tujuan keputusan, ketersediaan label, bentuk keluaran, serta tindakan pemasaran yang mungkin dilakukan, kemudian evaluasi informasi tambahan yang diperlukan sebelum metode ditetapkan.

b. Pilihan Berganda

1. Sebuah organisasi ingin mengetahui kelompok pelanggan yang terbentuk dari data tanpa label untuk menyesuaikan layanan. Pendekatan awal yang paling sesuai adalah...
    a. Regresi untuk menghasilkan nilai permintaan
    b. Klasterisasi untuk membentuk kelompok berdasarkan kemiripan
    c. Klasifikasi untuk memperkirakan kategori yang sudah ditentukan
    d. Pola asosiasi untuk menandai transaksi menyimpang

e. Deteksi anomali untuk memperkirakan nilai kontinu

2. Pernyataan yang paling tepat mengenai hubungan KDD dan penambangan data adalah...
   a. KDD hanya mencakup penerapan algoritma, sedangkan penambangan data mencakup interpretasi
   b. KDD dan penambangan data adalah istilah yang selalu memiliki cakupan sama
   c. Penambangan data dilakukan sebelum data dipilih dalam proses KDD
   d. Penambangan data merupakan tahap analitis di dalam proses KDD yang lebih luas
   e. KDD dimulai setelah seluruh pola selesai ditafsirkan

3. Sebuah tim menemukan kelompok pelanggan yang terbentuk dengan baik secara teknis, tetapi tidak mengetahui tindakan layanan yang dapat dilakukan berdasarkan kelompok tersebut. Kesimpulan paling tepat adalah...
   a. Pola tersebut belum tentu bernilai operasional karena belum dikaitkan dengan kebutuhan keputusan
   b. Model harus dianggap berhasil selama metrik teknisnya tinggi
   c. Kelompok tersebut otomatis dapat digunakan untuk menentukan kebijakan layanan
   d. Tim sebaiknya mengganti semua data dengan data yang lebih besar

e. Interpretasi tidak diperlukan jika algoritma sudah menghasilkan kelompok

4. Organisasi memiliki transaksi yang sudah diberi label “berisiko” dan “tidak berisiko”, lalu ingin memperkirakan label transaksi baru. Pendekatan yang sesuai adalah...
   a. Klasterisasi
   b. Regresi
   c. Pola asosiasi
   d. Deteksi anomali
   e. Klasifikasi

5. Dalam proses KDD, hasil penafsiran menunjukkan bahwa data yang dipilih tidak mewakili kebutuhan analisis. Tindakan yang paling sesuai adalah...
   a. Mempertahankan model dan mengabaikan ketidaksesuaian data
   b. Mengulang seluruh proses tanpa meninjau pertanyaan analisis
   c. Meninjau kembali pemilihan data atau rumusan pertanyaan dan melanjutkan proses secara iteratif
   d. Menganggap pola sebagai pengetahuan karena sudah terbentuk
   e. Mengganti interpretasi dengan penambahan jumlah algoritma

6. Pernyataan yang membedakan penambangan data dari pembelajaran mesin secara tepat adalah...

a. Penambangan data hanya mengelola basis data, sedangkan pembelajaran mesin hanya menyajikan laporan
b. Penambangan data berorientasi pada pencarian pengetahuan untuk kebutuhan tertentu, sedangkan pembelajaran mesin menyediakan pendekatan pembentukan model dari data
c. Pembelajaran mesin merupakan keseluruhan proses KDD, sedangkan penambangan data adalah salah satu jenis basis data
d. Penambangan data dan pembelajaran mesin tidak memiliki wilayah yang beririsan
e. Kecerdasan buatan hanya mencakup klasterisasi, sedangkan pembelajaran mesin hanya mencakup klasifikasi

7. Data transaksi akan dianalisis untuk menemukan produk yang kerap dibeli bersama dan digunakan dalam penataan promosi. Keluaran yang paling relevan adalah...
   a. Nilai prediksi permintaan setiap periode
   b. Label kelompok pelanggan
   c. Kategori risiko transaksi
   d. Aturan keterkaitan antarproduk
   e. Penanda pembacaan sensor yang menyimpang
8. Mengapa persiapan dan pemilihan data menjadi bagian penting dalam KDD?
   a. Data yang dipilih atau diubah secara tidak sesuai dapat menghasilkan pola yang tidak

menjawab kebutuhan atau mengaburkan informasi

b. Tahap tersebut memastikan semua algoritma menghasilkan pola yang sama
c. Persiapan data menggantikan kebutuhan untuk menilai hasil
d. Pemilihan data hanya diperlukan ketika tidak ada algoritma yang tersedia
e. Tahap tersebut menjamin hasil analisis dapat digunakan tanpa interpretasi

9. Tim puskesmas menggunakan klasterisasi catatan pasien untuk mendukung perencanaan program. Sikap yang paling tepat terhadap hasil pengelompokan adalah...

a. Menjadikannya keputusan klinis otomatis karena kelompok terbentuk dari data
b. Mengabaikan masukan tenaga kesehatan setelah model selesai
c. Menggunakan hasil untuk menentukan diagnosis individual tanpa pemeriksaan lain
d. Menganggap semua kelompok pasti relevan bagi alokasi layanan
e. Memeriksa hasil bersama pemahaman tenaga kesehatan dan menggunakannya sebagai dukungan perencanaan

10. Pelaku usaha memiliki data transaksi kerupuk rambak tanpa label sasaran pemasaran dan ingin memahami kelompok pembelinya. Pilihan awal yang paling sesuai adalah...

a. Klasifikasi, karena label sasaran selalu dapat diabaikan
b. Regresi, karena kelompok pelanggan harus berupa nilai kontinu
c. Klasterisasi, karena tujuan awalnya menemukan kelompok tanpa label
d. Deteksi anomali, karena seluruh pembeli dianggap menyimpang
e. Pola asosiasi, karena metode ini selalu membentuk segmen pelanggan

c. Soal Project / Studi Kasus

1. Sebuah jaringan toko kebutuhan harian ingin memperkirakan permintaan beberapa kategori produk untuk bulan berikutnya. Data transaksi tersedia selama dua tahun, tetapi sebagian cabang baru menggunakan pencatatan digital dalam enam bulan terakhir. Manajer operasional mengusulkan klasterisasi karena metode tersebut pernah digunakan tim lain, sedangkan tim pemasaran meminta segmentasi pelanggan untuk merancang promosi. Susun analisis kebutuhan yang membedakan kedua tujuan tersebut. Rekomendasikan pendekatan dan keluaran untuk setiap tujuan, jelaskan pemeriksaan data yang diperlukan, serta tentukan cara menilai apakah hasilnya dapat digunakan dalam keputusan operasional.
2. Puskesmas kabupaten mengusulkan pengelompokan catatan pasien untuk membantu merencanakan program layanan. Data berasal dari

beberapa fasilitas dengan kelengkapan pencatatan berbeda. Seorang pengelola ingin langsung menggunakan label kelompok untuk menentukan prioritas pasien, tetapi tenaga kesehatan mengingatkan bahwa kondisi pasien dipengaruhi konteks yang tidak seluruhnya tercatat. Rumuskan rancangan analisis yang menghubungkan kebutuhan program, pemilihan dan penyiapan data, klasterisasi, penilaian hasil, serta interpretasi bersama pemangku kepentingan. Evaluasi batas penggunaan hasil dan tindakan untuk mengurangi risiko keputusan yang tidak tepat.

BAB II
DATA DAN EKSPLORASI DATASET

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep dataset, objek, atribut, dan label untuk memetakan komponen data pada kasus Sistem Informasi.
2. Menentukan tipe, skala pengukuran, dan karakteristik data yang relevan sebagai dasar pemilihan teknik pengolahan dan analisis.
3. Memilih pendekatan pengelolaan data terstruktur, semi-terstruktur, dan tidak terstruktur sesuai bentuk serta kebutuhan analisis.
4. Menggunakan teknik Exploratory Data Analysis (EDA) dan visualisasi untuk memeriksa distribusi, nilai hilang, pencilan, serta hubungan antaratribut.
5. Mengimplementasikan langkah eksplorasi awal yang sesuai untuk data IoT, sensor, log, dan keamanan dengan mempertimbangkan karakteristik serta konteks **penggunaannya.**

B. Pendahuluan

Keputusan yang dihasilkan melalui data mining dapat keliru apabila data yang dianalisis tidak dipahami secara memadai sejak tahap awal. Kesalahan dalam mengenali satuan analisis, menafsirkan atribut, atau menganggap

suatu variabel sebagai label dapat mengubah makna pola yang ditemukan. Dalam Sistem Informasi, persoalan tersebut berpengaruh langsung terhadap analisis proses bisnis, evaluasi layanan, dan dukungan bagi pengambilan keputusan. Sebagai contoh, satu baris dalam dataset transaksi dapat merepresentasikan satu transaksi, sedangkan atributnya memuat waktu, nilai pembelian, atau kanal layanan; label, apabila tersedia, dapat menunjukkan kategori hasil yang hendak diprediksi. Pembedaan antara objek, atribut, dan label menjadi landasan untuk merumuskan masalah analisis serta memilih metode yang sesuai (Han, Kamber, dan Pei, 2012). Dengan demikian, eksplorasi dataset bukan sekadar pemeriksaan teknis, melainkan bagian dari penalaran analitis yang menghubungkan representasi data dengan pertanyaan organisasi.

Pemahaman terhadap tipe dan skala data juga menentukan operasi yang sah serta cara hasil analisis ditafsirkan. Kategori nominal, misalnya, tidak memiliki urutan bawaan, sedangkan data ordinal memiliki urutan yang tidak selalu menunjukkan jarak setara; pembedaan ini sejalan dengan pembahasan Stevens (1946) mengenai skala pengukuran. Data numerik, kategoris, waktu, dan teks juga menuntut perlakuan yang berbeda, termasuk dalam penanganan nilai hilang dan pemilihan ringkasan statistik. Selain itu, data dapat tersimpan dalam bentuk terstruktur,

semi-terstruktur, atau tidak terstruktur, sehingga rancangan pemeriksaan perlu menyesuaikan format dan konteks pembentukannya. Exploratory Data Analysis (EDA) membantu peneliti memeriksa pola, anomali, dan asumsi sebelum membangun model. Tukey (1977) menekankan eksplorasi sebagai cara untuk memperoleh pemahaman terhadap data, sementara visualisasi menyediakan sarana untuk mengamati distribusi maupun hubungan yang sulit dikenali melalui tabel semata.

Kebutuhan eksplorasi menjadi semakin beragam ketika organisasi mengolah data IoT, sensor, log, dan keamanan. Data sensor dapat tercatat berulang menurut waktu, memiliki frekuensi pengukuran yang tinggi, serta mengandung gangguan atau nilai yang hilang; karakteristik IoT seperti konektivitas dan pertukaran data antarpiranti juga memperluas skala serta keragaman sumber yang perlu dikelola (Gubbi et al., 2013). Log sistem merekam peristiwa operasional, sedangkan data keamanan dapat memuat jejak aktivitas yang perlu ditafsirkan dengan memperhatikan urutan waktu dan konteks kejadian. Bagi mahasiswa Sistem Informasi, kemampuan menerapkan pemeriksaan awal pada beragam sumber tersebut mendukung penilaian kualitas data dan pemilihan analisis yang relevan bagi kebutuhan organisasi. Bab ini membahas konsep dataset, objek, atribut, dan label; tipe, skala, dan karakteristik data; data terstruktur, semi-terstruktur, dan tidak terstruktur; EDA

dan visualisasi data; serta karakteristik data IoT, sensor, log, dan keamanan.

C. Konsep Dataset, Objek, Atribut, dan Label

Penambangan data dimulai dengan menetapkan apa yang dihitung sebagai satu unit pengamatan dan bagaimana unit tersebut direpresentasikan. Kumpulan data bukan sekadar berkas berisi baris dan kolom, melainkan representasi terpilih dari suatu fenomena untuk menjawab pertanyaan analitis. Pemahaman atas struktur ini membantu mencegah kesalahan ketika hasil analisis ditafsirkan atau digunakan untuk mendukung keputusan organisasi (Unknown, 2012; Kononenko, 2007).

Objek adalah unit yang diamati, sedangkan **atribut** adalah karakteristik yang dicatat pada objek tersebut. Dalam kumpulan data transaksi, satu objek dapat berupa satu transaksi, dengan atribut berupa waktu, nilai belanja, kanal pembayaran, dan identitas cabang. Jika unit analisisnya pelanggan, setiap objek mewakili seorang pelanggan dan atribut dapat merangkum frekuensi transaksi atau nilai pembelian dalam periode tertentu. Pemilihan unit analisis harus mengikuti pertanyaan yang hendak dijawab. Menggabungkan catatan transaksi dan pelanggan tanpa aturan agregasi, misalnya, dapat menyebabkan satu pelanggan muncul berkali-kali dan memengaruhi interpretasi pola.

Label adalah atribut sasaran yang nilainya telah diketahui pada data tertentu dan digunakan sebagai hasil yang hendak dipelajari dalam analisis terawasi. Label tidak selalu tersedia. Pada analisis segmentasi pelanggan tanpa kategori sasaran, kumpulan data dapat memuat objek dan atribut, tetapi tidak memiliki label yang ditetapkan sejak awal. Pembedaan tersebut berpengaruh pada perumusan tugas analisis dan pemilihan pendekatan, bukan semata-mata pada penamaan kolom (Unknown, 2012; Kononenko, 2007).

Sebagai ilustrasi di Indonesia, sebuah koperasi di Yogyakarta ingin mengidentifikasi transaksi yang perlu diperiksa lebih lanjut. Jika tujuan analisisnya adalah mengenali transaksi, objeknya ialah transaksi; atributnya dapat mencakup waktu, nilai, jenis pembayaran, serta cabang; sementara label “perlu diperiksa” hanya dapat digunakan apabila tersedia hasil pemeriksaan historis yang menjadi acuan. Label tidak boleh dibentuk hanya dari dugaan analis lalu diperlakukan seolah-olah merupakan fakta yang telah teramati.

Tabel berikut menunjukkan contoh rancangan kumpulan data transaksi. Nilai di dalamnya bersifat ilustratif dan digunakan untuk memperjelas hubungan antara objek, atribut, dan label.

Tabel 2.1 Contoh Representasi Objek dan Atribut Transaksi

ID transaksi	Waktu	Cabang	Nilai (Rp)	Kanal	Label pemeriksaan
T-001	09.15	Sleman	125.000	Tunai	Tidak
T-002	10.42	Bantul	780.000	Transfer	Ya
T-003	11.03	Sleman	210.000	Kartu	Tidak
T-004	13.28	Kulon Progo	1.250.000	Transfer	Ya
T-005	15.10	Bantul	95.000	Tunai	Tidak

Setiap baris merepresentasikan satu transaksi, sedangkan kolom selain ID menyediakan atribut yang dapat diperiksa. Kolom label hanya bermakna jika “Ya” dan “Tidak” merujuk pada kriteria pemeriksaan yang konsisten, bukan penilaian yang berubah-ubah.

Saat menerapkan rancangan tersebut, analis perlu memeriksa apakah setiap objek memiliki pengenal yang sesuai, apakah atribut memiliki satu makna yang jelas, serta apakah label tersedia dan relevan dengan tujuan analisis. ID transaksi, misalnya, berfungsi mengenali catatan, tetapi tidak otomatis layak diperlakukan sebagai ukuran perilaku. Atribut yang menggambarkan hasil setelah peristiwa yang hendak diprediksi juga berisiko membocorkan informasi sasaran. Pemeriksaan makna atribut dan asal label merupakan bagian dari pemahaman data sebelum prapemrosesan data dan pemodelan, yang akan dibahas pada bab berikutnya (García, 2015; Unknown, 2011).

D. Tipe, Skala, dan Karakteristik Data

Pemilihan cara meringkas dan mengolah data bergantung pada tipe nilai serta skala pengukurannya. Atribut nominal, misalnya, menunjukkan kategori tanpa urutan; atribut ordinal memiliki urutan, tetapi jarak antara tingkat yang berdekatan belum tentu setara. Skala interval memiliki jarak bermakna tanpa titik nol mutlak, sedangkan skala rasio memiliki titik nol yang memungkinkan perbandingan rasio. Pembedaan ini membantu menentukan operasi dan interpretasi yang layak diterapkan pada suatu atribut (Unknown, 2012; Setiawan, 2025).

Atribut kategoris menyatakan kelompok, seperti jenis pembayaran, status perangkat, atau kategori keluhan. Atribut numerik dapat berupa nilai diskret, seperti jumlah transaksi, maupun kontinu, seperti suhu atau durasi. Data waktu memiliki urutan kronologis dan perlu dipahami melalui zona waktu, ketelitian pencatatan, serta interval pengamatan. Data teks, seperti isi aduan pelanggan, membawa informasi dalam bentuk bahasa dan tidak dapat diperlakukan begitu saja sebagai bilangan. Prapemrosesan data perlu mempertahankan makna asli ketika nilai disandikan, diubah satuannya, atau diringkas (García, 2015).

Contoh penerapan di Indonesia dapat ditemukan pada sistem pemantauan suhu ruang penyimpanan obat di puskesmas. Suhu yang tercatat dalam derajat Celsius

merupakan data numerik dengan skala rasio, sedangkan status “normal” atau “perlu diperiksa” merupakan kategori nominal apabila tidak diberi urutan. Jika petugas memberi penilaian kondisi “baik”, “cukup”, dan “buruk”, nilainya bersifat ordinal: urutannya jelas, tetapi selisih antara “baik” dan “cukup” belum tentu sama dengan selisih antara “cukup” dan “buruk”.

Tabel berikut memperlihatkan jenis atribut dan operasi yang sesuai melalui contoh operasional. Pengelompokan ini membantu menentukan ringkasan awal tanpa menganggap semua nilai dapat diperlakukan dengan cara yang sama.

Tabel 2.2 Tipe Atribut dan Operasi yang Sesuai

Atribut	Contoh nilai	Tipe/skala	Ringkasan yang sesuai
Jenis pembayaran	Tunai, QRIS	Kategoris nominal	Frekuensi, modus
Tingkat layanan	Rendah, sedang, tinggi	Kategoris ordinal	Frekuensi, median urutan
Suhu ruang	24,5 °C	Numerik rasio	Rata-rata, median, rentang
Jumlah kunjungan	3 kali	Numerik diskret rasio	Jumlah, median, distribusi
Waktu pencatatan	08.30 WIB	Waktu	Urutan, selang waktu
Ulasan pelanggan	“Layanan cepat”	Teks	Pemeriksaan isi dan tema

Tabel tersebut menunjukkan bahwa tipe data membatasi pilihan ringkasan yang bermakna. Menghitung rata-rata

kode kategori pembayaran, misalnya, tidak menghasilkan ukuran yang dapat ditafsirkan; sebaliknya, median dapat digunakan pada urutan ordinal jika urutan kategorinya ditetapkan secara konsisten.

Karakteristik data lain yang perlu dicatat mencakup kelengkapan, ketelitian, konsistensi satuan, serta cara dan waktu data dikumpulkan. Dua kolom yang sama-sama numerik mungkin memiliki arti berlainan jika satuannya berbeda; demikian pula angka kosong dapat berarti belum diukur, tidak berlaku, atau gagal direkam. Sebelum membandingkan atribut atau membuat visualisasi, analis perlu mendokumentasikan definisi, satuan, rentang nilai yang diharapkan, dan makna kategori. Pemeriksaan kualitas lebih mendalam, termasuk penanganan nilai hilang dan inkonsistensi, dibahas pada Bab 4 (García, 2015; Setiawan, 2025).

E. Data Terstruktur, Semi-Terstruktur, dan Tidak Terstruktur

Bentuk penyimpanan memengaruhi cara data ditemukan, dibaca, dan disiapkan untuk analisis. Data terstruktur disusun menurut skema yang relatif tetap, umumnya dalam baris dan kolom dengan definisi atribut yang telah ditentukan. Data semi-terstruktur memiliki penanda atau aturan organisasi, tetapi tidak harus mengikuti susunan kolom yang seragam. Data tidak

terstruktur, seperti dokumen bebas, gambar, dan rekaman suara, tidak menyediakan tabel atribut siap pakai untuk keseluruhan isinya. Ketiga bentuk ini dapat hadir bersamaan dalam satu sistem informasi (Unknown, 2012; Unknown, 2011).

Data terstruktur lazim ditemukan dalam tabel transaksi, catatan persediaan, atau daftar pelanggan. Bentuk tersebut memudahkan penyaringan dan penggabungan catatan selama definisi kolom serta kunci penghubungnya dipahami. Data semi-terstruktur dapat dijumpai pada catatan aktivitas aplikasi yang memiliki penanda waktu dan jenis peristiwa, tetapi memuat atribut berbeda menurut peristiwa. Format seperti JSON dan XML memungkinkan komponen bersarang, sehingga satu catatan dapat mempunyai bidang yang tidak dimiliki catatan lain. Data tidak terstruktur mencakup isi surat elektronik, foto produk, dan percakapan layanan pelanggan. Agar dapat dianalisis, isinya biasanya perlu dipilih atau ditransformasikan sesuai pertanyaan penelitian (García, 2015).

Sebuah UMKM batik di Pekalongan dapat menggabungkan data penjualan pada tabel kasir, catatan pesanan dari aplikasi pesan, dan foto motif yang diunggah pelanggan. Tabel kasir termasuk terstruktur. Catatan dari aplikasi yang memiliki cap waktu serta informasi pengirim dan isi pesan dapat bersifat semi-terstruktur. Isi pesan dan foto merupakan bahan tidak terstruktur yang memerlukan

cara pengolahan berbeda. Menggabungkan semua sumber itu menuntut perhatian terhadap pengenal pesanan, waktu pencatatan, serta izin penggunaan data pelanggan.

Pemetaan berikut merangkum bentuk data yang mungkin ditemui dalam kegiatan tersebut beserta fokus pemeriksaan awal yang sesuai.

Tabel 2.3 Bentuk Data dan Fokus Pemeriksaan Awal

Bentuk	Contoh Indonesia	Organisasi	Fokus pemeriksaan
Terstruktur	Tabel transaksi toko batik	Kolom dan baris tetap	Tipe, satuan, duplikasi
Semi-terstruktur	Catatan pesan pesanan	Bidang bertanda, bervariasi	Kelengkapan bidang, waktu
Semi-terstruktur	Respons aplikasi berbentuk JSON	Kunci dan nilai bertingkat	Struktur, bidang bersarang
Tidak terstruktur	Foto motif batik	Berkas visual	Kualitas, konteks, hak akses
Tidak terstruktur	Ulasan pelanggan	Teks bebas	Bahasa, makna, konteks

Perbedaan bentuk data tidak menentukan tingkat kegunaannya, tetapi memengaruhi langkah untuk menyiapkan data. Tabel transaksi memungkinkan pemeriksaan langsung terhadap kolom, sedangkan isi pesan dan foto memerlukan penafsiran serta proses representasi yang sesuai sebelum dibandingkan atau dianalisis.

Dalam praktik, klasifikasi tersebut tidak selalu mutlak. Sebuah dokumen dapat memiliki metadata terstruktur, seperti tanggal dan pengirim, sekaligus isi teks yang tidak terstruktur. Karena itu, analis sebaiknya menilai komponen data, bukan sekadar memberi satu label pada seluruh berkas. Penyimpanan juga perlu mempertahankan hubungan antara data asli dan representasi hasil pengolahan agar penelusuran kembali tetap mungkin. Kajian big data menekankan bahwa keragaman sumber dan format menambah kebutuhan integrasi serta tata kelola, khususnya ketika data berasal dari berbagai layanan digital (Putri, 2025; Unknown, 2011).

F. Exploratory Data Analysis dan Visualisasi Data

Exploratory Data Analysis (EDA) adalah pemeriksaan awal terhadap kumpulan data untuk memahami bentuk, isi, dan pola yang mungkin relevan sebelum analisis lanjutan dilakukan. EDA bukan hanya penyajian grafik, melainkan rangkaian pemeriksaan yang menghubungkan pertanyaan analitis dengan kualitas dan karakteristik data. Dalam penerapannya, analis memeriksa struktur kolom, rentang nilai, distribusi, nilai hilang, kemungkinan pencilan, dan hubungan ant atribut. Ringkasan dan visualisasi membantu menemukan keadaan yang tidak tampak dari beberapa baris data saja (Setiawan, 2025; García, 2015).

Untuk menerapkan EDA, mahasiswa dapat memulai dengan menetapkan objek dan atribut yang akan diperiksa, lalu menghitung banyaknya catatan serta nilai unik dan kosong. Setelah itu, ringkasan numerik seperti nilai minimum, maksimum, median, dan rentang dapat membantu mendeteksi nilai yang tidak masuk akal. Atribut kategoris dapat diringkas melalui frekuensi dan proporsi. Pemeriksaan tersebut tetap harus mempertimbangkan satuan serta cara pengumpulan data; nilai ekstrem tidak otomatis salah, karena mungkin merepresentasikan kejadian yang jarang tetapi sah.

Visualisasi dipilih berdasarkan pertanyaan dan tipe atribut. Histogram atau grafik kepadatan dapat membantu melihat bentuk distribusi numerik; diagram kotak merangkum median dan sebaran serta membantu menandai nilai jauh dari pola umum. Diagram batang cocok untuk membandingkan jumlah kategori, sedangkan diagram pencar dapat memperlihatkan pola hubungan dua atribut numerik. Untuk data waktu, grafik garis mempertahankan urutan kronologis sehingga perubahan dapat diamati. Grafik yang dipilih secara tidak tepat dapat menyamarkan variasi atau memberi kesan perbandingan yang keliru; karena itu label, satuan, skala sumbu, dan rentang waktu perlu dicantumkan dengan jelas (Setiawan, 2025).

Sebagai contoh, pengelola layanan perpustakaan kampus di Bandung dapat menerapkan EDA pada catatan

peminjaman selama satu semester. Pengelola memeriksa apakah tanggal tercatat konsisten, menghitung transaksi per kategori koleksi, dan menampilkan jumlah peminjaman per minggu. Jika sebuah minggu tampak memiliki nilai sangat tinggi, catatan sumber perlu diperiksa sebelum menyimpulkan bahwa minat baca meningkat; nilai tersebut mungkin berkaitan dengan kegiatan peminjaman massal atau kesalahan pencatatan.

Hasil EDA sebaiknya dicatat sebagai temuan sementara, bukan kesimpulan sebab-akibat. Hubungan yang tampak pada diagram pencar, misalnya, belum membuktikan bahwa perubahan satu atribut menyebabkan perubahan atribut lain. Demikian pula, penghapusan nilai ekstrem tanpa pemeriksaan konteks dapat menghilangkan informasi yang berguna. EDA memberi dasar untuk menetapkan tindakan pengolahan selanjutnya, sedangkan prosedur pembersihan, transformasi, dan penanganan pencilan dibahas lebih rinci pada Bab 4 (García, 2015; Unknown, 2011).

Dalam konteks layanan publik di Indonesia, eksplorasi yang cermat dapat membantu membedakan perubahan pola nyata dari persoalan pencatatan. Karena itu, setiap grafik perlu dihubungkan kembali dengan definisi objek, periode pengamatan, dan sumber data. Pembaca dapat menerapkan urutan yang sama pada data akademik, penjualan, maupun operasional tanpa mengasumsikan bahwa satu jenis visualisasi sesuai untuk semua pertanyaan.

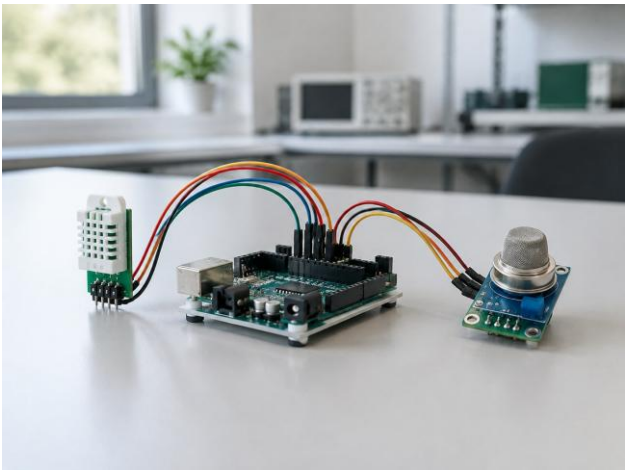
G. Karakteristik Data IoT, Sensor, Log, dan Data Keamanan

Data dari perangkat Internet untuk Segala (IoT), sensor, log, dan sistem keamanan mempunyai konteks pembentukan yang perlu diperiksa sebelum dianalisis. Sensor menghasilkan pengukuran terhadap besaran tertentu pada waktu tertentu, sementara log merekam peristiwa yang terjadi dalam aplikasi, perangkat, atau jaringan. Data keamanan dapat memuat aktivitas yang relevan untuk menelusuri kejadian, misalnya catatan autentikasi atau komunikasi jaringan. Penafsiran data tersebut memerlukan konteks sumber, waktu, identitas perangkat atau layanan, serta definisi setiap atribut (Dua, 2016).

Data sensor umumnya memiliki urutan waktu, satuan pengukuran, dan frekuensi pencatatan yang perlu diperhatikan. Gangguan pembacaan, keterlambatan pengiriman, atau catatan yang hilang dapat membuat rangkaian pengukuran tidak lengkap. Catatan dari beberapa sensor juga dapat sulit dibandingkan jika jam perangkat tidak selaras atau satuannya berbeda. Dalam contoh sistem pemantauan kualitas udara berbasis IoT, suhu, kelembapan, dan indikator gas dicatat melalui sensor dan dikirim ke platform pemantauan. Makna setiap nilai bergantung pada jenis sensor, satuan, waktu pembacaan, serta kondisi pemasangannya (Alfirdaus, 2024; Agung Suburdjati, 2026).

Log dan data keamanan menambahkan kebutuhan untuk memeriksa urutan peristiwa dan asal catatan. Satu kejadian yang sama dapat tercatat pada lebih dari satu komponen, sedangkan format atau zona waktu yang tidak seragam dapat menyulitkan penyelarasan. Catatan yang tidak muncul juga tidak selalu berarti aktivitas tidak terjadi; bisa jadi pencatatan gagal atau sumber tidak mengirimkan data. Dalam analisis keamanan, volume dan variasi catatan perlu dipahami bersama konteks operasional agar indikator yang tidak biasa tidak langsung dianggap sebagai serangan. Pengolahan data keamanan perlu mempertimbangkan karakteristik sumber serta perlindungan informasi sensitif (Dua, 2016; Amanda, 2025).

Perangkat sensor yang digunakan untuk mengamati lingkungan memperlihatkan bahwa nilai data memiliki hubungan langsung dengan alat, lokasi, dan kondisi pengukuran. Gambar berikut memperlihatkan contoh perangkat pemantauan yang dapat menghasilkan catatan sensor untuk ditinjau dalam proses eksplorasi.



Gambar 2.1 Perangkat Pemantauan Kualitas Udara Berbasis Sensor. Sumber: Ilustrasi, diolah penulis

Komponen yang terlihat mengingatkan bahwa data pengukuran bukan angka yang berdiri sendiri: interpretasinya bergantung pada sensor, pemasangan, dan waktu pencatatan. Saat menerapkan EDA, mahasiswa dapat memeriksa rentang nilai, satuan, cap waktu, kelengkapan catatan, serta konsistensi identitas perangkat sebelum membandingkan hasil antarperiode atau antarlokasi.

Pada sistem pemantauan kualitas udara di Indonesia, misalnya, mahasiswa dapat mengelompokkan catatan berdasarkan perangkat dan waktu, memeriksa apakah setiap atribut memiliki satuan, lalu membandingkan catatan yang hilang dengan jadwal pengiriman yang diharapkan.

Untuk log keamanan, pemeriksaan awal dapat menandai catatan yang format waktunya berbeda atau sumbernya tidak dikenali tanpa langsung menyimpulkan adanya serangan. Hasil eksplorasi seperti ini menjadi dasar untuk merancang prapemrosesan data yang sesuai; analisis rinci data sensor dan deteksi ancaman akan dibahas pada bab-bab berikutnya (Alfirdaus, 2024; Imron, 2023; Dua, 2016).

H. Rangkuman

Analisis data yang dapat dipercaya berawal dari penetapan unit pengamatan dan cara merepresentasikannya. Dataset bukan sekadar tabel, melainkan representasi terpilih dari fenomena yang disusun untuk menjawab pertanyaan tertentu. Objek adalah unit yang diamati, atribut menggambarkan karakteristiknya, dan label merupakan sasaran yang nilainya diketahui untuk dipelajari dalam analisis terawasi. Karena itu, pemilihan unit analisis harus konsisten dengan tujuan: catatan transaksi berbeda dari ringkasan pelanggan, dan pencampuran keduanya tanpa aturan agregasi dapat mengubah makna pola. Label juga harus bersumber dari hasil pengamatan yang konsisten; dugaan analis bukan pengganti fakta historis. Pengenal seperti ID berguna untuk membedakan catatan, tetapi belum tentu menjadi atribut yang bermakna bagi analisis. Atribut yang baru diketahui setelah hasil terjadi bahkan dapat membocorkan sasaran yang hendak diprediksi.

Tipe dan skala atribut menentukan ringkasan serta operasi yang dapat ditafsirkan. Kategori nominal sesuai diringkas dengan frekuensi atau modus, sedangkan atribut ordinal memiliki urutan sehingga median urutan dapat digunakan jika kategorinya ditetapkan secara konsisten. Data numerik dapat berupa diskret atau kontinu; data waktu perlu mempertimbangkan urutan, zona waktu, dan ketelitian pencatatan; sementara teks memerlukan pemeriksaan isi atau tema. Mengubah kategori menjadi angka tidak serta-merta membuat rata-ratanya bermakna. Definisi atribut, satuan, rentang yang diharapkan, kelengkapan, dan makna nilai kosong perlu diperiksa sebelum perbandingan atau visualisasi.

Data juga berbeda dalam bentuk penyimpanannya. Data terstruktur mengikuti skema baris dan kolom yang relatif tetap, data semi-terstruktur memiliki penanda atau aturan organisasi tanpa susunan seragam, dan data tidak terstruktur—seperti dokumen, gambar, atau suara—tidak menyediakan tabel atribut siap pakai untuk keseluruhan isinya. Sistem informasi dapat memuat ketiganya sekaligus. Memahami representasi, kualitas, dan konteks pengumpulan data membantu analis memilih langkah eksplorasi yang sesuai, menghindari kesimpulan keliru, serta menyiapkan data secara bertanggung jawab sebelum prapemrosesan dan pemodelan.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana perubahan unit analisis dari transaksi menjadi pelanggan dapat memengaruhi representasi objek, atribut, dan interpretasi pola. Rumuskan aturan agregasi yang perlu ditetapkan agar catatan transaksi berulang tidak menyebabkan kesimpulan tentang perilaku pelanggan menjadi bias.
2. Evaluasi kelayakan label “perlu diperiksa” untuk membangun model analisis transaksi. Jelaskan bagaimana asal label, konsistensi kriteria pemeriksaan, dan waktu pencatatan atribut dapat memengaruhi keandalan analisis serta risiko kebocoran informasi sasaran.
3. Bandingkan cara meringkas atribut nominal, ordinal, numerik, waktu, dan teks. Gunakan contoh konkret untuk menjelaskan mengapa satu operasi statistik tidak selalu sesuai bagi semua tipe data dan bagaimana pilihan ringkasan dapat memengaruhi interpretasi.
4. Telaah dampak perbedaan data terstruktur, semi-terstruktur, dan tidak terstruktur terhadap proses eksplorasi dataset. Rumuskan strategi awal untuk menggabungkan ketiganya dalam analisis Sistem Informasi tanpa mengabaikan skema, makna, dan konteks asal data.
5. Argumentasikan mengapa pencatatan definisi atribut, satuan, rentang nilai, dan makna nilai kosong perlu dilakukan sebelum visualisasi atau pemodelan. Sertakan contoh tentang bagaimana kelalaian terhadap

salah satu aspek tersebut dapat menghasilkan keputusan organisasi yang keliru.

b. Pilihan Berganda

1. Sebuah analisis bertujuan mempelajari kebiasaan belanja setiap pelanggan selama satu tahun. Rancangan unit analisis yang paling sesuai adalah...
    a. Setiap transaksi, tanpa menghubungkan transaksi milik pelanggan yang sama
    b. Setiap pelanggan, dengan atribut yang merangkum transaksinya selama periode tersebut
    c. Setiap baris pada berkas, tanpa memeriksa isi dan identitasnya
    d. Setiap cabang, dengan nilai transaksi sebagai label pelanggan
    e. Setiap jenis pembayaran, dengan transaksi sebagai atribut
2. Dalam dataset historis, hasil pemeriksaan transaksi belum dicatat secara konsisten. Tindakan paling tepat terhadap kolom “perlu diperiksa” adalah...
    a. Mengisi seluruh nilainya dengan “Tidak” agar dataset lengkap
    b. Menetapkan label berdasarkan dugaan analis dan memperlakukannya sebagai fakta
    c. Menggunakan ID transaksi sebagai pengganti label
    d. Menunda penggunaan label sampai sumber dan kriteria hasil pemeriksaan dapat dipastikan

e. Mengubah nilai label menjadi angka agar dapat langsung dimodelkan

3. Kolom ID transaksi terutama digunakan untuk...
    a. Membedakan dan mengenali catatan transaksi
    b. Mengukur besarnya perilaku belanja
    c. Menggantikan atribut sasaran yang tidak tersedia
    d. Menentukan urutan tingkat layanan
    e. Menghitung rata-rata transaksi pelanggan
4. Untuk merangkum atribut nominal “jenis pembayaran”, pilihan yang paling bermakna adalah...
    a. Rata-rata kode kategori
    b. Selisih nilai kategori tertinggi dan terendah
    c. Frekuensi setiap kategori dan modus
    d. Median angka hasil pengodean kategori
    e. Rasio antara kode QRIS dan kode tunai
5. Penilaian kondisi “baik”, “cukup”, dan “buruk” termasuk atribut ordinal karena...
    a. Nilainya selalu memiliki jarak yang sama
    b. Kategorinya menunjukkan urutan, tetapi jarak antartingkat belum tentu setara
    c. Nilainya memiliki titik nol mutlak
    d. Kategorinya tidak memiliki hubungan urutan
    e. Nilainya selalu berupa pengukuran kontinu
6. Mengapa rata-rata kode kategori pembayaran umumnya tidak menghasilkan ringkasan yang dapat ditafsirkan?

a. Karena kategori nominal tidak memiliki urutan atau jarak numerik bermakna
b. Karena rata-rata hanya berlaku untuk data waktu
c. Karena kategori pembayaran selalu bersifat ordinal
d. Karena semua data kategoris tidak dapat dihitung frekuensinya
e. Karena kode numerik tidak dapat disimpan dalam dataset

7. Atribut yang baru diketahui setelah hasil yang hendak diprediksi terjadi berisiko...
   a. Mengubah data numerik menjadi data teks
   b. Menghilangkan kebutuhan untuk menentukan objek
   c. Membuat seluruh label menjadi tidak terpakai
   d. Membocorkan informasi sasaran ke dalam proses analisis
   e. Mengubah data terstruktur menjadi semi-terstruktur

8. Manakah contoh data semi-terstruktur?
   a. Tabel transaksi dengan kolom tetap dan tipe data yang ditentukan
   b. Foto produk tanpa atribut pendamping
   c. Dokumen teks bebas tanpa penanda organisasi
   d. Rekaman suara percakapan pelanggan

e. Berkas dengan penanda dan aturan organisasi, tetapi susunan isinya tidak seragam seperti tabel tetap

9. Dalam menyiapkan data waktu untuk eksplorasi, hal yang perlu diperiksa adalah...
    a. Modus kode waktu
    b. Apakah waktu dapat diubah menjadi label tanpa kriteria
    c. Zona waktu, ketelitian pencatatan, dan interval pengamatan
    d. Rata-rata nama bulan yang dicatat
    e. Apakah setiap waktu memiliki kategori nominal
10. Dua atribut numerik menggunakan satuan berbeda. Langkah awal yang paling tepat sebelum membandingkan nilainya adalah...
    a. Menghapus atribut dengan rentang lebih kecil
    b. Memeriksa definisi atribut dan menyamakan atau mendokumentasikan satuannya
    c. Menganggap angka yang lebih besar selalu menunjukkan kondisi lebih penting
    d. Mengubah kedua atribut menjadi kategori tanpa mempertimbangkan makna
    e. Menghitung rata-rata keduanya lalu langsung membandingkan hasilnya

c. Soal Project / Studi Kasus

1. Sebuah koperasi di Yogyakarta hendak mengidentifikasi transaksi yang perlu diperiksa. Sistemnya menyimpan

ID transaksi, waktu, cabang, nilai belanja, kanal pembayaran, serta status pemeriksaan. Namun, sebagian cabang memakai kriteria pemeriksaan berbeda; beberapa status diperbarui beberapa hari setelah transaksi. Tim juga menemukan atribut "hasil tindak lanjut" yang hanya tersedia setelah pemeriksaan selesai. Manajer meminta model awal untuk menentukan transaksi yang perlu ditinjau setiap hari. Telaah rancangan dataset tersebut. Tentukan unit analisis, atribut yang dapat digunakan, kelayakan label, risiko kebocoran sasaran, dan pemeriksaan data yang perlu dilakukan sebelum pemodelan. Berikan rekomendasi kepada manajer.

2. Dinas kesehatan kabupaten ingin mengeksplorasi kondisi penyimpanan obat di sejumlah puskesmas. Data yang tersedia meliputi tabel suhu dengan waktu pencatatan dan kode ruang, berkas semi-terstruktur dari perangkat pemantau, serta catatan teks petugas tentang kondisi ruang. Sebagian suhu dicatat dalam satuan berbeda, ada waktu tanpa zona yang jelas, dan kolom kosong belum dibedakan antara alat rusak dan pengukuran yang tidak berlaku. Kepala dinas ingin membandingkan kondisi antarpuskesmas dan menetapkan prioritas pemeriksaan. Rumuskan strategi eksplorasi dan integrasi data, termasuk penanganan tipe atribut, satuan, waktu, nilai kosong, serta batas interpretasi sebelum data dipakai untuk keputusan.

BAB 3
METODOLOGI DAN SIKLUS DATA MINING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan tahapan proses Knowledge Discovery in Databases untuk mengubah data menjadi pengetahuan yang relevan bagi kebutuhan organisasi.
2. Menggunakan metodologi CRISP-DM untuk menyusun alur kerja proyek data mining sesuai tujuan dan konteks bisnis.
3. Menentukan kebutuhan bisnis serta karakteristik dan kualitas data sebagai dasar perumusan masalah analitik.
4. Menerapkan tahapan persiapan data, pemodelan, dan evaluasi dengan memilih teknik yang sesuai terhadap masalah dan kriteria keberhasilan.
5. Mengimplementasikan hasil pemodelan ke dalam lingkungan operasional serta menentukan langkah pemantauan dan pemeliharaan siklus hidup model machine learning.

B. Pendahuluan

Organisasi mengumpulkan data transaksi, interaksi pelanggan, dan proses operasional dalam jumlah besar,

tetapi ketersediaan data tidak dengan sendirinya menghasilkan keputusan yang lebih baik. Data yang tidak sesuai dengan kebutuhan analisis, mengandung nilai yang hilang, atau diproses tanpa tujuan yang jelas dapat menghasilkan pola yang menyesatkan. Karena itu, data mining perlu dipahami sebagai rangkaian kerja yang sistematis, bukan sekadar penerapan algoritma terhadap sekumpulan data. Dalam kerangka Knowledge Discovery in Databases (KDD), proses penemuan pengetahuan mencakup pemilihan data, prapemrosesan, transformasi, data mining, serta interpretasi dan evaluasi hasil (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Kerangka tersebut menegaskan bahwa mutu pengetahuan yang diperoleh bergantung pada keputusan di sepanjang proses, termasuk penetapan masalah dan penilaian kegunaan hasil. Bagi mahasiswa Sistem Informasi, pemahaman ini diperlukan untuk menghubungkan sumber daya data dan metode komputasional dengan proses, kebutuhan, serta tujuan organisasi.

Salah satu metodologi yang membantu mengorganisasikan pekerjaan tersebut adalah CRISP-DM, yang membagi proyek data mining ke dalam enam fase: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment (CRISP-DM Consortium, 2000; Shearer, 2000). Fase-fase itu tidak semata-mata membentuk urutan linear. Temuan pada

tahap pemodelan, misalnya, dapat menunjukkan bahwa data belum memadai sehingga proses perlu kembali ke tahap persiapan atau pemahaman data. Pada tahap pemahaman bisnis, tim merumuskan sasaran organisasi dan menerjemahkannya menjadi masalah analitik yang dapat ditangani. Selanjutnya, pemahaman data mencakup pemeriksaan sumber, struktur, kualitas, dan pola awal data. Persiapan data mengubah data mentah menjadi masukan yang layak untuk pemodelan, sedangkan pemodelan menerapkan teknik analitik yang sesuai dengan jenis masalah. Evaluasi kemudian menguji bukan hanya kinerja teknis model, melainkan juga kesesuaiannya dengan sasaran bisnis yang telah ditetapkan. Sebagai contoh, model untuk mengidentifikasi pelanggan yang berisiko berhenti berlangganan perlu dinilai berdasarkan ketepatan identifikasi sekaligus kegunaannya bagi strategi retensi.

Hasil analisis baru memberikan nilai operasional apabila dapat diterapkan secara bertanggung jawab dalam proses organisasi. Tahap deployment mencakup penyajian hasil atau integrasi model ke dalam sistem dan alur kerja, sementara pemantauan setelah penerapan diperlukan untuk mendeteksi penurunan kinerja ketika pola data maupun kondisi bisnis berubah. Dengan demikian, siklus hidup model machine learning berlanjut setelah model pertama kali dirilis: kinerja dan relevansi model perlu ditinjau, kemudian model diperbarui atau dibangun kembali

ketika bukti operasional menunjukkan kebutuhan tersebut. Perspektif siklus hidup ini membantu mahasiswa Sistem Informasi mempertimbangkan keterkaitan antara teknologi, kualitas data, proses bisnis, dan tata kelola hasil analitik. Bab ini membahas proses Knowledge Discovery in Databases, metodologi CRISP-DM, pemahaman bisnis dan data, persiapan data, pemodelan dan evaluasi, serta deployment dan siklus hidup model machine learning.

C. Proses Knowledge Discovery in Databases

Pembahasan dataset dan eksplorasi pada Bab 2 menjadi titik awal untuk memahami bagaimana data dapat diolah menjadi pengetahuan yang dapat digunakan. Kerangka Knowledge Discovery in Databases (KDD) menempatkan penambangan data sebagai bagian dari proses yang lebih luas: hasil analisis bergantung pada pemilihan data, kualitas prapemrosesan, transformasi, serta penafsiran temuan. Dengan demikian, keberhasilan proyek tidak cukup ditentukan oleh algoritma yang digunakan, tetapi juga oleh kesesuaian setiap tahap dengan tujuan analisis (Jain, 2023; Kononenko, 2007).

> 📖 **DEFINISI**
> **Knowledge Discovery in Databases (KDD)** adalah proses sistematis untuk menemukan pengetahuan yang valid, baru, berguna, dan dapat dipahami dari data melalui serangkaian tahap pengolahan dan penafsiran.

Proses KDD dapat diterapkan sebagai rangkaian tahap yang saling berkaitan. Pemilihan data menetapkan sumber dan bagian data yang relevan dengan pertanyaan analitik. Prapemrosesan mengurangi gangguan seperti data tidak konsisten atau tidak lengkap. Transformasi menyesuaikan representasi data agar sesuai dengan analisis, misalnya dengan membentuk fitur turunan dari catatan transaksi. Penambangan data kemudian menggunakan algoritma untuk menemukan pola atau membangun model. Tahap terakhir, interpretasi dan evaluasi, memeriksa apakah temuan memiliki makna dan dapat digunakan dalam konteks masalah. Tahap-tahap tersebut merupakan kerangka kerja konseptual, bukan prosedur yang selalu dijalankan satu kali secara linear; hasil pemeriksaan dapat mendorong tim kembali ke tahap sebelumnya (Unknown, 2011; Unknown, 2019).

Perbedaan antara KDD dan penambangan data perlu diperhatikan dalam pelaksanaan proyek. Penambangan data merujuk pada penerapan teknik komputasional untuk menemukan pola atau membentuk model, sedangkan KDD meliputi keseluruhan proses yang mengubah data mentah menjadi pengetahuan yang ditafsirkan. Sebuah algoritma klasterisasi, misalnya, dapat menghasilkan kelompok pelanggan, tetapi kelompok tersebut belum menjadi pengetahuan yang berguna sebelum diuji kestabilannya, diberi makna berdasarkan konteks organisasi, dan

dihubungkan dengan kemungkinan tindakan. Kerangka penambangan data menekankan perlunya mengaitkan teknik dengan pemahaman data dan tujuan penggunaan hasil (Kononenko, 2007; Jain, 2023).

Urutan kerja KDD dapat digunakan untuk merencanakan proyek yang konkret. Tabel berikut menunjukkan penerapan tahap-tahap tersebut pada analisis transaksi sebuah koperasi di Yogyakarta. Data dan situasi pada tabel merupakan ilustrasi pembelajaran, bukan hasil penelitian terhadap koperasi tertentu.

Tabel 3.1 Penerapan Tahap KDD pada Analisis Transaksi Koperasi

Tahap	Aktivitas	Hasil kerja	Pertanyaan pemeriksaan
Pemilihan	Memilih transaksi anggota	Data transaksi terpilih	Apakah rentang waktu sesuai?
Prapemrosesan	Memeriksa nilai kosong dan duplikasi	Data lebih konsisten	Apakah cacat data mengubah cakupan?
Transformasi	Membentuk ringkasan belanja anggota	Fitur analisis	Apakah fitur mewakili kebutuhan?
Penambangan data	Menemukan kelompok pola belanja	Hasil pengelompokan	Apakah hasil cukup stabil?
Interpretasi	Menelaah ciri tiap kelompok	Makna kelompok	Apakah makna masuk akal bagi koperasi?

Pemanfaatan	Menyusun usulan tindak lanjut	Rekomendasi operasional	Apakah tindakan dapat dijalankan?

Tabel tersebut memperlihatkan bahwa setiap tahap menghasilkan keluaran yang menjadi masukan bagi tahap selanjutnya. Pemeriksaan makna hasil dapat pula mengungkap masalah pada pemilihan atau transformasi data, sehingga proses perlu diulang sebelum rekomendasi digunakan.



Gambar 3.1 Rangkaian Proses KDD. Sumber: Diolah penulis

Rangkaian pada gambar menekankan hubungan antartahap sekaligus kemungkinan untuk kembali meninjau keputusan sebelumnya. Pada kasus koperasi, misalnya, hasil

pengelompokan yang sulit ditafsirkan dapat mendorong pemeriksaan ulang cara transaksi dipilih atau diringkas.

Pengetahuan yang diperoleh perlu dinilai dengan mempertimbangkan kualitas bukti dan konteks penggunaannya. Pola yang tampak menarik secara statistik belum tentu berguna bagi organisasi apabila tidak menjawab pertanyaan awal, tidak dapat dijelaskan kepada pengguna, atau didasarkan pada data yang tidak mewakili keadaan sebenarnya. Karena itu, tim perlu mencatat alasan pemilihan data, keputusan transformasi, dan batasan hasil. Praktik dokumentasi tersebut memudahkan peninjauan proses dan mengurangi risiko rekomendasi dibangun dari interpretasi yang keliru (Unknown, 2011; Solimun, 2023).

D. Metodologi CRISP-DM

Kerangka KDD menjelaskan lingkup proses penemuan pengetahuan, sedangkan metodologi CRISP-DM membantu tim mengorganisasikan proyek data mining ke dalam fase kerja yang dapat dikelola. Dalam konteks Sistem Informasi, metodologi ini memberi ruang bagi pembahasan kebutuhan organisasi, kondisi data, pengembangan model, dan penerapan hasil sebagai bagian dari satu proyek. Pelaksanaan yang terstruktur mendukung keterkaitan antara pilihan analitik dan tujuan penggunaan hasil (Unknown, 2011; Kononenko, 2007).

CRISP-DM umumnya membagi proyek ke dalam enam fase: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Fase-fase tersebut bukan tahapan yang harus dilalui sekali tanpa kembali. Sebagai contoh, apabila evaluasi menunjukkan bahwa model tidak menjawab kebutuhan operasional, tim dapat memperbaiki rumusan tujuan bisnis. Jika hasil pemodelan tidak memadai karena fitur yang tersedia kurang informatif, pekerjaan dapat kembali ke pemahaman atau persiapan data. Sifat iteratif ini membantu tim memperlakukan hasil sementara sebagai dasar pemeriksaan dan perbaikan, bukan sebagai keluaran final secara otomatis (Unknown, 2011; Unknown, 2019).

Penerapan metodologi tersebut sebaiknya menghasilkan artefak kerja pada tiap fase. Pada pemahaman bisnis, artefak dapat berupa pernyataan masalah dan kriteria keberhasilan. Pemahaman data menghasilkan inventaris sumber dan catatan kualitas. Persiapan data mendokumentasikan kumpulan data analitik beserta keputusan transformasinya. Fase pemodelan menghasilkan model dan catatan pengaturan eksperimen. Evaluasi membandingkan hasil dengan kriteria yang telah disepakati, sedangkan deployment merencanakan cara penyajian atau integrasi hasil. Artefak yang terdokumentasi membuat keputusan proyek lebih mudah ditinjau oleh analis, pemilik

proses, dan pengelola sistem (Kononenko, 2007; Solimun, 2023).

Penerapan di Indonesia dapat dilihat melalui ilustrasi proyek prediksi kadar karbon dioksida di sebuah ruang kelas di Bandung menggunakan data sensor suhu, kelembapan, dan CO_2. Tim terlebih dahulu perlu menetapkan siapa yang akan menggunakan hasil prediksi dan keputusan apa yang akan didukung, seperti pengaturan ventilasi. Setelah itu, tim memeriksa ketersediaan dan konsistensi data sensor sebelum membandingkan model. Studi pemodelan kadar CO_2 berbasis sensor dan perangkat ESP32 menunjukkan bentuk konteks IoT yang dapat menjadi pertimbangan dalam merancang kebutuhan data dan lingkungan penerapan; rincian metodologinya tetap perlu disesuaikan dengan tujuan proyek yang sedang dikerjakan (Agung Suburdjati, 2026).

Agar perpindahan antarfase dapat dikelola, tim dapat menetapkan keluaran minimum dan pertanyaan kendali untuk setiap fase. Tabel berikut menyajikan contoh penerapannya pada proyek ilustratif tersebut.

Tabel 3.2 Keluaran Fase CRISP-DM untuk Proyek Sensor CO_2

Fase	Keluaran utama	Pertanyaan kendali
Pemahaman bisnis	Tujuan dan pengguna hasil	Keputusan apa yang akan didukung?

Pemahaman data	Inventaris dan profil data	Apakah sensor menyediakan data yang relevan?
Persiapan data	Kumpulan data analitik	Apakah catatan waktu dan nilai sensor layak?
Pemodelan	Model kandidat	Apakah teknik sesuai dengan sasaran?
Evaluasi	Catatan hasil dan batasan	Apakah kriteria bisnis terpenuhi?
Deployment	Rencana penggunaan	Siapa memantau hasil operasional?

Keluaran pada tabel membantu proyek menjaga kesinambungan antara fase. Jika model belum memenuhi kriteria, tim dapat menggunakan catatan fase sebelumnya untuk menentukan bagian yang perlu ditinjau ulang, bukan sekadar mengganti algoritma.

CRISP-DM menyediakan struktur kerja, tetapi tidak menggantikan komunikasi antarpihak atau penilaian terhadap konteks organisasi. Pemilik proses perlu memastikan bahwa masalah analitik memang berkaitan dengan keputusan nyata, sementara tim teknis menjelaskan batasan data dan model secara dapat dipahami. Pada proyek Sistem Informasi, hubungan tersebut juga perlu mempertimbangkan alur kerja dan sistem yang akan menggunakan hasil. Metodologi menjadi efektif ketika fase-fasenya dipakai untuk menyelaraskan tujuan, bukti, dan pelaksanaan proyek, bukan sekadar sebagai daftar administrasi (Unknown, 2019; Dua, 2016).

E. Business Understanding dan Data Understanding

Fase pemahaman bisnis menerjemahkan kebutuhan organisasi menjadi masalah yang dapat ditangani melalui analisis data. Tim perlu mengidentifikasi pihak yang membutuhkan hasil, keputusan yang hendak didukung, batasan pelaksanaan, serta ukuran keberhasilan proyek. Tujuannya bukan menganggap bahwa setiap masalah organisasi harus diselesaikan dengan model pembelajaran mesin, melainkan memastikan bahwa penggunaan analitik relevan dengan kebutuhan dan sumber daya yang tersedia (Kononenko, 2007; Unknown, 2011).

Rumusan masalah bisnis yang terlalu luas menyulitkan pemilihan data dan evaluasi hasil. Sebagai contoh, tujuan "meningkatkan layanan pelanggan" belum menjelaskan layanan mana yang bermasalah atau bagaimana perubahan akan dinilai. Rumusan yang lebih operasional dapat menyatakan bahwa sebuah perusahaan transportasi di Jakarta ingin mengenali pelanggan yang berisiko mengurangi penggunaan layanan agar petugas dapat menyusun tindak lanjut. Pernyataan tersebut menghubungkan kelompok pengguna hasil, tindakan yang dipertimbangkan, dan kebutuhan analitik. Namun, tim tetap perlu menyepakati batas waktu, sumber data yang sah untuk digunakan, dan konsekuensi apabila prediksi keliru.

Setelah tujuan bisnis dijelaskan, tim menyusun rumusan analitik yang sesuai. Pertanyaan tentang perkiraan jumlah pelanggan termasuk masalah prediksi nilai, sedangkan pertanyaan tentang apakah pelanggan akan berhenti menggunakan layanan termasuk masalah klasifikasi. Jika organisasi belum memiliki label yang dapat dipercaya dan ingin menemukan kelompok perilaku, klasterisasi dapat dipertimbangkan. Pilihan tugas analitik harus mengikuti bentuk keputusan dan data yang tersedia, bukan ditentukan semata-mata oleh popularitas algoritma. Segmentasi pelanggan berbasis pola konsumsi, misalnya, menunjukkan bagaimana hasil pengelompokan dapat dikaitkan dengan perancangan strategi pemasaran, tetapi kebermanfaatannya tetap bergantung pada interpretasi dan tindakan organisasi (Iqbal, 2025; Priandini, 2025).

Fase pemahaman data memeriksa apakah data yang tersedia benar-benar mendukung rumusan masalah. Pemeriksaan mencakup asal data, struktur, satuan, rentang waktu, cakupan objek, label jika tersedia, serta batasan akses dan penggunaan. Eksplorasi awal dapat menemukan pola umum atau persoalan kualitas, tetapi tidak seharusnya langsung dianggap sebagai bukti sebab-akibat. Bab 2 telah membahas eksplorasi dataset; dalam fase ini, eksplorasi digunakan untuk memeriksa kesesuaian data terhadap sasaran bisnis. Untuk proyek keamanan siber, misalnya, data log perlu dipahami berdasarkan jenis peristiwa yang

dicatat dan konteks penggunaannya sebelum dipakai sebagai masukan analitik (Dua, 2016).

Dua fase ini perlu dihubungkan melalui kriteria yang dapat diperiksa. Tabel berikut memberikan contoh pemeriksaan untuk proyek retensi pelanggan pada perusahaan transportasi di Indonesia. Seluruh butir merupakan rancangan ilustratif yang perlu disesuaikan dengan keadaan organisasi.

Tabel 3.3 Pemeriksaan Awal Tujuan Bisnis dan Data

Aspek	Pertanyaan pemeriksaan	Contoh keputusan
Pengguna hasil	Siapa yang bertindak berdasarkan hasil?	Tim layanan pelanggan
Keputusan	Tindakan apa yang akan didukung?	Menentukan prioritas tindak lanjut
Unit analisis	Apa yang diprediksi?	Pelanggan per periode
Sumber data	Catatan apa yang tersedia?	Riwayat penggunaan layanan
Cakupan waktu	Apakah rentangnya relevan?	Tetapkan periode sesuai siklus bisnis
Kualitas awal	Apakah identitas dan waktu konsisten?	Periksa duplikasi dan catatan kosong

Tabel tersebut menunjukkan bahwa kecocokan data harus diperiksa bersama rancangan keputusan. Data yang banyak belum tentu memadai apabila tidak dapat dihubungkan dengan unit analisis atau rentang waktu yang sesuai.

✎ **CONTOH**
Sebuah perusahaan transportasi di Jakarta ingin memprioritaskan pelanggan untuk program retensi. Tim menetapkan bahwa hasil analisis akan digunakan oleh petugas layanan, lalu memeriksa apakah riwayat transaksi memiliki identitas pelanggan dan penanda waktu yang konsisten. Jika catatan kontak pelanggan tidak lengkap, tim mencatat keterbatasan tersebut dan membahas apakah proyek dapat dilanjutkan dengan data yang tersedia. Keputusan untuk melanjutkan atau memperbaiki pengumpulan data dibuat sebelum tim membangun model.

Contoh itu memperlihatkan bahwa pemahaman bisnis dan data berjalan beriringan. Rumusan tujuan dapat berubah ketika pemeriksaan data menemukan keterbatasan, tetapi perubahan tersebut perlu dicatat dan disepakati bersama pihak terkait. Pendekatan berbasis data juga memerlukan perhatian terhadap makna fitur, kualitas pengukuran, serta potensi dampak keputusan bagi pengguna. Karena itu, tim sebaiknya menghindari penggunaan atribut yang tidak memiliki alasan analitik atau tidak sesuai dengan tujuan layanan (Liu, 2018; Solimun, 2023).

F. Data Preparation, Modeling, dan Evaluation

Setelah tujuan bisnis dan kondisi data dipahami, tim menyiapkan kumpulan data yang dapat digunakan untuk pemodelan. Persiapan mencakup pemilihan catatan dan fitur yang relevan, pemeriksaan konsistensi, penggabungan sumber, serta pengubahan format data sesuai kebutuhan analisis. Keputusan pada tahap ini dapat memengaruhi hasil model; karena itu, tim perlu mencatat sumber data dan langkah yang dilakukan agar proses dapat ditinjau ulang. Teknik pembersihan, transformasi, dan rekayasa fitur dibahas lebih mendalam pada Bab 4 (Liu, 2018; Unknown, 2011).

Persiapan data harus mempertahankan hubungan antara fitur dan sasaran analisis. Dalam ilustrasi prediksi kebutuhan pemeliharaan pompa air di sebuah fasilitas pengolahan di Surabaya, data sensor dapat disusun berdasarkan waktu pengamatan, sedangkan catatan pemeliharaan dapat digunakan untuk memeriksa kejadian yang hendak diprediksi. Tim perlu memastikan bahwa informasi yang digunakan sebagai fitur memang tersedia sebelum waktu prediksi. Jika informasi setelah kerusakan terjadi ikut dimasukkan, model dapat tampak akurat pada pengujian tetapi tidak mencerminkan situasi penggunaan sebenarnya. Data sensor dan perangkat IoT menghadirkan

kebutuhan untuk menelaah waktu pencatatan dan konteks pengukuran secara cermat (Agung Suburdjati, 2026).

Pemodelan menerapkan teknik analitik yang sesuai dengan tugas dan karakteristik data. Tim dapat membandingkan beberapa model kandidat, tetapi perbandingan hanya bermakna jika setiap model dilatih dan diuji dengan rancangan yang sebanding. Pemilihan model juga tidak berhenti pada angka kinerja: kecepatan pemrosesan, kebutuhan penjelasan, dan kesesuaian terhadap alur kerja dapat memengaruhi keputusan penerapan. Pembahasan algoritma klasifikasi, regresi, dan klasterisasi secara terperinci tersedia pada bab-bab berikutnya; dalam bab ini, fokusnya adalah hubungan antara pemilihan teknik dan rancangan proyek (Kononenko, 2007; Unknown, 2019).

Evaluasi memeriksa apakah model memenuhi kriteria yang sudah ditetapkan. Untuk klasifikasi, kesalahan pada tiap kelas perlu dipertimbangkan, terutama ketika konsekuensi salah deteksi tidak sama. Pada sistem yang menandai kerusakan pompa, kegagalan mengenali kerusakan dan peringatan keliru dapat menimbulkan dampak operasional yang berbeda. Oleh sebab itu, satu nilai kinerja saja tidak selalu cukup untuk mendukung keputusan. Pengukuran khusus, seperti confusion matrix untuk klasifikasi dan MAE atau RMSE untuk regresi, akan dibahas lebih lanjut pada Bab 6 dan Bab 7. Pemisahan data evaluasi

juga harus dirancang agar kinerja yang diukur memberikan gambaran yang wajar tentang penggunaan pada data baru (Unknown, 2011; Kononenko, 2007).

Pada proyek pompa di Surabaya, tim dapat mencatat bagaimana data disiapkan, model kandidat dibentuk, dan hasil diperiksa berdasarkan kriteria yang disepakati dengan pengelola fasilitas. Tabel berikut menyajikan contoh catatan evaluasi kualitatif, bukan hasil eksperimen atau klaim kinerja suatu model tertentu.

Tabel 3.4 Catatan Evaluasi Ilustratif untuk Prediksi Gangguan Pompa

Aspek	Pertanyaan	Tindak lanjut
Kesesuaian data	Apakah waktu sensor konsisten?	Periksa urutan pencatatan
Kesesuaian fitur	Apakah fitur tersedia sebelum prediksi?	Singkirkan informasi masa depan
Kinerja teknis	Apakah kesalahan sesuai batas proyek?	Tinjau hasil pada data evaluasi
Dampak kesalahan	Apa akibat alarm keliru?	Sepakati ambang tindak lanjut
Kegunaan operasional	Siapa menanggapi hasil?	Tetapkan prosedur pemeriksaan
Reproduksibilitas	Dapatkah eksperimen diulang?	Simpan catatan proses

Catatan tersebut menghubungkan kualitas teknis dengan kondisi penggunaan. Model sebaiknya belum diteruskan ke penerapan apabila data evaluasi tidak

mencerminkan kondisi operasional atau tindak lanjut atas prediksi belum ditetapkan.

Evaluasi dapat menghasilkan tiga keputusan praktis: melanjutkan model ke tahap deployment, memperbaiki bagian tertentu, atau menghentikan pendekatan karena tidak sesuai kebutuhan. Jika hasil belum memenuhi kriteria, tim dapat kembali meninjau data, fitur, atau rumusan masalah sebelum mencoba algoritma lain. Praktik ini mencegah pemilihan model hanya karena satu metrik terlihat baik, tanpa mempertimbangkan mutu data dan kegunaan hasil (Unknown, 2019; Dua, 2016).

G. Deployment dan Siklus Hidup Model Machine Learning

Deployment adalah tahap ketika hasil proyek disiapkan untuk digunakan dalam lingkungan operasional. Bentuknya dapat berupa laporan analitik, dasbor, layanan prediksi, atau integrasi model dengan sistem yang telah digunakan organisasi. Bentuk penerapan harus sesuai dengan keputusan yang hendak didukung serta kemampuan pengguna dalam menindaklanjuti hasil. Model yang berkinerja baik pada pengujian belum tentu bermanfaat apabila hasilnya tidak sampai kepada pihak yang tepat atau tidak terhubung dengan proses kerja (Unknown, 2011; Kononenko, 2007).

Sebelum penerapan, tim perlu menyepakati tanggung jawab operasional: siapa yang menerima keluaran model, kapan hasil diperiksa, dan tindakan apa yang boleh dilakukan berdasarkan prediksi. Untuk sistem sensor IoT di sebuah pabrik di Cikarang, model yang memperkirakan kebutuhan pemeriksaan mesin dapat menyampaikan peringatan kepada petugas pemeliharaan. Petugas tetap perlu menilai kondisi mesin melalui prosedur yang berlaku, khususnya ketika konsekuensi prediksi salah dapat mengganggu keselamatan atau operasi. Pengembangan sistem analitik untuk konteks sensor perlu mempertimbangkan keterhubungan model dengan perangkat dan lingkungan operasional, bukan hanya ketepatan prediksinya (Agung Suburdjati, 2026).

Siklus hidup model berlanjut setelah penerapan awal. Tim memantau apakah model masih menerima data dengan struktur yang sesuai, apakah keluaran dihasilkan pada waktu yang diperlukan, dan apakah kinerja tetap memenuhi kriteria. Kondisi bisnis dapat berubah; sensor dapat diganti; pengguna dapat mengubah cara mereka berinteraksi dengan sistem. Perubahan semacam itu berpotensi mengurangi relevansi model. Pemantauan dapat mengungkap gejala yang perlu diperiksa, tetapi penetapan bahwa model harus diperbarui tetap memerlukan analisis terhadap konteks dan bukti operasional (Unknown, 2019; Unknown, 2026).

Pemeliharaan model dapat mencakup pemeriksaan data masukan, peninjauan kesalahan, pembaruan komponen perangkat lunak, dan pelatihan ulang ketika data baru yang relevan tersedia. Pelatihan ulang bukan tindakan otomatis yang selalu memperbaiki model; data baru juga perlu diperiksa kualitas serta keterwakilannya sebelum digunakan. Tim sebaiknya menyimpan versi model dan catatan perubahan agar hasil operasional dapat dikaitkan dengan model yang menghasilkan prediksi. Dokumentasi membantu penelusuran ketika terjadi gangguan dan mendukung komunikasi antara pengembang, pemilik proses, serta pengelola sistem (Dua, 2016; Solimun, 2023).

Dalam organisasi Indonesia, siklus tersebut dapat diterapkan pada sistem pemantauan perangkat di pabrik atau gedung perkantoran. Pengelola dapat menetapkan jadwal untuk memeriksa perubahan pola sensor, mencatat tindak lanjut atas peringatan, serta meninjau kembali kriteria penggunaan model. Tabel berikut merangkum contoh pengelolaan siklus hidup pada sistem pemantauan mesin di Cikarang.

Tabel 3.5 Pemantauan Siklus Hidup Model di Lingkungan Operasional

Waktu	Pemeriksaan	Penanggung jawab	Tindakan
Sebelum rilis	Kesesuaian keluaran dan alur kerja	Tim proyek	Uji alur penggunaan

Saat berjalan	Kelengkapan data masukan	Pengelola sistem	Periksa sumber data
Berkala	Pola kesalahan dan umpan balik	Analis dan pengguna	Tinjau kasus bermasalah
Saat ada perubahan	Struktur data atau perangkat	Pengelola teknis	Nilai dampak perubahan
Sebelum pembaruan	Mutu data terbaru	Tim data	Validasi data calon pelatihan
Setelah pembaruan	Hasil model dan catatan versi	Tim proyek	Dokumentasikan rilis

Pemeriksaan pada tabel memberi organisasi dasar untuk menelusuri masalah dan menentukan tindak lanjut. Jadwal serta penanggung jawab perlu disesuaikan dengan tingkat risiko dan kecepatan perubahan pada lingkungan penerapan.

Keputusan untuk memperbarui, mengganti, atau menghentikan model sebaiknya mempertimbangkan bukti kinerja dan kebutuhan organisasi. Apabila proses bisnis berubah sehingga keluaran tidak lagi mendukung keputusan yang sama, memperbaiki model saja mungkin tidak memadai; tujuan proyek dapat perlu dirumuskan ulang. Siklus hidup ini menegaskan bahwa pemeliharaan model mencakup aspek teknis, proses kerja, dan tata kelola, bukan sekadar pelatihan ulang algoritma (Unknown, 2019; Dua, 2016).

H. Rangkuman

Penemuan pengetahuan dari data memerlukan proses yang menghubungkan tujuan analisis, mutu data, teknik komputasional, dan konteks penggunaan. Kerangka Knowledge Discovery in Databases (KDD) mencakup pemilihan data, prapemrosesan, transformasi, penambangan data, interpretasi dan evaluasi, serta pemanfaatan hasil. Penambangan data hanyalah satu bagian dari rangkaian tersebut: algoritma dapat membentuk model atau menemukan pola, tetapi pola belum menjadi pengetahuan yang berguna sebelum diperiksa validitasnya, ditafsirkan sesuai konteks, dan dihubungkan dengan tindakan. Setiap tahap menghasilkan keluaran yang mendukung tahap berikutnya, tetapi evaluasi dapat mengungkap masalah yang mengharuskan tim meninjau kembali keputusan sebelumnya. Karena itu, alur KDD bersifat iteratif, bukan prosedur linear yang dijalankan sekali.

CRISP-DM menyediakan struktur untuk mengelola proyek data mining melalui enam fase: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Fase-fase ini mengaitkan kebutuhan organisasi dengan pengembangan dan penerapan model. Jika model tidak menjawab kebutuhan operasional atau data yang tersedia tidak mendukung analisis, tim dapat kembali merumuskan tujuan, meninjau data, atau

memperbaiki persiapan. Artefak seperti pernyataan masalah, inventaris sumber data, catatan transformasi, dokumentasi eksperimen, dan rencana penerapan membuat keputusan proyek dapat ditinjau oleh analis, pemilik proses, serta pengelola sistem.

Dalam praktik Sistem Informasi, keberhasilan tidak cukup dinilai dari kinerja algoritma. Tim perlu memastikan data mewakili masalah, fitur sesuai kebutuhan, hasil cukup stabil dan dapat ditafsirkan, serta rekomendasi mungkin dijalankan. Ilustrasi koperasi menunjukkan hubungan antara transaksi terpilih, pengelompokan pola belanja, pemaknaan kelompok, dan tindak lanjut operasional. Ilustrasi prediksi CO_2 ruang kelas menegaskan perlunya menetapkan pengguna dan keputusan yang didukung sebelum menilai data sensor dan membandingkan model. Dokumentasi alasan pemilihan data, batasan analisis, dan keputusan transformasi memperkuat akuntabilitas serta membantu mencegah tindakan organisasi bertumpu pada interpretasi yang keliru.

I. Latihan Mahasiswa

a. Essay

1. Analisis perbedaan cakupan KDD dan penambangan data melalui contoh pengelompokan pelanggan koperasi. Jelaskan mengapa keluaran algoritma belum dapat langsung dianggap sebagai pengetahuan, serta

tentukan pemeriksaan dan penafsiran yang diperlukan sebelum hasil digunakan untuk menyusun tindakan.

2. Evaluasi sifat iteratif CRISP-DM dengan menggunakan proyek prediksi kadar CO_2 ruang kelas. Uraikan kondisi yang dapat mendorong tim kembali dari evaluasi ke pemahaman bisnis atau persiapan data, lalu jelaskan bagaimana perubahan tersebut menjaga kesesuaian model dengan keputusan operasional.
3. Rumuskan artefak minimum untuk setiap fase CRISP-DM dalam proyek analisis transaksi koperasi. Hubungkan setiap artefak dengan keputusan yang perlu ditinjau oleh analis, pemilik proses, atau pengelola sistem agar alur kerja dapat dipertanggungjawabkan.
4. Bandingkan penilaian hasil data mining berdasarkan kinerja statistik dengan penilaian berdasarkan kegunaan organisasi. Terapkan perbandingan tersebut pada hasil pengelompokan pelanggan yang tampak stabil, tetapi sulit ditafsirkan, dan tentukan langkah evaluasi yang seharusnya dilakukan.
5. Argumentasikan peran dokumentasi keputusan pemilihan data, transformasi, dan interpretasi dalam siklus data mining. Jelaskan bagaimana dokumentasi tersebut membantu menemukan sumber kesalahan, menilai keterbatasan hasil, dan mencegah rekomendasi yang tidak sesuai konteks organisasi.

b. Pilihan Berganda

1. Sebuah algoritma menghasilkan tiga kelompok pelanggan koperasi dari data transaksi. Pernyataan yang paling tepat mengenai keluaran tersebut adalah...
    a. Keluaran itu sudah menjadi pengetahuan karena algoritma telah membentuk kelompok.
    b. Keluaran itu merupakan hasil penambangan data yang masih perlu dievaluasi dan ditafsirkan dalam konteks koperasi.
    c. Keluaran itu membuktikan bahwa data yang dipilih telah mewakili seluruh anggota.
    d. Keluaran itu dapat langsung digunakan sebagai rekomendasi tanpa pemeriksaan lanjutan.
    e. Keluaran itu menandakan bahwa seluruh tahapan KDD telah berlangsung secara linear.
2. Tim menemukan banyak catatan transaksi ganda sebelum membentuk fitur analisis. Aktivitas tersebut terutama termasuk tahap...
    a. Pemilihan data
    b. Prapemrosesan
    c. Transformasi
    d. Pemodelan
    e. Deployment
3. Dalam analisis transaksi, tim menghitung ringkasan belanja setiap anggota agar dapat digunakan untuk pengelompokan. Aktivitas ini paling tepat digolongkan sebagai...
    a. Pemilihan data, karena tim menetapkan anggota yang dianalisis

b. Prapemrosesan, karena tim menghapus ketidakkonsistenan data
c. Transformasi, karena tim membentuk representasi atau fitur untuk analisis
d. Interpretasi, karena tim memberi makna pada kelompok pelanggan
e. Evaluasi, karena tim membandingkan pola dengan kriteria keberhasilan

4. Setelah pola pengelompokan terbentuk, tim menilai bahwa kelompok-kelompok tersebut sulit dimaknai. Tindakan yang paling sesuai dengan sifat iteratif KDD adalah...

   a. Menggunakan hasil tersebut karena algoritma telah selesai dijalankan
   b. Menghapus seluruh data dan menghentikan proyek
   c. Meninjau kembali pemilihan atau transformasi data dan mengulangi analisis bila diperlukan
   d. Memindahkan hasil langsung ke deployment agar segera diuji pengguna
   e. Mengganti nama kelompok tanpa memeriksa karakteristik anggotanya

5. Manakah kondisi yang paling mendukung penilaian bahwa temuan KDD berguna bagi organisasi?

   a. Temuan berasal dari algoritma dengan kompleksitas paling tinggi.
   b. Pola memiliki jumlah kelompok yang paling banyak.

c. Hasil menjawab pertanyaan awal, dapat ditafsirkan, dan relevan untuk kemungkinan tindakan.
d. Model selesai dibangun dalam waktu paling singkat.
e. Seluruh fitur yang tersedia digunakan tanpa seleksi.

6. Pada proyek prediksi CO_2 ruang kelas, tim mendiskusikan siapa pengguna hasil prediksi dan keputusan ventilasi yang akan didukung. Diskusi tersebut terutama merupakan bagian dari fase...
   a. Pemahaman bisnis
   b. Pemahaman data
   c. Persiapan data
   d. Pemodelan
   e. Deployment

7. Inventaris sumber sensor dan catatan mengenai kelengkapan serta konsistensi datanya paling sesuai menjadi artefak fase...
   a. Pemahaman bisnis
   b. Pemahaman data
   c. Persiapan data
   d. Evaluasi
   e. Deployment

8. Tim membandingkan kinerja model dengan kriteria keberhasilan yang telah disepakati sebelum

eksperimen. Kegiatan ini terutama dilakukan dalam fase...

a. Pemahaman bisnis
b. Pemahaman data
c. Persiapan data
d. Evaluasi
e. Deployment

9. Setelah model dinilai layak, tim merencanakan cara menyajikan hasil prediksi kepada pengelola fasilitas kelas. Rencana tersebut terutama termasuk fase...

a. Pemahaman bisnis
b. Pemahaman data
c. Pemodelan
d. Evaluasi
e. Deployment

10. Model memiliki kinerja statistik yang baik, tetapi hasilnya tidak mendukung keputusan operasional yang ditetapkan pada awal proyek. Respons yang paling tepat adalah...

a. Menetapkan model sebagai hasil akhir karena kinerja statistiknya baik.
b. Menghapus catatan kriteria keberhasilan agar model dapat diterima.
c. Meninjau kesesuaian tujuan bisnis dan mempertimbangkan perbaikan pada fase terkait.
d. Menambah fitur secara acak tanpa memeriksa kebutuhan pengguna.

e. Menerapkan model terlebih dahulu dan menunda evaluasi kegunaannya.

c. Soal Project / Studi Kasus

1. Sebuah koperasi di Yogyakarta ingin mengelompokkan anggota berdasarkan transaksi selama dua tahun untuk merancang program promosi. Data berasal dari beberapa sistem kasir; sebagian transaksi tercatat ganda, format identitas anggota tidak konsisten, dan sejumlah anggota aktif hanya tercakup pada sebagian periode. Setelah dibuat, hasil pengelompokan menunjukkan satu kelompok dengan belanja tinggi, tetapi pengurus tidak memahami ciri kelompok lain. Pengurus meminta rekomendasi promosi segera, sementara analis belum mendokumentasikan alasan memilih rentang waktu dan fitur ringkasan belanja. Analisis masalah proyek ini menggunakan tahapan KDD. Tentukan pemeriksaan yang perlu dilakukan, bagian proses yang perlu ditinjau ulang, dan syarat sebelum rekomendasi diserahkan kepada pengurus.
2. Sebuah sekolah di Bandung menguji sistem prediksi kadar CO_2 ruang kelas menggunakan sensor suhu, kelembapan, dan CO_2. Model menghasilkan prediksi cukup akurat pada data uji, tetapi beberapa sensor tidak mengirim data pada jam istirahat. Pengelola fasilitas ingin memakai prediksi untuk mengatur ventilasi otomatis, sedangkan guru memerlukan penjelasan sederhana tentang kondisi yang memicu perubahan ventilasi. Tim belum menetapkan batas operasional yang aman atau cara menangani data sensor yang hilang, dan belum mencatat bagaimana

keberhasilan sistem akan dinilai setelah diterapkan. Evaluasi proyek tersebut dengan kerangka CRISP-DM. Rumuskan langkah perbaikan, artefak yang perlu dibuat, dan pertimbangan sebelum deployment.

BAB 4
DATA PREPROCESSING DAN FEATURE ENGINEERING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan teknik pembersihan data dan penanganan nilai hilang sesuai karakteristik data serta tujuan analisis.
2. Memilih dan menerapkan metode untuk menangani noise, outlier, duplikasi, dan inkonsistensi data.
3. Menghitung serta menerapkan normalisasi, standardisasi, encoding, dan transformasi data sesuai kebutuhan algoritma analisis.
4. Menggunakan teknik seleksi fitur, ekstraksi fitur, dan rekayasa fitur untuk menyiapkan data bagi pemodelan.
5. Mengimplementasikan tahapan preprocessing pada data sensor IoT dan log keamanan sistem dengan mempertimbangkan struktur serta konteks sumber datanya.

B. Pendahuluan

Data yang dikumpulkan organisasi jarang berada dalam kondisi siap dianalisis. Catatan transaksi dapat mengandung nilai yang hilang atau format tanggal yang tidak seragam, data pelanggan dapat tercatat lebih dari satu kali,

sedangkan sensor dan sistem keamanan dapat menghasilkan pengukuran menyimpang maupun log dengan struktur yang berbeda-beda. Jika masalah tersebut tidak ditangani, pola yang ditemukan melalui data mining berisiko merepresentasikan kesalahan pencatatan, bukan kondisi yang sebenarnya. Han, Kamber, dan Pei (2012) menempatkan pembersihan, integrasi, transformasi, dan reduksi data sebagai bagian mendasar dari proses penemuan pengetahuan karena kualitas data memengaruhi keandalan hasil analisis. Bagi mahasiswa Sistem Informasi, kemampuan menyiapkan data menjadi kompetensi penting untuk menghubungkan kebutuhan organisasi dengan proses analitik yang dapat dipertanggungjawabkan. Kompetensi tersebut tidak hanya mencakup penggunaan perangkat lunak, tetapi juga penilaian atas makna data, konsekuensi keputusan preprocessing, dan kesesuaiannya dengan tujuan analisis.

Preprocessing mencakup serangkaian keputusan yang saling berkaitan. Nilai hilang, misalnya, dapat dihapus atau diimputasi, tetapi pilihan tersebut perlu mempertimbangkan pola kehilangan dan dampaknya terhadap analisis; penggantian nilai secara sembarang dapat mengubah distribusi data (Little & Rubin, 2019). Demikian pula, outlier tidak selalu merupakan kesalahan: nilai ekstrem dapat menandai gangguan pencatatan, tetapi juga dapat menunjukkan kejadian penting yang perlu

diselidiki. Duplikasi dan inkonsistensi perlu ditangani dengan aturan yang mempertahankan identitas serta konteks rekaman. Setelah kualitas data diperiksa, transformasi seperti normalisasi, standardisasi, encoding, dan rekayasa fitur membantu menyelaraskan representasi data dengan kebutuhan metode analisis. Perbedaan skala antarf fitur, misalnya, dapat memengaruhi algoritma yang menggunakan jarak atau optimasi berbasis gradien, sehingga transformasi perlu dipilih berdasarkan karakteristik algoritma dan data (Kuhn & Johnson, 2013). Dengan demikian, preprocessing bukan prosedur mekanis yang diterapkan seragam, melainkan penerapan metode yang didasari pemahaman atas sumber dan tujuan penggunaan data.

Tantangan tersebut semakin nyata pada data sensor IoT dan log keamanan sistem. Data sensor dapat mengalir dalam frekuensi tinggi, memiliki jeda pengukuran, atau menunjukkan anomali akibat gangguan perangkat, sedangkan log keamanan memuat rekaman peristiwa yang beragam dan perlu dipertahankan konteks waktunya. Pedoman pengelolaan log menekankan pentingnya proses pengumpulan, penyimpanan, dan peninjauan yang konsisten agar rekaman dapat mendukung pemantauan serta investigasi (Kent & Souppaya, 2006). Karena itu, mahasiswa perlu menerapkan teknik preprocessing dengan memperhatikan karakteristik domain, bukan sekadar

mengejar data yang tampak seragam. Bab ini membahas data cleaning dan penanganan missing values; noise, outlier, duplikasi, dan inkonsistensi data; normalisasi, standardisasi, encoding, dan transformasi; feature selection, feature extraction, dan feature engineering; serta preprocessing data sensor IoT dan log keamanan sistem.

C. Data Cleaning dan Penanganan Missing Values

Data cleaning merupakan bagian prapemrosesan data yang memperbaiki atau menandai nilai tidak valid, format yang tidak seragam, dan informasi yang tidak lengkap. Tujuannya bukan sekadar menghasilkan kumpulan data tanpa sel kosong, melainkan menjaga agar setiap nilai yang digunakan tetap memiliki makna sesuai konteks pencatatannya. Karena itu, pembersihan perlu diawali dengan pemeriksaan tipe data, rentang nilai, aturan bisnis, dan hubungan antaratribut. Prinsip tersebut menempatkan keputusan cleaning sebagai bagian dari persiapan data yang harus disesuaikan dengan tujuan analisis, bukan sebagai tindakan otomatis yang berlaku sama pada semua kumpulan data (García, 2015; Luengo, 2020).

Nilai hilang dapat muncul karena alat pencatat gagal bekerja, responden tidak memberikan jawaban, atau suatu atribut memang tidak berlaku bagi objek tertentu. Pola dan penyebab kehilangan memengaruhi cara penanganannya.

Jika suatu kolom hampir seluruhnya kosong atau tidak dibutuhkan untuk analisis, kolom tersebut dapat dipertimbangkan untuk dihapus. Jika hanya sedikit baris yang kehilangan nilai dan penghapusannya tidak menimbulkan bias, penghapusan baris dapat menjadi pilihan. Namun, menghapus semua baris yang memiliki satu nilai kosong berisiko membuang banyak informasi, terutama ketika beberapa atribut memiliki kehilangan pada baris yang berbeda (Huffaker, 2018; García, 2015).

Imputasi mengganti nilai hilang dengan estimasi tertentu. Untuk atribut numerik, nilai rata-rata dapat digunakan ketika distribusi relatif simetris dan tidak banyak dipengaruhi nilai ekstrem, sedangkan median lebih tahan terhadap kemencengan distribusi. Atribut kategorikal dapat diisi dengan modus atau kategori khusus seperti “tidak diketahui”, apabila kategori tersebut memang berguna sebagai penanda. Pada data berurutan menurut waktu, imputasi dapat menggunakan nilai sebelum atau sesudahnya hanya jika asumsi kesinambungan pengukuran masuk akal. Pengisian dengan satu nilai tetap memang praktis, tetapi dapat mengurangi variasi dan mengaburkan pola asli; pilihan metode sebaiknya dicatat agar dampaknya dapat dievaluasi (García, 2015; Luengo, 2020).

Sebagai contoh, sebuah puskesmas di Yogyakarta menggabungkan data kunjungan pasien dari beberapa unit. Kolom tekanan darah yang kosong tidak semestinya

otomatis diisi dengan rata-rata seluruh pasien karena nilai tersebut bergantung pada konteks pemeriksaan dan kondisi pasien. Jika atribut itu menjadi sasaran analisis, tim dapat memeriksa apakah kekosongan terkonsentrasi pada unit atau periode tertentu, lalu memilih imputasi yang sesuai atau mempertahankan penanda kehilangan. Sebaliknya, kolom nomor telepon yang tidak diperlukan untuk analisis kunjungan dapat dikeluarkan dari kumpulan data analitik setelah kebutuhan penggunaan data dipastikan.

Urutan kerja yang konsisten membantu mencegah keputusan cleaning yang tidak terlacak. Tabel berikut merangkum contoh keputusan untuk beberapa kondisi data; nilai dan situasinya bersifat ilustratif.

Tabel 4.1 Contoh Penanganan Nilai Hilang

Atribut	Kondisi	Pilihan awal	Pemeriksaan
Suhu	Satu pembacaan kosong	Tandai atau interpolasi	Jarak waktu
Pendapatan	Beberapa nilai kosong	Median per kelompok	Distribusi kelompok
Jenis layanan	Kategori kosong	“Tidak diketahui”	Makna kategori
Nomor telepon	Banyak kosong	Keluarkan dari analisis	Relevansi fitur
Tekanan darah	Kosong pada unit tertentu	Selidiki sumber	Pola kehilangan

Tabel tersebut menunjukkan bahwa metode yang dipilih bergantung pada arti atribut dan pola kehilangan, bukan semata-mata jenis datanya. Sebelum menerapkan imputasi,

mahasiswa perlu membandingkan distribusi sebelum dan sesudah pengisian serta mencatat jumlah nilai yang diubah (Huffaker, 2018).

D. Noise, Outlier, Duplikasi, dan Inkonsistensi Data

Noise adalah variasi atau kesalahan pada data yang tidak merepresentasikan sinyal yang hendak dianalisis. Pada data numerik, noise dapat timbul dari gangguan alat, pembulatan, kesalahan input, atau perubahan kondisi pencatatan. Outlier adalah pengamatan yang tampak jauh dari sebagian besar data menurut ukuran tertentu. Keduanya tidak selalu sama: noise biasanya mencerminkan gangguan pengukuran atau pencatatan, sedangkan outlier hanya menyatakan posisi relatif suatu nilai. Nilai ekstrem dapat berasal dari kesalahan, tetapi bisa pula merekam kejadian nyata yang penting. Oleh karena itu, keputusan menghapus nilai ekstrem perlu didahului pemeriksaan konteks dan sumbernya (García, 2015; Blázquez-García, 2022).

Pemeriksaan awal dapat menggunakan ringkasan statistik, visualisasi distribusi, serta batas yang ditetapkan berdasarkan aturan domain. Untuk data numerik, rentang antarkuartil dapat membantu menandai nilai yang jauh dari bagian tengah distribusi, tetapi hasil penandaan tersebut bukan bukti bahwa nilai harus dibuang. Pada data

berurutan, lonjakan perlu diperiksa bersama waktu pengamatan dan nilai di sekitarnya karena perubahan tajam mungkin menandakan kerusakan sensor, tetapi mungkin pula merupakan peristiwa yang sah. Pendekatan terhadap deret waktu perlu mempertahankan urutan serta konteks temporal agar pembersihan tidak menghilangkan pola yang justru hendak ditemukan (Blázquez-García, 2022).

Duplikasi terjadi ketika satu kejadian atau entitas tercatat lebih dari sekali. Duplikasi dapat bersifat identik, misalnya dua baris dengan seluruh atribut sama, atau mendekati identik akibat variasi ejaan dan format. Penghapusan membutuhkan aturan identitas: transaksi yang memiliki nomor referensi sama mungkin merupakan salinan, tetapi dua transaksi dengan pelanggan dan waktu berdekatan belum tentu duplikat. Untuk data pelanggan, misalnya, nama yang berbeda tipis dapat merujuk pada orang yang sama atau dua orang berbeda. Aturan pencocokan sebaiknya menggunakan beberapa atribut, misalnya nomor identitas internal, alamat, dan tanggal transaksi, serta menyediakan pemeriksaan manual bagi pasangan yang meragukan (García, 2015; Luengo, 2020).

Inkonsistensi muncul ketika nilai yang semestinya memiliki arti sama disimpan dalam bentuk berbeda, atau ketika nilai melanggar aturan hubungan antarkolom. Contohnya, nama kabupaten dapat tercatat dengan singkatan dan ejaan berbeda; tanggal dapat memakai

format berbeda; atau jumlah barang bernilai negatif pada catatan penjualan yang tidak mengizinkan retur. Standardisasi format harus mengikuti aturan sumber data, bukan mengubah nilai secara membabi buta. Retur, pembatalan, dan koreksi perlu dibedakan dari kesalahan agar proses pembersihan tidak merusak makna transaksi (García, 2015).

Pada data transaksi UMKM batik di Pekalongan, satu pesanan yang masuk melalui aplikasi dan dicatat kembali oleh staf dapat tampak sebagai duplikasi. Tim perlu mencocokkan identitas pesanan, waktu, jumlah, dan status pembayaran sebelum menghapus salah satunya. Jika catatan menunjukkan transaksi sah yang dibatalkan lalu dibuat ulang, kedua rekaman dapat dipertahankan dengan status yang benar. Untuk pencilan harga, pemeriksaan katalog produk dan satuan transaksi lebih tepat daripada langsung memangkas nilai mahal, sebab kain batik dengan teknik khusus memang dapat memiliki harga jauh di atas produk reguler. Tindakan tersebut sejalan dengan prapemrosesan yang mempertahankan informasi relevan sambil memperbaiki kesalahan yang terverifikasi (Luengo, 2020).

E. Normalisasi, Standardisasi, Encoding, dan Transformasi

Setelah kualitas data diperiksa, atribut perlu disajikan dalam bentuk yang sesuai dengan algoritma dan tujuan analisis. Normalisasi dan standardisasi mengubah skala atribut numerik, sedangkan encoding mengubah kategori menjadi representasi yang dapat diproses. Transformasi juga dapat mencakup perubahan satuan, bentuk distribusi, atau struktur data. Langkah-langkah tersebut tidak membuat data menjadi "lebih benar" secara otomatis; setiap perubahan harus mempertahankan makna yang diperlukan dan diterapkan secara konsisten pada data yang kelak diproses oleh model (García, 2015; Luengo, 2020).

Normalisasi min–maks memetakan nilai ke rentang tertentu, biasanya dari 0 sampai 1. Untuk nilai x, batas bawah $x_{\min}$, dan batas atas $x_{\max}$, bentuk umumnya adalah:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Jika jarak antara nilai minimum dan maksimum kecil, perubahan ini dapat membantu menyetarakan skala atribut. Namun, nilai ekstrem dapat memperlebar rentang dan memampatkan sebagian besar pengamatan. Standardisasi mengubah data berdasarkan rata-rata μ dan simpangan baku σ:

$$z = \frac{x - \mu}{\sigma}$$

Hasil standardisasi memiliki pusat pada rata-rata nol dan skala berdasarkan simpangan baku, tetapi tidak membatasi nilai pada rentang tetap. Pemilihan di antara keduanya perlu memperhatikan algoritma yang digunakan, sebaran data, dan keberadaan nilai ekstrem (García, 2015; Luengo, 2020).

Encoding menerjemahkan kategori menjadi bentuk numerik. Pada kategori tanpa urutan, seperti jenis layanan, one-hot encoding dapat membentuk satu indikator untuk setiap kategori sehingga tidak menyiratkan adanya tingkatan. Pada kategori berurutan, seperti tingkat kepuasan rendah, sedang, dan tinggi, encoding ordinal dapat digunakan jika urutan itu bermakna. Memberi kode 1, 2, dan 3 pada kategori nominal tanpa alasan yang kuat berisiko membuat algoritma membaca hubungan urutan atau jarak yang sebenarnya tidak ada. Kategori langka juga perlu ditangani secara cermat agar tidak menyulitkan penerapan transformasi pada data baru (García, 2015).

Transformasi dapat digunakan untuk mengurangi kemencengan distribusi atau membentuk atribut yang lebih mudah ditafsirkan. Misalnya, jumlah transaksi yang sangat bervariasi antarpelanggan dapat diubah dengan fungsi logaritmik setelah memastikan nilai memenuhi syarat

transformasi. Satuan juga dapat diseragamkan: suhu yang tercatat dalam Celsius dan Fahrenheit tidak boleh digabung sebagai angka seolah-olah berada pada skala yang sama. Pada data UMKM kopi di Bandung, harga per kilogram, jumlah pesanan, dan jenis produk perlu ditangani berbeda: atribut numerik dapat diskalakan sesuai kebutuhan, sedangkan jenis produk dikodekan sebagai kategori.

Proses transformasi perlu menyimpan parameter yang diperoleh dari data pelatihan, seperti nilai minimum, maksimum, rata-rata, atau simpangan baku. Parameter itu kemudian digunakan kembali untuk mentransformasi data validasi dan data operasional, bukan dihitung ulang secara terpisah. Praktik ini menjaga konsistensi representasi dan menghindari penggunaan informasi dari data evaluasi ketika menyiapkan data pelatihan. Rincian pembagian data dan konsekuensinya bagi pemodelan akan dibahas lebih lanjut pada Bab 5; pada tahap ini, fokusnya adalah menerapkan transformasi yang sesuai dan mendokumentasikan aturan yang digunakan (Luengo, 2020).

F. Feature Selection, Feature Extraction, dan Feature Engineering

Feature selection, feature extraction, dan feature engineering merupakan tiga cara berbeda untuk menyiapkan fitur bagi pemodelan. Feature selection

memilih sebagian fitur yang sudah tersedia, feature extraction membentuk representasi baru dari sejumlah atribut, sedangkan feature engineering merancang atau mengubah fitur berdasarkan konteks masalah. Ketiganya bertujuan menyediakan masukan yang relevan dan dapat digunakan oleh algoritma, tetapi membutuhkan pertimbangan berbeda mengenai makna atribut, biaya pengumpulan, dan cara hasilnya ditafsirkan (Liu, 2018).

Seleksi fitur dapat dilakukan berdasarkan pengetahuan domain, hubungan statistik dengan sasaran analisis, atau kinerja fitur dalam prosedur pemilihan tertentu. Fitur dengan nilai konstan, banyak nilai hilang, atau informasi yang berulang mungkin tidak menambah manfaat analitis. Namun, hubungan yang kuat dengan sasaran tidak selalu berarti fitur itu layak digunakan: atribut dapat membocorkan informasi yang baru tersedia setelah peristiwa yang hendak diprediksi. Pada contoh prediksi pelanggan berhenti berlangganan, tanggal penutupan akun tidak dapat dipakai sebagai fitur prediksi sebelum keputusan berhenti terjadi. Seleksi juga perlu mempertimbangkan tujuan organisasi; fitur yang mudah diperoleh dan dapat dijelaskan bisa lebih sesuai daripada fitur yang sulit diverifikasi dalam proses bisnis (Liu, 2018).

Ekstraksi fitur merangkum atau memproyeksikan informasi dari sejumlah variabel menjadi representasi yang lebih ringkas. Pada citra, misalnya, tahap ekstraksi dapat

mengubah piksel menjadi ciri visual. Pada data teks, frekuensi kata atau representasi numerik lain dapat digunakan untuk merangkum dokumen; penerapan klasifikasi teks pada percakapan media sosial menunjukkan perlunya mengubah teks mentah menjadi bentuk yang dapat diproses lebih lanjut (Yuyun, 2021). Pilihan representasi berpengaruh terhadap informasi yang dipertahankan. Karena itu, hasil ekstraksi harus ditinjau berdasarkan tujuan analisis, bukan hanya berdasarkan jumlah fitur yang berhasil dikurangi (Liu, 2018).

Feature engineering merancang fitur yang menggambarkan konsep relevan dengan masalah. Dalam analisis pelanggan, tanggal transaksi dapat diolah menjadi jarak waktu sejak pembelian terakhir, jumlah transaksi, dan nilai belanja. Contoh pengelompokan pelanggan berdasarkan recency, frequency, dan monetary pada konteks perusahaan transportasi memperlihatkan bagaimana atribut transaksi diringkas menjadi ukuran yang berguna bagi segmentasi (Priandini, 2025). Dalam konteks toko batik di Yogyakarta, fitur “hari sejak pembelian terakhir” dapat membantu menggambarkan kebaruan hubungan pelanggan, sedangkan jumlah transaksi dalam tiga bulan terakhir dapat mencerminkan frekuensi. Periode dan definisi setiap fitur harus ditetapkan sesuai kebutuhan organisasi, bukan dipilih tanpa alasan.

Penerapan dapat dilakukan melalui urutan kerja yang terdokumentasi. Pertama, tentukan sasaran analisis dan batas waktu pengamatan. Kedua, periksa relevansi, kelengkapan, dan potensi kebocoran setiap atribut. Ketiga, pilih fitur yang tersedia secara sah pada saat keputusan harus dibuat. Keempat, bentuk fitur baru dengan aturan yang dapat diulang pada data baru. Kelima, evaluasi apakah fitur yang dihasilkan masih dapat ditafsirkan oleh pengguna. Prosedur tersebut membantu menjaga keselarasan antara representasi data dan keputusan yang didukung oleh model (Liu, 2018).

Di lingkungan Sistem Informasi, kualitas rekayasa fitur juga berkaitan dengan tata kelola. Nama fitur, satuan, sumber, waktu pembentukan, dan aturan perhitungannya perlu dicatat supaya hasil analisis dapat ditinjau dan diperbarui. Pengurangan fitur bukan tujuan tersendiri; fitur yang dibuang mungkin justru membawa informasi penting untuk kelompok pengguna tertentu. Karena itu, perubahan daftar fitur sebaiknya diuji pada kasus penggunaan yang dituju dan didokumentasikan bersama alasan pemilihannya (Liu, 2018).

G. Preprocessing Data Sensor IoT dan Log Keamanan Sistem

Data sensor Internet untuk Segala (IoT) dan log keamanan memiliki karakteristik yang menuntut

prapemrosesan kontekstual. Pembacaan sensor biasanya terikat pada perangkat, waktu, satuan, dan kondisi lingkungan, sedangkan log memuat peristiwa yang dicatat oleh beragam aplikasi atau perangkat. Sebelum menggabungkan sumber, setiap rekaman perlu memiliki identitas sumber dan penanda waktu yang dapat ditafsirkan. Penyelarasan waktu dan format bukan sekadar perapian kolom; kesalahan pada keduanya dapat membuat urutan kejadian keliru ketika data dianalisis (Luengo, 2020; Amanda, 2025).

Untuk data sensor, pemeriksaan awal mencakup rentang nilai yang masuk akal, interval pengukuran, nilai hilang, dan perubahan status perangkat. Pembacaan yang tidak wajar sebaiknya ditandai terlebih dahulu, bukan langsung dihapus, agar dapat dibandingkan dengan catatan perangkat atau sensor lain. Jika beberapa sensor mengukur besaran yang sama, identitas perangkat dan satuan masing-masing harus dipertahankan. Pada pemantauan kualitas udara di Indonesia, suhu, kelembapan, dan indikator gas dapat berasal dari modul berbeda; pencatatan sensor dan waktu pengukuran diperlukan agar nilai yang digabung benar-benar mewakili kondisi pada saat yang sama (Alfirdaus, 2024).

Pada log keamanan, variasi format dan sumber dapat menyulitkan analisis lintas sistem. Field seperti waktu, alamat sumber, identitas perangkat, jenis peristiwa, dan

hasil autentikasi dapat diseragamkan ke skema bersama, sambil menyimpan rekaman asli untuk penelusuran. Entri berulang tidak selalu duplikasi: sistem dapat mencatat upaya yang sama dari beberapa komponen, sehingga penghapusan perlu mempertimbangkan identitas peristiwa dan tujuan analisis. Konsistensi pengumpulan dan pengolahan membantu menjaga konteks ketika log digunakan untuk mendukung deteksi ancaman pada data bervolume besar (Amanda, 2025).

Tabel berikut memperlihatkan contoh data sensor hipotetis dari pemantauan kualitas udara di Bandung. Nilainya hanya digunakan untuk menunjukkan keputusan prapemrosesan dan bukan hasil pengukuran lapangan.

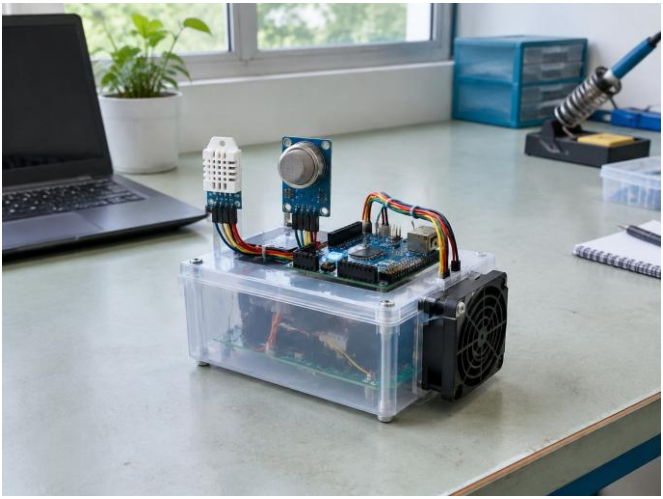
Tabel 4.2 Contoh Prapemrosesan Data Sensor

Waktu	Suhu (°C)	Kelembapan (%)	Gas (ppm)	Keputusan
09.00	29,1	68	410	Pertahankan
09.05	29,3	67	415	Pertahankan
09.10	—	67	418	Tandai suhu hilang
09.15	29,4	66	4.800	Verifikasi lonjakan
09.20	29,5	66	422	Pertahankan

Tabel tersebut memperlihatkan bahwa nilai hilang dan lonjakan memerlukan tindakan berbeda. Suhu kosong perlu ditandai atau ditangani dengan metode yang sesuai, sedangkan lonjakan gas harus diverifikasi terhadap kemampuan alat dan konteks pengukuran sebelum

dikoreksi atau dipertahankan (Alfirdaus, 2024; Blázquez-García, 2022).

Perangkat sensor yang digunakan pada pemantauan lapangan memberi konteks fisik bagi data yang kemudian diproses. Gambar berikut menampilkan contoh perangkat pengukuran kualitas udara yang relevan dengan pencatatan suhu, kelembapan, dan gas.



Gambar 4.1 Perangkat Sensor Pemantauan Kualitas Udara. Sumber: Ilustrasi, diolah penulis

Komponen fisik tersebut membantu menjelaskan mengapa catatan perangkat, satuan, dan waktu perlu dipertahankan bersama hasil pengukuran. Prapemrosesan yang mengabaikan konteks sumber dapat mengubah

pembacaan yang bermakna menjadi angka tanpa rujukan operasional. (Alfirdaus, 2024; Luengo, 2020).

Untuk aliran data yang terus bertambah, aturan pembersihan sebaiknya dapat dijalankan berulang dan menghasilkan jejak perubahan, termasuk nilai yang ditandai, baris yang dikeluarkan, serta alasan tindakannya. Pada data keamanan, akses terhadap rekaman mentah perlu dibatasi sesuai kebutuhan operasional karena log dapat memuat informasi sensitif. Dalam kedua konteks, hasil prapemrosesan perlu mempertahankan waktu dan identitas sumber agar dapat dipakai kembali untuk analisis lanjutan, tanpa menghilangkan informasi yang dibutuhkan untuk pemeriksaan insiden atau evaluasi model (Amanda, 2025; Luengo, 2020).

H. Rangkuman

Keandalan analisis data mining bergantung pada prapemrosesan yang menjaga kualitas sekaligus makna data. Data cleaning mencakup pemeriksaan tipe, rentang, aturan bisnis, serta hubungan antaratribut sebelum nilai diperbaiki atau ditandai. Nilai hilang perlu ditangani berdasarkan penyebab, pola, dan tujuan analisis: penghapusan baris atau kolom dapat dipertimbangkan, sedangkan imputasi dapat menggunakan rata-rata, median, modus, kategori khusus, atau nilai berdekatan pada deret waktu. Setiap pilihan memiliki dampak; pengisian seragam

dapat mengurangi variasi, sementara penghapusan berlebihan dapat membuang informasi atau menimbulkan bias. Karena itu, perubahan perlu dicatat dan distribusi data sebelum serta sesudah penanganan perlu dibandingkan.

Noise, outlier, duplikasi, dan inkonsistensi menuntut pemeriksaan yang berbeda. Noise mencerminkan gangguan pada sinyal, sedangkan outlier hanya menunjukkan posisi nilai yang jauh dari sebagian besar pengamatan; nilai ekstrem mungkin kesalahan, tetapi mungkin juga kejadian penting. Ringkasan statistik, visualisasi, batas domain, dan konteks waktu membantu menilai temuan tanpa menghapus informasi secara terburu-buru. Duplikasi harus ditetapkan melalui aturan identitas dan pencocokan beberapa atribut, bukan sekadar kemiripan. Standardisasi format dan pemeriksaan aturan bisnis juga harus membedakan kesalahan dari kejadian sah, seperti retur atau pembatalan.

Setelah kualitas data ditelaah, normalisasi dan standardisasi dapat menyelaraskan skala numerik, encoding menyajikan kategori dalam bentuk yang dapat diproses, dan transformasi menyesuaikan satuan, distribusi, atau struktur data. Langkah-langkah ini harus selaras dengan algoritma dan tujuan analisis karena tidak otomatis membuat data lebih benar. Dalam praktik Sistem Informasi, keputusan prapemrosesan memengaruhi hasil model, interpretasi, dan keputusan organisasi. Tim perlu menilai konteks sumber,

mempertahankan informasi relevan, mendokumentasikan perubahan, serta mengevaluasi konsekuensinya agar data analitik tetap dapat dipercaya dan digunakan secara bertanggung jawab.

I. Latihan Mahasiswa

a. Essay

1. Analisislah bagaimana pola kehilangan nilai pada data kunjungan pasien dapat memengaruhi pilihan antara penghapusan baris, imputasi, dan penandaan nilai hilang. Jelaskan pemeriksaan yang perlu dilakukan sebelum memilih metode serta risiko terhadap hasil analisis apabila keputusan dibuat tanpa mempertimbangkan konteks atribut.
2. Bandingkan penggunaan rata-rata, median, dan kategori "tidak diketahui" untuk mengisi nilai hilang pada atribut yang berbeda. Evaluasi kondisi distribusi, jenis atribut, dan tujuan analisis yang membuat setiap pilihan layak, termasuk dampak yang mungkin muncul pada variasi data.
3. Evaluasilah cara membedakan noise, outlier yang valid, dan kesalahan pengukuran pada data transaksi atau sensor. Rumuskan rangkaian pemeriksaan yang menggabungkan ringkasan statistik, visualisasi, aturan domain, serta konteks pencatatan sebelum nilai dipertahankan, diperbaiki, atau dikeluarkan.
4. Telaah risiko penghapusan duplikasi ketika data pelanggan atau transaksi memiliki variasi ejaan, waktu berdekatan, dan status pembatalan. Rumuskan aturan

pencocokan yang dapat mengurangi penghapusan keliru serta jelaskan kapan pemeriksaan manual perlu dilibatkan.

5. Argumentasikan bagaimana normalisasi, standardisasi, encoding, dan transformasi dapat dipilih untuk menyiapkan data bagi algoritma tertentu. Sertakan pertimbangan tentang makna atribut dan tujuan analisis yang mencegah perubahan representasi justru mengaburkan informasi.

b. Pilihan Berganda

1. Sebuah atribut tekanan darah memiliki banyak nilai kosong, terutama pada satu unit layanan. Apa langkah awal yang paling tepat sebelum memilih imputasi?
    a. Mengisi semua nilai kosong dengan rata-rata seluruh pasien
    b. Menghapus seluruh baris yang memiliki nilai kosong
    c. Memeriksa pola kehilangan dan proses pencatatan pada unit tersebut
    d. Mengganti semua nilai kosong dengan nol agar konsisten
    e. Menghapus atribut tanpa meninjau kebutuhan analisis
2. Kapan penggunaan median untuk mengimputasi atribut numerik paling dapat dipertanggungjawabkan?
    a. Ketika distribusi sangat menceng dan dipengaruhi nilai ekstrem

b. Ketika atribut kategorikal hanya memiliki dua kategori
c. Ketika setiap nilai kosong menandakan atribut tidak berlaku
d. Ketika data waktu harus selalu diisi dengan nilai berikutnya
e. Ketika seluruh nilai atribut diketahui memiliki distribusi seragam

3. Mengapa nilai yang ditandai sebagai outlier tidak semestinya langsung dihapus?
   a. Karena semua outlier pasti merupakan data yang benar
   b. Karena outlier selalu disebabkan kesalahan alat ukur
   c. Karena metode statistik tidak dapat menemukan nilai ekstrem
   d. Karena nilai ekstrem dapat merepresentasikan kejadian nyata yang penting
   e. Karena penghapusan outlier hanya berlaku untuk data kategorikal
4. Pada data sensor berurutan menurut waktu, lonjakan nilai tiba-tiba perlu diperiksa bersama waktu dan pengamatan di sekitarnya. Apa alasan utamanya?
   a. Agar urutan data dapat diacak sebelum pembersihan
   b. Agar perubahan sah tidak keliru dianggap sebagai gangguan
   c. Agar setiap lonjakan dapat diganti dengan nilai nol

d. Agar nilai setelah lonjakan selalu dihapus
e. Agar pemeriksaan konteks domain dapat dihindari

5. Dua transaksi memiliki pelanggan dan waktu berdekatan, tetapi nomor referensinya berbeda. Tindakan yang paling tepat adalah...
   a. Menghapus salah satunya karena pelanggan sama
   b. Menggabungkan keduanya tanpa memeriksa atribut lain
   c. Mempertahankan keduanya tanpa pemeriksaan apa pun
   d. Membandingkan identitas pesanan, jumlah, waktu, dan status transaksi
   e. Menghapus transaksi dengan jumlah barang lebih kecil

6. Dalam data penjualan, nilai jumlah barang negatif ditemukan. Apa pemeriksaan yang paling sesuai sebelum menandainya sebagai kesalahan?
   a. Memeriksa apakah aturan sumber data mengizinkan retur atau koreksi
   b. Mengubah seluruh nilai negatif menjadi nilai positif
   c. Menghapus semua transaksi dari tanggal yang sama
   d. Mengganti nilai tersebut dengan median jumlah barang
   e. Menghapus atribut jumlah barang dari analisis

7. Apa makna utama standardisasi format kategori, misalnya nama kabupaten yang ditulis dengan variasi ejaan?
    a. Menghapus kategori yang jarang muncul tanpa pemeriksaan
    b. Menyamakan representasi berdasarkan aturan sumber tanpa mengubah makna
    c. Mengubah semua kategori menjadi angka berurutan yang bermakna peringkat
    d. Mengganti kategori berbeda dengan kategori yang paling sering muncul
    e. Menghilangkan seluruh atribut kategorikal dari kumpulan data
8. Pilihan penanganan yang paling sesuai untuk atribut kategorikal dengan nilai hilang, ketika ketiadaan jawaban bermakna bagi analisis, adalah...
    a. Mengisi dengan rata-rata kategori
    b. Mengisi dengan median kategori
    c. Menggunakan kategori “tidak diketahui” setelah menilai maknanya
    d. Mengganti nilai hilang dengan kategori yang paling mahal
    e. Menghapus semua baris yang memiliki kategori tersebut
9. Apa tujuan utama normalisasi atau standardisasi dalam prapemrosesan?
    a. Menyelaraskan skala atribut numerik agar sesuai dengan kebutuhan algoritma

b. Mengubah semua atribut menjadi kategori
c. Menjamin tidak ada lagi nilai hilang dalam data
d. Menghapus seluruh perbedaan antarobservasi
e. Membuktikan bahwa semua nilai dalam data benar

10. Setelah imputasi dilakukan, tindakan evaluasi yang paling relevan adalah...
    a. Menghapus catatan metode agar data terlihat lebih sederhana
    b. Membandingkan distribusi sebelum dan sesudah serta mencatat nilai yang diubah
    c. Mengganti semua nilai yang tidak berubah dengan nilai imputasi
    d. Mengacak urutan data untuk menyamarkan perubahan
    e. Menganggap imputasi selalu meningkatkan kualitas analisis

c. Soal Project / Studi Kasus

1. Sebuah jaringan puskesmas menggabungkan catatan kunjungan dari empat unit untuk memprediksi kebutuhan tindak lanjut pasien. Kolom tekanan darah kosong terutama pada unit tertentu, sementara nomor telepon sering tidak tersedia. Tim menemukan beberapa nilai tekanan darah ekstrem dan perbedaan format tanggal. Sebagian data berasal dari periode ketika alat ukur sempat diperbaiki. Manajer meminta kumpulan data segera disiapkan, tetapi hasil analisis akan digunakan untuk mengatur jadwal layanan.

Analisis pola kehilangan dan sumber pencatatan, lalu rumuskan keputusan untuk setiap atribut. Jelaskan pemeriksaan outlier, perlakuan nomor telepon, standardisasi tanggal, serta dokumentasi dan evaluasi perubahan.

2. UMKM batik menggabungkan data pesanan dari situs web, aplikasi pesan, dan catatan staf. Beberapa pesanan memiliki nama pelanggan yang ejaannya sedikit berbeda, waktu pencatatan berdekatan, dan status pembayaran yang berubah setelah pembatalan. Harga beberapa kain jauh di atas produk lain karena teknik produksinya khusus. Sebelum membuat model untuk memperkirakan stok dan penjualan, tim juga harus mengubah kategori produk serta menyiapkan atribut numerik dengan skala berbeda. Evaluasilah risiko penghapusan duplikasi dan pencilan secara otomatis. Susun aturan pencocokan, pemeriksaan konteks transaksi, serta pilihan encoding dan penyelarasan skala yang sesuai dengan tujuan model.

BAB 5
MACHINE LEARNING UNTUK DATA MINING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep dan paradigma machine learning untuk merumuskan penyelesaian masalah data mining pada konteks Sistem Informasi.
2. Menentukan pendekatan supervised, unsupervised, atau semi-supervised learning yang sesuai dengan ketersediaan label dan tujuan analisis data.
3. Menggunakan pembagian data training, validation, dan testing serta cross-validation untuk mengevaluasi kinerja model secara tepat.
4. Menerapkan langkah diagnosis overfitting dan underfitting untuk menilai kemampuan generalisasi model pada data baru.
5. Memilih algoritma dan menerapkan optimasi model berdasarkan karakteristik data, tujuan analisis, dan hasil evaluasi.

B. Pendahuluan

Organisasi mengumpulkan data transaksi, interaksi pelanggan, operasi layanan, dan aktivitas digital dalam jumlah besar, tetapi akumulasi data tidak dengan sendirinya

menghasilkan keputusan yang lebih baik. Tantangan muncul ketika pola yang berguna harus dikenali secara konsisten, sementara hubungan antara variabel sering kompleks dan berubah dari waktu ke waktu. Machine learning menyediakan pendekatan untuk membangun model yang mempelajari pola dari data dan menggunakannya untuk membuat prediksi atau mengungkap struktur yang tidak langsung terlihat. Mitchell (1997) mendefinisikan pembelajaran mesin sebagai proses ketika kinerja program pada suatu tugas meningkat melalui pengalaman; definisi ini menekankan bahwa kualitas model bergantung pada data dan cara pembelajaran dirancang, bukan semata-mata pada pemilihan perangkat lunak. Dalam data mining, machine learning menjadi landasan untuk mengubah data operasional menjadi pengetahuan yang dapat mendukung klasifikasi risiko, segmentasi pelanggan, perkiraan permintaan, dan berbagai keputusan bisnis. Bagi mahasiswa Sistem Informasi, pemahaman ini diperlukan agar pemodelan tidak diperlakukan sebagai kegiatan teknis yang terpisah dari tujuan organisasi, proses bisnis, kualitas data, dan konsekuensi keputusan yang dihasilkan.

Penerapan machine learning dimulai dengan mencocokkan paradigma pembelajaran dengan masalah dan data yang tersedia. Pada supervised learning, model belajar dari contoh yang memiliki label, misalnya riwayat transaksi yang telah diklasifikasikan sebagai berisiko atau

tidak berisiko. Sebaliknya, unsupervised learning digunakan untuk menemukan struktur dalam data tanpa label, seperti kelompok pelanggan dengan pola pembelian serupa; semi-supervised learning dapat dimanfaatkan ketika hanya sebagian kecil observasi memiliki label. Perbedaan tersebut berpengaruh pada cara data dipersiapkan, algoritma yang dipilih, dan hasil yang dapat ditafsirkan (Bishop, 2006; Hastie, Tibshirani, dan Friedman, 2009). Evaluasi juga harus dirancang agar kinerja yang terukur mencerminkan kemampuan model menghadapi data baru. Pemisahan data training, validation, dan testing membantu membedakan proses pembelajaran, pemilihan model, dan pengukuran akhir, sedangkan cross-validation menyediakan estimasi kinerja yang lebih stabil ketika data terbatas (James et al., 2021). Tanpa rancangan evaluasi yang cermat, hasil yang tampak baik dapat mencerminkan kebocoran informasi atau kecocokan berlebihan terhadap data latihan, bukan pola yang dapat digeneralisasi.

Oleh karena itu, penerapan machine learning menuntut penalaran berulang: merumuskan tujuan, memeriksa karakteristik data, membangun model, mengevaluasi hasil, lalu menyesuaikan pilihan ketika bukti menunjukkan kinerja yang belum memadai. Overfitting terjadi ketika model terlalu menyesuaikan diri dengan rincian data training sehingga kinerjanya menurun pada data baru, sedangkan underfitting menunjukkan bahwa model belum menangkap

pola penting dalam data. Keduanya menggarisbawahi pentingnya keseimbangan antara kompleksitas model dan kemampuan generalisasi, yang dapat dikelola melalui pemilihan algoritma serta optimasi parameter berdasarkan evaluasi yang sahih (Hastie et al., 2009). Dalam konteks Sistem Informasi, keputusan tersebut juga perlu mempertimbangkan tujuan penggunaan, biaya kesalahan, dan keterbatasan data, sebab model yang akurat secara statistik belum tentu tepat untuk setiap proses organisasi. Bab ini membahas konsep dan paradigma machine learning, supervised, unsupervised, dan semi-supervised learning, pembagian data serta cross-validation, overfitting, underfitting, dan generalisasi model, serta pemilihan algoritma dan optimasi model.

C. Konsep dan Paradigma Machine Learning

Prapemrosesan data dan rekayasa fitur yang dibahas pada Bab 4 menyiapkan kumpulan data agar dapat digunakan dalam pemodelan. Tahap berikutnya adalah merumuskan bagaimana sistem belajar dari data dan bagaimana hasil pembelajaran tersebut akan digunakan. Dalam penambangan data, pembelajaran mesin menyediakan pendekatan untuk membangun model yang mengenali keteraturan pada data, kemudian menerapkan keteraturan itu pada kasus yang belum pernah diamati. Dengan demikian, model bukan sekadar kumpulan aturan yang ditulis secara manual, melainkan representasi pola

yang diperoleh melalui proses belajar dari contoh (Kononenko, 2007).

> **DEFINISI**
> **Pembelajaran mesin** adalah pendekatan komputasional yang membuat model menyesuaikan perilakunya berdasarkan pengalaman berupa data, sehingga kinerjanya pada suatu tugas dapat meningkat tanpa setiap aturan keputusan harus ditetapkan satu per satu oleh manusia (Kononenko, 2007).

Penerapan konsep tersebut bermula dari penentuan tugas yang hendak diselesaikan. Pada sistem informasi organisasi, tugas dapat berupa memperkirakan jumlah permintaan layanan, mengelompokkan pelanggan dengan kebiasaan transaksi serupa, atau menandai transaksi yang berpotensi menyimpang. Setiap tugas membutuhkan bentuk keluaran dan cara evaluasi yang berbeda. Prediksi nilai penjualan, misalnya, menghasilkan nilai numerik, sedangkan pengelompokan pelanggan menghasilkan kelompok yang perlu ditafsirkan lebih lanjut. Ketidakjelasan tujuan berisiko mendorong pemilihan model berdasarkan kemudahan penggunaan, bukan kesesuaian dengan kebutuhan organisasi (Unknown, 2011).

Paradigma pembelajaran menggambarkan jenis pengalaman yang digunakan model. Pengalaman itu dapat berupa contoh berlabel, struktur yang dicari langsung dari

data, atau gabungan sejumlah kecil contoh berlabel dengan banyak data yang belum berlabel. Pilihan paradigma berkaitan dengan ketersediaan label dan sifat pertanyaan analitis, bukan sekadar ukuran kumpulan data. Karena itu, tim analisis perlu memastikan siapa yang menetapkan label, seberapa dapat dipercaya label tersebut, dan tindakan apa yang akan diambil berdasarkan keluaran model (Unknown, 2019).

Di sebuah perguruan tinggi di Yogyakarta, misalnya, unit layanan akademik dapat merumuskan dua tujuan dari catatan konsultasi mahasiswa. Jika tujuannya memperkirakan apakah permohonan tertentu perlu ditindaklanjuti, tim memerlukan hasil tindak lanjut masa lalu sebagai acuan. Jika tujuannya menemukan kelompok pola keluhan untuk memperbaiki layanan, analisis dapat dimulai tanpa label kategori yang telah ditetapkan. Kedua pekerjaan menggunakan sumber data yang sama, tetapi menghasilkan pertanyaan, model, dan dasar evaluasi yang berbeda. Contoh ini menunjukkan mengapa masalah perlu dirumuskan sebelum algoritma dipilih.

Selain ketepatan prediksi, penerapan model perlu mempertimbangkan konteks operasionalnya. Pengguna keluaran model, frekuensi pembaruan, ketersediaan sumber daya komputasi, dan konsekuensi kesalahan dapat memengaruhi bentuk solusi yang layak. Model yang mampu mengenali pola pada catatan historis belum tentu berguna

apabila hasilnya tidak dapat diintegrasikan ke proses kerja atau tidak tersedia pada waktu pengambilan keputusan. Buku tentang pembelajaran mesin dan penambangan data menempatkan pemilihan representasi, algoritma, serta interpretasi hasil sebagai bagian yang berkaitan dalam suatu proses analitis (Kononenko, 2007; Unknown, 2019).

Sebagai penerapan, analis dapat menyusun pernyataan masalah yang memuat objek analisis, keluaran yang dibutuhkan, data yang tersedia, dan keputusan yang akan didukung. Rumusan seperti "mengidentifikasi transaksi yang perlu diperiksa lebih lanjut sebelum penutupan laporan harian" lebih operasional daripada "menerapkan algoritma tertentu pada transaksi". Rumusan pertama memberi dasar untuk menentukan label, fitur, ukuran evaluasi, dan kebutuhan operasional. Sesudah tujuan dan bentuk pengalaman belajar ditetapkan, pendekatan terawasi, tak terawasi, atau semi-terawasi dapat dipilih secara lebih terarah pada sub-bab berikutnya (Unknown, 2011).

D. Supervised, Unsupervised, dan Semi-Supervised Learning

Pilihan paradigma pembelajaran bergantung pada hubungan antara tujuan analisis dan ketersediaan label. Label adalah informasi yang menunjukkan keluaran atau kategori yang diharapkan untuk suatu observasi. Jika

organisasi memiliki catatan transaksi yang telah ditinjau dan diberi status, data tersebut dapat mendukung pembelajaran terawasi. Jika catatan hanya berisi atribut transaksi tanpa kategori, analisis dapat diarahkan untuk menemukan struktur atau kelompok melalui pembelajaran tak terawasi. Perbedaan ini memengaruhi cara model dibangun dan bagaimana hasilnya ditafsirkan (Kononenko, 2007; Unknown, 2011).

Dalam pembelajaran terawasi, model mempelajari hubungan antara fitur masukan dan label yang tersedia. Pendekatan ini dapat digunakan untuk mengklasifikasikan permohonan layanan ke kategori tindak lanjut atau memperkirakan nilai numerik seperti kebutuhan stok. Untuk menerapkannya, analis perlu memeriksa apakah label didefinisikan secara konsisten dan mewakili tujuan penggunaan. Jika status “berisiko” ditentukan dengan kriteria yang berbeda oleh unit-unit kerja, model dapat mempelajari ketidakkonsistenan pencatatan, bukan pola yang relevan. Oleh sebab itu, pemeriksaan mutu data dan label tetap menjadi bagian dari persiapan pemodelan (Unknown, 2011).

Pembelajaran tak terawasi bekerja tanpa label yang menjadi jawaban acuan. Salah satu penerapannya adalah klasterisasi, yaitu mengelompokkan observasi berdasarkan kemiripan fitur sehingga analis dapat memeriksa pola yang muncul. Sebuah koperasi di Jawa Tengah, misalnya, dapat

mengelompokkan anggota berdasarkan frekuensi transaksi, jenis layanan, dan pola waktu kunjungan. Kelompok yang terbentuk bukan otomatis kategori resmi atau bukti adanya tipe anggota yang tetap; analis perlu menilai apakah karakteristik kelompok tersebut masuk akal dan berguna bagi keputusan layanan. Pembahasan algoritma klasterisasi secara lebih mendalam disediakan pada Bab 8 (Kononenko, 2007).

Pembelajaran semi-terawasi dapat digunakan ketika sebagian observasi memiliki label, sedangkan sebagian besar lainnya belum diberi label. Pada layanan digital pemerintah daerah, misalnya, petugas mungkin telah memeriksa sebagian kecil tiket pengaduan dan mengategorikannya, sementara ribuan tiket lain belum ditinjau. Data berlabel dapat memberi petunjuk untuk memanfaatkan data yang lebih banyak, tetapi hasilnya tetap perlu diperiksa terhadap mutu label dan perbedaan pola antara data berlabel dan tidak berlabel. Jika contoh berlabel tidak mewakili keseluruhan kumpulan data, pembelajaran dari gabungan keduanya berisiko memperkuat gambaran yang bias (Unknown, 2019).

Perbedaan praktis ketiga paradigma dapat digunakan untuk merumuskan rencana pemodelan awal. Tabel berikut merangkum kebutuhan data, keluaran, dan contoh penggunaan yang dapat menjadi acuan saat memilih pendekatan.

Tabel 5.1 Kesesuaian Paradigma Pembelajaran dan Tujuan Analisis

Paradigma	Label	Keluaran	Contoh Indonesia	Pemeriksaan utama
Terawasi	Tersedia	Prediksi label atau nilai	Kategori pengaduan kota	Konsistensi label
Tak terawasi	Tidak tersedia	Kelompok atau pola	Segmen anggota koperasi	Makna kelompok
Semi-terawasi	Sebagian tersedia	Prediksi dengan dukungan data tanpa label	Kategori tiket layanan	Keterwakilan label
Terawasi	Tersedia	Estimasi nilai kontinu	Perkiraan kebutuhan stok UMKM	Ketepatan nilai prediksi

Tabel tersebut memperlihatkan bahwa jumlah label saja tidak menentukan pilihan; tujuan keluaran dan mutu informasi berlabel juga perlu diperiksa. Sebagai latihan penerapan, mahasiswa dapat memilih paradigma untuk sebuah layanan pengaduan di Surabaya dengan menimbang apakah label telah tersedia, keluaran yang diperlukan petugas, dan risiko salah mengategorikan laporan. (Kononenko, 2007; Unknown, 2011).

E. Training, Validation, Testing, dan Cross-Validation

Setelah paradigma pembelajaran dipilih, data perlu dibagi agar proses pengembangan model dan pengukuran kinerjanya tidak tercampur. Data pelatihan digunakan untuk mempelajari pola; data validasi membantu membandingkan pilihan model atau pengaturan parameter; sedangkan data pengujian menyediakan pemeriksaan akhir terhadap data yang tidak digunakan untuk kedua keperluan tersebut. Pemisahan fungsi ini mengurangi risiko menilai model hanya berdasarkan data yang telah dilihatnya selama pengembangan (Unknown, 2011; Unknown, 2019).

Untuk menerapkannya, analis dapat mengacak pembagian observasi apabila data tidak memiliki urutan waktu atau struktur kelompok yang perlu dipertahankan. Jika tujuan analisis adalah meramalkan periode berikutnya, pembagian acak dapat memberi gambaran evaluasi yang keliru karena data masa depan berpotensi masuk ke pelatihan. Dalam kondisi tersebut, data perlu disusun menurut waktu: periode terdahulu untuk pelatihan, periode berikutnya untuk validasi, dan periode paling akhir untuk pengujian. Pada data pelanggan, pembagian juga perlu mempertimbangkan agar catatan pelanggan yang sama tidak muncul di bagian pelatihan dan pengujian ketika hal itu dapat membocorkan informasi identitas atau kebiasaan individu.

Tabel berikut memperlihatkan contoh pembagian kumpulan data transaksi sebuah koperasi di Bandung. Angka pada tabel merupakan ilustrasi rancangan pembagian, bukan hasil eksperimen.

Tabel 5.2 Contoh Pembagian Data Transaksi Koperasi

Bagian data	Jumlah observasi	Proporsi	Fungsi
Pelatihan	700	70%	Mempelajari pola
Validasi	150	15%	Memilih pengaturan
Pengujian	150	15%	Evaluasi akhir
Total	1.000	100%	Seluruh data

Dalam contoh tersebut, data pengujian disisihkan sampai keputusan pengembangan selesai agar fungsinya sebagai evaluasi akhir tetap terjaga. Proporsi tersebut bukan ketentuan universal; ukuran data, ketidakseimbangan kelas, dan struktur observasi dapat menuntut rancangan berbeda. (Unknown, 2011).

Cross-validation membantu mengestimasi kinerja ketika jumlah data terbatas. Dalam validasi silang k-lipatan, data dibagi menjadi k bagian; setiap bagian digunakan bergiliran sebagai validasi, sementara bagian lainnya menjadi data pelatihan. Hasil evaluasi dari beberapa putaran kemudian diringkas untuk membandingkan model. Pendekatan ini memungkinkan pemanfaatan data yang lebih luas selama pengembangan, tetapi tidak menghilangkan kebutuhan akan data pengujian terpisah untuk evaluasi akhir apabila

rancangan tersebut digunakan. (Unknown, 2019; Hussin, 2022).

Pada penerapan di dinas layanan publik Kota Makassar, tim dapat memakai validasi silang untuk memilih pengaturan model klasifikasi tiket pengaduan, kemudian menyisihkan periode terbaru sebagai pengujian akhir. Pemisahan itu memberi gambaran apakah pola dari laporan terdahulu masih berguna untuk laporan yang datang kemudian. Gambar berikut merangkum peran berbeda pelatihan, validasi, dan pengujian serta penggunaan validasi silang dalam pengembangan model.



Gambar 5.1 Pembagian Data untuk Pengembangan dan Evaluasi Model. Sumber: Diolah penulis

Gambar tersebut menekankan bahwa validasi silang merupakan bagian dari proses pengembangan, sedangkan data pengujian akhir berfungsi sebagai pemeriksaan terpisah. Pembatasan akses terhadap data pengujian membantu menjaga agar estimasi kinerja tidak dipengaruhi keputusan yang berulang kali dibuat berdasarkan hasil pengujian. (Hussin, 2022).

Kebocoran data dapat terjadi apabila informasi yang seharusnya hanya tersedia setelah keputusan dibuat ikut digunakan dalam pelatihan atau validasi. Contohnya, model

untuk memprediksi keterlambatan pembayaran tidak boleh memakai kolom yang baru tercatat setelah keterlambatan terjadi. Prosedur pembagian juga perlu diterapkan setelah pemisahan data; langkah prapemrosesan yang mempelajari statistik dari seluruh data sebelum pembagian dapat membawa informasi dari data pengujian ke tahap pengembangan. Dengan menerapkan pemisahan yang sesuai dengan urutan waktu, kelompok, dan proses bisnis, hasil evaluasi lebih berguna untuk memperkirakan kinerja pada data baru (Unknown, 2011; Hussin, 2022).

F. Overfitting, Underfitting, dan Generalisasi Model

Kinerja model pada data pelatihan belum cukup untuk menunjukkan kemampuannya menghadapi observasi baru. Model dapat menangkap pola yang relevan, tetapi dapat pula menyesuaikan diri terlalu erat pada rincian khusus data pelatihan. Overfitting terjadi ketika penyesuaian itu mengganggu kinerja pada data yang tidak digunakan untuk pelatihan. Sebaliknya, underfitting terjadi ketika model terlalu sederhana atau proses belajarnya belum memadai untuk menangkap pola yang berguna. Kedua keadaan ini dinilai dengan membandingkan kinerja pada data pelatihan dan data validasi yang disusun secara tepat (Kononenko, 2007; Unknown, 2011).

Sebagai contoh, sebuah perusahaan distribusi di Jawa Barat membangun model untuk memperkirakan keterlambatan pengiriman berdasarkan jarak, waktu keberangkatan, jenis barang, dan kepadatan rute. Jika model sangat cocok dengan catatan pelatihan tetapi sering salah pada catatan validasi, analis perlu memeriksa kemungkinan overfitting. Jika model menunjukkan kinerja rendah pada kedua bagian data, pola masukan mungkin belum terwakili dengan baik, fitur mungkin kurang memadai, atau model belum sesuai untuk persoalan tersebut. Diagnosis ini tidak dapat ditentukan hanya dari satu nilai kinerja; analis perlu memeriksa data, pembagian, fitur, dan proses pelatihan secara bersama-sama (Unknown, 2019).

Generalisasi merujuk pada kemampuan model mempertahankan kinerja ketika diterapkan pada data yang belum pernah digunakan dalam pengembangan. Kemampuan tersebut bergantung pada kesesuaian data pelatihan dengan kondisi penggunaan, mutu pengukuran, serta perubahan proses bisnis. Jika pola pengiriman berubah saat musim mudik atau setelah perusahaan mengganti mitra logistik, data historis mungkin tidak lagi menggambarkan kondisi baru. Karena itu, kinerja yang baik pada pengujian historis harus ditafsirkan sesuai cakupan waktu dan populasi data yang diuji, bukan dianggap sebagai

jaminan untuk semua keadaan mendatang (Kononenko, 2007; Dua, 2016).

Diagnosis dapat dilakukan dengan menerapkan pemeriksaan berurutan. Pertama, bandingkan kinerja pelatihan dengan validasi; kesenjangan besar mengindikasikan bahwa model mungkin terlalu menyesuaikan diri pada data pelatihan. Kedua, jika keduanya sama-sama buruk, tinjau kembali relevansi fitur, mutu label, dan kesesuaian algoritma. Ketiga, periksa apakah ada kebocoran data atau perbedaan distribusi antara bagian data. Pada data keamanan sistem, misalnya, pola serangan yang berubah dapat menyebabkan model deteksi ancaman yang dahulu bekerja baik menjadi kurang andal pada lalu lintas jaringan terbaru; masalah ini tidak selalu diselesaikan dengan menambah kompleksitas model (Dua, 2016).

Beberapa penyesuaian dapat diterapkan setelah penyebab masalah dipahami. Untuk mengurangi overfitting, analis dapat membatasi kompleksitas model, menggunakan regularisasi jika tersedia, menambah contoh pelatihan yang relevan, atau memperbaiki pemisahan data. Untuk mengatasi underfitting, analis dapat meninjau fitur, memilih model yang lebih mampu mewakili pola, atau menyesuaikan proses pelatihan. Setiap perubahan perlu dinilai kembali menggunakan validasi, tanpa memilih pengaturan berdasarkan data pengujian akhir. Buku tentang

penambangan data dan pembelajaran mesin membahas hubungan antara kompleksitas model, proses pembelajaran, dan mutu evaluasi sebagai pertimbangan dalam membangun model yang berguna (Unknown, 2011; Unknown, 2019).

Sebagai penerapan di perusahaan distribusi tersebut, mahasiswa dapat membandingkan kinerja pelatihan dan validasi untuk dua rancangan model, lalu memeriksa penyebab perbedaan sebelum mengubah pengaturan. Jika selisih besar muncul hanya pada rute tertentu, pemeriksaan dapat diarahkan pada jumlah contoh, perubahan jadwal, atau pencatatan kondisi lalu lintas di rute itu. Langkah ini membedakan diagnosis berbasis bukti dari sekadar memilih model dengan skor pelatihan tertinggi.

G. Pemilihan Algoritma dan Optimasi Model

Pemilihan algoritma dimulai dari karakteristik masalah, bukan dari popularitas metode. Analis perlu menetapkan bentuk keluaran yang dibutuhkan, ukuran keberhasilan, struktur kumpulan data, dan batasan penggunaan. Untuk mengelompokkan pola tanpa label, algoritma yang dirancang bagi prediksi label tidak langsung menjawab kebutuhan tersebut. Sebaliknya, jika organisasi mempunyai contoh berlabel dan memerlukan prediksi kategori, pendekatan terawasi lebih sesuai sebagai titik awal. Buku algoritma untuk penambangan data menekankan

pentingnya menghubungkan sifat masalah dengan metode yang digunakan, sementara sumber pembelajaran mesin dan penambangan data membahas keterkaitan antara representasi data, algoritma, dan interpretasi keluaran (Unknown, 2019; Kononenko, 2007).

Kesesuaian algoritma juga dipengaruhi oleh bentuk data dan kebutuhan operasional. Jumlah observasi dan fitur, jenis variabel, keberadaan nilai hilang, ketidakseimbangan kelas, serta kebutuhan akan keluaran yang mudah dijelaskan dapat mengubah pilihan yang layak. Pada perusahaan ritel di Surabaya, misalnya, tim yang hendak menandai transaksi untuk diperiksa dapat mempertimbangkan kecepatan pemrosesan, proporsi transaksi yang berhasil ditandai, dan kemampuan petugas memahami alasan suatu transaksi diperiksa. Model dengan skor keseluruhan tinggi belum tentu berguna apabila terlalu banyak transaksi normal ditandai atau hasilnya sulit dipakai dalam prosedur kerja.

Optimasi model adalah proses menyetel pilihan atau parameter dengan menggunakan data pengembangan dan ukuran evaluasi yang telah ditetapkan. Proses ini dapat mencakup pemilihan fitur, penentuan konfigurasi algoritma, dan penyesuaian ambang keputusan ketika keluaran model berupa skor. Agar perbandingan adil, analis perlu menjaga agar perubahan yang diuji tidak sekaligus mengubah banyak unsur tanpa pencatatan. Hasil

pengaturan terbaik pada data validasi kemudian dikonfirmasi pada data pengujian yang tidak digunakan untuk memilih konfigurasi (Unknown, 2011; Hussin, 2022).

Sebagai penerapan, tim ritel Surabaya dapat membandingkan beberapa konfigurasi model menggunakan pembagian data yang sama. Tim terlebih dahulu menetapkan ukuran yang mencerminkan biaya kesalahan dalam proses pemeriksaan, kemudian mencatat kinerja validasi dan waktu pemrosesan setiap konfigurasi. Jika data transaksi berisi sangat banyak fitur, rekayasa fitur dan pemilihan fitur dapat membantu membentuk masukan yang lebih sesuai bagi pemodelan. Buku tentang rekayasa fitur menguraikan peran fitur dalam analitik dan pembelajaran mesin, sehingga keputusan mengenai representasi data perlu dipertimbangkan bersama pemilihan algoritma (Liu, 2018).

Optimasi tidak berarti mencari angka kinerja tertinggi tanpa mempertimbangkan cara angka itu diperoleh. Pengulangan pemilihan konfigurasi berdasarkan data pengujian akan menjadikan data tersebut bagian dari proses pengembangan, sehingga evaluasi akhirnya kehilangan sifat independen. Demikian pula, peningkatan kinerja pada data historis perlu dipertimbangkan bersama kestabilan hasil dan kemampuan sistem menjalankan model pada lingkungan organisasi. Pada pemantauan keamanan, misalnya, model perlu menangani volume log yang masuk

sesuai tenggat pemeriksaan; pilihan yang lambat dapat mengurangi kegunaan meskipun hasil evaluasinya tampak baik. Pertimbangan penerapan pada keamanan sistem dibahas dalam konteks pemanfaatan penambangan data dan pembelajaran mesin untuk cybersecurity (Dua, 2016).

Untuk menerapkan pemilihan dan optimasi secara tertib, mahasiswa dapat menyusun matriks keputusan berisi tujuan analisis, ketersediaan label, karakteristik data, ukuran evaluasi, batasan operasional, dan hasil validasi. Matriks itu memudahkan perbandingan alternatif tanpa menganggap satu algoritma selalu unggul pada semua masalah. Keputusan akhir dapat didasarkan pada gabungan hasil evaluasi dan kelayakan penggunaan, lalu diverifikasi menggunakan data pengujian yang disisihkan. Cara ini menghubungkan kebutuhan Sistem Informasi dengan bukti empiris pemodelan, bukan hanya dengan kecenderungan memilih metode yang sedang populer.

H. Rangkuman

Pemodelan machine learning dimulai dengan merumuskan masalah analitis secara jelas, bukan dengan langsung memilih algoritma. Rumusan yang baik menetapkan objek analisis, keluaran yang dibutuhkan, data yang tersedia, dan keputusan yang akan didukung. Kejelasan tujuan membantu menentukan fitur, ukuran evaluasi, serta kebutuhan operasional, seperti waktu

penyediaan hasil dan integrasinya ke alur kerja organisasi. Paradigma pembelajaran dipilih berdasarkan tujuan tersebut dan ketersediaan label. Pembelajaran terawasi menggunakan contoh berlabel untuk mempelajari hubungan antara fitur dan keluaran, baik berupa kategori maupun nilai numerik. Keberhasilannya bergantung pada konsistensi dan kesesuaian label dengan tujuan penggunaan; label yang tidak seragam dapat membuat model mereproduksi ketidakkonsistenan pencatatan. Pembelajaran tak terawasi mencari struktur dalam data tanpa label acuan, misalnya mengelompokkan anggota koperasi berdasarkan pola transaksi. Kelompok yang terbentuk perlu ditafsirkan dan dinilai manfaatnya, bukan langsung dianggap sebagai kategori yang resmi atau tetap. Pembelajaran semi-terawasi memanfaatkan sebagian data berlabel bersama data yang belum berlabel. Pendekatan ini dapat membantu ketika pelabelan mahal atau memerlukan peninjauan manusia, tetapi keterwakilan dan mutu contoh berlabel tetap harus diperiksa agar bias tidak diperkuat. Ketiga paradigma menghasilkan keluaran dan memerlukan pemeriksaan yang berbeda: konsistensi label, makna kelompok, atau keterwakilan label. Karena itu, jumlah data maupun ketersediaan label saja tidak cukup untuk menentukan pendekatan. Analis juga perlu mempertimbangkan pengguna keluaran, sumber daya komputasi, frekuensi pembaruan, serta konsekuensi kesalahan. Model yang mengenali pola historis belum tentu

berguna jika hasilnya terlambat, sulit diintegrasikan, atau tidak mendukung tindakan yang tepat. Dalam praktik Sistem Informasi, pemilihan pendekatan yang bertanggung jawab menghubungkan tujuan organisasi, mutu data, bentuk pengalaman belajar, evaluasi, interpretasi, dan konteks keputusan.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana perbedaan tujuan antara memperkirakan kebutuhan stok dan menemukan segmen pelanggan memengaruhi pilihan paradigma pembelajaran, bentuk keluaran, dan cara menilai hasil model. Gunakan hubungan antara tujuan, ketersediaan label, dan kebutuhan interpretasi dalam jawaban Anda.
2. Bandingkan pembelajaran terawasi, tak terawasi, dan semi-terawasi untuk pengelompokan tiket pengaduan pemerintah daerah ketika hanya sebagian kecil tiket telah ditinjau petugas. Evaluasi informasi tambahan yang diperlukan sebelum tim menetapkan pendekatan yang paling sesuai.
3. Evaluasi risiko penggunaan model terawasi apabila label “berisiko” ditetapkan dengan kriteria berbeda oleh beberapa unit kerja. Rumuskan pemeriksaan terhadap label dan data yang perlu dilakukan sebelum pemodelan, serta jelaskan dampak yang mungkin muncul jika masalah tersebut diabaikan.

4. Telaah bagaimana keterbatasan komputasi, frekuensi pembaruan, pengguna keluaran, dan konsekuensi kesalahan dapat mengubah kelayakan model untuk mendukung keputusan organisasi. Gunakan satu contoh keputusan operasional dan jelaskan pertimbangan teknis serta organisasionalnya.
5. Argumentasikan mengapa tujuan analitis perlu dirumuskan sebelum algoritma dipilih. Susun unsur-unsur pernyataan masalah untuk mendeteksi transaksi yang perlu diperiksa sebelum penutupan laporan harian, termasuk hubungan unsur tersebut dengan data dan evaluasi.

b. Pilihan Berganda

1. Sebuah organisasi memiliki catatan permohonan layanan dengan kategori tindak lanjut yang telah ditetapkan secara konsisten. Tujuan analisisnya adalah memprediksi kategori bagi permohonan baru. Paradigma yang paling sesuai adalah...
    a. Pembelajaran tak terawasi karena kategori harus ditemukan dari fitur
    b. Pembelajaran terawasi karena model belajar dari fitur dan kategori yang tersedia
    c. Pembelajaran semi-terawasi karena semua data berlabel
    d. Klasterisasi karena keluaran yang dibutuhkan berupa kelompok kemiripan
    e. Pembelajaran tak terawasi karena hasil prediksi belum diketahui

2. Koperasi ingin mengelompokkan anggota berdasarkan frekuensi transaksi dan waktu kunjungan, tetapi belum memiliki kategori anggota sebagai acuan. Langkah yang paling tepat adalah...
    a. Melatih model terawasi dengan menetapkan label dari hasil klaster
    b. Menganggap setiap kelompok yang terbentuk sebagai kategori resmi
    c. Menggunakan label tindak lanjut dari organisasi lain sebagai jawaban
    d. Menerapkan pembelajaran tak terawasi lalu menilai makna dan kegunaan kelompok
    e. Menunda analisis sampai seluruh anggota memperoleh label manual
3. Sebuah tim memiliki sedikit tiket pengaduan berlabel dan ribuan tiket tanpa label. Sebelum memakai pembelajaran semi-terawasi, pemeriksaan utama yang perlu dilakukan adalah...
    a. Memastikan contoh berlabel cukup mewakili pola dalam seluruh kumpulan data
    b. Menghapus semua tiket yang tidak memiliki label
    c. Menganggap label pada contoh kecil pasti benar dan berlaku untuk semua tiket
    d. Memastikan seluruh tiket tanpa label berasal dari periode yang sama
    e. Mengubah semua fitur menjadi label sebelum analisis

4. Beberapa unit menetapkan status “berisiko” dengan kriteria yang berbeda. Jika masalah ini diabaikan dalam pembelajaran terawasi, model berisiko...
    a. Menghasilkan kelompok tanpa hubungan dengan label
    b. Menghapus perbedaan definisi secara otomatis
    c. Mengubah tugas klasifikasi menjadi estimasi nilai kontinu
    d. Menghasilkan model yang selalu benar pada data baru
    e. Mempelajari ketidakkonsistenan pencatatan, bukan pola yang dimaksud
5. Pernyataan masalah “mengidentifikasi transaksi yang perlu diperiksa sebelum penutupan laporan harian” lebih berguna daripada “menerapkan algoritma tertentu” karena pernyataan tersebut...
    a. Menjamin model tidak akan menghasilkan kesalahan
    b. Menghilangkan kebutuhan untuk memeriksa mutu data
    c. Menghubungkan keluaran model dengan keputusan serta membantu menentukan label dan evaluasi
    d. Memastikan seluruh transaksi harus diproses secara manual
    e. Menentukan algoritma terbaik tanpa mempertimbangkan konteks

6. Model analisis catatan historis memiliki ketepatan tinggi, tetapi hasilnya tersedia setelah petugas mengambil keputusan. Evaluasi yang paling tepat adalah...
    a. Model sudah layak karena ketepatan prediksi merupakan satu-satunya ukuran
    b. Kegunaan operasionalnya terbatas karena hasil tidak tersedia pada waktu keputusan
    c. Model harus langsung diganti dengan klasterisasi
    d. Catatan historis harus dihapus agar keputusan tidak terlambat
    e. Keterlambatan hasil tidak relevan jika data yang digunakan berukuran besar
7. Tim ingin memperkirakan jumlah kebutuhan stok UMKM untuk periode berikutnya menggunakan catatan historis dengan nilai kebutuhan yang telah tercatat. Keluaran yang paling sesuai untuk tugas tersebut adalah...
    a. Kelompok anggota berdasarkan kemiripan transaksi
    b. Pola tanpa label yang belum ditafsirkan
    c. Kategori layanan yang ditetapkan petugas
    d. Estimasi nilai numerik kebutuhan stok
    e. Daftar fitur yang tidak berkaitan dengan keputusan
8. Model semi-terawasi menghasilkan prediksi buruk pada tiket pengaduan dari wilayah tertentu. Sebagian

besar contoh berlabel berasal dari wilayah lain. Tindakan evaluatif paling relevan adalah...

a. Memeriksa apakah data berlabel mewakili variasi wilayah dan pola tiket
b. Menganggap jumlah data tanpa label sudah cukup untuk menghilangkan bias
c. Menyimpulkan wilayah tersebut tidak memiliki pola pengaduan
d. Menghapus semua tiket dari wilayah yang prediksinya buruk
e. Mengganti seluruh label dengan hasil prediksi model

9. Dua tim menggunakan catatan konsultasi mahasiswa yang sama. Tim pertama memperkirakan tindak lanjut berdasarkan hasil kasus terdahulu, sedangkan tim kedua mencari kelompok pola keluhan tanpa kategori acuan. Penjelasan terbaik atas perbedaan pendekatan mereka adalah...

a. Sumber data yang sama mengharuskan kedua tim memakai algoritma yang sama
b. Tim kedua harus menggunakan label hasil tindak lanjut tim pertama
c. Perbedaan tujuan dan ketersediaan label menghasilkan pertanyaan analitis serta dasar evaluasi yang berbeda
d. Tim pertama tidak dapat menggunakan data historis karena keluarannya belum diketahui
e. Pengelompokan otomatis selalu lebih tepat daripada prediksi kategori

10. Dalam menilai hasil klasterisasi anggota koperasi, tindakan analis yang paling tepat adalah...
    a. Menyebut setiap klaster sebagai tipe anggota yang permanen
    b. Mengukur keberhasilan hanya dari jumlah klaster yang dihasilkan
    c. Memberi label resmi pada klaster tanpa meninjau karakteristiknya
    d. Menganggap klaster terbesar sebagai kelompok paling berguna
    e. Memeriksa karakteristik kelompok dan menilai relevansinya bagi keputusan layanan

c. Soal Project / Studi Kasus

1. Unit layanan akademik sebuah perguruan tinggi menerima ribuan tiket konsultasi setiap semester. Petugas telah mengategorikan sekitar 8 persen tiket, tetapi definisi kategori “akademik mendesak” belum seragam antarfakultas. Pimpinan ingin menggunakan tiket lama dan baru untuk membantu menentukan laporan mana yang perlu ditinjau lebih dahulu. Data mencakup isi tiket, fakultas, waktu pengiriman, dan hasil tindak lanjut; akses petugas terhadap dasbor juga berbeda-beda. Rumuskan pendekatan pemodelan awal dengan mempertimbangkan kelayakan paradigma, mutu serta keterwakilan label, keluaran yang dibutuhkan, evaluasi, dan integrasi ke alur kerja. Jelaskan pemeriksaan yang harus mendahului penerapan model.

2. Sebuah koperasi di Jawa Tengah ingin mengurangi keterlambatan pemeriksaan transaksi sebelum laporan harian ditutup. Catatan transaksi memuat nilai, waktu, jenis layanan, dan unit pemroses, tetapi belum memiliki label transaksi menyimpang. Tim keuangan dapat meninjau sejumlah kecil transaksi setiap hari; pemeriksaan penuh akan menunda laporan. Manajer meminta hasil yang dapat dipakai petugas pada hari yang sama, sementara anggaran komputasi terbatas dan kesalahan menandai transaksi normal dapat membebani tim. Rumuskan rencana analisis yang mencakup perumusan masalah, pilihan paradigma awal, pemanfaatan hasil peninjauan, evaluasi, dan langkah operasional untuk mengendalikan risiko kesalahan.

BAB 6
KLASIFIKASI DATA

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan tahapan klasifikasi untuk mengubah data berlabel menjadi model yang dapat digunakan dalam pengambilan keputusan pada sistem informasi.
2. Menggunakan Decision Tree dan Random Forest untuk membangun model klasifikasi berdasarkan karakteristik data dan kebutuhan analisis.
3. Menerapkan Naïve Bayes dan K-Nearest Neighbor untuk mengklasifikasikan data dengan mempertimbangkan asumsi serta karakteristik masing-masing metode.
4. Menggunakan Support Vector Machine untuk menentukan batas pemisah kelas pada data yang dianalisis.
5. Menghitung dan menafsirkan nilai confusion matrix, precision, recall, F1-score, dan AUC untuk mengevaluasi kinerja model klasifikasi.

B. Pendahuluan

Organisasi menghasilkan data transaksi, interaksi pelanggan, dan aktivitas operasional dalam jumlah besar, tetapi data tersebut baru bernilai bagi pengambilan

keputusan apabila dapat diolah menjadi informasi yang relevan. Salah satu kebutuhan yang kerap muncul ialah memperkirakan kategori suatu objek berdasarkan pola pada data sebelumnya, misalnya mengidentifikasi transaksi yang berpotensi menjadi fraud, menentukan pelanggan yang berisiko berhenti berlangganan, atau mengelompokkan permohonan layanan sesuai tingkat urgensinya. Klasifikasi merupakan teknik pembelajaran terawasi yang membangun model dari data berlabel, kemudian menggunakan model tersebut untuk memprediksi kelas pada data baru (Han, Kamber, dan Pei, 2012). Bagi mahasiswa Sistem Informasi, pemahaman klasifikasi diperlukan karena metode ini menghubungkan pengelolaan data dengan proses bisnis dan keputusan organisasi. Hasil klasifikasi dapat mendukung otomasi layanan, pemantauan risiko, dan penentuan prioritas tindak lanjut, tetapi manfaatnya bergantung pada kesesuaian metode, mutu data, serta ketepatan evaluasi.

Penerapan klasifikasi tidak cukup dilakukan dengan memilih algoritma dan menjalankannya pada seluruh data. Proses yang andal mencakup penentuan masalah dan kelas sasaran, penyiapan serta pemilihan fitur, pembagian data untuk pelatihan dan pengujian, pembangunan model, dan evaluasi kinerja. Pemisahan data tersebut membantu mengukur kemampuan model menghadapi kasus yang belum pernah dipelajari, sedangkan pemeriksaan terhadap

data yang tidak seimbang dan kebocoran informasi mencegah kesimpulan evaluatif yang menyesatkan. Beragam algoritma menawarkan cara berbeda dalam mengenali pola: Decision Tree membentuk aturan yang relatif mudah ditafsirkan, sementara Random Forest menggabungkan banyak pohon untuk meningkatkan kestabilan prediksi (Quinlan, 1986; Breiman, 2001). Naïve Bayes memanfaatkan teorema Bayes dengan asumsi independensi bersyarat ant fitur, sedangkan K-Nearest Neighbor menetapkan kelas berdasarkan kedekatan suatu data dengan tetangga terdekatnya (Domingos dan Pazzani, 1997; Cover dan Hart, 1967). Perbedaan prinsip ini berpengaruh pada kebutuhan persiapan data, interpretasi hasil, dan kecocokan metode terhadap konteks masalah.

Support Vector Machine membangun batas keputusan dengan memaksimalkan margin antarkelas dan dapat menggunakan fungsi kernel untuk menangani pola pemisahan yang tidak linear (Cortes dan Vapnik, 1995). Kinerja model klasifikasi selanjutnya perlu ditinjau melalui ukuran yang sesuai dengan konsekuensi kesalahan dalam konteks penerapan. Confusion matrix merangkum prediksi benar dan salah untuk setiap kelas; dari matriks tersebut, precision, recall, dan F1-score dapat digunakan untuk menilai aspek kinerja yang berbeda, sementara AUC merangkum kemampuan model membedakan kelas pada berbagai ambang keputusan (Fawcett, 2006). Sebagai

contoh, sistem deteksi fraud mungkin lebih mengutamakan recall agar transaksi berisiko tidak luput, tetapi peningkatan deteksi juga dapat menambah pemeriksaan terhadap transaksi yang sebenarnya sah. Karena itu, pemilihan metrik harus mempertimbangkan dampak operasional, bukan sekadar mengejar satu angka tertinggi. Bab ini membahas konsep dan tahapan klasifikasi, Decision Tree dan Random Forest, Naïve Bayes dan K-Nearest Neighbor, Support Vector Machine, serta evaluasi klasifikasi menggunakan confusion matrix, precision, recall, F1-score, dan AUC.

C. Konsep dan Tahapan Klasifikasi

Pembahasan klasifikasi melanjutkan alur pembelajaran terawasi yang telah diuraikan pada Bab 5. Dalam klasifikasi, kumpulan data pelatihan memuat fitur dan label kelas yang telah diketahui. Algoritma mempelajari hubungan antara keduanya untuk membentuk model, lalu model tersebut digunakan untuk menetapkan kelas bagi objek baru. Kelas dapat berupa kategori biner, seperti fraud atau bukan fraud, maupun lebih dari dua kategori, seperti tingkat urgensi rendah, sedang, dan tinggi. Dengan demikian, klasifikasi menghasilkan keputusan kategoris, bukan perkiraan nilai kontinu sebagaimana dibahas pada Bab 7 (Kononenko, 2007; Unknown, 2011).

Tahapan klasifikasi dimulai dengan merumuskan tujuan operasional dan kelas sasaran. Pada sistem informasi,

tujuan perlu dinyatakan dalam bentuk keputusan yang dapat ditindaklanjuti: misalnya, transaksi yang ditandai berisiko harus diperiksa oleh petugas, sedangkan transaksi lain dapat diproses secara normal. Definisi kelas perlu jelas agar pelabelan antarpengguna konsisten. Setelah itu, fitur dipilih berdasarkan ketersediaan, relevansi terhadap sasaran, dan waktu kemunculannya. Fitur yang baru diketahui setelah keputusan terjadi tidak boleh digunakan untuk memprediksi keputusan sebelumnya karena dapat menimbulkan kebocoran informasi. Penyiapan dan transformasi fitur mengikuti prinsip prapemrosesan data yang dibahas pada Bab 4 (Liu, 2018; Unknown, 2019).

Kumpulan data kemudian dipisahkan menjadi data pelatihan dan data pengujian, dengan data validasi bila diperlukan untuk memilih konfigurasi model. Pemisahan harus dilakukan sebelum proses yang mempelajari parameter dari data, seperti menghitung nilai pengganti untuk data hilang atau menentukan skala normalisasi, agar informasi data pengujian tidak masuk ke tahap pelatihan. Pada masalah dengan kelas tidak seimbang, pembagian berstrata dapat membantu menjaga proporsi kelas di setiap bagian data. Model dilatih pada data pelatihan, konfigurasi dipilih menggunakan validasi, lalu kinerja akhir diukur pada data pengujian yang belum digunakan dalam pengembangan model (Unknown, 2011; Kononenko, 2007).

Setelah model dibangun, ambang keputusan dan jenis kesalahan perlu dipertimbangkan sesuai konsekuensi proses bisnis. Sebagai contoh, jika sebuah koperasi simpan pinjam di Yogyakarta ingin menandai pengajuan yang berisiko gagal bayar, kegagalan mengenali pemohon berisiko dapat memiliki biaya berbeda dari menandai pemohon layak sebagai berisiko. Model yang digunakan untuk memberi sinyal kepada analis juga dapat memiliki kebutuhan berbeda dari model yang otomatis menolak pengajuan. Oleh sebab itu, tujuan klasifikasi tidak berhenti pada pencapaian akurasi; hasilnya harus dapat mendukung tindakan yang telah ditetapkan organisasi (Dua, 2016; Unknown, 2011).

Tahapan utama dan keluaran yang perlu diperiksa dapat dirangkum sebagai berikut. Urutan ini membantu memastikan bahwa pembangunan model tetap terhubung dengan tujuan penggunaan, bukan hanya pelaksanaan algoritma.

Tabel 6.1 Tahapan Klasifikasi dan Keluaran

Tahap	Aktivitas utama	Keluaran	Pemeriksaan
Perumusan masalah	Menetapkan sasaran dan kelas	Definisi kelas	Arti label konsisten
Persiapan data	Memilih fitur dan menangani data	Data siap latih	Tidak ada kebocoran

Pembagian data	Menyusun data latih dan uji	Partisi data	Proporsi kelas terjaga
Pelatihan	Menyesuaikan algoritma	Model klasifikasi	Proses tercatat
Evaluasi	Mengukur kesalahan model	Metrik evaluasi	Sesuai risiko penggunaan
Penerapan	Menghubungkan model ke proses	Prediksi operasional	Hasil dapat ditindaklanjuti

Tabel tersebut menunjukkan bahwa evaluasi dan penerapan merupakan bagian dari rangkaian klasifikasi, bukan kegiatan terpisah dari pemodelan. Kesalahan pada definisi kelas atau pembagian data dapat membuat hasil evaluasi tidak mencerminkan kinerja model dalam penggunaan sebenarnya.

D. Decision Tree dan Random Forest

Decision Tree membentuk aturan klasifikasi dengan membagi data secara bertahap berdasarkan nilai fitur. Setiap simpul keputusan menguji suatu kondisi, cabang menunjukkan hasil pengujian, dan simpul akhir menetapkan kelas. Karena struktur ini menyerupai rangkaian pertanyaan, aturan yang terbentuk dapat ditelusuri oleh pengguna. Pemisahan fitur biasanya dipilih untuk menghasilkan kelompok yang semakin terpisah menurut kelas, misalnya menggunakan ukuran entropi dan information gain. Proses pemilihan dan pemangkasan pohon memengaruhi kompleksitas aturan serta risiko model terlalu menyesuaikan diri dengan data pelatihan (Unknown, 2011; Unknown, 2019).

Salah satu ukuran ketidakmurnian simpul adalah entropi. Jika proporsi kelas ke-i pada suatu simpul adalah p_i, entropi dihitung sebagai berikut.

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Pemisahan yang baik menurunkan ketidakmurnian pada simpul-simpul hasil dibandingkan dengan simpul asal. Namun, pohon yang terus diperluas dapat menghasilkan banyak aturan yang hanya cocok dengan contoh pelatihan. Pembatasan kedalaman, jumlah minimum data pada simpul, atau pemangkasan pohon dapat mengurangi kompleksitas tersebut. Aturan yang lebih ringkas umumnya lebih mudah diperiksa, tetapi penyederhanaan berlebihan juga dapat menghilangkan pola yang berguna (Kononenko, 2007; Unknown, 2011).

Random Forest menggabungkan banyak Decision Tree yang dilatih menggunakan variasi sampel dan fitur. Pengambilan sampel ulang pada data pelatihan menghasilkan pohon-pohon yang berbeda, sedangkan pemilihan sebagian fitur pada tiap pemisahan membantu mengurangi ketergantungan antarpohon. Untuk klasifikasi, keluaran gabungan umumnya ditentukan melalui suara mayoritas atau agregasi probabilitas. Penggabungan tersebut dapat membuat hasil lebih stabil daripada

mengandalkan satu pohon, walaupun interpretasi aturan individual menjadi kurang langsung. Kinerja juga bergantung pada pengaturan jumlah pohon, jumlah fitur yang dipertimbangkan, dan karakteristik data (Religia, 2021; Unknown, 2019).

Perbedaan penerapan kedua metode terlihat pada kebutuhan untuk menjelaskan alasan prediksi. Misalnya, perusahaan ritel di Bandung dapat memakai pohon tunggal untuk menyusun aturan awal mengenai pelanggan yang berpotensi berhenti berbelanja, sehingga petugas dapat mengikuti jalur keputusan tertentu. Jika data pelanggan mencakup banyak fitur dan hubungan yang rumit, Random Forest dapat dipakai sebagai pembanding untuk memperoleh prediksi gabungan yang lebih stabil. Religia (2021) membahas optimasi algoritma pada Random Forest untuk klasifikasi data pemasaran bank, sehingga memberi contoh bahwa pengaturan model merupakan bagian dari penerapan, bukan sekadar memilih nama algoritma.

Pemilihan antara kedua metode sebaiknya mempertimbangkan tujuan analisis, ukuran data, serta kebutuhan penjelasan. Decision Tree memudahkan inspeksi aturan, tetapi pohon yang dalam dapat sulit dipahami dan peka terhadap perubahan data. Random Forest lebih sesuai ketika kestabilan prediksi menjadi perhatian, tetapi keputusan gabungan perlu dijelaskan dengan analisis tambahan jika organisasi harus menyampaikan alasan di

balik keputusan kepada pengguna atau pengawas. Dalam kedua kasus, fitur tetap perlu diperiksa agar tidak memasukkan informasi yang tidak tersedia pada saat prediksi dilakukan (Liu, 2018; Dua, 2016).

E. Naïve Bayes dan K-Nearest Neighbor

Naïve Bayes mengklasifikasikan objek dengan menerapkan teorema Bayes dan memperkirakan peluang suatu kelas berdasarkan fitur yang diamati. Untuk fitur X dan kelas C, peluang posterior berbanding lurus dengan peluang kelas dan peluang fitur pada kelas tersebut. Asumsi "naïve" menyatakan bahwa fitur bersifat independen secara bersyarat setelah kelas diketahui. Asumsi ini menyederhanakan perhitungan, tetapi tidak selalu mencerminkan hubungan nyata antarf fitur. Bentuk model yang sederhana dapat berguna ketika pemrosesan perlu dilakukan secara efisien atau data pelatihan terbatas (Kononenko, 2007; Unknown, 2011).

$$P(C \mid X) \propto P(C) \prod_{j=1}^{n} P(x_j \mid C)$$

Jenis Naïve Bayes disesuaikan dengan representasi fitur. Multinomial Naïve Bayes, misalnya, lazim digunakan pada fitur berupa frekuensi kemunculan kata. Dalam penerapan klasifikasi sentimen berbahasa Indonesia, teks ulasan atau unggahan dapat direpresentasikan sebagai frekuensi kata,

kemudian model memperkirakan kelas sentimen berdasarkan pola pada data berlabel. Yuyun (2021) membahas penggunaan Multinomial Naïve Bayes untuk klasifikasi sentimen pemerintah terhadap penanganan Covid-19 menggunakan data Twitter. Contoh ini menunjukkan keterkaitan pemilihan varian metode dengan bentuk data, bukan sekadar label masalah.

K-Nearest Neighbor (KNN) menentukan kelas suatu objek berdasarkan kelas tetangga terdekatnya dalam ruang fitur. Jarak Euclidean untuk dua objek x dan z dengan n fitur numerik dapat dihitung sebagai berikut.

$$d(x, z) = \sqrt{\sum_{j=1}^{n} (x_j - z_j)^2}$$

Nilai k menentukan banyaknya tetangga yang dipakai dalam pemungutan suara. K yang terlalu kecil dapat membuat keputusan peka terhadap noise, sedangkan K yang lebih besar dapat mencampurkan objek dari wilayah kelas berbeda. Karena jarak bergantung pada skala fitur, fitur dengan rentang nilai besar dapat mendominasi perhitungan apabila data belum dinormalisasi atau distandardisasi. Pemilihan ukuran jarak, skala fitur, dan nilai k perlu dilakukan menggunakan data pelatihan dan validasi, bukan dengan melihat hasil data pengujian (Unknown, 2011; Liu, 2018).

Perbedaan kebutuhan komputasi juga memengaruhi penerapan. Naïve Bayes membangun ringkasan probabilistik dari data pelatihan dan dapat menghasilkan prediksi tanpa membandingkan setiap objek baru dengan seluruh data. Sebaliknya, KNN menyimpan data pelatihan dan menghitung kedekatan ketika prediksi diminta; hal ini dapat membebani sistem saat kumpulan data besar. Sebagai konteks Indonesia, toko grosir di Tasikmalaya dapat menggunakan KNN untuk memperkirakan kategori permintaan produk berdasarkan pola transaksi sebelumnya, sementara Naïve Bayes dapat diterapkan pada data teks ulasan pelanggan. Budiawati (2026) mengangkat penerapan KNN untuk prediksi penjualan produk terlaris pada toko grosir, yang relevan sebagai contoh pemanfaatan metode dalam konteks usaha.

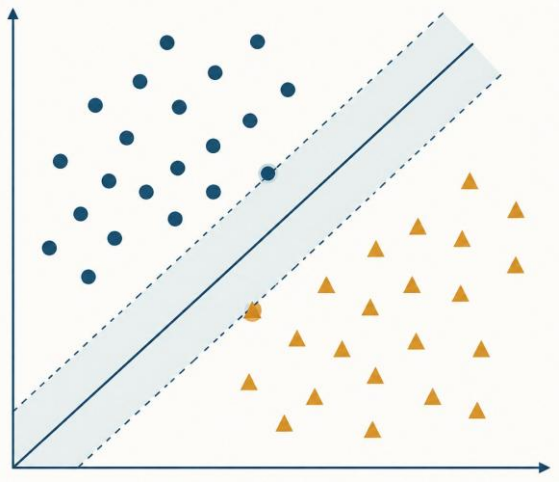
Penerapan kedua metode memerlukan perhatian pada kualitas fitur dan representasi data. KNN dapat memberikan hasil yang berbeda ketika fitur yang tidak relevan ikut dihitung dalam jarak, sedangkan Naïve Bayes dapat terpengaruh ketika asumsi independensi tidak sesuai dengan struktur fitur. Untuk teks, pemilihan representasi kata dan penanganan kata yang jarang muncul perlu diperhitungkan; untuk data numerik, skala fitur dan nilai hilang perlu ditangani sebelum pengukuran jarak. Karena itu, uji pada data yang tidak digunakan untuk pelatihan tetap diperlukan, serta ukuran evaluasi dipilih menurut

dampak kesalahan pada penggunaan akhir (Liu, 2018; Unknown, 2019).

F. Support Vector Machine

Support Vector Machine (SVM) membangun batas keputusan yang memisahkan kelas dengan memaksimalkan margin, yaitu jarak antara batas keputusan dan contoh terdekat dari masing-masing kelas. Contoh yang berada paling dekat dengan batas disebut support vector karena posisinya menentukan letak batas pemisah. Pada pemisahan linear, keputusan dapat dinyatakan melalui fungsi berbentuk $w^{\top}x + b = 0$, dengan w menentukan arah batas dan b menentukan pergeserannya. Margin yang lebih lebar dapat membantu model menggeneralisasi pola, meskipun hasilnya juga dipengaruhi oleh karakteristik data dan pengaturan parameter (Kononenko, 2007; Unknown, 2019).

Gambaran batas keputusan dan margin membantu menjelaskan bagaimana SVM membedakan kelas pada ruang fitur. Ilustrasi berikut menampilkan titik-titik dari dua kelas serta garis pemisahnya.



Gambar 6.1 Batas Keputusan dan Margin SVM. Sumber: Ilustrasi, diolah penulis

Titik yang menyentuh garis margin menandai contoh yang paling berpengaruh terhadap posisi batas pemisah. Jika data tidak dapat dipisahkan secara linear, SVM dapat menggunakan fungsi kernel untuk merepresentasikan hubungan yang lebih kompleks tanpa harus membuat transformasi fitur secara eksplisit. Pemilihan kernel, parameter regularisasi, dan parameter kernel memengaruhi bentuk batas keputusan. Karena itu, konfigurasi perlu disesuaikan menggunakan data validasi agar model tidak terlalu sederhana atau terlalu mengikuti pola khusus dalam data pelatihan (Unknown, 2011; Unknown, 2019).

SVM dapat digunakan untuk klasifikasi biner maupun multikelas, meskipun penanganan multikelas umumnya dibangun melalui penggabungan sejumlah keputusan biner. Dalam penerapan, data perlu disiapkan secara cermat: skala fitur numerik perlu diseragamkan karena model sensitif terhadap perbedaan skala, sedangkan fitur kategoris harus direpresentasikan dalam bentuk yang sesuai. Penggunaan kernel juga dapat meningkatkan kebutuhan komputasi, terutama ketika jumlah contoh pelatihan besar. Dengan demikian, SVM cocok dipertimbangkan ketika batas antarkelas perlu dimodelkan dengan margin yang jelas, tetapi biaya pelatihan dan kebutuhan interpretasi tetap harus diperhitungkan (Liu, 2018; Dua, 2016).

Konteks penerapan di Indonesia dapat dilihat pada klasifikasi data perkawinan anak di Maluku Utara. Kahar (2021) membahas klasifikasi data tersebut menggunakan kernel regression dan SVM. Contoh ini menegaskan perlunya menghubungkan tujuan analisis dengan pemilihan metode serta memperlakukan data secara bertanggung jawab, terutama ketika hasil klasifikasi menyangkut kelompok masyarakat. Pada penggunaan lain, seperti penyaringan aktivitas jaringan yang mencurigakan, SVM dapat menjadi salah satu model pembanding; pembahasan penerapan analitik keamanan secara lebih luas diuraikan pada Bab 11 (Dua, 2016).

G. Evaluasi Klasifikasi dengan Confusion Matrix, Precision, Recall, F1-Score, dan AUC

Evaluasi klasifikasi mengukur kesesuaian prediksi dengan label aktual pada data yang tidak digunakan untuk melatih model. Confusion matrix merangkum hasil tersebut dengan memisahkan prediksi benar dan salah menurut kelas. Untuk klasifikasi biner, empat komponennya ialah true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). Penentuan kelas positif harus dinyatakan terlebih dahulu karena makna metrik bergantung pada kelas yang dianggap sebagai sasaran utama, misalnya transaksi fraud dalam sistem pemantauan bank (Unknown, 2011; Dua, 2016).

Precision mengukur proporsi prediksi positif yang benar, sedangkan recall mengukur proporsi kasus positif aktual yang berhasil ditemukan. F1-score merupakan rerata harmonis precision dan recall, sehingga nilainya menurun ketika salah satu dari kedua metrik tersebut rendah. Rumusnya sebagai berikut.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Contoh berikut menggunakan hasil pengujian hipotetis pada sistem penandaan transaksi di sebuah bank di Indonesia. Dari 300 transaksi, 45 merupakan transaksi fraud menurut pemeriksaan, sementara sistem menandai 51 transaksi sebagai fraud. Rincian prediksinya ditampilkan dalam tabel.

Tabel 6.2 Confusion Matrix Klasifikasi Transaksi

Hasil aktual	Prediksi fraud	Prediksi bukan fraud	Jumlah
Fraud	36	9	45
Bukan fraud	15	240	255
Jumlah	51	249	300

Berdasarkan tabel, precision adalah $36/51 \approx 0{,}706$ atau $70{,}6\%$, sedangkan recall adalah $36/45 = 0{,}80$ atau 80%. F1-score menjadi sekitar 75%. Artinya, sistem menemukan delapan dari setiap sepuluh transaksi fraud dalam data uji, tetapi sebagian transaksi yang ditandai tetap bukan fraud. Nilai-nilai tersebut membantu menjelaskan jenis kesalahan yang perlu ditangani, bukan sekadar menyatakan bahwa model “baik” atau “buruk”.

Pemilihan metrik mengikuti dampak operasional kesalahan. Jika transaksi fraud yang lolos menimbulkan kerugian besar, organisasi dapat mengutamakan recall; namun, kenaikan recall dapat meningkatkan jumlah transaksi sah yang perlu diperiksa. Jika sumber daya pemeriksaan terbatas, precision juga menjadi pertimbangan karena menunjukkan seberapa sering

penandaan positif benar. F1-score berguna ketika precision dan recall sama-sama relevan, tetapi tidak menggambarkan seluruh biaya kesalahan dan tidak memperhitungkan true negative secara langsung. Karena itu, interpretasi perlu mengaitkan metrik dengan kapasitas petugas, toleransi risiko, serta proses tindak lanjut (Dua, 2016; Unknown, 2011).

AUC, atau luas di bawah kurva ROC, merangkum kemampuan model membedakan kelas pada berbagai ambang keputusan. Kurva ROC membandingkan true positive rate dengan false positive rate ketika ambang berubah. Berbeda dari precision, recall, dan F1-score yang dapat dihitung pada satu ambang tertentu, AUC memberi gambaran kinerja pemeringkatan pada rentang ambang. Nilai AUC yang lebih tinggi menunjukkan pemisahan kelas yang lebih baik secara umum, tetapi tidak menentukan ambang operasional terbaik. Pada data dengan kelas sangat tidak seimbang, AUC perlu dibaca bersama confusion matrix dan metrik lain agar hasil evaluasi tidak menyamarkan banyaknya kesalahan positif palsu (Unknown, 2011; Kononenko, 2007).

H. Rangkuman

Klasifikasi merupakan pembelajaran terawasi yang menggunakan data berfitur dan berlabel untuk membentuk model, kemudian menetapkan kelas pada objek baru.

Hasilnya berupa kategori, seperti fraud atau bukan fraud, berbeda dari prediksi nilai kontinu. Pembangunan model dimulai dengan merumuskan keputusan operasional dan mendefinisikan kelas secara konsisten. Fitur dipilih berdasarkan relevansi, ketersediaan, dan waktu kemunculannya; informasi yang baru tersedia setelah keputusan dibuat dapat menimbulkan kebocoran jika digunakan untuk memprediksi keputusan tersebut. Data dibagi menjadi bagian pelatihan, validasi bila diperlukan, dan pengujian. Pembagian dilakukan sebelum transformasi yang mempelajari parameter dari data, sementara pembagian berstrata membantu menjaga proporsi kelas. Model dilatih pada data pelatihan, dipilih atau disetel menggunakan validasi, lalu dinilai pada data uji yang belum digunakan dalam pengembangan.

Evaluasi klasifikasi tidak cukup bertumpu pada akurasi. Ambang keputusan dan jenis kesalahan perlu disesuaikan dengan konsekuensi proses bisnis, misalnya membedakan biaya luput mengenali pemohon berisiko dari biaya menandai pemohon layak sebagai berisiko. Karena itu, prediksi harus terhubung dengan tindakan yang dapat dijalankan organisasi. Decision Tree membentuk aturan melalui pemisahan fitur; entropi mengukur ketidakmurnian simpul, sedangkan information gain membantu memilih pemisahan. Pohon yang terlalu kompleks berisiko menyesuaikan diri berlebihan pada data pelatihan, sehingga

pembatasan kedalaman atau pemangkasan dapat dipertimbangkan. Struktur aturan pohon relatif mudah ditelusuri, tetapi pohon dalam dapat sulit dipahami dan peka terhadap perubahan data.

Random Forest menggabungkan banyak pohon yang dibangun dari variasi sampel dan fitur, lalu menghasilkan kelas melalui suara mayoritas atau agregasi probabilitas. Cara ini dapat menstabilkan prediksi, meskipun alasan di balik keputusan gabungan tidak semudah aturan satu pohon untuk diperiksa. Naïve Bayes memperkirakan peluang kelas berdasarkan fitur dengan asumsi independensi bersyarat antarf fitur setelah kelas diketahui. Asumsi tersebut menyederhanakan perhitungan, tetapi perlu dinilai kesesuaiannya dengan data. Dalam praktik Sistem Informasi, pilihan metode perlu mempertimbangkan karakteristik data, kebutuhan penjelasan, risiko kesalahan, dan cara hasil model digunakan dalam proses organisasi.

I. Latihan Mahasiswa

a. Essay

1. Analisis alur pembangunan model klasifikasi untuk mendeteksi transaksi berisiko. Jelaskan bagaimana definisi kelas, pemilihan fitur berdasarkan waktu ketersediaannya, pembagian data, dan prapemrosesan dapat memengaruhi keabsahan evaluasi serta kelayakan penerapan model.

2. Bandingkan Decision Tree dan Random Forest untuk mendukung keputusan retensi pelanggan pada perusahaan ritel. Evaluasi perbedaan keduanya dalam keterlacakan alasan prediksi, kestabilan hasil, kompleksitas aturan, dan kebutuhan organisasi untuk menjelaskan keputusan kepada petugas.
3. Evaluasi penggunaan akurasi sebagai ukuran utama model yang menandai pengajuan pinjaman berisiko gagal bayar. Analisis bagaimana biaya kesalahan yang berbeda dan pilihan ambang keputusan seharusnya memengaruhi rancangan evaluasi serta tindakan yang dilakukan setelah prediksi diperoleh.
4. Telaah hubungan entropi, information gain, dan pemangkasan dalam pembentukan Decision Tree. Rumuskan strategi untuk mengendalikan kompleksitas pohon tanpa menghilangkan pola yang bermanfaat, serta jelaskan bagaimana strategi tersebut dapat dinilai pada data yang tidak digunakan untuk melatih model.
5. Argumentasikan kapan Naïve Bayes layak dipertimbangkan untuk klasifikasi dalam Sistem Informasi. Analisis manfaat asumsi independensi bersyarat bagi penyederhanaan perhitungan, keterbatasannya ketika fitur saling berhubungan, dan pemeriksaan yang perlu dilakukan sebelum model dipakai.

b. Pilihan Berganda

1. Sebuah tim menghitung nilai pengganti data hilang dan parameter normalisasi menggunakan seluruh data

sebelum membagi data latih dan data uji. Perbaikan yang paling tepat adalah...

a. Menghapus seluruh fitur yang memiliki data hilang.
b. Menghitung ulang parameter hanya pada data uji.
c. Membagi data terlebih dahulu, lalu mempelajari parameter prapemrosesan dari data latih.
d. Menggabungkan data latih dan uji setelah normalisasi.
e. Menggunakan data uji sebagai data validasi untuk memilih model.

2. Manakah sasaran yang paling sesuai dengan tugas klasifikasi?

a. Memperkirakan jumlah penjualan bulan depan dalam satuan unit.
b. Menetapkan kategori urgensi tiket sebagai rendah, sedang, atau tinggi.
c. Menghitung rata-rata waktu layanan setiap pelanggan.
d. Mengurutkan pelanggan berdasarkan total belanja tanpa menetapkan kelas.
e. Memperkirakan nilai saldo rekening pada akhir tahun.

3. Pada data dengan proporsi kelas yang tidak seimbang, tujuan utama pembagian berstrata adalah...

a. Menghilangkan kelas minoritas dari data uji.
b. Menambah jumlah data tanpa mengubah distribusi fitur.

c. Membuat data pelatihan selalu seimbang sempurna.
d. Menjaga proporsi kelas relatif terwakili pada bagian data yang dibentuk.
e. Memastikan seluruh anggota kelas minoritas masuk ke data pelatihan.

4. Dalam pembentukan Decision Tree, information gain terutama digunakan untuk...
   a. Mengukur biaya penerapan model di organisasi.
   b. Memilih pemisahan yang menurunkan ketidakmurnian kelas pada simpul hasil.
   c. Menentukan jumlah pohon dalam Random Forest.
   d. Menghitung peluang posterior pada Naïve Bayes.
   e. Menetapkan ambang keputusan berdasarkan biaya kesalahan.

5. Sebuah Decision Tree sangat dalam dan memiliki banyak aturan yang hanya cocok dengan data pelatihan. Tindakan yang paling relevan untuk mengendalikan masalah tersebut adalah...
   a. Menambahkan fitur yang baru diketahui setelah keputusan dibuat.
   b. Mengukur kinerja hanya pada data pelatihan.
   c. Menghapus data uji sebelum evaluasi.
   d. Memperbanyak cabang agar setiap contoh memiliki aturan khusus.
   e. Membatasi kedalaman atau memangkas pohon.

6. Ciri utama Random Forest dalam klasifikasi adalah...

a. Menggunakan satu pohon penuh tanpa variasi data atau fitur.
b. Menetapkan kelas hanya dengan menghitung entropi data uji.
c. Menggabungkan prediksi banyak pohon yang dibangun melalui variasi sampel dan fitur.
d. Mengasumsikan semua fitur independen tanpa bergantung pada kelas.
e. Menghasilkan nilai kontinu sebagai keluaran utama klasifikasi.

7. Asumsi “naïve” pada Naïve Bayes menyatakan bahwa...
   a. Fitur independen secara bersyarat setelah kelas diketahui.
   b. Setiap kelas memiliki jumlah contoh yang sama.
   c. Seluruh fitur harus berupa kategori biner.
   d. Kelas sasaran tidak perlu ditentukan saat pelatihan.
   e. Data uji dapat digunakan untuk menghitung parameter model.

8. Dalam model penandaan pengajuan pinjaman, salah mengenali pemohon berisiko dan salah menandai pemohon layak memiliki konsekuensi berbeda. Implikasi yang paling tepat adalah...
   a. Akurasi selalu cukup untuk menentukan penggunaan model.
   b. Semua prediksi positif harus otomatis ditolak.
   c. Ambang keputusan harus selalu ditetapkan pada nilai yang sama.

d. Metrik dan ambang keputusan perlu dipilih dengan mempertimbangkan biaya setiap jenis kesalahan.
e. Evaluasi hanya perlu dilakukan setelah model digunakan secara operasional.

9. Menggunakan fitur yang baru tercatat setelah keputusan kredit dibuat untuk memprediksi keputusan tersebut berisiko menyebabkan...
   a. Pembagian data berstrata.
   b. Kebocoran informasi.
   c. Pemangkasan pohon.
   d. Agregasi probabilitas.
   e. Independensi bersyarat.

10. Agar evaluasi akhir mencerminkan kemampuan model pada data yang belum digunakan dalam pengembangan, data pengujian seharusnya...
   a. Dipakai berulang kali untuk memilih fitur dan konfigurasi model.
   b. Digabungkan dengan data pelatihan sebelum semua tahap prapemrosesan.
   c. Digunakan untuk menetapkan aturan kelas setelah evaluasi.
   d. Dipakai sebagai data pelatihan tambahan sebelum pengukuran akhir.
   e. Disimpan terpisah dari proses pengembangan dan digunakan untuk pengukuran akhir.

c. Soal Project / Studi Kasus

1. Sebuah koperasi simpan pinjam di Yogyakarta ingin menggunakan klasifikasi untuk menandai pengajuan yang berisiko gagal bayar agar diperiksa analis. Data historis berisi label hasil pembayaran, pendapatan, masa kerja, jumlah pinjaman, dan status tunggakan yang dicatat beberapa bulan setelah pengajuan disetujui. Kegagalan mengenali pemohon berisiko dapat menambah kerugian, sedangkan menandai pemohon layak sebagai berisiko dapat memperlambat layanan. Pimpinan sedang mempertimbangkan Decision Tree yang mudah ditelusuri dan Random Forest yang menggabungkan banyak pohon. Analisis definisi kelas, fitur yang layak, pembagian dan evaluasi data, serta pilihan model dan tindak lanjut yang disarankan.
2. Perusahaan ritel di Bandung hendak mengidentifikasi pelanggan yang berpotensi berhenti berbelanja untuk menentukan siapa yang menerima program retensi. Tim memiliki riwayat transaksi, keluhan, respons terhadap promosi, dan label berhenti atau tetap aktif. Sebagian data promosi tercatat setelah pelanggan berhenti berbelanja. Manajer meminta aturan yang dapat dipahami petugas, tetapi tim analitik juga mengkhawatirkan ketidakstabilan satu pohon ketika data berubah. Rancang pendekatan klasifikasi yang menangani risiko kebocoran dan menilai kompromi antara keterjelasan aturan dan kestabilan prediksi. Sertakan cara evaluasi serta batas penggunaan hasil model.

BAB 7
REGRESI DAN PREDICTIVE ANALYTICS

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep regresi dan predictive analytics untuk merumuskan masalah prediksi pada konteks sistem informasi.
2. Menggunakan regresi linear sederhana dan berganda untuk memodelkan hubungan antara variabel prediktor dan variabel respons.
3. Memilih dan menerapkan pendekatan regresi nonlinear atau regularisasi sesuai karakteristik data dan tujuan pemodelan.
4. Menghitung serta menggunakan MAE, MSE, RMSE, dan R^2 untuk mengevaluasi kinerja model regresi.
5. Mengimplementasikan proses prediksi berbasis data sensor dan time-series IoT dengan mempertimbangkan karakteristik data temporal.

B. Pendahuluan

Organisasi menghasilkan data dalam jumlah besar melalui transaksi digital, sistem operasional, dan perangkat yang terhubung. Tantangan analitis tidak berhenti pada penyimpanan atau penyajian data, tetapi mencakup pemanfaatannya untuk memperkirakan nilai yang belum

diketahui, seperti permintaan produk, durasi layanan, konsumsi energi, atau suhu mesin pada periode berikutnya. Regresi menyediakan pendekatan untuk memodelkan hubungan antara satu atau lebih variabel prediktor dan suatu variabel respons, sedangkan predictive analytics menggunakan model tersebut untuk menghasilkan perkiraan yang mendukung keputusan. James et al. (2021) menjelaskan bahwa regresi merupakan salah satu metode utama dalam pembelajaran terawasi karena dapat digunakan untuk mengestimasi respons numerik sekaligus menafsirkan hubungan dalam data. Bagi mahasiswa Sistem Informasi, pemahaman ini relevan karena prediksi perlu diintegrasikan dengan proses bisnis, infrastruktur data, dan kebutuhan pengambilan keputusan organisasi.

Penerapan regresi menuntut lebih dari sekadar menjalankan algoritma. Analis perlu menerjemahkan persoalan bisnis menjadi variabel yang dapat diukur, menyiapkan data yang sesuai, memilih bentuk model, lalu menafsirkan hasilnya dengan mempertimbangkan keterbatasan asumsi. Regresi linear sederhana membantu memodelkan hubungan antara satu prediktor dan respons, sedangkan regresi linear berganda memperluas analisis dengan melibatkan beberapa prediktor secara simultan. Apabila pola hubungan tidak memadai dijelaskan oleh garis lurus atau model terlalu kompleks terhadap data, regresi nonlinear dan regularisasi dapat menjadi pilihan. Evaluasi

juga harus dilakukan secara terukur: MAE, MSE, dan RMSE merangkum besarnya kesalahan prediksi dengan karakteristik yang berbeda, sementara R^2 menunjukkan proporsi variasi respons yang dapat dijelaskan oleh model. Seperti ditekankan Kuhn dan Johnson (2013), pemilihan model dan evaluasi yang cermat diperlukan agar kinerja yang tampak baik tidak semata-mata mencerminkan kecocokan berlebihan terhadap data pelatihan.

Kebutuhan tersebut semakin nyata pada sistem yang mengandalkan sensor dan Internet of Things (IoT). Data sensor umumnya direkam berurutan menurut waktu, dapat mengandung gangguan pengukuran, nilai hilang, dan perubahan pola, serta tidak selalu memenuhi asumsi bahwa setiap observasi saling bebas. Karena itu, pemodelan prediksi perlu memperhatikan urutan temporal dan pemisahan data pelatihan serta pengujian agar informasi masa depan tidak bocor ke dalam proses pembentukan model. Hyndman dan Athanasopoulos (2021) menekankan pentingnya evaluasi berdasarkan urutan waktu dalam peramalan, sehingga pengujian lebih mencerminkan penggunaan model pada kondisi nyata. Bab ini membahas konsep regresi dan predictive analytics, regresi linear sederhana dan berganda, regresi nonlinear dan regularisasi, evaluasi menggunakan MAE, MSE, RMSE, dan R^2, serta penerapan prediksi pada data sensor dan time-series IoT.

C. Konsep Regresi dan Predictive Analytics

Pembahasan klasifikasi pada Bab 6 berfokus pada penentuan kelas atau kategori, sedangkan regresi digunakan ketika keluaran yang dicari berupa nilai numerik. Perbedaan ini mengubah cara masalah dirumuskan, model dibangun, dan hasil ditafsirkan. Dalam sistem informasi, nilai yang diperkirakan dapat berupa jumlah transaksi, waktu penyelesaian layanan, kebutuhan persediaan, konsumsi energi, atau suhu perangkat pada waktu mendatang. Regresi menjadi salah satu pendekatan pembelajaran terawasi karena model mempelajari hubungan antara fitur dan respons dari contoh yang telah memiliki nilai target (Kononenko, 2007).

> **DEFINISI**
> **Regresi** adalah pendekatan pemodelan yang mengestimasi hubungan antara satu atau lebih variabel prediktor dan variabel respons numerik, kemudian menggunakan hubungan tersebut untuk memperkirakan nilai respons pada observasi baru.

Predictive analytics merupakan rangkaian kegiatan yang memanfaatkan data, metode statistik, dan model komputasional untuk menghasilkan perkiraan yang dapat mendukung keputusan. Prosesnya dimulai dengan menerjemahkan kebutuhan organisasi menjadi pertanyaan prediksi yang terukur. Analis menetapkan apa yang

diprediksi, untuk siapa atau objek apa prediksi dibuat, serta kapan hasil itu harus tersedia. Langkah tersebut mencegah pemilihan algoritma mendahului pemahaman masalah. Dalam pemanfaatan data prediktif di bidang klinis, misalnya, mutu dan kesesuaian data serta konteks penggunaan hasil merupakan bagian dari persoalan yang harus ditangani, bukan sekadar urusan teknis setelah model selesai (Fernando & Jain, 2022).

Model prediksi tidak otomatis menjelaskan sebab-akibat. Koefisien yang menunjukkan hubungan antara dua variabel dapat berguna untuk memperkirakan respons, tetapi hubungan tersebut belum membuktikan bahwa perubahan pada satu variabel menyebabkan perubahan pada variabel lain. Pemisahan tujuan prediksi dari tujuan penjelasan membantu organisasi menghindari keputusan yang terlalu jauh dari bukti. Untuk kebutuhan operasional, model sebaiknya dinilai berdasarkan ketepatan prediksi pada data yang belum digunakan saat pelatihan serta kegunaan hasilnya dalam proses keputusan (Kononenko, 2007).

Di Indonesia, sebuah usaha batik di Pekalongan dapat menggunakan catatan transaksi, musim penjualan, jumlah promosi, dan jumlah produk yang tersedia untuk memperkirakan penjualan minggu berikutnya. Prediksi tersebut dapat membantu pemilik menentukan kebutuhan bahan dan jadwal produksi. Namun, hasilnya perlu

ditafsirkan bersama informasi kontekstual, seperti periode menjelang hari raya atau perubahan saluran penjualan. Jika kondisi pasar berubah, pola yang dipelajari dari data sebelumnya mungkin tidak lagi menggambarkan situasi terkini.

Penerapan predictive analytics juga mencakup alur kerja yang terhubung: penentuan sasaran, pemeriksaan data, pembentukan fitur, pelatihan model, evaluasi, dan penggunaan hasil. Prapemrosesan data serta pemilihan fitur dapat memengaruhi informasi yang tersedia bagi algoritma; fitur yang tidak relevan atau kualitas pencatatan yang buruk dapat mengurangi manfaat pemodelan (Liu, 2018). Sesudah model diterapkan, organisasi perlu mencatat prediksi dan hasil aktual agar kinerjanya dapat diperiksa kembali. Dengan demikian, keluaran model tidak diperlakukan sebagai kepastian, melainkan sebagai estimasi yang memiliki tingkat kesalahan dan bergantung pada data serta konteks penggunaannya.

D. Simple dan Multiple Linear Regression

Regresi linear memodelkan perubahan nilai respons sebagai kombinasi linear dari satu atau lebih fitur. Pada regresi linear sederhana, hanya satu fitur digunakan sebagai prediktor. Bentuk modelnya adalah:

$$\hat{y} = \beta_0 + \beta_1 x$$

Dalam persamaan tersebut, $\hat{y}$ adalah nilai respons hasil estimasi, x adalah fitur, β_0 adalah intersep, dan β_1 adalah koefisien kemiringan. Koefisien menunjukkan perubahan rata-rata nilai prediksi untuk setiap kenaikan satu satuan pada x, dengan syarat bentuk hubungan linear cukup sesuai untuk konteks data. Pemodelan regresi memerlukan penentuan respons dan prediktor yang bermakna bagi tujuan analisis, bukan sekadar memilih kolom yang tersedia (Kononenko, 2007).

Regresi linear berganda melibatkan beberapa fitur secara simultan. Bentuk umumnya dapat ditulis sebagai:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Setiap koefisien menggambarkan hubungan antara satu fitur dan respons dengan fitur lain dalam model dianggap tetap. Penafsiran tersebut perlu hati-hati ketika prediktor saling berkorelasi kuat atau ketika terdapat variabel penting yang tidak tercatat. Rekayasa fitur dan pemilihan representasi variabel membantu mengubah data mentah menjadi masukan yang lebih berguna bagi pemodelan (Liu, 2018).

Tabel berikut menyajikan data ilustratif penjualan mingguan sebuah usaha batik di Pekalongan. Nilai promosi dan penjualan sengaja disusun mengikuti hubungan linear

sederhana agar perhitungan koefisien dapat ditunjukkan secara transparan.

Tabel 7.1 Data Ilustratif Promosi dan Penjualan Batik

Minggu	Promosi (juta Rp)	Penjualan (juta Rp)
1	1	23
2	2	26
3	3	29
4	4	32
5	5	35

Data tersebut sesuai dengan persamaan ilustratif $\hat{y} = 20 + 3x$. Intersep sebesar 20 menunjukkan nilai prediksi ketika pengeluaran promosi bernilai nol dalam model, sedangkan koefisien 3 menunjukkan kenaikan prediksi penjualan sebesar Rp3 juta untuk setiap tambahan Rp1 juta promosi. Pola ini merupakan ilustrasi matematis, bukan bukti bahwa promosi selalu menyebabkan kenaikan penjualan sebesar itu.

Pada penerapan nyata, koefisien umumnya diestimasi dengan meminimalkan jumlah kuadrat residual, yaitu selisih antara respons aktual dan respons prediksi. Regresi berganda dapat menambahkan fitur seperti jumlah hari operasional, musim, atau jumlah pesanan daring untuk menghasilkan model yang lebih sesuai dengan proses bisnis. Namun, penambahan fitur tidak otomatis memperbaiki prediksi. Setiap fitur harus tersedia pada saat prediksi dibuat, didefinisikan secara konsisten, dan tidak membawa

informasi dari masa depan. Prinsip pemodelan data serta interpretasi hubungan antarkomponen perlu disesuaikan dengan tujuan analisis dan karakteristik data (Solimun dkk., 2023).

Sebelum koefisien digunakan dalam keputusan, kinerja model perlu diuji pada data yang tidak digunakan untuk mengestimasi koefisien. Pemeriksaan pola residual juga membantu mengenali ketidaksesuaian, seperti pola melengkung atau sebaran kesalahan yang makin besar pada nilai prediksi tertentu. Jika pola hubungan tidak memadai dijelaskan oleh garis lurus, bentuk fungsi lain atau pendekatan regularisasi dapat dipertimbangkan sebagaimana dibahas pada sub-bab berikutnya.

E. Regresi Nonlinear dan Regularisasi

Hubungan antara fitur dan respons tidak selalu dapat direpresentasikan dengan garis lurus. Dalam konteks operasional, kenaikan penggunaan sumber daya dapat mula-mula meningkatkan hasil secara cepat, lalu memberikan tambahan hasil yang semakin kecil. Regresi nonlinear digunakan ketika bentuk hubungan tersebut memerlukan fungsi selain hubungan linear sederhana. Salah satu cara membentuk model yang lebih lentur adalah menambahkan suku berpangkat, misalnya:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$$

Walaupun memuat suku kuadrat, model tersebut tetap linear terhadap koefisien β_0, β_1, dan β_2. Perbedaan ini penting: nonlinearitas pada fitur tidak selalu berarti bahwa algoritma harus mengestimasi koefisien melalui prosedur nonlinear. Transformasi fitur, termasuk pembentukan interaksi dan suku berpangkat, merupakan bagian dari rekayasa fitur yang perlu dirancang sesuai pengetahuan domain (Liu, 2018).

Sebagai ilustrasi di Indonesia, penggunaan listrik sebuah fasilitas pengolahan hasil pertanian di Jawa Timur dapat berubah secara tidak linear terhadap volume produksi. Pada tingkat produksi rendah, tambahan produksi mungkin hanya memerlukan sedikit energi; ketika mesin mendekati kapasitasnya, kebutuhan pendinginan atau pemrosesan tambahan dapat meningkat lebih tajam. Analis dapat menguji bentuk linear dan bentuk polinomial pada data historis, kemudian membandingkan kesalahan prediksinya pada data pengujian. Bentuk yang lebih rumit tidak dipilih hanya karena tampak mampu mengikuti seluruh titik data.

Model dengan terlalu banyak fitur atau suku polinomial dapat menyesuaikan diri secara berlebihan terhadap data pelatihan. Regularisasi mengendalikan kompleksitas dengan memberikan penalti pada besarnya koefisien. Ridge menambahkan penalti kuadrat, sedangkan Lasso menggunakan jumlah nilai absolut koefisien. Secara umum, fungsi objektif Ridge dapat dinyatakan sebagai:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Parameter λ mengatur kekuatan penalti. Nilai λ yang lebih besar mendorong koefisien menjadi lebih kecil, tetapi dapat pula membuat model terlalu sederhana jika penalti berlebihan. Pada Lasso, penalti nilai absolut dapat mendorong sebagian koefisien menjadi nol sehingga fitur tertentu tidak berkontribusi dalam model. Kedua pendekatan menyediakan cara berbeda untuk mengendalikan kompleksitas, bukan jaminan bahwa model akan selalu lebih akurat (Kononenko, 2007).

Sebelum menerapkan regularisasi, skala fitur perlu diperiksa karena penalti dapat memperlakukan koefisien secara tidak seimbang ketika satu fitur berskala sangat besar dan fitur lain berskala kecil. Penskalaan sebaiknya dipelajari dari data pelatihan saja, lalu diterapkan tanpa menghitung ulang parameter menggunakan data pengujian. Praktik ini mencegah informasi pengujian memengaruhi pembentukan model. Prapemrosesan yang konsisten merupakan bagian dari penyiapan data untuk analisis, bukan langkah terpisah yang boleh diabaikan (García dkk., 2015).

Pemilihan bentuk nonlinear dan tingkat regularisasi dilakukan melalui validasi yang sesuai dengan struktur data.

Pada data transaksi tanpa urutan temporal, pembagian dapat mengikuti rancangan validasi yang dibahas pada Bab 5. Untuk data berurutan waktu, pembagian acak berisiko mencampurkan informasi periode mendatang ke pelatihan; penanganan khusus untuk data tersebut diuraikan pada sub-bab G. Model yang dipilih harus menyeimbangkan ketepatan prediksi, kompleksitas, dan kemudahan penerapan pada sistem informasi.

F. Evaluasi Model dengan MAE, MSE, RMSE, dan R^2

Evaluasi regresi mengukur seberapa jauh prediksi model dari nilai aktual. Ukuran yang dipilih memengaruhi cara kesalahan diringkas dan ditafsirkan. Untuk data uji dengan n observasi, galat pada observasi ke-i dapat ditulis sebagai $e_i = y_i - \hat{y}_i$, dengan y_i sebagai nilai aktual dan $\hat{y}_i$ sebagai nilai prediksi. MAE menghitung rata-rata nilai absolut galat, sedangkan MSE menghitung rata-rata kuadrat galat. RMSE merupakan akar kuadrat MSE sehingga satuannya kembali sama dengan satuan respons (Unknown, 2012).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{RMSE} = \sqrt{\text{MSE}}$$

MAE mudah dibaca sebagai rata-rata besar kesalahan dalam satuan target. MSE memberi bobot lebih besar pada galat yang besar karena galat dikuadratkan, sedangkan RMSE mempertahankan karakteristik tersebut dengan satuan yang lebih mudah dibandingkan dengan target. Karena karakteristiknya berbeda, pemilihan metrik perlu mempertimbangkan biaya kesalahan dalam proses bisnis, bukan hanya memilih angka yang paling kecil (Setiawan, 2025).

Contoh berikut memakai prediksi kebutuhan bahan baku mingguan sebuah UMKM makanan di Bandung. Nilai aktual dan prediksi dinyatakan dalam kilogram, sementara selisih bertanda menunjukkan arah kesalahan.

Tabel 7.2 Perhitungan Galat Prediksi Kebutuhan Bahan

Minggu	Aktual (kg)	Prediksi (kg)	Galat (kg)	Galat kuadrat
1	10	11	-1	1
2	12	10	2	4
3	13	14	-1	1
4	15	13	2	4

Dari empat observasi tersebut, MAE adalah $1{,}5$ kg, MSE adalah $2{,}5$ kg ^{2}, dan RMSE sekitar $1{,}58$ kg. Nilai ini menunjukkan ringkasan besarnya galat untuk kumpulan observasi contoh; penafsiran operasionalnya bergantung

pada toleransi persediaan dan dampak kekurangan atau kelebihan bahan.

Koefisien determinasi R^2 membandingkan variasi residual model dengan variasi nilai aktual terhadap rata-ratanya. Rumus yang umum digunakan adalah:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Pada tabel, rata-rata aktual adalah $12{,}5$ kg, jumlah kuadrat residual adalah 10, dan jumlah kuadrat penyimpangan aktual dari rata-ratanya adalah 17. Dengan demikian, R^2 sekitar $0{,}41$. Nilai itu menyatakan proporsi variasi yang dijelaskan oleh model relatif terhadap tolok ukur rata-rata pada data contoh, bukan persentase akurasi prediksi dan bukan bukti sebab-akibat (Kononenko, 2007).

Metrik harus dihitung pada data yang tidak digunakan untuk melatih model agar estimasi kinerja lebih mencerminkan pemakaian pada kasus baru. Dua model dapat memiliki MAE yang mirip tetapi RMSE berbeda apabila salah satunya menghasilkan beberapa kesalahan ekstrem. Karena itu, pelaporan lebih dari satu metrik dapat membantu organisasi melihat pola risiko dengan lebih lengkap. Untuk peramalan temporal, proses pemisahan data harus mempertahankan urutan waktu agar evaluasi

tidak memanfaatkan informasi masa depan (Unknown, 2012).

G. Prediksi Berbasis Data Sensor dan Time-Series IoT

Data sensor dalam Internet untuk Segala (IoT) biasanya tercatat bersama penanda waktu dan identitas perangkat. Urutan tersebut membedakannya dari kumpulan data yang setiap barisnya dapat diperlakukan sebagai observasi tanpa keterkaitan temporal. Prediksi pada data seperti ini dapat ditujukan untuk memperkirakan nilai sensor pada menit atau jam berikutnya, mengestimasi konsumsi energi harian, atau memperkirakan waktu hingga suatu perangkat membutuhkan pemeliharaan. Perancangan model perlu menyesuaikan tujuan, rentang waktu prediksi, dan ketersediaan data pada saat keputusan dibuat.

Sebuah sistem pemantauan suhu ruang penyimpanan ikan di Pelabuhan Ratu, Jawa Barat, misalnya, dapat merekam suhu setiap beberapa menit. Untuk memperkirakan suhu periode selanjutnya, analis dapat membentuk fitur dari nilai terdahulu, seperti suhu satu langkah sebelumnya, suhu pada beberapa interval terdahulu, jam pengukuran, serta ringkasan bergerak. Pendekatan ini mengubah informasi waktu menjadi masukan bagi model regresi. Data dan fitur yang digunakan harus tersedia pada saat prediksi diterbitkan, sedangkan

pencatatan waktu perlu konsisten agar urutan observasi tidak keliru.

Ilustrasi berikut menunjukkan sensor suhu yang dipasang pada sistem pendingin penyimpanan. Posisi sensor pada fasilitas nyata membantu menjelaskan bahwa pembacaan merupakan hasil pengukuran suatu proses fisik, bukan angka yang terlepas dari kondisi perangkat.



Gambar 7.1 Penempatan Sensor Suhu pada Ruang Penyimpanan Ikan. Sumber: Ilustrasi, diolah penulis

Penempatan sensor menentukan bagian proses yang diwakili oleh data. Jika sensor berada dekat pintu atau aliran udara keluar, pembacaannya mungkin berbeda dari suhu di antara produk yang disimpan; interpretasi prediksi perlu memperhatikan konfigurasi tersebut.

Salah satu contoh penerapan lokal adalah penelitian prediksi kadar CO_2 berdasarkan suhu dan kelembapan menggunakan sensor DHT22 dan MQ135 yang diintegrasikan dengan ESP32. Kasus tersebut memperlihatkan pemanfaatan beberapa pembacaan sensor sebagai prediktor bagi nilai numerik kualitas udara, meskipun implementasi praktis tetap perlu memperhatikan karakteristik alat, kalibrasi, dan mutu pencatatan (Agung Suburdjati dkk., 2026). Untuk peramalan yang memiliki pola musiman, penerapan Holt-Winters pada sistem business intelligence transportasi menunjukkan relevansi metode deret waktu dalam mendukung perkiraan jumlah pelanggan, meski konteks dan variabel targetnya berbeda dari sensor fisik (Priandini dkk., 2025).

Pemisahan data pelatihan dan pengujian perlu mengikuti waktu: periode terdahulu digunakan untuk pelatihan dan periode sesudahnya untuk pengujian. Jika data diacak, model dapat memperoleh pola dari periode mendatang yang belum tersedia dalam penggunaan sebenarnya. Validasi bergulir, yaitu melatih model pada jendela waktu terdahulu lalu menguji pada jendela berikutnya, dapat digunakan untuk menilai kestabilan prediksi pada beberapa periode. Pemantauan kualitas deret waktu juga perlu mencakup nilai hilang, gangguan pengukuran, perubahan pola, dan jeda komunikasi perangkat; persoalan anomali pada deret waktu menjadi salah satu aspek yang perlu

dibedakan dari tugas prediksi itu sendiri (Blázquez-García, 2022).

Dalam lingkungan IoT, prediksi baru bermanfaat jika hasilnya dapat dihubungkan dengan tindakan yang sesuai. Perkiraan suhu yang mendekati batas operasional dapat memicu pemeriksaan perangkat, sedangkan estimasi konsumsi energi dapat membantu penjadwalan beban. Organisasi perlu menyimpan waktu prediksi, nilai yang diperkirakan, nilai aktual ketika tersedia, dan versi model agar kesalahan dapat ditelusuri. Pemisahan antara prediksi, ambang keputusan, dan tindakan operasional menjaga agar estimasi model tidak diperlakukan sebagai kepastian tentang kondisi fisik yang akan terjadi.

H. Rangkuman

Regresi digunakan ketika keluaran yang ingin diperkirakan berupa nilai numerik, berbeda dari klasifikasi yang menetapkan kelas atau kategori. Dalam predictive analytics, organisasi terlebih dahulu merumuskan sasaran prediksi, objek yang diprediksi, dan waktu ketersediaan hasil. Data kemudian diperiksa, fitur dibentuk, model dilatih dan dievaluasi, lalu hasilnya digunakan serta dipantau. Prediksi merupakan estimasi yang mengandung kesalahan dan bergantung pada mutu data serta konteks; hubungan yang dipelajari model juga tidak dengan sendirinya membuktikan sebab-akibat.

Regresi linear sederhana menghubungkan satu fitur dengan respons melalui intersep dan koefisien kemiringan. Pada ilustrasi penjualan batik, model $\hat{y} = 20 + 3x$ memperkirakan kenaikan penjualan sebesar Rp3 juta untuk setiap tambahan Rp1 juta promosi. Angka tersebut menggambarkan pola dalam data ilustratif, bukan bukti bahwa promosi selalu menyebabkan kenaikan penjualan sebesar itu. Regresi linear berganda menggunakan beberapa fitur sekaligus; setiap koefisien ditafsirkan dengan fitur lain dalam model dianggap tetap. Penafsiran perlu mempertimbangkan korelasi antarfitur, kemungkinan variabel penting yang tidak tercatat, dan ketersediaan fitur pada saat prediksi dibuat.

Koefisien regresi umumnya ditentukan dengan meminimalkan jumlah kuadrat residual, yaitu selisih antara nilai aktual dan prediksi. Pengujian pada data yang tidak dipakai saat pelatihan menilai kemampuan model menghadapi observasi baru, sementara pola residual dapat mengungkap ketidaksesuaian seperti hubungan melengkung atau sebaran kesalahan yang membesar. Jika garis lurus tidak memadai, transformasi fitur seperti suku berpangkat dapat membentuk hubungan yang lebih lentur. Model dengan suku kuadrat, misalnya, nonlinear terhadap fitur tetapi tetap linear terhadap koefisiennya. Regularisasi dapat dipertimbangkan sebagai pendekatan pemodelan, tetapi penambahan fitur atau kerumitan model tidak

otomatis meningkatkan kinerja. Bagi praktisi Sistem Informasi, pemilihan model perlu sejalan dengan kebutuhan keputusan, kualitas pencatatan, dan perubahan kondisi operasional; prediksi serta hasil aktual perlu dicatat agar kinerja dapat diperiksa kembali.

I. Latihan Mahasiswa

a. Essay

1. Analisis perbedaan tujuan regresi dan klasifikasi dalam sistem informasi, lalu rumuskan contoh keputusan operasional yang tepat untuk masing-masing pendekatan. Jelaskan pula mengapa model regresi yang akurat tetap tidak cukup untuk menyimpulkan hubungan sebab-akibat antara fitur dan respons.
2. Evaluasi penggunaan model $\hat{y} = 20 + 3x$ untuk memperkirakan penjualan usaha batik ketika pengeluaran promosi meningkat. Telaah makna intersep dan koefisien, batas penafsiran berdasarkan data ilustratif, serta informasi tambahan yang diperlukan sebelum pemilik menjadikan hasil model sebagai dasar keputusan produksi.
3. Bandingkan regresi linear sederhana dan berganda dalam konteks prediksi penjualan mingguan. Analisis bagaimana korelasi kuat antarfitur, variabel penting yang tidak tercatat, dan ketersediaan fitur pada waktu prediksi dapat memengaruhi penafsiran koefisien maupun kegunaan model.

4. Telaah bagaimana pembagian data pelatihan dan pengujian serta pemeriksaan residual saling melengkapi dalam evaluasi regresi. Rumuskan tindakan yang sebaiknya diambil apabila model tampak akurat pada data pelatihan, tetapi pola residual menunjukkan kelengkungan dan kesalahan membesar pada nilai prediksi tertentu.
5. Argumentasikan bagaimana organisasi sebaiknya menangani model prediksi setelah diterapkan ketika kondisi pasar berubah. Sintesislah peran pencatatan prediksi dan hasil aktual, konteks penggunaan, mutu data, serta pemantauan kinerja dalam menentukan apakah model masih layak mendukung keputusan.

b. Pilihan Berganda

1. Sebuah sistem memperkirakan waktu penyelesaian layanan dalam satuan menit. Pendekatan yang paling sesuai adalah...
    a. Klasifikasi, karena setiap layanan harus dimasukkan ke kategori tetap
    b. Regresi, karena respons yang diperkirakan berupa nilai numerik
    c. Pengelompokan, karena waktu layanan merupakan data berurutan
    d. Asosiasi, karena sistem menghubungkan fitur dengan hasil
    e. Perangkingan, karena durasi layanan memiliki urutan

2. Pada model ilustratif $\hat{y} = 20 + 3x$, dengan x adalah promosi dalam juta rupiah dan $\hat{y}$ adalah penjualan dalam juta rupiah, interpretasi koefisien 3 yang paling tepat adalah...
    a. Setiap tambahan Rp1 juta promosi pasti menyebabkan penjualan naik Rp3 juta
    b. Penjualan selalu bernilai Rp3 juta ketika promosi tidak dilakukan
    c. Prediksi penjualan bertambah Rp3 juta untuk kenaikan promosi Rp1 juta menurut model
    d. Promosi menjelaskan tepat 3 persen variasi penjualan
    e. Penjualan aktual tidak mungkin berbeda dari nilai prediksi model
3. Mengapa suatu model perlu dievaluasi pada data yang tidak digunakan untuk mengestimasi koefisien?
    a. Agar jumlah fitur dalam model bertambah secara otomatis
    b. Agar hubungan sebab-akibat antarvariabel dapat dibuktikan
    c. Agar kesalahan pada data pelatihan dapat dihilangkan seluruhnya
    d. Agar kemampuan prediksi pada observasi baru dapat dinilai
    e. Agar intersep model selalu bernilai nol
4. Sebuah fitur berisi jumlah promosi yang baru diketahui setelah minggu penjualan berakhir. Fitur tersebut tidak

layak digunakan untuk memprediksi penjualan minggu itu terutama karena...

a. Fitur numerik tidak dapat dipakai dalam regresi
b. Informasinya belum tersedia pada saat prediksi seharusnya dibuat
c. Fitur tersebut pasti memiliki korelasi sempurna dengan respons
d. Regresi berganda hanya menerima fitur yang berbentuk kategori
e. Fitur tersebut akan mengubah respons numerik menjadi kelas

5. Dalam regresi linear berganda, koefisien suatu fitur paling tepat ditafsirkan sebagai...

a. Perubahan prediksi respons terkait kenaikan satu satuan fitur dengan fitur lain dalam model dianggap tetap
b. Besarnya pengaruh kausal fitur tanpa dipengaruhi keadaan lain
c. Persentase seluruh variasi respons yang dijelaskan oleh fitur tersebut
d. Perubahan respons aktual yang pasti terjadi pada setiap observasi
e. Nilai prediksi respons ketika semua fitur lainnya bernilai nol

6. Model dengan bentuk $\hat{y} = \beta_o + \beta_1 x + \beta_2 x^2$ paling tepat digambarkan sebagai...

a. Model klasifikasi karena respons dapat berubah mengikuti x

b. Model nonlinear terhadap koefisien karena mengandung pangkat dua
c. Model linear terhadap koefisien, tetapi memungkinkan pola nonlinear terhadap x
d. Model yang tidak dapat diestimasi menggunakan data numerik
e. Model yang selalu lebih akurat daripada regresi linear sederhana

7. Pada grafik residual, kesalahan tampak membesar saat nilai prediksi meningkat. Temuan ini terutama menunjukkan bahwa...
   a. Semua prediksi model pasti mengalami bias positif
   b. Respons seharusnya diubah menjadi variabel kategori
   c. Intersep harus selalu dihapus dari persamaan
   d. Sebaran kesalahan mungkin tidak konstan pada berbagai tingkat prediksi
   e. Model telah membuktikan hubungan kausal antara fitur dan respons

8. Sebuah perusahaan menemukan bahwa promosi dan jumlah produk tersedia berkorelasi kuat. Dalam menafsirkan koefisien regresi berganda, perusahaan sebaiknya...
   a. Menganggap setiap koefisien sebagai pengaruh kausal yang terpisah
   b. Memperhatikan bahwa korelasi antarfitur dapat menyulitkan penafsiran koefisien

c. Menghapus respons dari analisis agar koefisien lebih stabil
d. Menambahkan semua kolom yang tersedia tanpa pemeriksaan
e. Menggunakan data pelatihan saja untuk menilai ketepatan prediksi

9. Kinerja model yang baik pada data historis tidak menjamin kinerja yang sama ketika pola pasar berubah. Tindakan yang paling sesuai adalah...
   a. Memperlakukan prediksi model sebagai nilai yang pasti
   b. Menghentikan pencatatan hasil aktual setelah model diterapkan
   c. Memantau prediksi dan hasil aktual serta meninjau kinerja dalam konteks terbaru
   d. Menghapus fitur kontekstual agar model tidak berubah
   e. Menganggap perubahan pasar sebagai bukti kesalahan pencatatan

10. Urutan kegiatan yang paling sesuai untuk penerapan predictive analytics adalah...
    a. Memilih algoritma, membuat keputusan, menetapkan sasaran, lalu memeriksa data
    b. Membentuk fitur, menerapkan model, menetapkan respons, lalu mengevaluasi hasil
    c. Menilai hasil aktual, memilih fitur, merumuskan sasaran, lalu melatih model

d. Menetapkan sasaran, memeriksa data, membentuk fitur, melatih dan mengevaluasi model, lalu menggunakan hasil
e. Mengumpulkan semua kolom, membuat prediksi, lalu menentukan masalah yang hendak dijawab

c. Soal Project / Studi Kasus

1. Sebuah usaha batik di Pekalongan menggunakan catatan mingguan untuk merencanakan produksi. Tim sistem informasi mengusulkan model yang memprediksi penjualan dari pengeluaran promosi dan jumlah produk tersedia. Jumlah produk tersedia saat ini dicatat pada akhir minggu, sedangkan target prediksi adalah penjualan minggu berikutnya. Data historis hanya mencakup satu periode menjelang hari raya, dan pemilik menganggap koefisien promosi sebagai bukti bahwa promosi selalu menaikkan penjualan. Rumuskan sasaran prediksi dan fitur yang layak digunakan, lalu analisis risiko kebocoran informasi, keterbatasan data, dan penafsiran kausal. Tentukan cara evaluasi dan pemantauan yang mendukung keputusan persediaan.
2. Sebuah fasilitas pengolahan hasil pertanian di Jawa Timur hendak memperkirakan konsumsi listrik berdasarkan volume produksi. Analisis awal menunjukkan tambahan produksi berkaitan dengan kenaikan listrik, tetapi kenaikannya tampak makin kecil pada volume tinggi. Model garis lurus memiliki kesalahan yang membesar pada tingkat produksi tertentu. Manajer ingin segera memakai model untuk

menentukan anggaran energi, sedangkan data operasional mencatat perubahan mesin dan jam kerja secara tidak konsisten. Evaluasi apakah regresi linear sederhana memadai, rumuskan pemeriksaan data dan alternatif pemodelan yang patut diuji, serta jelaskan rancangan evaluasi dan pemantauan sebelum hasil digunakan dalam keputusan.

BAB 8
CLUSTERING DAN SEGMENTASI DATA

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep dan pendekatan clustering untuk mengelompokkan data berdasarkan karakteristik dan tujuan analisis.
2. Menghitung penempatan objek ke dalam cluster menggunakan algoritma K-Means serta menafsirkan hasil pengelompokannya.
3. Menggunakan hierarchical clustering dengan pilihan ukuran jarak dan metode penggabungan yang sesuai untuk membentuk serta membaca dendrogram.
4. Menentukan parameter DBSCAN untuk mengidentifikasi cluster berbasis kepadatan dan membedakan anggota cluster dari noise.
5. Mengimplementasikan alur clustering pada data IoT, network traffic, dan aktivitas sistem serta menafsirkan hasilnya untuk mendukung kebutuhan Sistem Informasi.

B. Pendahuluan

Organisasi menghasilkan data dari beragam proses, mulai dari transaksi pelanggan dan log aplikasi hingga perangkat Internet of Things (IoT) dan lalu lintas jaringan.

Data tersebut sering kali belum memiliki label yang menunjukkan kategori atau pola yang relevan bagi pengambilan keputusan. Clustering menyediakan pendekatan analisis tanpa label untuk mengelompokkan objek berdasarkan kemiripan karakteristik, sehingga pola yang sebelumnya tersembunyi dapat ditelaah secara sistematis. Jain (2010) menjelaskan bahwa clustering merupakan persoalan mendasar dalam analisis data, tetapi tidak memiliki satu definisi operasional atau algoritma yang berlaku untuk semua konteks. Hasil pengelompokan dipengaruhi oleh representasi data, ukuran kemiripan, algoritma, dan parameter yang digunakan. Karena itu, clustering bukan sekadar menjalankan perangkat lunak, melainkan proses pemodelan yang menuntut pemilihan metode sesuai sifat data dan tujuan analisis.

Pemahaman tersebut relevan bagi mahasiswa Sistem Informasi karena hasil clustering dapat menghubungkan data operasional dengan kebutuhan perancangan layanan dan keputusan organisasi. Segmentasi pelanggan, misalnya, dapat membantu membedakan kelompok pengguna berdasarkan pola transaksi, sedangkan pengelompokan catatan jaringan dapat mendukung penelusuran perilaku lalu lintas yang tidak lazim. Pada lingkungan IoT, data sensor yang terus mengalir dapat membentuk kelompok pola operasi, sementara log aktivitas sistem dapat digunakan untuk mengenali perilaku proses yang serupa. Namun,

interpretasi hasil harus mempertimbangkan konteks sumber data: skala fitur yang berbeda dapat mengubah perhitungan jarak, noise dapat mengaburkan struktur kelompok, dan cluster yang terbentuk belum tentu memiliki makna bisnis atau operasional tanpa validasi. Han, Kamber, dan Pei (2012) menekankan bahwa pemilihan teknik data mining perlu mempertimbangkan karakteristik data dan tujuan penerapan; prinsip ini mendasari penggunaan clustering secara bertanggung jawab dalam analisis Sistem Informasi.

Bab ini membahas penerapan beberapa pendekatan clustering melalui konsep dan pendekatan clustering sebagai landasan untuk memahami pembentukan kelompok. Pembahasan K-Means berfokus pada penentuan centroid, pengalokasian objek, serta pengulangan proses hingga kondisi berhenti tercapai, mengikuti dasar algoritma yang diperkenalkan MacQueen (1967). Hierarchical clustering dibahas melalui penggabungan atau pemisahan kelompok dan interpretasi dendrogram, sedangkan DBSCAN memperkenalkan pengelompokan berbasis kepadatan yang dapat membedakan titik noise (Ester et al., 1996). Selanjutnya, prinsip-prinsip tersebut diterapkan pada data IoT, network traffic, dan aktivitas sistem dengan memperhatikan persiapan data, pemilihan parameter, serta penafsiran hasil. Dengan alur ini, mahasiswa dapat

menghubungkan pilihan algoritma dengan struktur data dan kebutuhan analisis yang dihadapi.

C. Konsep dan Pendekatan Clustering

Prapemrosesan data dan rekayasa fitur pada Bab 4 menyediakan dasar agar data dapat digunakan dalam pemodelan. Salah satu pendekatan pembelajaran tak terawasi yang dapat diterapkan setelah tahap tersebut adalah klasterisasi, yaitu pengelompokan objek berdasarkan kemiripan tanpa menggunakan label kelas yang telah ditentukan. Tujuan klasterisasi bukan sekadar membagi data menjadi kelompok, melainkan menghasilkan struktur kelompok yang dapat ditafsirkan sesuai kebutuhan analisis. Karena definisi kemiripan bergantung pada fitur, ukuran jarak, dan algoritma, satu kumpulan data dapat menghasilkan pengelompokan berbeda ketika pilihan-pilihan tersebut berubah (Unknown, 2012; Kononenko, 2007).

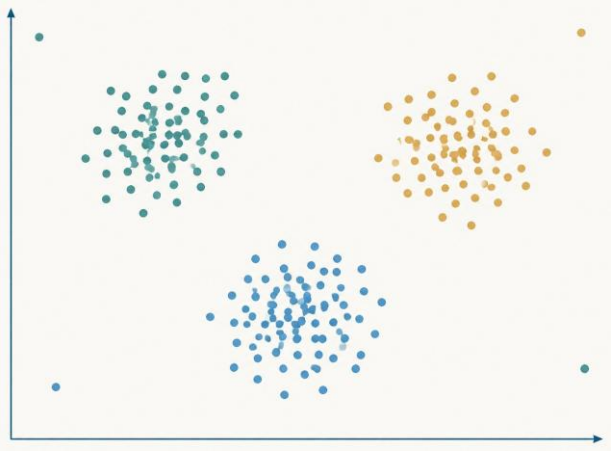
Objek dalam kumpulan data direpresentasikan sebagai titik dalam ruang fitur. Pada data numerik, kemiripan lazim dihitung melalui ukuran jarak, seperti jarak Euclidean atau Manhattan. Jarak Euclidean peka terhadap skala fitur; perbedaan suhu beberapa derajat dapat tampak kecil dibandingkan volume transaksi ribuan rupiah, meskipun keduanya sama-sama penting bagi analisis. Karena itu, fitur numerik umumnya perlu dinormalisasi atau distandardisasi

sebagaimana dibahas pada Bab 4. Untuk data kategorikal atau teks, ukuran kemiripan perlu disesuaikan dengan bentuk representasinya. Penggunaan ukuran yang tidak cocok dapat mengubah hubungan antartitik dan mengarahkan algoritma pada kelompok yang kurang bermakna (García, 2015; Unknown, 2012).

Pendekatan klasterisasi dapat dibedakan menurut cara membentuk kelompok. Metode berbasis partisi, seperti K-Means, membagi data ke dalam sejumlah kelompok yang ditentukan sebelumnya. Metode hierarkis membangun susunan kelompok bertingkat, sedangkan metode berbasis kepadatan mengidentifikasi wilayah dengan titik-titik yang berdekatan dan dapat menandai titik terpencil sebagai noise. Pilihan pendekatan sebaiknya mempertimbangkan bentuk sebaran data, ukuran kumpulan data, keberadaan noise, dan kebutuhan untuk menjelaskan hasil. Segmentasi pelanggan berbasis RFM, misalnya, dapat digunakan perusahaan transportasi untuk membedakan pola pelanggan berdasarkan kebaruan, frekuensi, dan nilai transaksi; pemaknaan tiap kelompok tetap memerlukan pengetahuan mengenai operasi dan layanan perusahaan (Priandini, 2025).

Gambar berikut menggambarkan gagasan dasar pengelompokan titik berdasarkan kedekatan pada ruang fitur. Warna dan jarak antartitik membantu memperlihatkan bahwa hasil klasterisasi merupakan

struktur yang dibentuk dari representasi data, bukan label yang secara otomatis menyatakan kategori bisnis tertentu.



Gambar 8.1 Kelompok Titik pada Ruang Fitur. Sumber: Ilustrasi, diolah penulis

Titik yang tampak berdekatan dalam gambar dapat dianggap mirip menurut fitur dan ukuran jarak yang digunakan. Namun, interpretasi sebagai segmen tertentu—misalnya pelanggan aktif atau perangkat dengan pola operasi sejenis—harus didukung pemeriksaan karakteristik masing-masing kelompok.

Hasil klasterisasi perlu diperiksa dari sisi teknis dan kontekstual. Secara teknis, analis dapat menilai kekompakan anggota dalam kelompok dan keterpisahannya dari kelompok lain, lalu membandingkan hasil ketika jumlah

kelompok atau parameter diubah. Secara kontekstual, profil fitur tiap kelompok perlu ditelaah untuk memastikan bahwa perbedaan yang ditemukan relevan terhadap tujuan analisis. Klaster yang secara matematis terpisah belum tentu berguna bagi keputusan organisasi jika tidak berkaitan dengan tindakan yang dapat dilakukan. Pendekatan tersebut sejalan dengan penggunaan klasterisasi sebagai sarana eksplorasi pola, bukan sebagai pengganti penilaian terhadap konteks data (Unknown, 2012; Kononenko, 2007).

D. K-Means Clustering

K-Means merupakan algoritma berbasis partisi yang membagi objek ke dalam K kelompok. Setiap kelompok diwakili oleh centroid, yaitu titik pusat yang dihitung dari rata-rata fitur anggotanya. Algoritma dimulai dengan menetapkan K centroid awal, menempatkan setiap objek pada centroid terdekat, lalu menghitung ulang centroid berdasarkan anggota kelompok saat itu. Proses penempatan dan pembaruan diulang hingga keanggotaan tidak berubah, perubahan centroid cukup kecil, atau batas iterasi tercapai. Sasaran perhitungannya adalah memperkecil jumlah kuadrat jarak objek terhadap centroid kelompoknya (Unknown, 2012).

Misalkan kumpulan data berisi objek x_i dan centroid kelompok k adalah μ_k. Fungsi objektif K-Means dapat dituliskan sebagai berikut.

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Nilai J mengukur variasi dalam kelompok berdasarkan jarak kuadrat. Karena centroid awal dapat memengaruhi hasil, algoritma dapat dijalankan beberapa kali dengan inisialisasi berbeda, kemudian hasilnya dibandingkan. Fitur juga perlu disiapkan dengan skala yang sebanding; jika satu fitur memiliki rentang jauh lebih besar, kontribusinya terhadap jarak dapat mendominasi fitur lain. Pengaruh skala dan pemilihan representasi perlu dipertimbangkan sebagai bagian dari prapemrosesan data (García, 2015).

Contoh berikut memakai empat objek dengan dua fitur numerik dan dua centroid awal yang telah ditetapkan. Jarak dihitung menggunakan Euclidean, kemudian setiap objek ditempatkan pada centroid terdekat. Data ini merupakan ilustrasi perhitungan, bukan hasil pengukuran dari perusahaan tertentu.

Tabel 8.1 Penempatan Objek pada Iterasi Awal K-Means

Objek	Fitur x	Fitur y	Jarak C1	Jarak C2
P1	1	1	0,00	5,00
P2	1	2	1,00	4,47
P3	4	4	4,24	1,00
P4	5	4	5,00	0,00

P1 dan P2 ditempatkan pada kelompok C1, sedangkan P3 dan P4 ditempatkan pada C2. Setelah centroid diperbarui,

pusat kelompok pertama menjadi $(1,1,5)$ dan pusat kelompok kedua menjadi $(4,5,4)$; iterasi berikutnya menguji kembali penempatan berdasarkan pusat baru.

Dalam penerapan di Indonesia, segmentasi pelanggan perusahaan transportasi dapat menggunakan fitur RFM untuk membedakan pelanggan yang sering bertransaksi dari pelanggan yang sudah lama tidak menggunakan layanan. Studi Priandini, Asfi, dan Magdalena menggunakan K-Means berbasis RFM dalam sistem business intelligence untuk mendukung strategi pengelolaan pelanggan, sehingga hasil kelompok dapat menjadi masukan bagi perancangan tindak lanjut yang berbeda (Priandini, 2025). Studi segmentasi pelanggan berdasarkan kepribadian dan pola konsumsi juga menunjukkan konteks penggunaan K-Means untuk menyusun kelompok pelanggan, dengan interpretasi yang bergantung pada fitur yang dipilih (Iqbal, 2025).

K-Means paling sesuai ketika jumlah kelompok dapat ditentukan dan kelompok yang dicari relatif kompak menurut ukuran jarak yang digunakan. Algoritma ini memerlukan kehati-hatian terhadap pencilan karena nilai ekstrem dapat menggeser centroid. Selain itu, hasil klasterisasi tidak menyediakan nama segmen secara otomatis. Analis perlu memeriksa rata-rata fitur setiap kelompok dan menghubungkannya dengan tujuan,

misalnya menyusun prioritas layanan atau menyesuaikan penawaran bagi kelompok pelanggan tertentu (Unknown, 2012; Priandini, 2025).

E. Hierarchical Clustering

Hierarchical clustering membentuk struktur kelompok bertingkat yang dapat dilihat melalui dendrogram. Berbeda dari K-Means yang meminta jumlah kelompok sebelum proses pengelompokan, pendekatan hierarkis memungkinkan analis meninjau hubungan penggabungan atau pemisahan pada beberapa tingkat. Pada pendekatan agglomerative, setiap objek mula-mula menjadi kelompok tersendiri, kemudian pasangan kelompok yang paling dekat digabungkan secara bertahap. Pada pendekatan divisive, seluruh objek bermula sebagai satu kelompok, lalu kelompok tersebut dipecah menjadi bagian yang lebih kecil. Perbandingan K-Means dan agglomerative hierarchical clustering dalam konteks pengolahan data menunjukkan bahwa keduanya membentuk kelompok melalui mekanisme yang berbeda dan menghasilkan struktur keluaran yang berbeda pula (Hartono, 2025; Unknown, 2012).

Pada metode agglomerative, ukuran jarak antarkelompok ditentukan oleh aturan linkage. Single linkage memakai jarak terdekat antara anggota kedua kelompok, sedangkan complete linkage memakai jarak

terjauh. Average linkage menghitung rata-rata jarak antarpasangan anggota, sementara Ward memilih penggabungan yang menghasilkan kenaikan terkecil pada variasi dalam kelompok. Pilihan linkage memengaruhi urutan penggabungan dan bentuk dendrogram; oleh karena itu, metode tersebut perlu dipilih sesuai struktur data dan tujuan analisis, bukan semata-mata mengikuti pengaturan bawaan perangkat lunak (Unknown, 2012).

Dendrogram menampilkan tinggi penggabungan sebagai petunjuk mengenai jarak atau perubahan saat dua kelompok disatukan. Pemotongan dendrogram pada ketinggian tertentu menghasilkan sejumlah kelompok; semakin tinggi pemotongan, umumnya semakin sedikit kelompok yang terbentuk. Keputusan menentukan titik potong dapat dibantu dengan membandingkan perubahan jarak antarpenggabungan dan meninjau profil fitur kelompok yang dihasilkan. Interpretasi tetap harus mengacu pada konteks, sebab cabang yang terpisah jauh secara visual belum tentu merepresentasikan kategori operasional yang berguna (Hartono, 2025; Unknown, 2012).

Sebagai contoh di Indonesia, sebuah pemerintah daerah dapat menggunakan data indikator layanan pada kecamatan untuk mengeksplorasi kemiripan wilayah. Setelah fitur seperti waktu layanan dan volume permohonan disetarakan skalanya, pendekatan hierarkis dapat membantu menampilkan kecamatan yang bergabung

pada tingkat kemiripan berbeda. Hasilnya dapat digunakan sebagai bahan penelaahan pola pelayanan, tetapi tidak dengan sendirinya membuktikan bahwa kecamatan dalam satu kelompok memiliki kebutuhan kebijakan yang identik. Perbedaan kondisi lokal perlu diperiksa sebelum kelompok dijadikan dasar pengalokasian sumber daya.

Penerapan hierarkis juga menuntut perhatian terhadap ukuran kumpulan data dan pilihan jarak. Karena hubungan antarkelompok diperiksa berulang kali, proses dapat menjadi lebih berat ketika jumlah objek meningkat. Data campuran yang menggabungkan fitur numerik dan kategorikal memerlukan ukuran kemiripan yang sesuai, sementara fitur yang tidak relevan dapat menghasilkan penggabungan yang menyesatkan. Perbandingan dengan K-Means pada data yang sama dapat membantu menguji kestabilan interpretasi, sebagaimana perbandingan algoritma klasterisasi digunakan untuk memahami perbedaan proses dan hasil pengelompokan (Hartono, 2025; García, 2015).

F. DBSCAN dan Density-Based Clustering

DBSCAN adalah pendekatan berbasis kepadatan yang membentuk kelompok dari wilayah dengan konsentrasi titik yang memadai. Algoritma ini menggunakan dua parameter utama: ε (epsilon), yaitu radius pencarian tetangga, dan MinPts, yaitu jumlah minimum titik yang diperlukan agar

suatu wilayah dianggap padat. Titik inti memiliki tetangga sebanyak atau lebih dari MinPts dalam radius ε. Titik yang berada dalam lingkungan titik inti dapat menjadi titik batas, sedangkan titik yang tidak terjangkau oleh kelompok padat dapat ditandai sebagai noise. Dengan demikian, DBSCAN tidak mengharuskan jumlah kelompok ditentukan terlebih dahulu (Unknown, 2012).

Proses penerapannya dimulai dengan memilih titik yang belum diperiksa dan mencari tetangga dalam radius ε. Jika jumlah tetangga memenuhi MinPts, titik tersebut menjadi titik inti dan kelompok diperluas melalui titik-titik yang kepadatannya terhubung. Jika syarat kepadatan tidak terpenuhi, titik dapat sementara ditandai sebagai noise; titik tersebut masih mungkin menjadi titik batas ketika ditemukan dalam jangkauan titik inti. Pengertian kepadatan lokal ini membuat DBSCAN berguna saat data memiliki noise atau kelompok dengan bentuk tidak bulat, berbeda dari asumsi partisi kompak yang lazim pada K-Means (Unknown, 2012).

Pemilihan parameter perlu disesuaikan dengan skala fitur dan sebaran data. Jika ε terlalu kecil, banyak titik dapat menjadi noise dan kelompok terpecah; jika terlalu besar, kelompok berbeda dapat bergabung. MinPts yang terlalu rendah dapat menghasilkan kelompok dari kepadatan yang lemah, sedangkan nilai terlalu tinggi dapat mengabaikan pola lokal yang sah. Menstandarkan fitur numerik sebelum

menghitung jarak membantu mencegah satuan tertentu mendominasi radius. Untuk data berdimensi tinggi atau kepadatan yang sangat bervariasi, satu pasangan parameter mungkin tidak menggambarkan semua pola secara memadai (García, 2015; Unknown, 2012).

Dalam konteks Indonesia, pengelola jaringan kampus di Yogyakarta dapat menerapkan DBSCAN pada ringkasan sesi koneksi berdasarkan durasi, volume data, dan jumlah permintaan. Kelompok padat dapat menunjukkan pola penggunaan jaringan yang berulang, sementara sesi yang berada di luar kelompok dapat diprioritaskan untuk pemeriksaan lebih lanjut. Titik noise tidak otomatis berarti serangan; perubahan jadwal kuliah, pemeliharaan layanan, atau perilaku pengguna yang sah juga dapat menghasilkan pola yang jarang. Analisis keamanan perlu menggabungkan hasil pengelompokan dengan konteks operasional dan prosedur pemeriksaan ancaman yang dibahas lebih lanjut pada Bab 11 (Dua, 2016).

Kekuatan DBSCAN dalam mengenali titik yang tidak termasuk kelompok padat menjadikannya berguna untuk eksplorasi data lalu lintas jaringan, tetapi hasilnya tetap dipengaruhi mutu fitur dan parameter. Jika pola normal memiliki kepadatan berbeda antarsegmen jaringan, algoritma dapat mengelompokkan satu segmen dengan baik tetapi menandai banyak titik pada segmen lain sebagai noise. Karena itu, analis sebaiknya menguji beberapa

kombinasi parameter, memeriksa jumlah anggota kelompok dan noise, serta meninjau contoh catatan yang mewakili tiap hasil. Penggunaan seperti ini menempatkan DBSCAN sebagai alat untuk menemukan struktur dan kandidat pemeriksaan, bukan sebagai keputusan keamanan otomatis (Dua, 2016; Unknown, 2012).

G. Clustering Data IoT, Network Traffic, dan Aktivitas Sistem

Penerapan klasterisasi pada data operasional dimulai dengan menentukan unit analisis dan jangka waktu pengamatan. Pada Internet untuk Segala (IoT), unit dapat berupa pembacaan sensor per interval, ringkasan kondisi perangkat, atau agregat pengukuran per jam. Untuk network traffic, unit analisis dapat berupa sesi koneksi, sedangkan pada aktivitas sistem unitnya dapat berupa rangkaian catatan log yang diringkas per pengguna atau proses. Penentuan unit tersebut memengaruhi makna kelompok: pengelompokan pembacaan individual menjelaskan pola pengukuran, sementara pengelompokan agregat waktu dapat menggambarkan keadaan operasi yang lebih luas (Unknown, 2012; Dua, 2016).

Prapemrosesan dan pemilihan fitur harus mempertimbangkan sumber data. Data sensor dapat memuat suhu, kelembapan, atau indikator gas; beberapa nilai perlu diringkas agar pola operasi perangkat dapat

dibandingkan. Sistem monitoring kualitas udara berbasis IoT di Indonesia, misalnya, mengumpulkan pengukuran suhu, kelembapan, dan gas menggunakan sensor, sehingga fitur tersebut dapat menjadi dasar eksplorasi pola kondisi lingkungan setelah kualitas data diperiksa (Alfirdaus, 2024). Pada sistem monitoring kateter berbasis IoT, data yang dikumpulkan memiliki konteks klinis khusus, sehingga pengelompokan harus dilakukan dengan mempertimbangkan batas penggunaan dan interpretasi oleh petugas yang berwenang (Banjarnahor, 2022).

Network traffic dan log aktivitas memerlukan fitur yang merangkum perilaku tanpa menghilangkan konteks penting. Fitur sesi dapat mencakup durasi, jumlah paket, volume data, atau waktu terjadinya koneksi, sedangkan log aplikasi dapat diringkas berdasarkan frekuensi kejadian, jenis operasi, dan rentang waktu. Kolom identitas sensitif sebaiknya tidak langsung diperlakukan sebagai fitur numerik; representasi, akses, dan perlindungan data perlu dipertimbangkan sebelum pemodelan. Untuk konteks keamanan, klasterisasi dapat membantu mengungkap pola aktivitas serupa, tetapi tidak menggantikan analisis ancaman atau evaluasi keamanan yang dibahas pada Bab 11 dan Bab 12 (Dua, 2016).

Alur penerapan berikut dapat digunakan sebagai kerangka kerja yang dapat disesuaikan dengan sumber data. Tahapannya menghubungkan tujuan operasional dengan

persiapan fitur dan pemeriksaan hasil, tanpa menetapkan satu algoritma untuk semua kasus.

Tabel 8.2 Alur Penerapan Klasterisasi pada Data Operasional

Tahap	Data IoT	Network traffic	Aktivitas sistem
Unit analisis	Pembacaan per interval	Sesi koneksi	Ringkasan log
Fitur contoh	Suhu, kelembapan	Durasi, volume	Frekuensi kejadian
Persiapan	Cek sensor dan skala	Ringkas per sesi	Kelompokkan per waktu
Algoritma awal	K-Means atau DBSCAN	DBSCAN atau hierarkis	K-Means atau hierarkis
Pemeriksaan hasil	Profil kondisi sensor	Tinjau sesi noise	Tafsir pola aktivitas

Tabel tersebut menunjukkan bahwa pemilihan algoritma mengikuti karakteristik data dan pertanyaan operasional, bukan sekadar ketersediaan perangkat lunak. Misalnya, DBSCAN dapat dipertimbangkan ketika noise menjadi perhatian, sedangkan hierarkis berguna untuk menelaah hubungan bertingkat antarpola.

Pada penerapan nyata, hasil klasterisasi perlu disajikan bersama ringkasan profil kelompok, rentang waktu, dan batasan analisis. Operator dapat memeriksa apakah satu kelompok mencerminkan operasi normal, kondisi khusus, atau perubahan sistem yang telah direncanakan. Pada perangkat IoT, perubahan sensor atau penempatan perangkat dapat mengubah distribusi data; pada jaringan, pembaruan layanan dapat membentuk pola koneksi baru.

Oleh sebab itu, kelompok yang terbentuk pada satu periode tidak seharusnya dianggap tetap berlaku tanpa pemeriksaan ulang. Pengujian lintas periode dan konfirmasi oleh pemilik proses membantu memastikan bahwa pola yang ditemukan tetap sesuai dengan kondisi lapangan (Alfirdaus, 2024; Dua, 2016).

H. Rangkuman

Klasterisasi merupakan pendekatan pembelajaran tak terawasi yang mengelompokkan objek berdasarkan kemiripan tanpa label kelas yang telah ditentukan. Hasilnya bergantung pada representasi fitur, ukuran kemiripan, dan algoritma yang dipilih; perubahan pada unsur-unsur tersebut dapat menghasilkan kelompok berbeda. Pada data numerik, jarak Euclidean dan Manhattan lazim digunakan, tetapi skala fitur perlu diperhatikan. Tanpa normalisasi atau standardisasi, fitur dengan rentang nilai besar dapat mendominasi perhitungan jarak. Data kategorikal dan teks juga memerlukan representasi serta ukuran kemiripan yang sesuai. Tiga pendekatan utama memiliki karakteristik berbeda: metode berbasis partisi, seperti K-Means, membentuk sejumlah kelompok yang ditentukan sebelumnya; metode hierarkis menyusun kelompok secara bertingkat; sementara metode berbasis kepadatan mengenali wilayah dengan titik-titik berdekatan dan dapat menandai titik terpencil sebagai noise. K-Means menetapkan centroid awal, menempatkan objek pada

centroid terdekat, lalu memperbarui pusat kelompok berdasarkan rata-rata fitur anggotanya. Pengulangan berlangsung sampai keanggotaan atau posisi centroid cukup stabil, dengan sasaran memperkecil jumlah kuadrat jarak objek terhadap centroid kelompok. Inisialisasi, skala fitur, dan pencilan dapat memengaruhi hasil sehingga perlu dipertimbangkan dalam penerapan. Dalam Sistem Informasi, segmentasi pelanggan menggunakan fitur RFM—kebaruan, frekuensi, dan nilai transaksi—dapat membantu organisasi merancang tindak lanjut yang berbeda. Namun, algoritma tidak otomatis memberi nama atau makna bisnis pada kelompok. Analis perlu memeriksa profil fitur tiap kelompok, menilai kekompakan dan keterpisahannya, serta membandingkan hasil ketika jumlah kelompok atau parameter diubah. Interpretasi juga harus mempertimbangkan pengetahuan operasional dan tujuan organisasi. Dengan demikian, klasterisasi berfungsi sebagai sarana eksplorasi pola dan pendukung keputusan, bukan pengganti evaluasi kontekstual. Kelompok yang terpisah secara matematis baru bernilai praktis apabila karakteristiknya dapat ditafsirkan dan dihubungkan dengan tindakan yang relevan.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana perbedaan skala antara nilai transaksi dan frekuensi pembelian dapat memengaruhi

hasil K-Means pada data pelanggan. Rumuskan langkah prapemrosesan yang sesuai dan jelaskan bagaimana analis dapat memeriksa apakah perubahan tersebut menghasilkan segmentasi yang lebih bermakna.

2. Bandingkan metode berbasis partisi, hierarkis, dan kepadatan untuk mengelompokkan data operasional yang memiliki bentuk sebaran beragam serta sejumlah pencilan. Evaluasi pendekatan yang paling sesuai jika organisasi belum mengetahui jumlah kelompok dan perlu membedakan titik terpencil dari kelompok utama.
3. Evaluasi pengaruh centroid awal terhadap hasil K-Means dengan merujuk pada proses penempatan objek, pembaruan centroid, dan fungsi objektifnya. Argumentasikan strategi yang dapat digunakan analis untuk mengurangi risiko memilih hasil yang bergantung pada inisialisasi tertentu.
4. Telaah mengapa kekompakan dan keterpisahan klaster belum cukup untuk membuktikan bahwa segmentasi pelanggan bermanfaat bagi organisasi. Rumuskan pemeriksaan tambahan yang menghubungkan profil kelompok, tujuan analisis, dan tindakan bisnis yang mungkin dilakukan.
5. Analisis penggunaan fitur RFM untuk mengelompokkan pelanggan perusahaan transportasi. Evaluasi bagaimana kombinasi kebaruan, frekuensi, dan nilai transaksi dapat mendukung tindak lanjut berbeda, serta jelaskan batas interpretasi yang perlu dijaga sebelum kelompok diberi nama segmen bisnis.

b. Pilihan Berganda

1. Mengapa fitur numerik umumnya perlu dinormalisasi atau distandardisasi sebelum menghitung jarak Euclidean?
    a. Agar seluruh objek memperoleh label kelas yang sama
    b. Agar fitur dengan rentang besar tidak mendominasi perhitungan jarak
    c. Agar jumlah kelompok dapat ditentukan secara otomatis
    d. Agar pencilan selalu dikeluarkan dari kumpulan data
    e. Agar centroid tidak perlu diperbarui
2. Dalam K-Means, bagaimana centroid baru suatu kelompok dihitung setelah objek ditempatkan?
    a. Dengan memilih objek yang paling jauh dari pusat lama
    b. Dengan mengambil nilai maksimum setiap fitur
    c. Dengan menghitung rata-rata fitur anggota kelompok
    d. Dengan menghapus anggota yang jaraknya paling besar
    e. Dengan menghitung rata-rata seluruh objek tanpa memperhatikan kelompok
3. Apa sasaran fungsi objektif K-Means yang dijelaskan dalam bab?

a. Memaksimalkan jarak antartitik dalam setiap kelompok
b. Meminimalkan jumlah kelompok yang terbentuk
c. Memaksimalkan jumlah fitur yang digunakan
d. Memperkecil jumlah kuadrat jarak objek terhadap centroid kelompoknya
e. Menghilangkan seluruh titik pencilan dari data

4. Seorang analis ingin menemukan wilayah data yang padat sekaligus menandai titik terpencil sebagai noise. Pendekatan yang paling sesuai adalah metode berbasis....
   a. partisi
   b. centroid
   c. kepadatan
   d. standardisasi
   e. hierarki kelas

5. Pernyataan yang tepat mengenai hierarchical clustering adalah....
   a. Jumlah kelompok harus selalu ditentukan sebelum proses pengelompokan
   b. Struktur kelompok bertingkat dapat ditinjau melalui dendrogram
   c. Semua kelompok diwakili oleh rata-rata fitur sebagaimana pada K-Means
   d. Setiap objek harus ditempatkan pada centroid terdekat

e. Metode ini hanya dapat digunakan pada data berlabel

6. Apa konsekuensi yang mungkin terjadi jika satu fitur memiliki rentang nilai jauh lebih besar daripada fitur lain dalam perhitungan jarak?
   a. Fitur tersebut dapat mendominasi pengukuran kemiripan
   b. Fitur lain otomatis dihapus dari data
   c. Semua centroid akan berada pada titik yang sama
   d. Jumlah kelompok bertambah secara otomatis
   e. Objek akan memperoleh label kelas yang benar
7. Mengapa K-Means dapat menghasilkan pengelompokan berbeda ketika dijalankan kembali?
   a. Algoritma selalu mengganti ukuran jarak pada setiap iterasi
   b. Setiap iterasi menghapus fitur yang telah digunakan
   c. Centroid awal yang berbeda dapat memengaruhi hasil
   d. K-Means menentukan label berdasarkan konteks bisnis
   e. Data numerik berubah menjadi kategorikal pada iterasi akhir
8. Perusahaan ingin mengelompokkan pelanggan berdasarkan kebaruan, frekuensi, dan nilai transaksi. Kombinasi fitur yang digunakan dikenal sebagai....
   a. PCA

b. RFM
c. Euclidean
d. centroid
e. dendrogram

9. Penilaian teknis terhadap hasil klasterisasi dapat mencakup pemeriksaan....
   a. kekompakan anggota kelompok dan keterpisahan antarkelompok
   b. nama segmen yang ditetapkan oleh algoritma
   c. kesesuaian warna pada visualisasi kelompok
   d. jumlah label kelas yang tersedia
   e. banyaknya tindakan pemasaran yang telah dilakukan

10. Mengapa kelompok hasil klasterisasi tidak boleh langsung dianggap sebagai kategori bisnis yang pasti?
    a. Algoritma klasterisasi selalu menghasilkan satu kelompok
    b. Setiap kelompok hanya dapat berisi data kategorikal
    c. Hasil kelompok perlu ditafsirkan berdasarkan profil fitur dan konteks organisasi
    d. Klasterisasi tidak dapat digunakan untuk eksplorasi pola
    e. Ukuran jarak selalu menghasilkan kelompok yang sama

c. Soal Project / Studi Kasus

1. Sebuah perusahaan transportasi ingin menyusun program retensi pelanggan dari riwayat transaksi selama satu tahun. Tim memiliki data kebaruan transaksi, frekuensi perjalanan, dan nilai belanja, tetapi rentang nilai belanja jauh lebih besar daripada dua fitur lainnya. Manajer meminta empat segmen yang dapat langsung diberi nama dan ditindaklanjuti. Analis juga menemukan beberapa akun dengan transaksi bernilai sangat tinggi, tetapi jarang menggunakan layanan. Rancang alur analisis yang mencakup prapemrosesan, pemilihan dan penerapan metode, pemeriksaan hasil, serta cara merumuskan tindak lanjut tanpa menganggap nama segmen muncul otomatis dari algoritma.
2. Unit pengelola fasilitas kampus mengumpulkan data sensor suhu dan getaran dari mesin selama beberapa bulan. Sebagian mesin membentuk pola operasi yang relatif teratur, sementara beberapa catatan menunjukkan nilai ekstrem yang diduga berasal dari kerusakan atau gangguan sensor. Tim belum mengetahui jumlah pola operasi yang ada dan ingin menghindari agar titik ekstrem menggeser pusat kelompok. Telaah pilihan metode klasterisasi yang relevan, tentukan langkah persiapan data, dan rumuskan cara mengevaluasi apakah kelompok yang ditemukan berguna untuk pemeliharaan tanpa mengabaikan kemungkinan kesalahan pengukuran.

BAB 9
ASSOCIATION RULE DAN PATTERN MINING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep frequent pattern dan association rule untuk mengidentifikasi keterkaitan antarelemen dalam data transaksi.
2. Menghitung nilai support, confidence, dan lift serta menggunakan metrik tersebut untuk menilai kekuatan aturan asosiasi.
3. Menyelesaikan proses pencarian frequent itemset menggunakan algoritma Apriori berdasarkan ambang minimum support yang ditetapkan.
4. Memilih dan menerapkan FP-Growth atau sequential pattern mining sesuai karakteristik data dan tujuan analisis.
5. Mengimplementasikan pendekatan pattern mining pada kasus data IoT dan keamanan sistem dengan mempertimbangkan konteks serta batasan interpretasi hasil.

B. Pendahuluan

Organisasi mengumpulkan data transaksi, log aktivitas, dan aliran sensor dalam jumlah besar, tetapi nilai analitis

data tersebut tidak selalu tampak melalui ringkasan statistik biasa. Pola keterkaitan yang berulang dapat membantu menjelaskan produk yang kerap dibeli bersamaan, urutan layanan yang digunakan pelanggan, atau kombinasi kejadian yang mendahului anomali pada sistem. Association rule dan pattern mining menyediakan pendekatan untuk menemukan keteraturan semacam itu secara sistematis. Agrawal, Imieliński, dan Swami (1993) memperkenalkan pencarian aturan asosiasi pada basis data transaksi, dengan tujuan mengungkap hubungan antarelemen tanpa harus merumuskan hipotesis hubungan secara terperinci sejak awal. Bagi mahasiswa Sistem Informasi, pendekatan ini relevan karena menghubungkan pengelolaan data dengan kebutuhan operasional dan pengambilan keputusan, seperti perancangan promosi, pengaturan persediaan, dan pemantauan layanan digital. Namun, aturan yang ditemukan perlu ditafsirkan secara cermat: kemunculan dua item secara bersamaan menunjukkan asosiasi, bukan bukti bahwa satu item menyebabkan item lainnya.

Penerapan pattern mining bertumpu pada kemampuan membedakan pola yang cukup sering muncul dari pola yang hanya terjadi sesekali. Support mengukur frekuensi kemunculan suatu itemset dalam keseluruhan transaksi, sedangkan confidence menunjukkan proporsi transaksi yang memuat konsekuen di antara transaksi yang memuat anteseden. Lift membandingkan kekuatan asosiasi tersebut

dengan frekuensi dasar konsekuen, sehingga membantu menilai apakah hubungan yang tampak melampaui kebetulan akibat popularitas item. Ketiga ukuran ini perlu digunakan secara terpadu karena confidence tinggi, misalnya, belum tentu menunjukkan asosiasi yang kuat apabila konsekuennya memang sangat umum. Pada tingkat komputasi, Apriori memanfaatkan sifat bahwa setiap subset dari frequent itemset juga harus frequent untuk memangkas kandidat, tetapi pembangkitan kandidat berulang dapat membebani pemrosesan data besar. Han, Pei, dan Yin (2000) mengusulkan FP-Growth yang memadatkan basis data ke dalam struktur FP-tree dan menambang pola tanpa pembangkitan kandidat eksplisit. Untuk data yang memiliki dimensi waktu, sequential pattern mining memperluas analisis dengan memperhatikan urutan kemunculan item, bukan hanya koeksistensinya.

Ragam data yang kini dikelola organisasi memperluas penerapan metode tersebut sekaligus meningkatkan tuntutan kehati-hatian. Data Internet of Things (IoT), misalnya, dapat merekam urutan pembacaan sensor, perubahan status perangkat, dan kejadian jaringan; pola berulang di dalamnya dapat mendukung pemeliharaan prediktif atau membantu mengenali rangkaian aktivitas yang mencurigakan. Dalam keamanan sistem, kombinasi peristiwa pada log dapat digunakan untuk menyusun indikator awal gangguan, tetapi pola yang ditemukan harus

divalidasi agar lonjakan aktivitas normal tidak keliru diklasifikasikan sebagai serangan. Dengan demikian, proses analisis mencakup pemilihan ambang frekuensi, pemilihan algoritma, pemeriksaan kualitas data, dan penafsiran hasil sesuai konteks bisnis maupun teknis. Bab ini membahas konsep frequent pattern dan association rule, ukuran support, confidence, dan lift, algoritma Apriori, FP-Growth dan sequential pattern mining, serta penerapan pattern mining pada data IoT dan keamanan sistem.

C. Konsep Frequent Pattern dan Association Rule

Pembahasan pada bab sebelumnya telah memperkenalkan klasterisasi untuk menemukan kelompok data tanpa label. Bab ini beralih pada tujuan yang berbeda: menemukan kombinasi elemen yang berulang dan hubungan keterjadiannya di dalam kumpulan data. Analisis tersebut berguna ketika organisasi belum menetapkan hipotesis hubungan tertentu, tetapi ingin menelusuri keteraturan yang dapat mendukung keputusan operasional (Unknown, 2012; Kononenko, 2007).

Frequent pattern adalah pola yang muncul berulang dalam kumpulan data dan memenuhi ambang frekuensi yang ditentukan. Dalam data transaksi, elemen yang dianalisis disebut *item*, sedangkan gabungan satu atau lebih item disebut *itemset*. Sebuah transaksi dapat memuat

sejumlah item, misalnya satu keranjang belanja berisi roti, susu, dan telur. Itemset dinyatakan frequent apabila jumlah atau proporsi transaksi yang memuatnya mencapai minimum support. Ambang ini berfungsi sebagai penyaring: pola yang terlalu jarang tidak diteruskan ke tahap analisis aturan, meskipun mungkin menarik untuk kasus tertentu (Unknown, 2011; Unknown, 2019).

Association rule atau aturan asosiasi menyatakan keterkaitan bersyarat antara dua kelompok item. Bentuk umumnya adalah $X \rightarrow Y$, dengan X sebagai anteseden dan Y sebagai konsekuen; kedua kelompok tersebut tidak tumpang tindih. Aturan itu menunjukkan bahwa transaksi yang memuat X juga cenderung memuat Y, bukan bahwa X menyebabkan Y. Pembedaan ini penting karena keputusan bisnis yang didasarkan pada asosiasi tidak boleh ditafsirkan sebagai hubungan sebab-akibat tanpa rancangan pengujian tambahan (Unknown, 2012; Kononenko, 2007).

Sebagai ilustrasi di Indonesia, sebuah UMKM roti di Bandung dapat mengubah catatan penjualan harian menjadi transaksi per nota. Apabila roti tawar dan selai berulang kali muncul bersama, pola itu dapat dipertimbangkan untuk penataan rak atau paket promosi. Penelitian penerapan Apriori untuk menyusun kombinasi menu makanan bergizi memperlihatkan bagaimana pola keterkaitan item dapat digunakan dalam konteks pengambilan keputusan menu; hasilnya tetap perlu

dipertimbangkan bersama kebutuhan gizi dan kebijakan penyusunan menu, bukan diterapkan hanya karena frekuensi kemunculan (Angga Wijaya, 2026).

Pola asosiasi dapat disajikan sebagai kombinasi item yang sering muncul, kemudian dikembangkan menjadi aturan yang dapat dibaca pengelola. Sebelum aturan dibentuk, kualitas representasi transaksi perlu dipastikan: satu baris harus mewakili unit analisis yang konsisten, kode item tidak boleh tercampur, dan duplikasi transaksi perlu ditangani. Keputusan tentang apakah transaksi mewakili nota, kunjungan, atau pelanggan akan mengubah interpretasi pola. Karena itu, hasil mining harus selalu dibaca sesuai definisi unit dan cakupan data yang dipakai (Unknown, 2011; Kononenko, 2007).

Visualisasi berikut menunjukkan konteks nyata pengumpulan data transaksi di gerai ritel, yaitu saat barang dipindai dan dicatat sebagai bagian dari satu transaksi. Adegan tersebut menghubungkan konsep itemset dengan sumber data yang lazim digunakan dalam analisis asosiasi.

Pemindaian barang menjelaskan bagaimana item dapat tercatat sebagai satu kelompok transaksi tanpa harus menganggap keterkaitannya sebagai sebab-akibat. Data transaksi yang terbentuk menjadi bahan untuk menghitung frekuensi pola dan menilai aturan pada bagian berikutnya.

D. Support, Confidence, dan Lift

Frequent pattern dan aturan asosiasi memerlukan ukuran kuantitatif agar pola tidak dinilai hanya berdasarkan kesan. Support, confidence, dan lift menjawab pertanyaan yang berbeda: seberapa luas pola muncul dalam seluruh data, seberapa sering konsekuen ditemukan ketika anteseden hadir, dan apakah keterkaitan itu lebih kuat daripada kemunculan dasar konsekuen. Ketiganya lazim dibaca bersama karena satu metrik saja dapat memberi gambaran yang menyesatkan (Unknown, 2012; Unknown, 2019).

Support untuk itemset X adalah proporsi transaksi yang memuat seluruh item dalam X.

$$\text{support}(X) = \frac{\text{jumlah transaksi yang memuat } X}{\text{jumlah seluruh transaksi}}$$

Untuk aturan $X \rightarrow Y$, confidence dihitung sebagai proporsi transaksi yang memuat X dan Y di antara transaksi yang memuat X.

$$\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

Lift membandingkan confidence aturan dengan support konsekuennya.

$$\text{lift}(X \rightarrow Y) = \frac{\text{confidence}(X \rightarrow Y)}{\text{support}(Y)}$$

Lift lebih dari satu menunjukkan asosiasi positif relatif terhadap frekuensi dasar konsekuen; nilai satu menunjukkan bahwa kemunculan bersama tidak melampaui tingkat yang diharapkan dari frekuensi tersebut, sedangkan nilai di bawah satu menunjukkan asosiasi negatif. Interpretasi itu tidak dengan sendirinya membuktikan kausalitas. Dalam penerapan, ambang minimum support membantu membatasi pola yang layak diuji, sementara confidence dan lift membantu menyaring serta menafsirkan aturan yang dihasilkan (Unknown, 2012; Unknown, 2019).

Tabel berikut menggunakan lima transaksi ilustratif dari sebuah toko bahan pangan di Yogyakarta. Tabel ini menampilkan isi tiap transaksi agar support item dan pasangan item dapat diturunkan langsung dari data.

Tabel 9.1 Transaksi Ilustratif Toko Bahan Pangan

Transaksi	Roti	Susu	Telur
T1	Ya	Ya	Ya
T2	Ya	Ya	Tidak
T3	Ya	Tidak	Ya
T4	Tidak	Ya	Ya
T5	Ya	Ya	Ya

Dari lima transaksi, roti muncul empat kali dan susu muncul empat kali, sehingga support masing-masing adalah $4/5 = 0{,}80$. Roti dan susu muncul bersama dalam tiga

transaksi, sehingga support pasangan itu adalah $3/5 = 0{,}60$. Confidence aturan roti menuju susu adalah $3/4 = 0{,}75$, sedangkan support susu sebesar $0{,}80$ menghasilkan lift $0{,}75/0{,}80 = 0{,}9375$. Jadi, walaupun roti dan susu cukup sering dibeli bersama, aturan tersebut tidak menunjukkan asosiasi positif dibandingkan frekuensi dasar susu. Contoh ini menegaskan bahwa confidence yang tampak tinggi belum cukup untuk menyatakan suatu aturan kuat (Unknown, 2011; Unknown, 2012).

Pada data penjualan di Indonesia, misalnya transaksi menu sebuah kantin kampus di Surabaya, pengelola dapat memakai support untuk mengetahui apakah pasangan menu cukup sering muncul untuk ditindaklanjuti. Confidence dapat membantu memperkirakan proporsi pembeli satu menu yang juga memilih menu lain, sedangkan lift memberi konteks terhadap popularitas menu konsekuen. Penerapan Apriori pada penentuan kombinasi menu bergizi menunjukkan relevansi pola transaksi untuk mendukung keputusan menu, tetapi interpretasi tetap memerlukan pertimbangan konteks dan tujuan penyusunan menu (Angga Wijaya, 2026).

E. Algoritma Apriori

Algoritma Apriori mencari frequent itemset secara bertahap, mulai dari item tunggal, lalu pasangan item, kemudian kombinasi berukuran lebih besar. Dasar

efisiensinya ialah sifat anti-monotonik: jika suatu itemset frequent, semua subset-nya juga harus frequent. Sebaliknya, apabila suatu subset tidak mencapai minimum support, gabungan yang memuat subset tersebut tidak mungkin frequent. Sifat ini memungkinkan algoritma memangkas kandidat sebelum menghitung seluruh kombinasi pada tahap berikutnya (Unknown, 2019; Unknown, 2012).

Proses Apriori umumnya mencakup empat kegiatan. Pertama, hitung support semua item tunggal dan simpan yang memenuhi ambang. Kedua, gabungkan itemset frequent berukuran k untuk membentuk kandidat berukuran $k + 1$. Ketiga, pangkas kandidat apabila salah satu subset berukuran k tidak frequent. Keempat, pindai transaksi untuk menghitung support kandidat yang tersisa. Proses berhenti saat tidak ada lagi kandidat yang memenuhi minimum support. Setelah frequent itemset diperoleh, aturan asosiasi dibentuk dari pemisahan itemset menjadi anteseden dan konsekuen, lalu dinilai dengan confidence dan, bila diperlukan, lift (Unknown, 2011; Unknown, 2019).

Sebagai contoh, sebuah koperasi mahasiswa di Makassar mencatat 100 transaksi dan menetapkan minimum support sebesar 20%. Jika kopi muncul pada 35 transaksi, roti pada 30 transaksi, dan keduanya bersama pada 24 transaksi, maka keduanya lolos sebagai frequent itemset tunggal maupun pasangan. Kandidat {kopi, roti, gula} hanya dapat

dipertimbangkan jika seluruh pasangan subset-nya, yaitu {kopi, roti}, {kopi, gula}, dan {roti, gula}, sudah memenuhi ambang support. Jika salah satu pasangan gagal, kandidat tiga item tersebut dipangkas tanpa perlu menghitung frekuensinya dalam seluruh transaksi. Prinsip ini mengurangi pekerjaan yang tidak diperlukan pada kumpulan kandidat (Unknown, 2012; Unknown, 2019).

Penerapan Apriori pada penyusunan menu makanan bergizi merupakan contoh konteks Indonesia yang menunjukkan bagaimana frequent itemset dapat membantu merangkum kombinasi menu yang muncul dalam data. Ketika menerapkannya, analis perlu memastikan definisi transaksi dan pengodean item konsisten, lalu memilih minimum support berdasarkan tujuan. Ambang terlalu rendah menghasilkan banyak kandidat yang jarang, sedangkan ambang terlalu tinggi dapat menghilangkan kombinasi yang relevan. Oleh sebab itu, pengelola sebaiknya menguji beberapa ambang yang masuk akal dan menafsirkan pola bersama pengetahuan domain (Angga Wijaya, 2026; Unknown, 2011).

Kelemahan utama Apriori muncul ketika jumlah item dan panjang transaksi meningkat. Pada setiap tingkat, kandidat perlu dibangkitkan dan basis data dipindai kembali untuk menghitung support. Banyaknya kombinasi dapat tumbuh cepat, khususnya saat minimum support rendah. Analisis awal terhadap ukuran data, kepadatan transaksi, dan

kapasitas pemrosesan membantu menentukan apakah Apriori sesuai. Untuk data yang mengandung banyak pola berulang, FP-Growth dapat menjadi alternatif karena membangun struktur pohon dan menghindari pembangkitan kandidat eksplisit (Unknown, 2019; Unknown, 2012).

F. FP-Growth dan Sequential Pattern Mining

FP-Growth dirancang untuk menemukan frequent pattern dengan mengurangi kebutuhan membangkitkan kandidat berulang seperti pada Apriori. Algoritma ini memindai kumpulan data untuk memperoleh frekuensi item, mengurutkan item yang memenuhi minimum support, lalu menyusun transaksi ke dalam FP-tree. Awalan transaksi yang sama dapat berbagi cabang sehingga struktur tersebut merangkum bagian data yang berulang. Setelah pohon terbentuk, algoritma menelusuri basis pola bersyarat untuk mengekstraksi frequent itemset (Unknown, 2019; Unknown, 2012).

Misalnya, sebuah toko oleh-oleh di Yogyakarta memiliki banyak nota yang mencakup kombinasi bakpia, kopi, dan teh. Jika sebagian besar nota berbagi awalan item yang sama setelah pengurutan berdasarkan frekuensi, FP-tree menyimpan awalan itu satu kali dengan informasi jumlah kemunculannya. Penambangan kemudian membentuk pola bersyarat bagi item tertentu, sehingga pola kombinasi dapat

ditemukan tanpa membuat seluruh kandidat satu demi satu. Struktur ini dapat mengurangi pemindaian kandidat, tetapi kebutuhan memori dan efektivitasnya tetap dipengaruhi oleh ukuran data serta seberapa banyak transaksi berbagi item (Unknown, 2011; Unknown, 2019).

Frequent itemset hanya menangkap koeksistensi item, bukan urutan waktu. Sequential pattern mining digunakan ketika urutan kejadian menjadi bagian dari pertanyaan analisis. Sebuah urutan dapat dinyatakan sebagai rangkaian item atau peristiwa yang muncul pada waktu berbeda dalam satu sekuens, misalnya pelanggan melihat produk, menambahkannya ke keranjang, lalu menyelesaikan pembelian. Analisis harus menetapkan batas sekuens, jarak waktu yang relevan, serta minimum support urutan agar hasilnya dapat ditafsirkan secara konsisten (Unknown, 2012; Unknown, 2019).

Pada layanan digital Indonesia, urutan tindakan pengguna aplikasi belanja dapat ditelaah untuk mengidentifikasi rangkaian aktivitas yang sering mendahului transaksi atau penghentian proses pembelian. Demikian pula, pola unggahan dan interaksi media sosial dapat berubah ketika data baru terus masuk. Kajian tentang penambangan pola inkremental pada media sosial membahas tantangan volume dan kecepatan data, serta kebutuhan memperbarui pola tanpa selalu memproses ulang keseluruhan data. Kajian tersebut relevan untuk

memahami bahwa pemilihan strategi harus memperhitungkan sifat pembaruan kumpulan data, bukan hanya bentuk pola yang dicari (Al Riyami, 2026; KRYVENCHUK, 2022).

Pilihan antara FP-Growth dan sequential pattern mining ditentukan terutama oleh struktur pertanyaan. Jika tujuannya menemukan kombinasi item yang muncul bersama tanpa memperhatikan urutan, FP-Growth dapat digunakan untuk menambang frequent itemset. Jika urutan peristiwa memiliki makna, sequential pattern mining lebih sesuai. Dalam praktik, kedua metode memerlukan prapemrosesan data dan definisi unit analisis yang jelas; hasilnya tidak otomatis menjelaskan motif pengguna atau membuktikan hubungan sebab-akibat. Buku tentang algoritma penambangan data dan pembelajaran mesin memberikan landasan untuk memahami perbedaan strategi pencarian pola tersebut (Unknown, 2019; Unknown, 2012).

G. Pattern Mining pada Data IoT dan Keamanan Sistem

Pada data Internet untuk Segala (IoT), peristiwa dapat berasal dari pembacaan sensor, perubahan status perangkat, atau komunikasi antarnode. Pattern mining dapat membantu menemukan kombinasi kejadian yang berulang, misalnya suhu meningkat bersama perubahan kelembapan pada rentang waktu tertentu. Sebelum

penambangan, data perlu dibentuk menjadi unit transaksi atau sekuens, misalnya satu jendela waktu, satu siklus operasi, atau satu sesi perangkat. Pemilihan unit tersebut mengubah pola yang terdeteksi, sehingga harus disesuaikan dengan proses yang diamati (Unknown, 2012; Unknown, 2019).

Contoh keranjang belanja membantu memperkenalkan *item* dan *itemset* sebelum pembahasan beralih ke frekuensi kemunculannya dalam kumpulan transaksi. Gambar ini ditempatkan pada sub-bab tentang frequent pattern dan association rule.



Gambar 9.1 Item dalam Satu Transaksi Belanja. Sumber: Ilustrasi, diolah penulis

Keranjang tersebut memperlihatkan tiga *item* yang membentuk satu *itemset* dalam sebuah transaksi. Dengan membandingkan banyak transaksi serupa, analis dapat menilai apakah kombinasi item itu cukup sering muncul untuk disebut *frequent pattern*.

Contoh penerapan di Indonesia dapat menggunakan sistem pemantauan kualitas udara berbasis IoT. Pembacaan suhu, kelembapan, dan indikator gas dapat dikelompokkan ke dalam kategori seperti rendah, sedang, atau tinggi untuk membentuk item yang dapat ditambang. Pola berulang antara kondisi sensor dan status perangkat dapat membantu petugas memeriksa kondisi yang patut diperhatikan, tetapi kategorisasi harus mempertahankan makna teknis data. Penelitian sistem pemantauan kualitas udara berbasis sensor menunjukkan penggunaan perangkat IoT untuk merekam indikator lingkungan secara real-time; pola hasil mining tetap perlu ditafsirkan berdasarkan kalibrasi sensor dan konteks lokasi (Alfirdaus, 2024).

Dalam keamanan sistem, log dapat diubah menjadi item peristiwa, misalnya autentikasi gagal, akses ke layanan tertentu, atau perubahan hak akses. Frequent pattern dapat menandai kombinasi kejadian yang sering muncul dalam aktivitas normal, sedangkan urutan peristiwa dapat membantu menelusuri rangkaian yang mendahului insiden. Namun, peristiwa yang sering muncul tidak otomatis aman, dan pola yang jarang tidak otomatis berbahaya. Hasil mining

perlu diperiksa bersama konteks operasional, aturan akses, serta bukti lain sebelum menjadi dasar tindakan keamanan (Dua, 2016; Amanda, 2025).

Untuk menerapkan metode ini, analis dapat membentuk transaksi dari jendela waktu atau sesi, menentukan item yang merepresentasikan peristiwa, lalu menetapkan minimum support berdasarkan volume dan tujuan pemantauan. Pada jaringan dengan beban bervariasi, ambang yang sama mungkin tidak tepat untuk semua layanan atau periode. Evaluasi hasil dapat mencakup pemeriksaan contoh transaksi pendukung aturan, perbandingan pola pada periode berbeda, dan verifikasi oleh pengelola sistem. Pendekatan ini mengurangi risiko menganggap korelasi sesaat sebagai indikator ancaman yang pasti (Dua, 2016; Amanda, 2025).

Pola konsumsi energi pada jaringan listrik pintar juga dapat dianalisis untuk mendukung efisiensi dan keandalan. Kajian pattern mining pada data energi smart grid menempatkan penemuan pola sebagai bagian dari upaya memahami data berskala besar dan mengidentifikasi pola yang bermakna. Sementara itu, sistem IoT pemantauan kateter menunjukkan bagaimana perangkat dapat mengirimkan informasi pemantauan kondisi secara berkelanjutan; analisis pola pada sistem semacam ini harus memperhatikan keselamatan, privasi, dan konsekuensi jika alarm keliru. Karena data operasional dapat memuat

informasi sensitif, pengelolaan akses dan perlindungan data perlu menjadi bagian dari rancangan analisis, bukan tambahan setelah pola ditemukan (Hassan, 2025; Banjarnahor, 2022; Dua, 2016).

H. Rangkuman

Frequent pattern adalah kombinasi item yang muncul berulang dalam kumpulan transaksi dan memenuhi ambang minimum support. Itemset yang lolos penyaringan dapat dikembangkan menjadi association rule berbentuk $X \rightarrow Y$, yang menyatakan kecenderungan kemunculan konsekuen Y ketika anteseden X hadir. Aturan tersebut menggambarkan asosiasi, bukan sebab-akibat, sehingga tidak cukup menjadi dasar untuk menyatakan bahwa satu item menyebabkan pembelian item lain. Penafsiran juga bergantung pada kualitas data dan definisi unit analisis: transaksi dapat mewakili nota, kunjungan, atau pelanggan, dan pilihan ini memengaruhi pola yang ditemukan. Sebelum analisis, kode item, duplikasi, serta konsistensi pencatatan perlu diperiksa.

Support mengukur proporsi seluruh transaksi yang memuat suatu itemset. Confidence mengukur proporsi transaksi yang memuat X dan Y di antara transaksi yang memuat X, sedangkan lift membandingkan confidence dengan support konsekuen. Karena menjawab pertanyaan berbeda, ketiganya perlu dibaca bersama. Confidence yang

tinggi belum tentu menunjukkan asosiasi positif apabila konsekuen memang populer; lift di atas satu menunjukkan asosiasi positif relatif terhadap frekuensi dasar konsekuen, nilai satu menunjukkan tingkat yang sebanding, dan nilai di bawah satu menunjukkan asosiasi negatif. Tidak satu pun dari ukuran ini membuktikan kausalitas.

Apriori mencari frequent itemset secara bertahap, dari item tunggal menuju kombinasi yang lebih besar. Sifat anti-monotonik menjadi dasar pemangkasan: jika suatu subset tidak frequent, itemset yang memuat subset tersebut tidak mungkin frequent. Algoritma membangkitkan kandidat, memangkas kandidat yang tidak memenuhi syarat subset, lalu menghitung support kandidat tersisa sampai tidak ada lagi itemset frequent. Aturan kemudian dapat dibentuk dari frequent itemset dan dinilai menggunakan confidence serta lift. Dalam praktik Sistem Informasi, hasilnya dapat mendukung keputusan seperti penataan rak, promosi, atau perancangan kombinasi menu, tetapi rekomendasi perlu mempertimbangkan tujuan organisasi dan konteks operasional. Analisis pola membantu menemukan keteraturan yang berguna; keputusan tetap memerlukan penilaian manusia atas mutu data, dampak, dan batas interpretasinya.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana perubahan unit analisis dari nota menjadi pelanggan dapat mengubah makna frequent pattern dan association rule pada data penjualan UMKM. Jelaskan pula pemeriksaan kualitas data yang perlu dilakukan sebelum membandingkan hasil dari kedua definisi unit tersebut.
2. Evaluasi aturan roti → susu pada tabel transaksi dengan menggunakan support pasangan, confidence, dan lift. Rumuskan kesimpulan yang layak disampaikan kepada pengelola toko serta batas interpretasi yang harus dipertahankan ketika hasil digunakan untuk merancang promosi.
3. Bandingkan peran minimum support dan confidence dalam proses pencarian pola dengan Apriori. Telaah bagaimana pemilihan ambang yang terlalu tinggi atau terlalu rendah memengaruhi kandidat, pola yang ditemukan, dan kegunaan hasil bagi pengambil keputusan.
4. Argumentasikan mengapa frequent itemset tidak dengan sendirinya membuktikan adanya hubungan sebab-akibat. Susun rancangan tindak lanjut yang dapat digunakan organisasi untuk menguji apakah promosi suatu item benar-benar meningkatkan pembelian item lain.
5. Rumuskan alur penerapan Apriori untuk data transaksi kantin kampus, mulai dari penyiapan data hingga interpretasi aturan. Sertakan peran sifat anti-monotonik, support, confidence, dan lift, serta satu pertimbangan konteks yang perlu menyertai rekomendasi menu.

b. Pilihan Berganda

1. Sebuah itemset dinyatakan frequent apabila....
   a. seluruh item di dalamnya memiliki harga yang sama
   b. jumlah atau proporsi transaksi yang memuatnya mencapai minimum support
   c. confidence setiap aturan yang dibentuk darinya bernilai satu
   d. item-itemnya selalu dibeli dalam urutan yang sama
   e. itemset tersebut memiliki lift lebih dari satu
2. Pada lima transaksi, roti dan susu muncul bersama dalam tiga transaksi. Berapakah support itemset {roti, susu}?
   a. 0,20
   b. 0,40
   c. 0,60
   d. 0,75
   e. 0,80
3. Jika roti muncul dalam empat transaksi dan roti serta susu muncul bersama dalam tiga transaksi, confidence aturan roti → susu adalah....
   a. 0,60
   b. 0,70
   c. 0,75
   d. 0,80

e. 1,00

4. Confidence aturan roti → susu adalah 0,75 dan support susu adalah 0,80. Interpretasi lift sebesar 0,9375 yang paling tepat adalah....
   a. roti menyebabkan penurunan pembelian susu
   b. asosiasi roti dengan susu positif dan lebih kuat daripada frekuensi dasar susu
   c. kemunculan roti dan susu tidak melampaui frekuensi dasar susu
   d. setiap pembeli roti juga membeli susu
   e. susu muncul dalam 93,75% seluruh transaksi
5. Dalam Apriori, apabila suatu pasangan item tidak memenuhi minimum support, kesimpulan yang tepat mengenai kandidat tiga item yang memuat pasangan tersebut adalah....
   a. kandidat tersebut harus dihitung karena berukuran lebih besar
   b. kandidat tersebut tidak mungkin frequent dan dapat dipangkas
   c. kandidat tersebut pasti memiliki lift di bawah satu
   d. kandidat tersebut tetap frequent jika confidence-nya tinggi
   e. kandidat tersebut hanya dapat dipangkas setelah aturan dibentuk
6. Seorang analis memperoleh confidence tinggi untuk aturan $X \rightarrow Y$. Mengapa ia masih perlu memeriksa lift?
   a. Lift menentukan jumlah item dalam anteseden

b. Lift menunjukkan apakah confidence melampaui frekuensi dasar konsekuen
c. Lift mengubah nilai support menjadi jumlah transaksi
d. Lift membuktikan bahwa anteseden menyebabkan konsekuen
e. Lift mengidentifikasi transaksi yang mengalami duplikasi

7. Dua tim menganalisis catatan penjualan yang sama, tetapi tim pertama mendefinisikan transaksi sebagai nota dan tim kedua sebagai pelanggan. Mengapa hasil mereka dapat berbeda?
   a. Support selalu bernilai sama untuk data yang sumbernya sama
   b. Unit analisis memengaruhi item yang dikelompokkan sebagai satu transaksi
   c. Definisi unit analisis hanya memengaruhi visualisasi hasil
   d. Apriori mensyaratkan satu transaksi untuk setiap pelanggan
   e. Confidence tidak dapat dihitung dari data pelanggan

8. Langkah yang sesuai dengan proses Apriori setelah frequent itemset berukuran k diperoleh adalah....
   a. menghapus seluruh item tunggal dari data
   b. membentuk kandidat berukuran $k + 1$, memangkas kandidat yang memiliki subset tidak frequent, lalu menghitung support

c. membentuk aturan dari semua kombinasi tanpa memeriksa support
d. mengurutkan transaksi berdasarkan confidence
e. menghentikan pencarian karena itemset berukuran lebih besar tidak dapat frequent

9. Dalam analisis transaksi, minimum support terutama digunakan untuk....
   a. menentukan apakah anteseden menyebabkan konsekuen
   b. menyaring itemset berdasarkan frekuensi kemunculannya dalam seluruh transaksi
   c. mengukur popularitas konsekuen hanya pada transaksi yang memuat anteseden
   d. memastikan setiap aturan memiliki lift lebih dari satu
   e. menetapkan unit analisis yang paling sesuai
10. Sebuah toko menemukan bahwa pasangan menu sering muncul bersama. Sebelum menjadikan temuan itu sebagai rekomendasi menu, pengelola sebaiknya....
   a. menganggap frekuensi bersama sebagai bukti bahwa satu menu menyebabkan pilihan menu lain
   b. mengabaikan kualitas pencatatan selama support cukup tinggi
   c. mempertimbangkan tujuan menu dan konteks operasional, serta tidak menafsirkan asosiasi sebagai kausalitas
   d. memilih pasangan dengan confidence tertinggi tanpa melihat metrik lain

e. menaikkan minimum support sampai hanya satu aturan tersisa

c. Soal Project / Studi Kasus

1. Sebuah UMKM roti di Bandung mencatat penjualan selama enam bulan. Sistem kasir mengekspor satu baris untuk setiap item yang terjual, tetapi nomor nota kadang kosong dan beberapa transaksi tercatat dua kali setelah sinkronisasi. Pemilik ingin mengetahui kombinasi produk untuk penataan rak dan paket promosi. Ia mengusulkan analisis langsung atas setiap baris karena jumlah datanya lebih besar, lalu meminta tim memilih aturan dengan confidence tertinggi. Namun, tim belum menetapkan apakah unit analisisnya nota atau pelanggan, dan belum memeriksa apakah pembelian grosir mencerminkan pola pelanggan biasa. Analisis langkah penyiapan data, pemilihan unit transaksi, penggunaan support, confidence, dan lift, serta rekomendasikan cara menyampaikan hasil tanpa klaim sebab-akibat.
2. Kantin sebuah kampus di Surabaya ingin menyusun paket makan siang yang menarik sekaligus mendukung pilihan menu bergizi. Data transaksi memuat kombinasi makanan dan minuman, tetapi satu mahasiswa dapat membeli beberapa kali dalam sehari. Pengelola menetapkan minimum support yang tinggi agar laporan ringkas, kemudian berencana memilih aturan dengan confidence tertinggi untuk menentukan paket. Sejumlah menu populer muncul pada hampir semua transaksi, sementara menu sayur memiliki frekuensi lebih rendah. Tim analitik diminta menilai apakah

rancangan tersebut memadai dan membuat rekomendasi yang dapat dipertanggungjawabkan. Evaluasi definisi unit transaksi dan ambang, jelaskan kegunaan lift, serta rumuskan validasi dan pertimbangan gizi yang perlu dilakukan sebelum menerapkan paket.

BAB 10
DATA MINING UNTUK INTERNET OF THINGS (IOT)

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep arsitektur IoT untuk memetakan hubungan antara perangkat sensor, jaringan komunikasi, platform pemrosesan, dan aplikasi dalam ekosistem data IoT.
2. Menggunakan prosedur akuisisi dan pengelolaan data sensor untuk menangani karakteristik data yang bervariasi, berkecepatan tinggi, dan berjumlah besar.
3. Menentukan teknik streaming data dan real-time data mining yang sesuai dengan kebutuhan analisis serta batas waktu respons suatu aplikasi IoT.
4. Memilih pendekatan edge, fog, atau cloud analytics berdasarkan kebutuhan latensi, kapasitas komputasi, konektivitas, dan tata kelola data.
5. Mengimplementasikan alur analitik IoT untuk mendukung predictive maintenance melalui pemanfaatan data sensor dan hasil analisis prediktif.

B. Pendahuluan

Perangkat Internet of Things (IoT) menghasilkan aliran data berkelanjutan dari beragam lingkungan, mulai dari

mesin produksi dan jaringan energi hingga kendaraan dan perangkat kesehatan. Data tersebut dapat membantu organisasi mengenali kondisi operasional, mendeteksi anomali, dan merespons perubahan secara cepat. Namun, nilai analitisnya tidak muncul secara otomatis. Perbedaan format, frekuensi pengiriman, kualitas pengukuran, serta keterbatasan konektivitas dan kapasitas komputasi menimbulkan persoalan dalam pengumpulan dan pemrosesan data. Atzori, Iera, dan Morabito (2010) menjelaskan IoT sebagai perpaduan teknologi yang menghubungkan objek fisik dengan layanan dan jaringan komunikasi, sedangkan Gubbi et al. (2013) menekankan peran pengolahan data dalam membangun aplikasi IoT yang cerdas. Dengan demikian, data mining untuk IoT perlu menghubungkan pemahaman terhadap sumber data dengan pemilihan metode analisis dan penempatan komputasi yang tepat.

Bagi mahasiswa Sistem Informasi, kajian ini relevan karena penerapan IoT melibatkan lebih dari sekadar perangkat dan algoritma. Sistem informasi berbasis IoT harus menyelaraskan proses bisnis, infrastruktur teknologi, kebutuhan pengguna, serta kebijakan pengelolaan data. Sebagai contoh, pada fasilitas manufaktur, sensor getaran dan suhu dapat mengirimkan data kondisi mesin secara berkala. Data itu perlu diperoleh dan dikelola secara konsisten, diproses dengan mempertimbangkan kebutuhan

respons, lalu diterjemahkan menjadi informasi yang mendukung keputusan pemeliharaan. Analisis aliran data menghadirkan tantangan tersendiri karena data terus datang dan model tidak selalu dapat menunggu seluruh kumpulan data tersedia; karakteristik tersebut menjadi perhatian dalam kajian data stream (Aggarwal, 2007). Selain itu, keputusan mengenai pemrosesan di edge, fog, atau cloud memengaruhi latensi, penggunaan jaringan, dan lokasi penyimpanan data. Shi et al. (2016) menunjukkan bahwa komputasi edge dapat mendekatkan pemrosesan ke sumber data, sehingga pilihan arsitektur perlu disesuaikan dengan kebutuhan aplikasi, bukan ditetapkan secara seragam.

Bab ini membahas penerapan data mining pada ekosistem IoT dengan orientasi pada pemecahan persoalan analitis yang konkret. Pembahasan dimulai dari arsitektur IoT dan ekosistem datanya, termasuk hubungan antara perangkat, jaringan, platform, dan aplikasi. Selanjutnya, dibahas akuisisi serta pengelolaan data sensor, kemudian teknik streaming data dan real-time data mining untuk menghasilkan analisis dalam batas waktu operasional. Bab ini juga menguraikan pemilihan edge, fog, dan cloud analytics berdasarkan kebutuhan pemrosesan, sebelum menerapkan analitik IoT pada predictive maintenance dan intelligent IoT analytics. Rangkaian pembahasan tersebut menempatkan proses data mining sebagai bagian dari

rancangan sistem informasi yang dapat mengubah pengamatan sensor menjadi dukungan keputusan yang relevan dan tepat waktu.

C. Arsitektur IoT dan Ekosistem Data IoT

Pembahasan data sensor dan pemilihan metode analisis pada bab-bab sebelumnya menjadi landasan untuk memahami bagaimana data bergerak dalam sistem Internet untuk Segala (IoT). Arsitektur IoT menghubungkan objek fisik yang melakukan pengukuran dengan jaringan komunikasi, platform pemrosesan, dan aplikasi yang menggunakan hasil analisis. Bagi penambangan data, pemetaan hubungan ini membantu menentukan sumber data, konteks pengukuran, serta titik yang tepat untuk menjalankan analitik (Jain, 2023; Unknown, 2011).

Ekosistem data IoT dapat dipahami melalui beberapa lapisan yang saling bergantung. Lapisan perangkat mencakup sensor, aktuator, dan pengendali yang mengubah keadaan fisik menjadi data digital. Lapisan konektivitas membawa data melalui jaringan lokal atau jaringan luas, sementara lapisan platform menerima, menyimpan, dan memprosesnya. Pada lapisan aplikasi, hasil pemrosesan disajikan sebagai informasi untuk pemantauan atau pengambilan keputusan. Sistem monitoring kualitas udara berbasis sensor dan ESP8266, misalnya, mengirimkan pengukuran suhu, kelembapan, dan gas ke platform

pemantauan; alur ini menunjukkan bahwa keluaran sensor perlu dipadukan dengan konektivitas serta layanan aplikasi agar dapat digunakan (Alfirdaus, 2024).

Tabel berikut merangkum fungsi setiap lapisan dan keputusan analitis yang perlu dipertimbangkan ketika menerapkan sistem IoT. Pemisahan tersebut bersifat konseptual; pada implementasi, beberapa fungsi dapat dijalankan oleh perangkat atau platform yang sama.

Tabel 10.1 Lapisan Arsitektur dan Peran Data IoT

Lapisan	Komponen umum	Data atau fungsi	Keputusan analitis
Perangkat	Sensor, aktuator	Pengukuran dan respons	Frekuensi sampling
Konektivitas	Wi-Fi, seluler	Pengiriman data	Protokol dan kapasitas
Platform	Gerbang, layanan data	Penerimaan dan penyimpanan	Validasi dan agregasi
Analitik	Mesin pemrosesan	Ekstraksi pola dan prediksi	Lokasi pemrosesan
Aplikasi	Dasbor, sistem operasi	Informasi dan tindak lanjut	Bentuk peringatan

Tabel memperlihatkan bahwa keputusan pada satu lapisan memengaruhi lapisan lain. Frekuensi sampling yang tinggi dapat memperkaya gambaran perubahan kondisi, tetapi sekaligus menambah lalu lintas jaringan dan kebutuhan penyimpanan. Karena itu, rancangan arsitektur perlu mempertimbangkan kebutuhan informasi aplikasi,

kemampuan perangkat, dan kapasitas platform secara bersamaan.

Visualisasi berikut menekankan hubungan antara unsur fisik, konektivitas, pengolahan, dan pemanfaatan data, bukan sekadar bentuk perangkat tunggal. Susunan tersebut dapat digunakan sebagai kerangka awal untuk memetakan aliran data pada suatu kasus organisasi.



Gambar 10.1 Ekosistem Data IoT pada Fasilitas Produksi.
Sumber: Diolah penulis

Infografik membantu mahasiswa menelusuri asal data hingga pemanfaatannya dalam proses organisasi. Dalam penerapan nyata, setiap sambungan juga perlu dipetakan terhadap tanggung jawab pengelolaan, kualitas data, dan kebutuhan layanan.

Ketika menerapkan kerangka tersebut pada fasilitas pengolahan pangan di Jawa Barat, mahasiswa dapat memetakan sensor suhu ruang produksi sebagai sumber data, jaringan lokal sebagai jalur komunikasi, dan aplikasi pemantauan sebagai pengguna informasi. Selanjutnya, tujuan analisis perlu ditentukan: apakah sistem hanya menampilkan kondisi, memberikan peringatan ketika suhu melewati batas operasional, atau mendukung prediksi gangguan peralatan. Pemisahan tujuan ini membantu memilih data dan fitur yang benar-benar diperlukan tanpa menganggap bahwa seluruh data harus disimpan atau dianalisis dengan cara yang sama (Jain, 2023; Alfirdaus, 2024).

Pemetaan arsitektur juga mendukung penelusuran masalah. Jika nilai sensor tidak muncul pada aplikasi, penyebabnya mungkin berada pada perangkat, konektivitas, atau layanan penerimaan data. Jika data tersedia tetapi tidak sesuai konteks, identitas perangkat, satuan, dan waktu pengukuran perlu diperiksa. Dengan demikian, arsitektur bukan hanya rancangan teknis, melainkan kerangka untuk menghubungkan kualitas pengamatan, proses organisasi, dan kebutuhan penambangan data (Unknown, 2011).

D. Akuisisi dan Pengelolaan Data Sensor

Setelah sumber dan jalur data dipetakan, langkah berikutnya adalah memastikan bahwa data sensor dapat dikumpulkan dan dikelola secara konsisten. Akuisisi mencakup pengambilan nilai dari sensor, pencatatan identitas perangkat dan waktu, serta pengiriman data ke media penerimaan. Pengelolaan mencakup pemeriksaan mutu, penyimpanan, pengaturan akses, dan penyiapan data untuk analisis. Tahapan ini menentukan apakah kumpulan data cukup dapat dipercaya untuk mendukung pemodelan, selaras dengan prinsip prapemrosesan data yang dibahas pada Bab 4 (García, 2015; Unknown, 2011).

Sensor menghasilkan pengukuran yang dipengaruhi oleh karakteristik perangkat dan kondisi penggunaannya. Karena itu, setiap rekaman sebaiknya menyertakan identitas sensor, waktu pengukuran, satuan, lokasi atau konteks pengukuran, serta status perangkat jika tersedia. Penanda waktu memungkinkan pengurutan data dan penggabungan pengukuran dari beberapa sumber. Identitas dan satuan membantu mencegah nilai dari sensor berbeda diperlakukan seolah-olah setara. Studi pemantauan kualitas udara berbasis sensor menunjukkan penggunaan beberapa komponen pengukuran dalam satu sistem, sehingga pengelolaan konteks setiap nilai menjadi bagian dari interpretasi data (Alfirdaus, 2024).

Pada tahap pengumpulan, prosedur perlu disesuaikan dengan tujuan aplikasi. Sistem pemantauan kateter berbasis IoT, misalnya, memiliki konteks penggunaan yang berbeda dari pemantauan kualitas udara: data berkaitan dengan kondisi layanan kesehatan dan perlu ditangani sesuai kebutuhan pemantauan pasien. Contoh tersebut menunjukkan bahwa rancangan akuisisi tidak cukup berfokus pada keberhasilan mengirim nilai; sistem juga perlu memastikan bahwa data dapat dikaitkan dengan objek pemantauan yang benar dan tersedia bagi pengguna yang berwenang (Banjarnahor, 2022).

Pengelolaan mutu dapat diterapkan sebagai serangkaian pemeriksaan operasional. Pemeriksaan pertama menguji apakah nilai berada dalam rentang yang masuk akal bagi sensor dan konteksnya. Pemeriksaan berikutnya menandai nilai yang hilang, rekaman duplikat, urutan waktu yang tidak konsisten, atau perubahan pola pengiriman yang tidak biasa. Data yang gagal dalam pemeriksaan tidak harus langsung dihapus: sebagian dapat ditandai untuk dianalisis, sedangkan rekaman yang tidak layak dapat dikeluarkan dari proses pelatihan. Pendekatan tersebut konsisten dengan kebutuhan pembersihan dan transformasi sebelum penambangan data (García, 2015).

Dalam aplikasi pemantauan udara di lingkungan perkotaan Indonesia, mahasiswa dapat menetapkan prosedur yang mencatat nilai sensor beserta waktu, lokasi

perangkat, dan satuannya. Jika satu perangkat berhenti mengirimkan data sementara perangkat lain tetap aktif, catatan tersebut membantu membedakan gangguan konektivitas dari perubahan kondisi udara. Sistem kemudian dapat mempertahankan rekaman yang masih valid, memberi penanda pada data yang diragukan, dan meneruskan informasi operasional kepada pengelola. Contoh ini menegaskan bahwa dokumentasi konteks pengukuran sama pentingnya dengan pengumpulan nilai numerik.

Penyimpanan juga perlu mengikuti pola penggunaan data. Data mentah berguna untuk penelusuran dan pemrosesan ulang, sedangkan data yang telah diringkas dapat mendukung pemantauan jangka panjang dengan kebutuhan ruang lebih rendah. Kebijakan retensi, pencadangan, dan akses ditentukan berdasarkan kebutuhan operasional serta tata kelola organisasi. Dengan membedakan data mentah, data hasil prapemrosesan, dan data fitur, tim dapat menelusuri bagaimana suatu hasil analitik dibentuk dan memperbaiki prosedur ketika ditemukan perubahan mutu sensor (García, 2015; Banjarnahor, 2022).

E. Streaming Data dan Real-Time Data Mining

Data sensor yang dikirim terus-menerus membentuk aliran, bukan kumpulan data yang selalu tersedia secara

utuh sebelum analisis dimulai. Karakteristik ini menuntut prosedur pemrosesan yang dapat menerima rekaman baru, memperbarui ringkasan, dan menghasilkan keluaran dalam batas waktu tertentu. Konsep aliran data juga muncul pada penambangan data dengan volume dan kecepatan tinggi, ketika pemrosesan ulang seluruh data untuk setiap pembaruan menjadi tidak efisien (Al Riyami, 2026; Unknown, 2011).

Untuk menerapkan penambangan data waktu nyata, kebutuhan respons harus dirumuskan terlebih dahulu. Pada pemantauan mesin, peringatan untuk kondisi yang berpotensi merusak peralatan mungkin perlu disampaikan dalam waktu singkat. Sebaliknya, ringkasan konsumsi energi harian dapat dihitung dengan jeda lebih longgar. Penentuan batas waktu tersebut membantu memilih ukuran jendela pengamatan, frekuensi pembaruan, dan kompleksitas algoritma. Sistem prediksi kadar CO_2 yang memanfaatkan sensor dan perangkat ESP32 menunjukkan bagaimana data lingkungan dapat digunakan sebagai masukan untuk analisis prediktif, meskipun rancangan waktu respons tetap perlu disesuaikan dengan tujuan aplikasi (Agung Suburdjati, 2026).

Pemrosesan berbasis jendela mengelompokkan data menurut rentang waktu atau jumlah rekaman. Jendela waktu, misalnya, dapat digunakan untuk menghitung rerata suhu setiap beberapa menit, sedangkan jendela

berdasarkan jumlah rekaman dapat merangkum sejumlah pengukuran terakhir. Pemilihan jendela memengaruhi kecepatan respons dan tingkat kehalusan ringkasan. Jendela yang terlalu panjang dapat menunda respons terhadap perubahan cepat; jendela yang terlalu pendek dapat membuat keluaran lebih peka terhadap gangguan sesaat. Strategi pembaruan bertahap relevan untuk aliran yang terus bertambah karena informasi baru dapat diproses tanpa selalu mengulang seluruh analisis terdahulu (Al Riyami, 2026).

Pada fasilitas pengolahan hasil pertanian di Sulawesi Selatan, data suhu dan kelembapan dapat diterima secara berkala untuk memantau kondisi penyimpanan. Mahasiswa dapat menerapkan aturan pemrosesan yang menghitung ringkasan dalam jendela pendek dan membandingkannya dengan batas operasional yang ditentukan pengelola. Jika kondisi melampaui batas, aplikasi meneruskan peringatan; jika tidak, sistem memperbarui ringkasan tanpa mengirim notifikasi berulang. Contoh ini memperlihatkan bahwa real-time tidak semata-mata berarti seketika, tetapi merujuk pada kesesuaian waktu keluaran dengan kebutuhan keputusan.

Pemrosesan aliran harus menangani rekaman yang terlambat, tidak berurutan, atau tidak diterima akibat gangguan komunikasi. Sistem dapat menggunakan penanda waktu pengukuran, bukan hanya waktu data tiba, untuk

menjaga urutan kejadian. Rekaman yang datang terlambat dapat dimasukkan dalam koreksi ringkasan jika kebijakan aplikasi mengizinkannya. Ketika koneksi terputus, perangkat atau gerbang dapat menyimpan data sementara agar pengiriman dapat dilanjutkan setelah konektivitas pulih. Pengaturan tersebut mencegah jeda komunikasi langsung ditafsirkan sebagai perubahan kondisi fisik.

Penerapan analitik waktu nyata perlu membedakan pemantauan berbasis aturan dan pemodelan prediktif. Aturan sederhana dapat menghasilkan peringatan ketika nilai melewati ambang, sedangkan model dapat memperkirakan kecenderungan atau kelas kondisi berdasarkan beberapa fitur. Pemilihan model mempertimbangkan kebutuhan latensi dan kemampuan pemrosesan, selain ukuran evaluasi yang dibahas pada Bab 6 dan Bab 7. Hasil penelitian tentang penemuan pola inkremental pada aliran data menunjukkan relevansi pembaruan bertahap ketika data terus bertambah, tetapi manfaatnya tetap bergantung pada kesesuaian metode dengan tujuan operasional (Al Riyami, 2026; Agung Suburdjati, 2026).

F. Edge, Fog, dan Cloud Analytics untuk IoT

Penempatan analitik menentukan seberapa jauh data harus berpindah sebelum menghasilkan keluaran. Analitik edge menjalankan pemrosesan pada perangkat atau dekat

sumber data; fog menyediakan lapisan perantara seperti gerbang lokal; sedangkan cloud menawarkan layanan terpusat untuk pemrosesan dan penyimpanan berskala lebih luas. Ketiganya dapat digunakan secara terpisah atau dikombinasikan. Pemilihan lokasi analitik merupakan bagian dari rancangan pengolahan data dan fusi informasi, bukan keputusan yang semata-mata ditentukan oleh kecanggihan teknologi (Jain, 2023; Unknown, 2012).

Analitik edge sesuai ketika aplikasi memerlukan respons cepat atau koneksi ke pusat tidak selalu tersedia. Perangkat dapat menghitung ringkasan atau menerapkan aturan sederhana sebelum mengirimkan keluaran. Namun, kemampuan komputasi, memori, dan energi perangkat dapat membatasi kompleksitas model yang dijalankan. Fog dapat menggabungkan data dari beberapa perangkat di lokasi yang sama, sehingga proses yang lebih berat daripada kemampuan sensor dapat dilakukan dekat sumber. Cloud berguna untuk menyimpan data lintas lokasi, menggabungkan kumpulan data lebih besar, dan menjalankan analisis yang membutuhkan sumber daya terpusat (Unknown, 2012; Unknown, 2011).

Keputusan penempatan dapat diterapkan dengan membandingkan kebutuhan aplikasi. Tabel berikut menyajikan pertimbangan utama yang dapat digunakan mahasiswa saat memetakan pemrosesan pada sistem IoT.

Tabel 10.2 Pertimbangan Penempatan Analitik IoT

Lokasi	Latensi	Konektivitas	Kapasitas	Contoh keluaran
Edge	Rendah	Dapat terbatas	Terbatas	Peringatan lokal
Fog	Rendah-menengah	Jaringan lokal	Menengah	Agregasi area
Cloud	Bergantung jaringan	Internet diperlukan	Dapat diskalakan	Analisis lintas lokasi

Tabel memperlihatkan bahwa tidak ada satu lokasi yang unggul untuk seluruh kebutuhan. Keputusan perlu menghubungkan batas waktu respons, ketersediaan koneksi, kapasitas perangkat, dan cakupan data yang diperlukan oleh aplikasi.

Pada instalasi pompa air di wilayah pertanian Nusa Tenggara Timur, pengendali lokal dapat mendeteksi kondisi yang memerlukan tindakan segera meskipun koneksi internet terputus. Gerbang fog dapat menggabungkan pengukuran beberapa pompa dalam satu area, sedangkan cloud dapat menyimpan data lintas lokasi untuk membandingkan pola pemakaian dan kebutuhan perawatan. Pembagian tugas seperti ini merupakan contoh penerapan, bukan ketentuan yang harus selalu diikuti. Sistem pemantauan IoT di Indonesia memperlihatkan beragam susunan perangkat dan platform; rancangan perlu mengikuti tujuan layanan dan kondisi operasional di lapangan (Alfirdaus, 2024; Banjarnahor, 2022).

Penempatan data juga memengaruhi tata kelola dan keamanan. Pemrosesan lokal dapat mengurangi pengiriman data mentah, sedangkan penyimpanan terpusat memudahkan penggabungan dan pemeliharaan kumpulan data. Namun, pilihan tersebut perlu disertai pengaturan hak akses, mekanisme perlindungan, dan pencatatan pemrosesan. Pada lingkungan yang melibatkan data sensitif, organisasi dapat mempertimbangkan analitik terfederasi sebagai salah satu pendekatan untuk mengurangi kebutuhan pemusatan data; penerapannya harus dinilai berdasarkan rancangan sistem dan kebutuhan perlindungan data (B., 2026).

Untuk menerapkan keputusan arsitektur, mahasiswa dapat mulai dengan menetapkan keluaran yang harus tersedia, lalu mencatat toleransi latensi, ketergantungan konektivitas, serta ukuran data yang akan diproses. Model yang hanya membutuhkan perhitungan ringkas dapat ditempatkan di edge, sementara pelatihan berkala atau analisis historis dapat dilakukan pada cloud. Pendekatan gabungan menuntut pembagian tanggung jawab yang jelas: versi model, format fitur, dan waktu pembaruan perlu dikelola agar keluaran lokal tetap konsisten dengan analisis terpusat (Unknown, 2011; Jain, 2023).

G. Predictive Maintenance dan Intelligent IoT Analytics

Predictive maintenance memanfaatkan data kondisi peralatan untuk memperkirakan kebutuhan pemeliharaan sebelum gangguan berkembang menjadi kerusakan operasional. Berbeda dari pemeliharaan terjadwal semata, pendekatan prediktif menggunakan pola pengukuran dan riwayat kondisi untuk mendukung penentuan prioritas pemeriksaan. Data sensor seperti suhu, getaran, tekanan, atau arus listrik dapat menjadi masukan model, tetapi kegunaannya bergantung pada mutu pencatatan dan kesesuaiannya dengan kejadian kerusakan yang hendak diprediksi (Unknown, 2011; Agung Suburdjati, 2026).

Alur penerapan dimulai dengan menentukan keputusan yang ingin didukung. Organisasi dapat memperkirakan apakah mesin akan memerlukan pemeriksaan dalam periode tertentu, mengelompokkan tingkat kondisi, atau memperkirakan perubahan suatu indikator. Setelah tujuan ditetapkan, data sensor dipasangkan dengan catatan operasi dan riwayat pemeliharaan. Fitur yang relevan dapat berupa nilai terbaru, perubahan antarpengukuran, rerata dalam jendela waktu, atau variasi pengukuran. Pemilihan fitur dan penyiapan data mengikuti prinsip rekayasa fitur yang dibahas pada Bab 4, sedangkan pemilihan algoritma dan evaluasi mengikuti Bab 5 hingga Bab 7 (García, 2015; Liu, 2018).

Sebagai contoh penerapan di Indonesia, sebuah pabrik pengolahan kelapa sawit di Riau dapat menggunakan sensor suhu dan getaran pada motor konveyor. Data kondisi dipasangkan dengan catatan inspeksi, penggantian komponen, dan waktu henti mesin. Tim kemudian membangun model untuk mengidentifikasi pola yang mendahului kebutuhan pemeriksaan, lalu membandingkan prediksi dengan catatan pemeliharaan yang tidak digunakan dalam pelatihan. Penelitian berbasis IoT yang memprediksi kadar CO_2 menggunakan data sensor menunjukkan penerapan data pengukuran sebagai masukan bagi model prediktif, meskipun tujuan dan variabel pada kasus tersebut berbeda dari pemeliharaan mesin (Agung Suburdjati, 2026).

Keberhasilan model tidak cukup dinilai dari ketepatan prediksi pada kumpulan data uji. Dalam konteks pemeliharaan, prediksi yang terlambat dapat menyebabkan gangguan, sedangkan peringatan palsu yang terlalu sering dapat memicu pemeriksaan yang tidak diperlukan. Karena itu, evaluasi perlu mempertimbangkan konsekuensi operasional dari tiap jenis kesalahan dan kemampuan tim merespons keluaran model. Hasil analitik sebaiknya disajikan sebagai dukungan keputusan, bukan pengganti penilaian teknisi. Pendekatan prediktif dalam penambangan data menuntut hubungan yang jelas antara tujuan prediksi, data yang tersedia, dan keputusan yang akan diambil (Unknown, 2011; Kurniati, 2021).

Intelligent IoT analytics memperluas penerapan ini dengan menggabungkan pemantauan, prediksi, dan pembaruan analisis ketika data baru diterima. Model dapat berjalan pada gerbang lokal untuk menandai perubahan mendesak, sedangkan data historis dikirim ke platform untuk evaluasi dan pelatihan ulang berkala. Pola pada sistem energi pintar juga menunjukkan peluang analisis data skala besar untuk mendukung efisiensi dan keandalan, walaupun penentuan tindakan tetap memerlukan konteks proses di lapangan (Hassan, 2025). Dengan alur tersebut, mahasiswa dapat mengimplementasikan sistem yang menghubungkan pengukuran sensor dengan keputusan inspeksi, memeriksa hasilnya terhadap kejadian nyata, dan memperbarui model ketika pola operasi berubah (Agung Suburdjati, 2026; Unknown, 2011).

H. Rangkuman

Data IoT mengalir dari objek fisik menuju proses organisasi melalui perangkat, konektivitas, platform, analitik, dan aplikasi. Sensor mengubah kondisi lingkungan atau peralatan menjadi pengukuran digital, sedangkan aktuator dapat menjalankan respons. Jaringan lokal atau luas membawa data ke gerbang dan platform untuk diterima, disimpan, serta diproses. Hasil analitik kemudian ditampilkan melalui aplikasi, misalnya dasbor pemantauan atau sistem peringatan. Pembagian lapisan ini bersifat

konseptual karena beberapa fungsi dapat berada pada perangkat atau platform yang sama.

Keputusan pada satu lapisan memengaruhi kinerja lapisan lain. Frekuensi sampling yang tinggi dapat memberikan gambaran perubahan kondisi secara lebih rinci, tetapi meningkatkan lalu lintas jaringan dan kebutuhan penyimpanan. Karena itu, rancangan sistem perlu disesuaikan dengan tujuan aplikasi, kemampuan perangkat, kapasitas jaringan, dan kebutuhan analisis. Sistem yang sekadar menampilkan kondisi, memberi peringatan saat melewati ambang, atau memprediksi gangguan memerlukan data dan pengolahan yang berbeda. Pemetaan arsitektur juga membantu menelusuri sumber masalah, mulai dari sensor dan konektivitas hingga layanan penerimaan dan aplikasi.

Agar layak dianalisis, data sensor harus dikumpulkan beserta konteks yang memadai. Identitas perangkat, waktu pengukuran, satuan, lokasi, dan status perangkat membantu memastikan bahwa nilai dapat ditafsirkan serta dibandingkan dengan benar. Pengelolaan mutu mencakup pemeriksaan rentang nilai, data hilang, duplikasi, ketidakkonsistenan urutan waktu, dan pola pengiriman yang tidak biasa. Rekaman yang meragukan dapat ditandai untuk diperiksa, sementara data yang tidak layak dapat dikeluarkan dari pelatihan model. Konteks penggunaan juga menentukan rancangan akuisisi dan siapa yang berwenang

mengakses informasi, terutama ketika data berkaitan dengan layanan kesehatan.

Bagi Sistem Informasi, pengelolaan IoT bukan sekadar menghubungkan sensor ke aplikasi. Organisasi perlu menyelaraskan tujuan analisis, mutu dan konteks data, kapasitas infrastruktur, tanggung jawab pengelolaan, serta tindak lanjut operasional. Dengan pemetaan dan pemeriksaan yang konsisten, data sensor dapat menjadi dasar pemantauan, peringatan, maupun prediksi yang lebih dapat dipercaya.

I. Latihan Mahasiswa

a. Essay

1. Analisis hubungan antara frekuensi sampling sensor, kapasitas jaringan, kebutuhan penyimpanan, dan tujuan aplikasi pada fasilitas pengolahan pangan. Rumuskan pertimbangan yang dapat digunakan untuk memilih frekuensi sampling yang memadai tanpa mengumpulkan data berlebihan.
2. Bandingkan kebutuhan data dan alur analitik untuk sistem yang hanya menampilkan suhu ruang produksi dengan sistem yang memperingatkan operator ketika suhu melewati batas dan sistem yang memprediksi gangguan peralatan. Jelaskan dampak perbedaan tujuan tersebut terhadap arsitektur dan pengelolaan data.

3. Evaluasi konsekuensi jika rekaman sensor disimpan tanpa identitas perangkat, penanda waktu, satuan, atau lokasi pengukuran. Telaah bagaimana kekurangan konteks itu dapat memengaruhi penggabungan data, interpretasi pola, penelusuran gangguan, dan keandalan model penambangan data.
4. Telaah prosedur pemeriksaan mutu untuk data sensor yang mencakup nilai di luar rentang, data hilang, duplikasi, urutan waktu tidak konsisten, dan perubahan pola pengiriman. Rumuskan cara membedakan rekaman yang perlu ditandai untuk pemeriksaan dari rekaman yang layak dikeluarkan dari pelatihan model.
5. Argumentasikan mengapa pemetaan arsitektur IoT dapat digunakan sekaligus untuk merancang layanan analitik dan menelusuri masalah operasional. Gunakan contoh kegagalan data pada aplikasi pemantauan untuk menjelaskan peran perangkat, konektivitas, platform, analitik, dan aplikasi.

b. Pilihan Berganda

1. Sensor suhu mengirimkan pengukuran melalui Wi-Fi ke gerbang, lalu data diterima dan disimpan oleh layanan platform sebelum ditampilkan pada dasbor. Keputusan tentang pemeriksaan validitas nilai dan agregasi data paling tepat ditempatkan pada lapisan....
    a. Perangkat
    b. Konektivitas
    c. Platform

d. Aplikasi
e. Aktuator

2. Sebuah sistem meningkatkan frekuensi sampling untuk menangkap perubahan suhu yang cepat. Dampak yang paling perlu dipertimbangkan bersama manfaat tersebut adalah....
    a. Berkurangnya kebutuhan untuk mencatat waktu pengukuran
    b. Bertambahnya lalu lintas jaringan dan kebutuhan penyimpanan
    c. Hilangnya kebutuhan validasi data pada platform
    d. Berkurangnya kebutuhan menentukan tujuan aplikasi
    e. Terjaminnya ketepatan seluruh nilai sensor
3. Dua sensor mengukur parameter yang sama, tetapi menggunakan satuan dan lokasi berbeda. Informasi yang paling membantu mencegah interpretasi keliru adalah....
    a. Identitas perangkat, satuan, waktu, dan lokasi pengukuran
    b. Nama aplikasi yang menampilkan hasil pengukuran
    c. Frekuensi peringatan yang dikirim kepada operator
    d. Jenis visualisasi yang digunakan pada dasbor
    e. Jumlah pengguna yang membuka platform

4. Dalam pemeriksaan mutu, ditemukan beberapa nilai sensor yang tidak masuk akal, tetapi belum diketahui apakah penyebabnya kerusakan sensor atau kondisi khusus. Tindakan yang paling tepat adalah....
    a. Menghapus seluruh data dari sensor tersebut
    b. Mengganti nilai dengan rata-rata seluruh sensor
    c. Menandai rekaman untuk pemeriksaan sebelum menentukan penggunaannya
    d. Memasukkan semua nilai langsung ke pelatihan model
    e. Mengubah satuan pengukuran agar nilai tampak wajar
5. Data sensor tidak muncul di aplikasi pemantauan. Langkah penelusuran awal yang paling sesuai dengan pemetaan arsitektur adalah....
    a. Mengganti metode analitik tanpa memeriksa aliran data
    b. Memeriksa perangkat, konektivitas, dan layanan penerimaan data secara berurutan
    c. Menghapus data lama dari aplikasi
    d. Menaikkan frekuensi sampling seluruh sensor
    e. Mengubah ambang peringatan pada dasbor
6. Sistem IoT digunakan untuk memantau pasien melalui perangkat layanan kesehatan. Keputusan pengelolaan yang paling sesuai dengan konteks tersebut adalah....
    a. Memberikan akses data kepada seluruh pengguna platform

b. Memusatkan perhatian hanya pada keberhasilan pengiriman nilai
c. Mengaitkan data dengan objek pemantauan yang benar dan membatasi akses bagi pengguna berwenang
d. Menghapus identitas perangkat untuk menyederhanakan penyimpanan
e. Mengabaikan waktu pengukuran selama nilai sensor tersedia

7. Organisasi ingin memprediksi gangguan mesin, bukan hanya menampilkan kondisi saat ini. Perubahan rancangan yang paling relevan adalah....
   a. Menentukan data dan fitur yang mendukung tujuan prediksi serta lokasi pemrosesannya
   b. Menghilangkan pencatatan identitas sensor
   c. Mengganti aplikasi pemantauan dengan konektivitas saja
   d. Menganggap seluruh data sensor otomatis berguna untuk model
   e. Mengabaikan pemeriksaan mutu karena model dapat mengoreksi semua kesalahan
8. Dua aliran data sensor akan digabungkan untuk menganalisis perubahan kondisi dari waktu ke waktu. Langkah yang paling tepat sebelum penggabungan adalah....
   a. Memastikan penanda waktu, identitas perangkat, dan satuan dapat ditafsirkan secara konsisten

a. Menghapus nilai yang berbeda dari sensor pertama
b. Mengurutkan data berdasarkan nama aplikasi
c. Menggabungkan nilai tanpa memeriksa lokasi pengukuran
d. Menyimpan hanya nilai terbaru tanpa konteks

9. Pemrosesan analitik pada gerbang lokal dipertimbangkan untuk sistem yang harus merespons cepat dan memiliki konektivitas terbatas. Evaluasi yang paling tepat adalah....
    a. Pemrosesan lokal selalu lebih akurat daripada pemrosesan platform
    b. Pemrosesan lokal dapat mengurangi kebutuhan pengiriman dan latensi, tetapi perlu mempertimbangkan kemampuan perangkat
    c. Pemrosesan lokal menghilangkan kebutuhan penyimpanan data
    d. Pemrosesan lokal membuat identitas dan waktu sensor tidak diperlukan
    e. Pemrosesan lokal menjamin tidak ada data hilang
10. Catatan pengiriman menunjukkan waktu yang tidak berurutan dan beberapa rekaman dengan isi sama. Pemeriksaan mutu yang paling langsung diperlukan adalah....
    a. Pemeriksaan satuan dan warna dasbor
    b. Pemeriksaan duplikasi dan konsistensi urutan waktu
    c. Pemeriksaan jumlah pengguna aplikasi

d. Pemeriksaan kapasitas aktuator
e. Pemeriksaan bentuk peringatan yang ditampilkan

c. Soal Project / Studi Kasus

1. Sebuah fasilitas pengolahan pangan memasang sensor suhu di beberapa ruang produksi. Nilai dikirim melalui jaringan lokal ke gerbang IoT, lalu ke platform dan dasbor supervisor. Sistem saat ini mengambil data setiap beberapa detik, tetapi jaringan sering padat dan biaya penyimpanan meningkat. Supervisor membutuhkan peringatan cepat ketika suhu melewati batas, sedangkan manajer juga ingin mengenali pola yang mendahului gangguan mesin. Beberapa catatan tidak memiliki satuan yang seragam dan ada sensor yang sesekali berhenti mengirimkan data. Analisis arsitektur dan mutu data sistem tersebut. Rumuskan rancangan sampling, pemeriksaan, penyimpanan, dan analitik yang sesuai dengan kedua kebutuhan organisasi.
2. Sebuah dinas kota menggabungkan data dari sensor kualitas udara yang dipasang pada lokasi berbeda. Dasbor menampilkan nilai yang kadang tidak sejalan dengan pengamatan petugas lapangan. Pemeriksaan awal menemukan perbedaan satuan, jam perangkat yang tidak seragam, rekaman berulang, dan jeda pengiriman dari beberapa sensor. Pimpinan meminta hasil yang dapat digunakan untuk menentukan lokasi pemantauan prioritas, tetapi ingin menghindari kesimpulan yang menyesatkan. Telaah kemungkinan sumber masalah pada perangkat, konektivitas,

platform, dan aplikasi. Susun prosedur validasi serta rekomendasi pengelolaan data sebelum analisis digunakan untuk mendukung keputusan.

BAB 11
DATA MINING UNTUK KEAMANAN SISTEM INFORMASI

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep data mining untuk merumuskan kebutuhan analisis keamanan berdasarkan karakteristik data dan konteks ancaman pada sistem informasi.
2. Menggunakan teknik pengolahan data untuk menyiapkan log sistem dan network traffic sebagai masukan analisis keamanan.
3. Mengimplementasikan pendekatan machine learning untuk mendeteksi intrusi serta menentukan metrik yang sesuai guna mengevaluasi hasil deteksi.
4. Memilih metode analisis yang sesuai untuk mengidentifikasi indikasi malware, phishing, fraud, dan ancaman lain berdasarkan karakteristik data yang tersedia.
5. Menerapkan security analytics dan cyber threat intelligence untuk menghubungkan temuan analisis dengan konteks ancaman serta mendukung penentuan respons keamanan.

B. Pendahuluan

Sistem informasi menghasilkan jejak digital dalam jumlah besar melalui autentikasi, akses aplikasi, transaksi, komunikasi jaringan, dan perubahan konfigurasi. Jejak tersebut dapat menjadi sumber bukti untuk mendeteksi serangan, tetapi volumenya membuat pemeriksaan manual sulit dilakukan secara konsisten dan tepat waktu. Tantangan ini terlihat, misalnya, pada deteksi intrusi jaringan: pola serangan dapat berubah, sementara aktivitas berbahaya sering kali menyerupai penggunaan normal. Sommer dan Paxson (2010) menunjukkan bahwa penerapan machine learning pada deteksi intrusi perlu mempertimbangkan keterbatasan asumsi tentang data dan lingkungan operasional, bukan hanya kinerja model pada data uji. Karena itu, data mining dalam keamanan siber tidak sekadar mencari pola, tetapi menerapkan metode analisis secara cermat terhadap data yang bising, tidak seimbang, dan bergantung pada konteks.

Bagi mahasiswa Sistem Informasi, pemahaman tersebut menghubungkan kemampuan analitis dengan pengelolaan proses, teknologi, dan risiko organisasi. Keputusan keamanan dapat memengaruhi ketersediaan layanan, kerahasiaan data, kepatuhan, serta kepercayaan pengguna; hasil deteksi yang keliru pun dapat membebani tim operasional atau mengganggu layanan yang sah. Analisis log dan network traffic, misalnya, memerlukan pemilihan

atribut, penanganan data yang tidak lengkap, serta pemahaman atas pola aktivitas normal sebelum anomali dapat ditafsirkan. Pada tahap evaluasi, akurasi saja tidak selalu memadai, terutama ketika jumlah kejadian serangan jauh lebih kecil daripada aktivitas normal. Buczak dan Guven (2016) membahas beragam pendekatan machine learning untuk analisis keamanan siber sekaligus memperlihatkan pentingnya menyesuaikan metode dengan jenis data dan tujuan deteksi. Dengan demikian, penerapan data mining perlu disertai pertimbangan mengenai kualitas data, konsekuensi kesalahan klasifikasi, privasi, dan keterjelasan hasil bagi pengambil keputusan.

Bab ini membangun kemampuan untuk menerapkan analisis tersebut dalam alur kerja keamanan sistem informasi. Pembahasan dimulai dengan konsep data mining dalam cybersecurity, kemudian menguraikan pengolahan dan analisis log serta network traffic sebagai dasar ekstraksi informasi. Selanjutnya, mahasiswa mempelajari penerapan machine learning untuk intrusion detection, termasuk pemilihan pendekatan dan evaluasi hasilnya. Cakupan deteksi diperluas pada malware, phishing, fraud, dan ancaman lain yang memiliki karakteristik data berbeda. Pada bagian akhir, security analytics dan cyber threat intelligence dibahas sebagai cara mengaitkan keluaran analisis dengan konteks ancaman dan kebutuhan respons organisasi; pendekatan ini sejalan dengan panduan NIST

mengenai pertukaran informasi ancaman siber (NIST SP 800-150, 2016).

C. Konsep Data Mining dalam Cybersecurity

Penambangan data untuk keamanan siber menerapkan teknik analisis terhadap jejak digital sistem informasi guna menemukan pola aktivitas, mengidentifikasi penyimpangan, dan membantu menentukan tindakan keamanan. Berbeda dari penggunaan penambangan data untuk memahami perilaku pelanggan atau memprediksi permintaan, analisis keamanan berhadapan dengan aktivitas yang sengaja disamarkan, data yang tidak seimbang, dan pola ancaman yang berubah. Oleh karena itu, hasil analisis perlu ditafsirkan dengan mempertimbangkan konteks aset, proses bisnis, dan dampak insiden terhadap organisasi (Dua, 2016; Jain, 2022).

Sumber data keamanan dapat mencakup catatan autentikasi, log aplikasi, rekaman firewall, lalu lintas jaringan, data endpoint, dan laporan insiden. Setiap sumber menggambarkan bagian berbeda dari aktivitas sistem. Catatan autentikasi, misalnya, dapat menunjukkan waktu masuk, identitas akun, alamat sumber, serta keberhasilan atau kegagalan akses. Lalu lintas jaringan menyediakan petunjuk tentang komunikasi antarperangkat, sedangkan log aplikasi dapat menunjukkan tindakan yang dilakukan setelah pengguna berhasil masuk. Penggabungan sumber

tersebut membantu membentuk gambaran peristiwa yang lebih utuh, tetapi memerlukan penyelarasan format, waktu, dan identitas entitas (Dua, 2016).

Tujuan analisis harus dirumuskan sebelum memilih algoritma. Organisasi dapat berfokus pada klasifikasi kejadian yang telah diketahui, pengelompokan aktivitas yang memiliki kemiripan, pencarian pola hubungan antarperistiwa, atau penentuan prioritas untuk pemeriksaan analis. Pilihan ini menentukan kebutuhan label, jenis fitur, jangka waktu pemrosesan, dan bentuk keluaran yang dapat digunakan. Sebagai contoh, model yang mengurutkan peringatan berdasarkan tingkat risiko membantu analis memusatkan perhatian, tetapi tidak dengan sendirinya membuktikan bahwa suatu aktivitas merupakan serangan (Jain, 2022).

Dalam konteks Indonesia, sebuah perguruan tinggi di Yogyakarta dapat menghubungkan log akses sistem akademik, jaringan kampus, dan layanan surat elektronik untuk menelusuri percobaan pengambilalihan akun. Kegagalan masuk yang berulang dari satu alamat sumber mungkin tampak mencurigakan, tetapi perlu dibandingkan dengan pola autentikasi pengguna, jadwal layanan, dan bukti lain. Jika mahasiswa sedang menggunakan jaringan luar kampus atau mengalami masalah kata sandi, lonjakan kegagalan tidak otomatis berarti serangan. Pembedaan ini

memperlihatkan bahwa nilai suatu pola bergantung pada konteks operasional, bukan hanya pada skor model.

Prapemrosesan data dan rekayasa fitur membantu mengubah catatan mentah menjadi masukan analisis yang lebih konsisten. Contoh fitur mencakup jumlah kegagalan autentikasi dalam jendela waktu, perubahan lokasi akses, rasio koneksi yang ditolak, atau banyaknya tujuan jaringan yang dihubungi satu perangkat. Pemilihan fitur harus mempertahankan informasi yang relevan sekaligus menghindari atribut yang tidak diperlukan. Pengembangan fitur yang baik membantu algoritma menangkap karakteristik aktivitas, tetapi tidak menggantikan pemeriksaan mutu data dan pengetahuan tentang sistem yang dianalisis (Liu, 2018).

Hasil penambangan data sebaiknya dipandang sebagai dukungan bagi proses pengambilan keputusan keamanan. Analis perlu memeriksa bukti yang mendasari peringatan, mencatat keputusan, dan menggunakan hasil pemeriksaan untuk memperbaiki label atau aturan analisis. Siklus tersebut membuat keluaran model dapat dikaitkan dengan tindakan organisasi, seperti meminta verifikasi pengguna, membatasi akses sementara, atau melakukan pemeriksaan endpoint. Dengan demikian, keberhasilan penerapan ditentukan tidak hanya oleh kemampuan model membedakan pola, tetapi juga oleh kesesuaian hasilnya

dengan kebutuhan respons dan risiko sistem (Dua, 2016; Jain, 2022).

D. Analisis Log dan Network Traffic

Analisis log dan lalu lintas jaringan menyediakan dasar untuk memahami aktivitas sistem sebelum suatu kejadian ditetapkan sebagai ancaman. Log umumnya merekam peristiwa yang dihasilkan sistem operasi, aplikasi, perangkat keamanan, atau layanan autentikasi. Lalu lintas jaringan merekam komunikasi antarperangkat, baik sebagai paket maupun ringkasan aliran komunikasi. Kedua sumber ini berbeda dalam tingkat rincian dan cara pencatatannya, sehingga perlu dipahami secara terpisah sebelum informasi lintas-sumber digabungkan (Dua, 2016).

Langkah awal adalah menetapkan ruang lingkup sumber, rentang waktu, dan tujuan pemeriksaan. Data kemudian diperiksa untuk memastikan format waktu, zona waktu, identitas perangkat, dan struktur bidang dapat dibandingkan. Log yang tidak mencatat waktu secara konsisten dapat menghasilkan urutan kejadian keliru; identitas alamat jaringan yang berubah juga dapat menyulitkan penautan aktivitas ke perangkat. Prapemrosesan data dapat mencakup penghapusan duplikasi, penanganan nilai kosong, normalisasi format, dan pencatatan asal data agar hasil analisis dapat ditelusuri kembali (Unknown, 2011).

Setelah itu, fitur dibentuk sesuai jenis kejadian yang hendak ditemukan. Untuk aktivitas autentikasi, fitur dapat berupa jumlah kegagalan masuk, proporsi kegagalan terhadap seluruh percobaan, serta variasi akun yang diakses dalam interval waktu tertentu. Untuk lalu lintas jaringan, fitur dapat mencakup jumlah koneksi, durasi aliran, volume data, dan keragaman tujuan komunikasi. Jendela waktu harus dipilih berdasarkan pola layanan: interval singkat dapat menonjolkan lonjakan, sedangkan interval lebih panjang membantu melihat perubahan kebiasaan. Fitur yang dirancang dengan tepat meningkatkan kegunaan data bagi pemodelan (Liu, 2018).

Tabel berikut memuat data sintetis yang menggambarkan bagaimana beberapa rekaman dapat diringkas menjadi fitur untuk pemeriksaan awal. Nilai ini bersifat ilustratif, bukan hasil pengukuran dari organisasi tertentu.

Tabel 11.1 Contoh Ringkasan Aktivitas Keamanan

Waktu	Sumber	Kegagalan Masuk	Tujuan Jaringan	Volume Keluar
09.00	10.2.4.18	1	3	2 MB
09.05	10.2.4.18	8	4	3 MB
09.10	10.2.4.18	14	5	4 MB
09.15	10.2.4.27	0	2	1 MB
09.20	10.2.4.18	2	19	86 MB

Baris pada pukul 09.05 dan 09.10 menunjukkan peningkatan kegagalan masuk dari sumber yang sama,

sedangkan catatan pukul 09.20 memperlihatkan kenaikan jumlah tujuan dan volume keluar. Perubahan tersebut dapat menjadi petunjuk untuk ditelaah, tetapi tidak cukup untuk menyimpulkan adanya serangan tanpa konteks tambahan.

Di sebuah kampus di Surabaya, tim teknologi informasi dapat menerapkan pendekatan serupa untuk meninjau log autentikasi portal mahasiswa dan ringkasan lalu lintas jaringan. Aktivitas yang tidak biasa perlu dicocokkan dengan jadwal penggunaan, pemeliharaan layanan, dan kepemilikan alamat jaringan. Dengan menghubungkan waktu dan identitas perangkat secara hati-hati, analis dapat mengurangi risiko menganggap kegiatan sah sebagai insiden atau melewatkan rangkaian kejadian yang tersebar di beberapa sumber (Dua, 2016).

E. Intrusion Detection Berbasis Machine Learning

Intrusion detection berbasis pembelajaran mesin menggunakan fitur dari aktivitas sistem atau jaringan untuk menandai kejadian yang berpotensi melanggar kebijakan keamanan. Sistem dapat menghasilkan klasifikasi seperti aktivitas normal dan mencurigakan, skor risiko, atau daftar kejadian untuk diperiksa analis. Keluaran tersebut merupakan indikasi berdasarkan pola dalam data, bukan kepastian bahwa intrusi telah terjadi. Penggunaan model

perlu mempertimbangkan karakteristik data, perubahan lingkungan, serta konsekuensi kesalahan deteksi (Dua, 2016; Amanda, 2025).

Pendekatan terawasi dapat diterapkan ketika kumpulan data memiliki label kejadian yang cukup tepercaya. Model dilatih menggunakan contoh aktivitas yang sudah ditandai, lalu digunakan untuk mengklasifikasikan data baru. Tantangan utamanya ialah ketimpangan kelas: kejadian normal biasanya jauh lebih banyak daripada kejadian intrusi, sementara variasi serangan baru mungkin belum terwakili dalam pelatihan. Pendekatan tak terawasi dapat membantu menemukan kelompok atau penyimpangan tanpa label lengkap, tetapi pola yang tidak biasa belum tentu berbahaya. Pilihan pendekatan sebaiknya mengikuti ketersediaan label dan tujuan operasional (Jain, 2022; Unknown, 2019).

Penerapan dapat dimulai dengan menentukan unit analisis, misalnya satu koneksi, satu sesi pengguna, atau kumpulan aktivitas dalam interval waktu. Fitur kemudian dipilih untuk mewakili perilaku yang relevan, seperti jumlah koneksi, durasi sesi, kegagalan autentikasi, atau variasi tujuan komunikasi. Data pelatihan, validasi, dan pengujian perlu dipisahkan sedemikian rupa agar kejadian yang sama tidak muncul pada lebih dari satu bagian dan memberi kesan kinerja yang terlalu tinggi. Perancangan fitur yang

mempertahankan informasi penting menjadi bagian utama dari mutu masukan model (Liu, 2018).

Evaluasi harus menyesuaikan biaya kesalahan. Positif palsu dapat memicu pemeriksaan yang tidak perlu, sedangkan negatif palsu dapat membuat aktivitas berbahaya luput dari perhatian. Karena itu, akurasi saja tidak memadai, terutama saat data tidak seimbang. Precision membantu menilai proporsi peringatan positif yang benar, recall menunjukkan proporsi kejadian positif yang berhasil ditemukan, dan F1-score merangkum keseimbangan kedua ukuran tersebut. Pemilihan ambang keputusan juga perlu mempertimbangkan kemampuan tim untuk meninjau peringatan (Dua, 2016).

Gambar berikut menggambarkan konteks operasional tempat keluaran deteksi intrusi ditinjau bersama informasi keamanan lain. Adegan ini menekankan bahwa hasil model menjadi masukan bagi pemeriksaan analis, bukan keputusan otomatis yang berdiri sendiri.



Gambar 11.1 Analis Meninjau Peringatan Intrusi. Sumber: Ilustrasi, diolah penulis

Adegan tersebut menunjukkan peran analis dalam memeriksa peringatan dengan memperhatikan konteks sistem dan prioritas organisasi. Alur peninjauan semacam ini membantu menghubungkan skor model dengan keputusan tindak lanjut, seperti meminta verifikasi atau menyelidiki perangkat terkait.

Sebagai contoh penerapan di Indonesia, pengelola jaringan rumah sakit di Bandung dapat menguji model pada lalu lintas jaringan yang telah dipisahkan dari sistem produksi. Jika model menandai komunikasi yang menyimpang, analis perlu memeriksa apakah hal tersebut berasal dari pemeliharaan, pembaruan perangkat lunak, atau aktivitas yang tidak sah. Uji terbatas dan peninjauan manusia memungkinkan organisasi memahami pola positif

palsu sebelum mengandalkan model dalam operasi harian (Amanda, 2025; Dua, 2016).

F. Malware, Phishing, Fraud, dan Threat Detection

Malware, phishing, dan fraud menghasilkan bukti digital dengan karakteristik berbeda, sehingga metode deteksinya tidak selalu dapat disamakan. Analisis malware dapat memanfaatkan atribut berkas, perilaku proses, perubahan sistem, atau komunikasi jaringan. Deteksi phishing dapat menggunakan ciri pesan, tautan, domain, dan konteks pengiriman. Analisis fraud berfokus pada pola transaksi atau penggunaan akun yang menyimpang dari ketentuan dan perilaku yang diharapkan. Penambangan data dapat membantu mengolah atribut tersebut menjadi dasar pengelompokan, klasifikasi, atau penentuan prioritas investigasi (Dua, 2016; Jain, 2022).

Untuk malware, fitur statis dapat diperoleh dari metadata berkas dan karakteristik struktur berkas, sedangkan fitur perilaku dapat mencakup tindakan proses, perubahan berkas, atau pola komunikasi. Fitur perilaku berguna ketika berkas yang diamati telah berubah bentuk, tetapi membutuhkan lingkungan pengamatan yang terisolasi agar aktivitas mencurigakan tidak berdampak pada sistem nyata. Dalam kasus phishing, data teks dapat diolah untuk mengekstraksi ciri pesan, sementara atribut

tautan dan domain dapat dipakai sebagai masukan tambahan. Rekayasa fitur perlu menjaga agar data sensitif tidak disimpan atau diproses melampaui kebutuhan analisis (Liu, 2018).

Deteksi fraud memiliki tantangan khusus karena tindakan yang tampak sah dapat menjadi bagian dari penyalahgunaan, dan pola transaksi normal berbeda antarjenis layanan. Model dapat menandai transaksi untuk pemeriksaan berdasarkan gabungan fitur, misalnya waktu, frekuensi, kanal, dan perubahan kebiasaan akun. Hasil tersebut sebaiknya diperlakukan sebagai sinyal risiko, bukan dasar tunggal untuk menolak transaksi atau menuduh pemilik akun. Keputusan perlu memperhatikan konteks operasional dan prosedur verifikasi yang berlaku (Dua, 2016).

Di Indonesia, misalnya, pengelola lokapasar dapat menyaring pesan yang dikirim melalui layanan pelanggan untuk menemukan tautan yang menyerupai halaman masuk palsu. Pesan dengan tautan mencurigakan dapat diprioritaskan untuk diperiksa, lalu ditindaklanjuti melalui prosedur pelaporan platform. Pada sebuah koperasi simpan pinjam di Jawa Tengah, pola transaksi yang tidak lazim juga dapat ditandai untuk verifikasi tambahan, tetapi petugas tetap perlu memeriksa riwayat dan alasan transaksi sebelum mengambil tindakan terhadap anggota.

Keempat jenis deteksi tersebut dapat dibedakan berdasarkan sumber data dan tindak lanjut yang diperlukan. Perbandingan berikut menyajikan titik awal untuk memilih fitur yang sesuai.

Tabel 11.2 Sumber Fitur dan Tindak Lanjut Deteksi

Ancaman	Sumber Data	Contoh Fitur	Tindak Lanjut
Malware	Berkas, endpoint	Perubahan proses	Pemeriksaan terisolasi
Phishing	Pesan, tautan	Pola domain	Tinjau dan laporkan
Fraud	Transaksi, akun	Frekuensi transaksi	Verifikasi transaksi
Akses ilegal	Log, jaringan	Kegagalan autentikasi	Periksa sesi akun
Eksfiltrasi	Lalu lintas jaringan	Volume keluar	Telusuri perangkat

Perbedaan sumber data menegaskan bahwa tidak ada satu kelompok fitur yang sesuai bagi semua ancaman. Pengelola sistem dapat menggabungkan beberapa sinyal, tetapi interpretasi dan tindakan tindak lanjut tetap disesuaikan dengan jenis risiko serta kebijakan organisasi (Dua, 2016; Jain, 2022).

G. Security Analytics dan Cyber Threat Intelligence

Security analytics menggabungkan data dari berbagai sumber keamanan untuk membantu organisasi memahami kejadian, pola risiko, dan perubahan aktivitas sistem.

Cakupannya dapat meliputi pemantauan peringatan, pengaitan catatan lintas-sumber, analisis tren, serta penilaian prioritas investigasi. Cyber threat intelligence menambahkan konteks mengenai ancaman, misalnya hubungan antarindikator dan informasi yang dapat membantu organisasi menafsirkan relevansi suatu kejadian. Informasi tersebut lebih bermanfaat jika dikaitkan dengan aset, layanan, dan kebutuhan operasional organisasi (Dua, 2016).

Proses analitik dapat dimulai dengan pengumpulan data log, peringatan perangkat keamanan, laporan insiden, dan indikator ancaman yang tersedia bagi organisasi. Data perlu dinormalisasi agar atribut yang memiliki makna sama dapat dibandingkan, sekaligus mempertahankan asal dan waktu pencatatannya. Selanjutnya, analis dapat menghubungkan indikator dengan aktivitas sistem, menilai tingkat keyakinan, dan menentukan apakah temuan memerlukan penyelidikan. Tanpa pengelolaan kualitas dan konteks, banyaknya indikator berisiko menambah beban pemeriksaan tanpa membantu keputusan (Jain, 2022).

Dalam penerapan di Indonesia, tim keamanan sebuah perusahaan logistik di Jakarta dapat menggabungkan peringatan dari firewall, catatan akses aplikasi pengiriman, dan laporan pengguna tentang pesan mencurigakan. Jika suatu akun menunjukkan pola akses yang tidak sesuai dengan aktivitas biasanya, analis dapat memeriksa apakah

terdapat perubahan akses ke layanan lain atau komunikasi yang tidak lazim. Informasi tentang ancaman membantu memberi konteks, tetapi keputusan seperti membatasi akun harus didukung bukti yang relevan dan mengikuti prosedur organisasi. Contoh ini memperlihatkan hubungan antara keluaran analitik dan tindakan operasional.

Agar temuan dapat digunakan lintas-unit, organisasi perlu menyepakati cara mencatat indikator, sumber informasi, tingkat keyakinan, waktu pengamatan, dan tindakan yang telah dilakukan. Informasi yang dibagikan harus cukup terstruktur untuk ditafsirkan penerima, tetapi tidak mengungkap data sensitif yang tidak diperlukan. Praktik ini mendukung koordinasi antara tim teknologi informasi, pengelola risiko, dan unit layanan. Dalam konteks sistem dengan data sensitif, pengaturan akses serta perlindungan data saat diproses juga perlu menjadi bagian dari rancangan analitik (Arman, 2026).

Hasil analitik sebaiknya diterjemahkan menjadi keputusan yang dapat dijalankan, seperti memprioritaskan pemeriksaan, meminta konfirmasi pengguna, memperbarui aturan pemantauan, atau membatasi akses secara sementara. Organisasi perlu merekam alasan keputusan dan hasil tindak lanjut agar dapat meninjau apakah peringatan tersebut berguna. Umpan balik dari analis dapat membantu memperbaiki kualitas label dan relevansi fitur, tetapi perubahan model tetap perlu diuji sebelum

diterapkan luas. Pendekatan ini membantu menjaga agar proses analitik responsif terhadap perubahan ancaman tanpa menjadikan setiap penyimpangan sebagai insiden terkonfirmasi (Dua, 2016; Liu, 2018).

Penerapan security analytics juga memerlukan batas penggunaan yang jelas. Log keamanan dapat memuat identitas pengguna, alamat jaringan, atau informasi aktivitas yang bersifat sensitif. Pengumpulan dan penggunaan data sebaiknya dibatasi pada kebutuhan yang telah ditetapkan, dengan pengaturan hak akses dan masa penyimpanan yang sesuai. Dengan demikian, organisasi dapat menggunakan hasil analisis untuk memperkuat respons keamanan tanpa mengabaikan perlindungan informasi dan akuntabilitas keputusan (Arman, 2026).

H. Rangkuman

Penambangan data untuk keamanan sistem informasi mengolah jejak digital guna mengenali pola aktivitas, menemukan penyimpangan, dan membantu menentukan prioritas pemeriksaan. Berbeda dari analisis yang berfokus pada perilaku pelanggan, analisis keamanan menghadapi aktivitas yang dapat disamarkan, ketimpangan jumlah kejadian normal dan ancaman, serta pola yang berubah. Karena itu, hasil model harus dibaca bersama konteks aset, proses bisnis, dan dampak potensial insiden. Sumber data seperti log autentikasi, log aplikasi, firewall, endpoint, dan

lalu lintas jaringan merekam bagian berbeda dari suatu peristiwa. Penggabungannya dapat menghasilkan gambaran yang lebih utuh, tetapi mensyaratkan penyelarasan waktu, format, dan identitas perangkat atau akun.

Tujuan analisis menentukan metode, fitur, dan bentuk keluaran yang sesuai. Klasifikasi memerlukan label yang dapat dipercaya, sedangkan pengelompokan atau pencarian hubungan antarperistiwa dapat membantu ketika tujuan pemeriksaan berbeda. Skor risiko dan daftar peringatan membantu mengurutkan pekerjaan analis, tetapi bukan bukti bahwa serangan telah terjadi. Sebelum pemodelan, data perlu diperiksa mutunya, dinormalisasi, diselaraskan, dan ditelusuri asalnya. Fitur seperti jumlah kegagalan masuk dalam jendela waktu, variasi akun yang diakses, jumlah tujuan komunikasi, serta volume lalu lintas keluar merangkum aktivitas agar dapat dianalisis. Pemilihan interval waktu dan fitur perlu mengikuti pola layanan yang diamati.

Analisis log dan lalu lintas jaringan dapat mengungkap perubahan yang patut ditelaah, tetapi lonjakan kegagalan masuk atau volume keluar juga mungkin berkaitan dengan penggunaan sah, pemeliharaan, atau perubahan jaringan. Sistem deteksi intrusi berbasis pembelajaran mesin dapat menghasilkan klasifikasi maupun skor untuk mendukung pemeriksaan, dengan kinerja yang dipengaruhi mutu label,

ketimpangan kelas, dan perubahan lingkungan. Penerapan yang bertanggung jawab menghubungkan peringatan dengan pemeriksaan bukti, pencatatan keputusan, dan tindakan yang sebanding dengan risiko, seperti verifikasi pengguna atau pembatasan akses sementara. Umpan balik dari pemeriksaan dapat memperbaiki label dan analisis berikutnya. Dengan demikian, manfaat penambangan data bergantung pada keterpaduan teknologi, pengetahuan operasional, dan keputusan manusia.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana penggabungan log autentikasi, log aplikasi, dan lalu lintas jaringan dapat memperjelas dugaan pengambilalihan akun. Jelaskan pula masalah penyelarasan waktu, identitas, dan konteks operasional yang dapat menyebabkan analis menghubungkan peristiwa secara keliru.
2. Bandingkan penggunaan klasifikasi terawasi dengan pengelompokan aktivitas ketika organisasi menyelidiki pola ancaman yang belum dikenal. Evaluasi kebutuhan data, keluaran, dan keterbatasan masing-masing pendekatan, lalu rumuskan kondisi yang mendukung pemilihan salah satunya.
3. Evaluasi fitur jumlah kegagalan masuk, variasi akun yang diakses, jumlah tujuan jaringan, dan volume data keluar untuk mendeteksi aktivitas mencurigakan. Telaah pengaruh jendela waktu dan pola layanan

terhadap interpretasi fitur, serta jelaskan mengapa indikator tersebut belum membuktikan adanya serangan.

4. Argumentasikan bagaimana organisasi sebaiknya menggunakan skor risiko dari sistem deteksi intrusi dalam alur respons keamanan. Rumuskan langkah pemeriksaan, pencatatan keputusan, dan tindakan yang proporsional agar manfaat model tetap terjaga tanpa memperlakukan setiap peringatan sebagai kepastian insiden.
5. Telaah risiko mutu data dan ketimpangan kelas dalam pengembangan model deteksi intrusi. Analisis dampaknya terhadap hasil klasifikasi dan prioritas pemeriksaan, kemudian rumuskan cara menggabungkan pemeriksaan data, pengetahuan sistem, dan umpan balik analis untuk menjaga kegunaan model.

b. Pilihan Berganda

1. Sebuah tim menggabungkan log autentikasi dan log aplikasi, tetapi waktu pada kedua sumber menggunakan zona waktu berbeda. Risiko utama jika perbedaan ini tidak ditangani adalah...
    a. Fitur jumlah kegagalan masuk otomatis menjadi lebih informatif
    b. Urutan kejadian dapat keliru sehingga hubungan antarperistiwa salah ditafsirkan
    c. Label kejadian akan selalu berubah menjadi tidak terawasi

d. Volume lalu lintas jaringan tidak dapat dihitung
e. Model hanya dapat digunakan untuk data historis

2. Tujuan utama menyusun fitur jumlah kegagalan masuk dalam jendela waktu tertentu adalah...
   a. Menghapus kebutuhan untuk memeriksa log mentah
   b. Memastikan semua kegagalan autentikasi merupakan serangan
   c. Merangkum pola aktivitas agar dapat dibandingkan dan dianalisis
   d. Mengidentifikasi identitas pengguna tanpa data pendukung
   e. Menggantikan pencatatan asal data

1. Sebuah model memberi skor risiko tinggi pada akun mahasiswa setelah banyak percobaan masuk gagal. Tindakan yang paling tepat adalah...
   a. Menyatakan akun telah diambil alih dan menghapusnya
   b. Mengabaikan peringatan karena model tidak memberikan kepastian
   c. Memeriksa bukti dan konteks sebelum menentukan tindakan keamanan
   d. Mengubah semua percobaan gagal menjadi label serangan
   e. Membatasi seluruh akses jaringan kampus tanpa pemeriksaan

2. Mengapa klasifikasi terawasi membutuhkan label kejadian yang dapat dipercaya?
    a. Label digunakan sebagai contoh untuk mempelajari kategori aktivitas
    b. Label menyamakan format waktu seluruh sumber data
    c. Label menghapus kebutuhan akan fitur
    d. Label memastikan ancaman tidak berubah
    e. Label membuat jumlah kejadian normal dan ancaman selalu seimbang
3. Manakah contoh fitur yang paling langsung menggambarkan keragaman komunikasi jaringan?
    a. Jumlah akun yang dibuat oleh organisasi
    b. Jumlah tujuan jaringan yang dihubungi satu perangkat
    c. Lama waktu sejak kata sandi terakhir diubah
    d. Jumlah pegawai yang memiliki hak administratif
    e. Jumlah laporan insiden pada tahun sebelumnya
4. Apa fungsi utama pencatatan asal data selama prapemrosesan?
    a. Menghindari penggunaan jendela waktu
    b. Membantu menelusuri kembali sumber informasi yang dianalisis
    c. Mengubah seluruh aktivitas menjadi data berlabel
    d. Menghilangkan kebutuhan normalisasi
    e. Memastikan setiap peringatan merupakan insiden

5. Dalam data keamanan yang jumlah kejadian normalnya jauh lebih besar daripada kejadian ancaman, masalah yang perlu diperhatikan adalah...
    a. Ketimpangan kelas yang dapat memengaruhi hasil pemodelan
    b. Ketiadaan seluruh fitur numerik
    c. Ketidakmungkinan menggabungkan berbagai sumber
    d. Kesamaan pola semua aktivitas
    e. Tidak perlunya pemeriksaan label
6. Mengapa interval waktu pendek dan panjang dapat memberikan gambaran berbeda?
    a. Interval pendek selalu menghasilkan klasifikasi yang benar
    b. Interval panjang tidak dapat digunakan untuk fitur jaringan
    c. Interval pendek dapat menonjolkan lonjakan, sedangkan interval panjang membantu melihat perubahan kebiasaan
    d. Interval panjang menghilangkan semua nilai kosong
    e. Interval pendek memastikan aktivitas sah tidak terdeteksi
7. Dalam analisis keamanan, peningkatan volume data keluar paling tepat diperlakukan sebagai...
    a. Bukti pasti terjadinya pencurian data
    b. Indikasi yang perlu diperiksa bersama konteks dan bukti lain

c. Alasan untuk menghapus catatan jaringan terkait
d. Label yang cukup untuk melatih model terawasi
e. Bukti bahwa perangkat sumber tidak mengalami masalah

8. Bagaimana umpan balik analis dapat mendukung siklus analisis keamanan?
   a. Menggantikan seluruh kebutuhan pengumpulan data
   b. Membantu memperbaiki label atau aturan berdasarkan hasil pemeriksaan
   c. Menjamin semua aktivitas baru pernah terlihat sebelumnya
   d. Menghilangkan kebutuhan mencatat keputusan
   e. Memastikan model tidak mengalami perubahan kinerja

c. Soal Project / Studi Kasus

1. Sebuah perguruan tinggi mengamati lonjakan kegagalan masuk ke portal akademik dari satu alamat sumber pada malam hari. Pada waktu yang hampir sama, lalu lintas keluar dari perangkat tersebut meningkat dan mencakup banyak tujuan jaringan. Tim keamanan mengetahui bahwa sebagian mahasiswa sedang mengikuti ujian daring dari luar kampus, sementara layanan autentikasi juga baru menjalani pemeliharaan. Log dari portal dan jaringan memakai zona waktu berbeda, dan alamat sumber mungkin digunakan bersama oleh beberapa pengguna. Susun rencana analisis yang menentukan pemeriksaan data,

fitur dan konteks yang perlu dibandingkan, serta batas bukti sebelum menetapkan insiden. Jelaskan tindakan sementara yang proporsional dan cara mencatat hasilnya.

2. Sebuah perusahaan layanan keuangan menguji model deteksi intrusi menggunakan data historis. Sebagian besar rekaman berlabel normal, sedangkan label ancaman berasal dari laporan insiden yang tidak selalu lengkap. Setelah pembaruan aplikasi dan perubahan pola kerja jarak jauh, model menghasilkan lebih banyak peringatan berisiko tinggi. Analis hanya mampu memeriksa sebagian peringatan setiap hari, dan manajemen meminta pemblokiran otomatis untuk semua peringatan tertinggi. Evaluasi risiko teknis dan operasional dari permintaan tersebut. Rancang alur pemeriksaan yang memanfaatkan skor, bukti lintas-sumber, umpan balik analis, dan tindakan yang sebanding dengan konsekuensi bagi layanan.

BAB 12
ANOMALY DETECTION PADA IOT DAN KEAMANAN SISTEM

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan konsep anomali dan outlier untuk membedakan penyimpangan data dari variasi normal pada konteks operasional sistem informasi.
2. Menghitung atau menafsirkan skor deteksi anomali menggunakan pendekatan statistik, berbasis jarak, dan berbasis kepadatan.
3. Menggunakan Local Outlier Factor dan Isolation Forest untuk mengidentifikasi pola data yang berpotensi menyimpang.
4. Memilih serta menerapkan metode deteksi anomali yang sesuai untuk data sensor dan perangkat Internet of Things (IoT) dengan mempertimbangkan karakteristik data dan keterbatasan perangkat.
5. Mengimplementasikan alur deteksi anomali pada data jaringan dan aktivitas pengguna untuk membantu mengenali indikasi gangguan atau ancaman keamanan.

B. Pendahuluan

Perangkat Internet of Things (IoT) menghasilkan aliran data berkelanjutan melalui sensor, perangkat tepi, jaringan,

dan layanan komputasi yang saling terhubung. Dalam lingkungan tersebut, perubahan suhu yang tidak wajar, lonjakan lalu lintas jaringan, atau pola akses pengguna yang menyimpang dapat menjadi petunjuk awal kerusakan perangkat, kesalahan konfigurasi, maupun aktivitas berbahaya. Namun, tidak setiap nilai ekstrem merupakan ancaman: pembacaan sensor dapat berubah karena kondisi lingkungan, sedangkan lonjakan jaringan dapat terjadi akibat proses bisnis yang sah. Karena itu, deteksi anomali perlu membedakan penyimpangan yang bermakna dari variasi normal dengan memperhatikan konteks, karakteristik data, dan konsekuensi keputusan. Chandola, Banerjee, dan Kumar (2009) menjelaskan bahwa anomali merupakan pola yang tidak sesuai dengan perilaku yang diharapkan dalam suatu konteks, sehingga penilaiannya tidak dapat dilepaskan dari model tentang kondisi normal.

Bagi mahasiswa Sistem Informasi, kemampuan menerapkan deteksi anomali diperlukan untuk menghubungkan analisis data dengan pengelolaan proses, risiko, dan keamanan organisasi. Sistem informasi yang memanfaatkan IoT dapat mendukung pemantauan aset, layanan publik, fasilitas industri, maupun infrastruktur kampus; pada saat yang sama, sistem tersebut menghadapi kendala berupa data bervolume besar, kualitas pengukuran yang beragam, keterbatasan komputasi perangkat, dan perubahan pola penggunaan. Pemilihan metode yang tidak

sesuai dapat menghasilkan terlalu banyak alarm palsu atau gagal mengenali penyimpangan yang relevan. Pendekatan statistik menilai ketidaklaziman berdasarkan distribusi data, sedangkan metode berbasis jarak dan kepadatan mempertimbangkan kedekatan atau lingkungan lokal suatu titik. Breunig et al. (2000), melalui Local Outlier Factor (LOF), menunjukkan bahwa kepadatan lokal dapat membantu mengenali objek yang menyimpang dibandingkan tetangganya; Liu, Ting, dan Zhou (2008) memperkenalkan Isolation Forest yang mengisolasi titik melalui pemisahan acak untuk mendeteksi anomali secara efisien. Pemahaman atas prinsip tersebut membantu mahasiswa menyesuaikan teknik analisis dengan bentuk data dan kebutuhan operasional, bukan sekadar menjalankan algoritma.

Penerapan deteksi anomali juga menuntut proses yang sistematis. Data perlu diperiksa kualitasnya, fitur yang relevan ditentukan, dan perilaku normal dipelajari sebelum hasil deteksi ditafsirkan. Pada data sensor, analisis dapat mempertimbangkan urutan waktu, rentang pengukuran, dan hubungan antarsensor; pada jaringan dan aktivitas pengguna, pola koneksi, frekuensi permintaan, waktu akses, serta perubahan kebiasaan dapat menjadi dasar pemeriksaan. Hasil algoritma tetap memerlukan validasi karena skor anomali menunjukkan tingkat penyimpangan, bukan bukti tunggal terjadinya serangan. Dengan demikian, keputusan tindak lanjut perlu menggabungkan hasil analisis,

konteks sistem, dan kebijakan keamanan. Bab ini membahas konsep anomali dan outlier, deteksi berbasis statistik, jarak, dan kepadatan, penggunaan Local Outlier Factor serta Isolation Forest, penerapan pada sensor dan perangkat IoT, serta deteksi anomali pada jaringan dan aktivitas pengguna.

C. Konsep Anomali dan Outlier

Prapemrosesan data pada Bab 4 membahas outlier sebagai salah satu masalah kualitas data; pada bab ini, perhatian beralih pada cara menentukan apakah penyimpangan tersebut mengandung informasi operasional. **Anomali** adalah pola yang berbeda dari perilaku yang diharapkan dalam konteks tertentu. **Outlier** adalah objek atau pengamatan yang terletak jauh dari sebagian besar data menurut ukuran tertentu. Kedua istilah berkaitan, tetapi tidak selalu dapat dipertukarkan: sebuah nilai ekstrem mungkin merupakan kesalahan pencatatan, kejadian langka yang sah, atau indikasi gangguan yang perlu diperiksa. Penilaian anomali pada data runtun waktu, misalnya, perlu memperhatikan urutan dan konteks pengamatan, bukan hanya besar-kecilnya nilai (Blázquez-García, 2022).

Penyimpangan dapat muncul dalam beberapa bentuk. Anomali titik terjadi ketika satu pengamatan menyimpang dari pengamatan lain, seperti satu pembacaan suhu yang

jauh lebih tinggi daripada rentang operasi. Anomali kontekstual hanya dianggap menyimpang dalam kondisi tertentu; suhu yang wajar pada siang hari, misalnya, mungkin tidak wajar untuk ruang penyimpanan yang harus dijaga dingin. Anomali kolektif terbentuk dari sekumpulan pengamatan yang secara bersama-sama menunjukkan pola tidak biasa, meskipun tiap nilai secara individual masih tampak normal. Lonjakan kecil yang terjadi terus-menerus pada lalu lintas sebuah perangkat dapat menjadi contoh pola kolektif. Pengelompokan bentuk anomali seperti ini membantu menentukan unit analisis dan fitur yang perlu disiapkan (Blázquez-García, 2022).

Dalam sistem informasi, deteksi anomali biasanya dimulai dengan mempelajari perilaku normal dari kumpulan data, lalu memberi skor penyimpangan pada pengamatan baru. Sebagian pendekatan menggunakan data berlabel dan membedakan kelas normal serta anomali; pendekatan lain mempelajari struktur data tanpa label atau hanya menggunakan sejumlah kecil contoh gangguan. Pemilihan pendekatan perlu mengikuti ketersediaan label dan kebutuhan operasional. Seperti dibahas pada Bab 5, pembelajaran tak terawasi dapat digunakan ketika label kejadian langka sulit diperoleh. Namun, hasil model tetap perlu diperiksa karena pola yang belum pernah diamati tidak otomatis berarti ancaman. Prinsip umum penambangan data menempatkan pemahaman data dan

tujuan analisis sebagai bagian dari pemilihan serta penggunaan model (Unknown, 2012).

Sebagai ilustrasi, sebuah fasilitas penyimpanan ikan di Indonesia mencatat suhu ruangan setiap jam. Nilai yang jauh dari kebiasaan dapat muncul akibat pintu yang terbuka, kerusakan pendingin, atau sensor yang tidak berfungsi. Tabel berikut menggunakan data buatan untuk menunjukkan bagaimana satu nilai menonjol terhadap pembacaan lain. Nilai ekstrem menjadi sinyal pemeriksaan, bukan kesimpulan tentang penyebabnya.

Tabel 12.1 Contoh Pembacaan Suhu dan Penandaan Penyimpangan

Waktu	Suhu (°C)	Catatan
08.00	20	Normal
09.00	21	Normal
10.00	19	Normal
11.00	20	Normal
12.00	20	Normal
13.00	30	Perlu diperiksa

Nilai 30 °C tampak berbeda dari pengamatan lain, tetapi keputusan operasional tetap memerlukan pemeriksaan terhadap kondisi ruangan, status pendingin, dan kualitas sensor. Dengan demikian, deteksi menandai kasus untuk ditindaklanjuti, bukan menggantikan diagnosis teknis.

Karena itu, keberhasilan deteksi tidak cukup diukur dari kemampuan menemukan nilai ekstrem. Organisasi perlu mempertimbangkan biaya alarm palsu, keterlambatan

mengenali gangguan, serta kemampuan petugas menindaklanjuti notifikasi. Prapemrosesan data juga berpengaruh: perbedaan satuan, nilai hilang, dan pencilan akibat kesalahan pencatatan dapat mengubah skor deteksi. Bab 4 telah membahas penyiapan data; untuk analisis anomali, langkah tersebut perlu dipadukan dengan konteks waktu, kelompok perangkat, dan kondisi operasi agar variasi normal tidak salah ditandai sebagai kejadian abnormal (García, 2015).

D. Statistical, Distance, dan Density-Based Detection

Tiga keluarga pendekatan yang umum digunakan untuk deteksi anomali menilai penyimpangan melalui distribusi statistik, jarak antarpengamatan, atau kepadatan lingkungan lokal. Ketiganya menghasilkan ukuran yang dapat dipakai untuk memeringkat pengamatan, tetapi makna skor dan asumsi dasarnya berbeda. Pemilihan metode sebaiknya mempertimbangkan bentuk data, skala fitur, dan pola normal yang hendak dikenali. Kerangka penambangan data umum membantu menempatkan pilihan algoritma dalam kaitannya dengan struktur kumpulan data dan tujuan analisis (Unknown, 2012).

Pendekatan statistik menilai seberapa tidak lazim suatu nilai berdasarkan distribusi. Jika data mendekati distribusi normal, skor baku dapat dihitung sebagai:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Di sini, x_i adalah nilai pengamatan, μ adalah rata-rata, dan σ adalah simpangan baku. Nilai absolut z_i yang besar menunjukkan jarak dari rata-rata dalam satuan simpangan baku. Metode ini praktis untuk pemantauan satu fitur dengan pola yang relatif stabil, tetapi sensitif terhadap asumsi distribusi, nilai ekstrem yang menggeser rata-rata, serta perubahan pola dari waktu ke waktu. Pada data yang miring atau memiliki banyak nilai ekstrem, ukuran berbasis median dan rentang antarkuartil dapat lebih sesuai. Evaluasi pola runtun waktu perlu pula mempertimbangkan tren dan musiman agar perubahan yang berulang tidak otomatis dianggap menyimpang (Blázquez-García, 2022).

Pendekatan berbasis jarak mengukur seberapa jauh suatu titik dari tetangganya atau dari kelompok data yang menjadi pembanding. Salah satu penerapan sederhana menghitung jarak ke tetangga ke-k; skor besar dapat menandakan bahwa titik tersebut berada jauh dari kumpulan pengamatan lain. Jarak Euclidean dapat dipakai pada fitur numerik, tetapi perbedaan skala dapat membuat fitur bersatuan besar mendominasi perhitungan. Karena itu, standardisasi atau normalisasi perlu diterapkan sesuai karakteristik fitur. Bab 4 menguraikan transformasi skala, sedangkan pembahasan K-nearest neighbor pada Bab 6 memberikan landasan tentang penggunaan tetangga

terdekat. Prinsip jarak dan pemrosesan fitur ini merupakan bagian dari teknik umum penambangan data (García, 2015; Unknown, 2012).

Pendekatan berbasis kepadatan membandingkan kerapatan di sekitar sebuah titik dengan kerapatan wilayah lain. Titik yang berada pada wilayah jarang dapat dianggap menyimpang, meskipun jaraknya terhadap tetangga tertentu tidak terlalu besar. Sifat ini berguna ketika kumpulan data memiliki kelompok dengan kepadatan berbeda. DBSCAN, yang dibahas pada Bab 8, menggunakan kepadatan untuk membentuk klaster dan mengenali titik yang tidak masuk ke kelompok; pendekatan deteksi berbasis kepadatan memanfaatkan gagasan kedekatan lokal dengan tujuan penilaian penyimpangan. Perbedaan kepadatan dapat menjadi petunjuk yang hilang bila hanya mengandalkan jarak global (Kononenko, 2007).

Misalnya, pada pemantauan mesin pengolahan pangan di Jawa Barat, kombinasi suhu, getaran, dan arus listrik dapat dinilai dengan lebih dari satu pendekatan. Skor statistik dapat membantu memeriksa lonjakan suhu, sedangkan jarak atau kepadatan dapat menandai kombinasi fitur yang jarang terjadi. Hasil dari beberapa metode tidak harus sama karena masing-masing membuat asumsi berbeda. Analisis komparatif perlu memeriksa apakah alarm bertepatan dengan perubahan proses, jadwal perawatan, atau gangguan sensor, bukan hanya menetapkan ambang

numerik secara otomatis. Prinsip prapemrosesan dan pemilihan fitur berpengaruh langsung terhadap interpretasi skor yang dihasilkan (García, 2015).

E. Local Outlier Factor dan Isolation Forest

Local Outlier Factor (LOF) dan Isolation Forest memberikan cara berbeda untuk menilai ketidaklaziman pengamatan. LOF mengukur penyimpangan relatif berdasarkan lingkungan lokal, sedangkan Isolation Forest menilai kemudahan suatu titik dipisahkan melalui pemisahan fitur secara bertahap. Kedua metode menghasilkan skor yang membantu memeringkat kandidat anomali, tetapi skornya tidak dengan sendirinya membuktikan adanya kerusakan atau serangan. Karena deteksi anomali pada data runtun waktu bergantung pula pada struktur data dan konteks, penerapan metode perlu disertai pemeriksaan pola pengamatan (Blázquez-García, 2022).

Pada LOF, setiap titik dibandingkan dengan sejumlah tetangga terdekat. Jika kepadatan lokal suatu titik lebih rendah daripada kepadatan tetangganya, titik itu memperoleh skor LOF yang lebih tinggi dan lebih mungkin diperlakukan sebagai kandidat outlier. Jumlah tetangga, yang lazim dinyatakan sebagai parameter k, memengaruhi cakupan lingkungan yang digunakan. Nilai k terlalu kecil dapat membuat hasil sensitif terhadap noise, sedangkan

nilai terlalu besar berpotensi mengaburkan penyimpangan lokal. Metode ini sesuai ketika kelompok normal memiliki variasi kepadatan, tetapi perhitungan jarak dapat menjadi mahal pada kumpulan data yang sangat besar. Fitur berskala berbeda juga perlu ditangani sebelum jarak dihitung, sejalan dengan prinsip prapemrosesan data (García, 2015).

Isolation Forest dibangun dari kumpulan pohon yang memisahkan pengamatan menggunakan pemilihan fitur dan titik pemisah secara acak. Titik yang terisolasi dengan sedikit pemisahan cenderung memperoleh skor anomali lebih tinggi karena jarang berada di wilayah padat. Pendekatan ini dapat digunakan pada data dengan banyak fitur dan tidak memerlukan perhitungan jarak terhadap seluruh pasangan titik. Namun, hasilnya tetap dipengaruhi oleh representasi fitur, ukuran sampel, serta proporsi anomali yang diasumsikan atau ditentukan dalam konfigurasi. Kerangka penambangan data menekankan bahwa parameter dan bentuk representasi data perlu disesuaikan dengan tujuan analisis, bukan dipilih terpisah dari konteks (Unknown, 2012).

Pertimbangkan kumpulan data operasional sebuah perusahaan logistik di Indonesia yang memuat waktu pemindaian paket, durasi proses, dan jumlah pemindaian ulang. Satu paket yang diproses sangat lambat mungkin bukan anomali jika terjadi di jalur dengan pemeriksaan

tambahan. LOF dapat membantu menemukan kasus yang berbeda dari paket-paket dengan karakteristik lokal serupa, sementara Isolation Forest dapat menyaring pengamatan yang mudah dipisahkan dari keseluruhan data. Petugas kemudian memeriksa apakah penyebabnya adalah gangguan perangkat pemindai, prosedur khusus, atau kesalahan pencatatan. Contoh ini menegaskan bahwa anomali relatif terhadap kelompok pembanding dan proses bisnisnya.

Dalam penerapan, kedua metode sebaiknya diuji pada data historis yang telah ditinjau, lalu dibandingkan melalui kualitas alarm dan beban pemeriksaan yang ditimbulkan. Sampel data normal dapat berasal dari periode operasi yang dipahami, sedangkan kejadian yang telah dikonfirmasi dapat digunakan untuk memeriksa apakah skor meningkat pada situasi relevan. Jika label kejadian jarang tersedia, evaluasi dapat melibatkan pemeriksaan ahli terhadap sampel skor tertinggi. Seperti pada metode penambangan data lainnya, pembersihan data dan pemilihan fitur memengaruhi keluaran model (García, 2015; Unknown, 2012). Hasil akhirnya perlu diterjemahkan menjadi prioritas pemeriksaan yang dapat dilaksanakan, bukan sekadar daftar angka.

F. Deteksi Anomali pada Sensor dan Perangkat IoT

Data sensor dan perangkat Internet untuk Segala (IoT) umumnya diterima berulang kali menurut urutan waktu, sehingga penyimpangan perlu ditafsirkan berdasarkan perilaku perangkat dan kondisi lingkungan. Satu pembacaan yang tidak wajar dapat timbul karena perubahan fisik, kerusakan sensor, gangguan komunikasi, atau kesalahan pencatatan. Analisis runtun waktu karena itu memperhatikan nilai, perubahan antarwaktu, dan pola berulang, bukan hanya ambang tetap (Blázquez-García, 2022).

Tahap penerapan dimulai dengan menentukan peristiwa yang hendak dideteksi, misalnya suhu melampaui rentang operasi atau data sensor berhenti masuk. Data perlu dilengkapi informasi waktu, identitas perangkat, satuan, lokasi, dan status koneksi. Selanjutnya, pemeriksaan kualitas dilakukan untuk menemukan nilai hilang, duplikasi, lompatan waktu, serta nilai yang tidak mungkin secara fisik. Bab 4 membahas langkah prapemrosesan tersebut; pada analisis sensor, pemeriksaan juga perlu mencakup sinkronisasi waktu dan konsistensi antarperangkat. Sistem pemantauan kualitas udara berbasis IoT yang menggunakan DHT11, MQ-135, dan ESP8266 menunjukkan konteks pengukuran suhu, kelembapan, dan gas yang dapat dipantau berkelanjutan (Alfirdaus, 2024).

Model deteksi dapat menggunakan ambang yang ditentukan oleh kebutuhan operasional, statistik jendela bergerak, atau metode berbasis jarak dan kepadatan. Ambang tunggal sederhana untuk diterapkan pada perangkat dengan sumber daya terbatas, tetapi dapat menghasilkan alarm berlebihan bila kondisi lingkungan berubah. Jendela bergerak membantu membandingkan pembacaan terkini dengan periode terdekat; untuk mengenali perubahan hubungan beberapa sensor, fitur gabungan dapat digunakan. Prediksi kadar CO_2 berdasarkan suhu dan kelembapan dengan sensor DHT22 dan MQ135 yang terhubung ke ESP32 memperlihatkan bahwa variabel sensor dapat dianalisis bersama, bukan selalu secara terpisah (Agung Suburdjati, 2026).

Penerapan pada fasilitas kesehatan memberi contoh pentingnya konteks dan konsekuensi. Sistem pemantauan kateter berbasis IoT dirancang untuk membantu pemantauan volume agar kapasitas penampungan tidak terlampaui (Banjarnahor, 2022). Jika data menunjukkan perubahan yang tajam, sistem perlu membedakan kejadian nyata dari sensor yang lepas atau komunikasi yang terputus. Alarm yang muncul harus disertai waktu, perangkat asal, nilai terkini, dan status konektivitas agar petugas dapat memeriksanya dengan tepat. Keputusan klinis tidak boleh diserahkan hanya kepada skor anomali; sistem berfungsi mendukung kewaspadaan dan pemeriksaan manusia.

Arsitektur pemrosesan juga menentukan metode yang layak digunakan. Perangkat tepi dapat menerapkan aturan ringan untuk mendeteksi nilai mustahil atau kehilangan koneksi, sedangkan analisis yang memerlukan riwayat panjang dan gabungan banyak sensor dapat dilakukan pada layanan pusat. Pengiriman data untuk analisis perlu memperhitungkan keterbatasan energi, kapasitas jaringan, dan kemampuan perangkat. Sistem pemantauan kualitas udara menunjukkan penggunaan mikrokontroler dan platform pemantauan sebagai bagian dari alur pengumpulan serta penyajian data (Alfirdaus, 2024). Tabel berikut merangkum pilihan analisis sesuai gejala yang diamati.

Tabel 12.2 Gejala Sensor dan Tindakan Analisis

Gejala	Kemungkinan sumber	Pemeriksaan awal	Respons
Nilai melonjak	Perubahan fisik, noise	Bandingkan sensor sekitar	Verifikasi kondisi
Data berhenti	Daya, jaringan	Periksa waktu terakhir	Pulihkan koneksi
Nilai datar lama	Sensor macet	Bandingkan pola historis	Periksa perangkat
Perbedaan antarsensor	Lokasi, kalibrasi	Tinjau posisi dan satuan	Kalibrasi ulang
Pola bergeser	Perubahan lingkungan	Bandingkan periode setara	Sesuaikan acuan

Tabel tersebut menekankan bahwa bentuk penyimpangan memberi petunjuk untuk memilih pemeriksaan awal, tetapi tidak cukup untuk memastikan

penyebabnya. Respons perlu menggabungkan skor model dengan kondisi perangkat dan prosedur operasional.

G. Deteksi Anomali pada Jaringan dan Aktivitas Pengguna

Pada jaringan dan aktivitas pengguna, anomali dapat terlihat sebagai perubahan volume koneksi, frekuensi permintaan, tujuan komunikasi, waktu akses, atau pola penggunaan akun. Satu perubahan tidak selalu menunjukkan serangan: pembaruan layanan, kegiatan bisnis musiman, dan perubahan jadwal kerja juga dapat menghasilkan perilaku yang berbeda dari kebiasaan. Analisis keamanan karena itu memeriksa penyimpangan dalam konteks aset, akun, dan proses organisasi. Penggunaan penambangan data dalam keamanan siber menempatkan analisis pola sebagai salah satu cara mendukung identifikasi aktivitas yang perlu diselidiki (Dua, 2016; Jain, 2022).

Alur kerja dapat dimulai dengan menggabungkan log autentikasi, catatan koneksi jaringan, dan informasi inventaris perangkat. Data kemudian disusun menjadi fitur yang dapat dibandingkan, seperti jumlah upaya masuk dalam selang waktu tertentu, banyaknya alamat tujuan, durasi sesi, atau perubahan lokasi akses. Tahap ini berhubungan dengan analisis log dan lalu lintas jaringan yang dibahas pada Bab 11. Setelah periode perilaku normal

ditetapkan, metode statistik, jarak, kepadatan, LOF, atau Isolation Forest dapat digunakan untuk menghasilkan skor dan memprioritaskan kasus. Prapemrosesan diperlukan agar perbedaan format waktu, alamat, dan kategori tidak mengacaukan perbandingan (García, 2015).

Pada sebuah perguruan tinggi di Indonesia, misalnya, akun administrasi yang biasanya digunakan pada jam kerja dapat tiba-tiba melakukan banyak autentikasi gagal pada malam hari dan mengakses beberapa layanan baru. Sistem dapat menandai rangkaian tersebut karena menyimpang dari kebiasaan akun, lalu menghubungkannya dengan informasi perangkat, alamat jaringan, dan jadwal pemeliharaan. Sebaliknya, lonjakan akses pada masa registrasi mahasiswa mungkin merupakan kegiatan sah. Karena itu, deteksi perlu menggabungkan anomali teknis dengan kalender operasional serta konfirmasi pengelola sistem. Sistem deteksi intrusi berbasis AI pada lingkungan big data dibahas sebagai respons terhadap tantangan volume dan kompleksitas ancaman, sekaligus menunjukkan perlunya pengolahan yang sesuai dengan skala data (Amanda, 2025).

Skor anomali dapat digunakan untuk menyusun antrean pemeriksaan, bukan untuk menetapkan bahwa suatu akun telah disalahgunakan. Tim keamanan dapat memeriksa bukti tambahan, seperti perubahan hak akses, pola alamat sumber, atau korelasi waktu dengan peristiwa lain. Dalam

konteks Indonesia, kajian identifikasi bahaya keamanan siber pada proses perizinan bahan nuklir daring menegaskan relevansi pemeriksaan risiko pada sistem layanan digital yang memproses informasi penting (Imron, 2023). Namun, sebuah anomali pada sistem tertentu tetap harus diverifikasi sesuai bukti dan konteks operasionalnya.

Gambar berikut menunjukkan lingkungan operasional tempat notifikasi jaringan dapat ditinjau bersama log dan status layanan. Adegan ini menekankan bahwa hasil deteksi memerlukan pemeriksaan oleh petugas, bukan hanya keluaran otomatis.



Gambar 12.1 Pemeriksaan Anomali di Pusat Operasi Jaringan. Sumber: Ilustrasi, diolah penulis

Adegan tersebut memperlihatkan hubungan antara pemantauan otomatis dan pemeriksaan manusia: skor

mengarahkan perhatian, sedangkan petugas menilai bukti, dampak, dan langkah respons yang sesuai.

Agar penerapan berkelanjutan, organisasi perlu memantau kualitas alarm, jumlah kasus yang terbukti relevan, dan waktu yang dibutuhkan untuk menindaklanjuti temuan. Pola penggunaan berubah ketika layanan diperbarui atau struktur organisasi bergeser; model yang tidak ditinjau dapat terus menghasilkan alarm yang tidak lagi bermakna. Penyesuaian harus dilakukan dengan memperhatikan akses terhadap log dan kerahasiaan informasi pengguna. Sumber keamanan siber dan deteksi intrusi menempatkan analitik data sebagai dukungan bagi pengenalan ancaman, tetapi hasil analisis tetap perlu ditempatkan dalam proses pengawasan keamanan organisasi (Dua, 2016; Amanda, 2025).

H. Rangkuman

Anomali adalah pola yang menyimpang dari perilaku yang diharapkan dalam konteks tertentu, sedangkan outlier merupakan pengamatan yang letaknya jauh dari sebagian besar data menurut ukuran tertentu. Nilai ekstrem belum tentu menandakan gangguan: penyebabnya dapat berupa kesalahan pencatatan, kejadian langka yang sah, atau masalah yang perlu ditindaklanjuti. Anomali dapat berbentuk titik, kontekstual, atau kolektif. Pembedaan ini membantu menentukan unit analisis, fitur, serta konteks

waktu dan operasi yang perlu diperiksa. Dalam sistem informasi, deteksi umumnya mempelajari pola normal lalu memberi skor penyimpangan pada data baru. Pendekatan dapat menggunakan label, tanpa label, atau sejumlah kecil contoh gangguan. Karena pola baru tidak otomatis merupakan ancaman, hasil deteksi perlu ditafsirkan dan diverifikasi oleh petugas.

Metode statistik, berbasis jarak, dan berbasis kepadatan menilai penyimpangan dengan asumsi berbeda. Skor baku mengukur jarak nilai dari rata-rata dalam satuan simpangan baku, tetapi sensitif terhadap distribusi dan nilai ekstrem. Metode jarak membandingkan kedekatan antartitik dan memerlukan perhatian pada skala fitur. Metode kepadatan menilai kelangkaan lokal sehingga dapat mengenali penyimpangan dalam data yang memiliki kelompok dengan kepadatan berbeda. LOF menilai suatu titik relatif terhadap kepadatan lingkungan lokalnya, sedangkan Isolation Forest menilai seberapa mudah titik dipisahkan melalui pemisahan fitur bertahap.

Pada data IoT, seperti suhu ruang penyimpanan atau kombinasi suhu, getaran, dan arus mesin, skor anomali merupakan sinyal pemeriksaan, bukan diagnosis penyebab. Prapemrosesan—termasuk penanganan nilai hilang, pencilan pencatatan, dan perbedaan skala—serta pemahaman tren, musim, kelompok perangkat, dan kondisi operasi menentukan mutu hasil. Evaluasi juga perlu

mempertimbangkan alarm palsu, keterlambatan deteksi, dan kapasitas petugas menindaklanjuti notifikasi. Dalam praktik keamanan sistem, keputusan yang bertanggung jawab menggabungkan keluaran model dengan pemeriksaan konteks dan bukti operasional sebelum menetapkan tindakan. Dengan demikian, deteksi anomali mendukung pemantauan dan pengambilan keputusan, tetapi tidak menggantikan verifikasi teknis maupun pertimbangan manusia.

I. Latihan Mahasiswa

a. Essay

1. Analisislah perbedaan anomali titik, kontekstual, dan kolektif dengan menggunakan contoh pemantauan suhu fasilitas penyimpanan ikan. Jelaskan bagaimana perbedaan bentuk anomali tersebut memengaruhi unit analisis, fitur yang digunakan, dan tindak lanjut operasional.
2. Bandingkan pendekatan statistik, berbasis jarak, dan berbasis kepadatan untuk mendeteksi penyimpangan pada data sensor yang memiliki fitur suhu, getaran, dan arus listrik. Pertimbangkan asumsi, pengaruh skala fitur, serta jenis pola yang berpotensi terlewat oleh masing-masing pendekatan.
3. Evaluasilah penggunaan skor baku untuk mendeteksi gangguan pada data suhu yang berubah mengikuti waktu dan kondisi operasi. Rumuskan langkah prapemrosesan dan pemeriksaan konteks yang

diperlukan agar variasi normal tidak keliru ditandai sebagai anomali.

4. Telaah perbedaan cara LOF dan Isolation Forest menilai ketidaklaziman pengamatan. Argumentasikan jenis karakteristik data yang membuat salah satu pendekatan lebih sesuai, serta jelaskan mengapa hasil kedua metode tetap perlu diperiksa sebelum menjadi dasar tindakan.
5. Rumuskan kerangka evaluasi sistem deteksi anomali untuk perangkat IoT yang mempertimbangkan alarm palsu, keterlambatan deteksi, kualitas data, dan kemampuan petugas menindaklanjuti notifikasi. Jelaskan bagaimana kerangka tersebut menghubungkan kinerja model dengan kebutuhan operasional organisasi.

b. Pilihan Berganda

1. Satu pembacaan suhu 30 °C tercatat di antara pengukuran 19–21 °C pada ruang penyimpanan ikan. Kesimpulan awal yang paling tepat adalah....
    a. Nilai tersebut membuktikan bahwa pendingin mengalami kerusakan.
    b. Nilai tersebut merupakan sinyal penyimpangan yang perlu diperiksa bersama kondisi ruangan dan sensor.
    c. Nilai tersebut pasti merupakan kesalahan pencatatan sehingga dapat langsung dihapus.
    d. Nilai tersebut hanya dapat disebut anomali jika tersedia label gangguan.

e. Nilai tersebut menunjukkan pola kolektif karena terjadi pada satu waktu.

2. Suhu yang wajar pada siang hari tetapi tidak wajar untuk ruang penyimpanan yang harus tetap dingin merupakan contoh....
   a. anomali titik
   b. anomali kolektif
   c. pencilan akibat skala fitur
   d. anomali kontekstual
   e. variasi yang selalu normal

3. Sekumpulan lonjakan kecil pada lalu lintas sebuah perangkat terjadi berulang kali, meskipun setiap lonjakan secara individual masih tampak normal. Bentuk penyimpangan tersebut paling tepat disebut....
   a. anomali kolektif
   b. anomali titik
   c. anomali kontekstual
   d. kesalahan standardisasi
   e. outlier statistik tunggal

4. Pada data yang mendekati distribusi normal, skor baku suatu pengamatan bernilai besar secara absolut. Makna yang paling tepat adalah....
   a. Pengamatan tersebut pasti berasal dari kesalahan sensor.
   b. Pengamatan tersebut berada dekat dengan rata-rata data.

c. Pengamatan tersebut jauh dari rata-rata dalam satuan simpangan baku.
d. Pengamatan tersebut memiliki kepadatan lokal yang tinggi.
e. Pengamatan tersebut pasti lebih mudah dipisahkan oleh Isolation Forest.

5. Sebuah fitur arus listrik memiliki rentang numerik jauh lebih besar daripada fitur getaran. Tim menggunakan jarak Euclidean pada kedua fitur tanpa transformasi skala. Risiko utamanya adalah....
   a. Fitur getaran akan selalu memiliki varians nol.
   b. Jarak akan berubah menjadi ukuran kepadatan lokal.
   c. Pengamatan tidak dapat dihitung jaraknya.
   d. Metode akan otomatis mengubah kedua fitur menjadi distribusi normal.
   e. Fitur arus listrik dapat mendominasi perhitungan jarak.

6. Mengapa pendekatan berbasis kepadatan dapat berguna ketika data memiliki kelompok dengan kepadatan berbeda?
   a. Pendekatan ini hanya mengandalkan jarak dari rata-rata global.
   b. Pendekatan ini dapat menilai kelangkaan suatu titik relatif terhadap lingkungan lokalnya.
   c. Pendekatan ini menghilangkan kebutuhan untuk memeriksa kualitas data.

d. Pendekatan ini memastikan semua titik dalam klaster merupakan data normal.
e. Pendekatan ini memerlukan label untuk setiap jenis gangguan.

7. Data perangkat IoT memiliki label normal dan gangguan yang sangat terbatas. Pilihan pendekatan yang paling sesuai dengan kondisi tersebut adalah....
   a. hanya menggunakan metode yang membutuhkan label lengkap
   b. menghapus seluruh data yang tidak memiliki label
   c. menetapkan semua pengamatan baru sebagai gangguan
   d. mempertimbangkan pembelajaran tak terawasi, lalu memeriksa hasil berdasarkan konteks
   e. menyimpulkan jenis gangguan dari skor model tanpa verifikasi

8. Pola suhu ruang penyimpanan naik setiap siang akibat kondisi operasi yang berulang, tetapi model selalu menandainya sebagai anomali. Perbaikan yang paling tepat adalah....
   a. memasukkan konteks waktu dan kondisi operasi ke dalam analisis
   b. menghapus seluruh pengamatan siang hari
   c. menaikkan ambang deteksi tanpa mengevaluasi pola data
   d. mengganti semua nilai siang dengan rata-rata harian

e. menganggap seluruh nilai ekstrem sebagai kesalahan sensor

9. Perbedaan mendasar antara LOF dan Isolation Forest adalah....
   a. LOF memerlukan label lengkap, sedangkan Isolation Forest hanya untuk data runtun waktu.
   b. LOF mengukur jarak dari rata-rata, sedangkan Isolation Forest menghitung simpangan baku.
   c. LOF menilai penyimpangan relatif terhadap kepadatan lokal, sedangkan Isolation Forest menilai kemudahan titik dipisahkan.
   d. LOF hanya memproses data kategorikal, sedangkan Isolation Forest hanya memproses data teks.
   e. LOF menghapus pencilan, sedangkan Isolation Forest memperbaiki nilai sensor.

10. Organisasi mengevaluasi sistem deteksi anomali yang menghasilkan banyak notifikasi. Pertimbangan operasional yang paling lengkap adalah....
   a. memilih model dengan skor tertinggi tanpa memeriksa alarm yang dihasilkan
   b. meningkatkan jumlah notifikasi agar tidak ada penyimpangan terlewat
   c. menilai hanya kemampuan model menemukan nilai ekstrem
   d. menghentikan pemeriksaan petugas setelah model menghasilkan skor

e. mempertimbangkan alarm palsu, keterlambatan deteksi, dan kemampuan petugas menindaklanjuti notifikasi

c. Soal Project / Studi Kasus

1. Sebuah perusahaan pengolahan ikan memasang sensor suhu dan kelembapan di beberapa ruang penyimpanan. Dalam satu bulan, sistem menandai banyak kenaikan suhu pada siang hari. Sebagian kejadian berlangsung singkat saat pintu dibuka untuk pemindahan barang, sementara kejadian lain bertepatan dengan sensor yang menunjukkan nilai tetap selama beberapa jam. Tim operasi menggunakan satu ambang suhu untuk semua ruang dan belum mencatat jadwal pemindahan barang. Manajer meminta rekomendasi sebelum sistem digunakan untuk mengirim peringatan otomatis kepada teknisi. Analisislah bentuk anomali yang mungkin muncul, data tambahan dan prapemrosesan yang dibutuhkan, pendekatan deteksi yang layak diuji, serta prosedur verifikasi sebelum alarm memicu tindakan.
2. Jaringan perangkat IoT sebuah pabrik mengirimkan data suhu, getaran, dan arus listrik. Tim keamanan mengamati bahwa beberapa perangkat menghasilkan kombinasi pembacaan yang jarang, tetapi nilainya masing-masing masih berada dalam rentang yang pernah tercatat. Data belum distandardisasi, label insiden hanya tersedia untuk sebagian kecil kejadian, dan petugas menerima terlalu banyak alarm dari sistem lama. Pimpinan meminta rancangan uji coba deteksi anomali yang tidak langsung memblokir

perangkat tanpa pemeriksaan. Evaluasilah pilihan metode, langkah persiapan dan validasi data, cara menilai alarm palsu serta keterlambatan deteksi, dan alur eskalasi yang menjaga keputusan tetap dapat diverifikasi.

BAB 13
ADVANCED MACHINE LEARNING DAN INTELLIGENT DATA MINING

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan pendekatan ensemble learning melalui bagging dan boosting untuk menyelesaikan permasalahan klasifikasi atau prediksi pada data.
2. Menggunakan Random Forest, Gradient Boosting, dan XGBoost dengan mempertimbangkan karakteristik data serta tujuan analisis.
3. Mengimplementasikan Artificial Neural Network atau metode deep learning pada permasalahan yang sesuai dengan struktur dan skala data.
4. Menggunakan teknik Explainable AI untuk menafsirkan hasil model dan menjelaskan kontribusi fitur terhadap prediksi.
5. Memilih serta menerapkan pendekatan AutoML untuk mendukung proses pemodelan, dengan mempertimbangkan kinerja, interpretabilitas, dan kebutuhan organisasi.

B. Pendahuluan

Organisasi menghasilkan data dalam jumlah besar dan beragam, tetapi nilai analitis data tersebut bergantung pada

kemampuan model untuk menangkap pola yang kompleks sekaligus menghasilkan prediksi yang dapat digunakan. Model tunggal, terutama ketika dibangun dengan asumsi sederhana, dapat menghadapi keterbatasan dalam menangani interaksi antarf fitur, ketidaklinieran, maupun variasi data. Ensemble learning mengatasi sebagian persoalan tersebut dengan menggabungkan beberapa model; bagging, misalnya, dirancang untuk mengurangi varians melalui pembentukan model dari sampel bootstrap, sedangkan boosting membangun model secara bertahap untuk memperbaiki kesalahan prediksi sebelumnya (Breiman, 1996; Friedman, 2001). Perkembangan metode ini mendorong penggunaan algoritma seperti Random Forest, Gradient Boosting, dan XGBoost dalam beragam tugas analitik. Bagi mahasiswa Sistem Informasi, pemahaman terhadap metode tersebut diperlukan untuk menerjemahkan kebutuhan organisasi—seperti memperkirakan permintaan, mengidentifikasi risiko pelanggan, atau mendeteksi transaksi tidak wajar—menjadi proses pemodelan yang tepat.

Selain metode ensemble, Artificial Neural Network dan deep learning menyediakan pendekatan untuk mempelajari representasi data yang kompleks, termasuk data berukuran besar dan tidak terstruktur. LeCun, Bengio, dan Hinton (2015) menjelaskan bahwa kemajuan deep learning ditopang oleh kemampuannya mempelajari representasi

bertingkat dari data. Namun, kinerja prediksi yang tinggi tidak dengan sendirinya menjamin bahwa model mudah dipahami atau layak digunakan dalam proses organisasi. Keputusan berbasis model dapat memengaruhi layanan, alokasi sumber daya, dan perlakuan terhadap individu; karena itu, pemahaman mengenai alasan di balik prediksi menjadi bagian dari penerapan yang bertanggung jawab. Ribeiro, Singh, dan Guestrin (2016), melalui pendekatan LIME, menunjukkan cara menjelaskan prediksi model kompleks dengan membangun penjelasan lokal yang dapat ditafsirkan. Dalam konteks Sistem Informasi, mahasiswa perlu menerapkan teknik interpretasi untuk menilai kontribusi fitur, mengenali keterbatasan model, serta menyampaikan hasil analisis kepada pemangku kepentingan.

Proses pengembangan model juga dapat melibatkan banyak pilihan terkait algoritma, prapemrosesan, dan konfigurasi parameter. AutoML berupaya mengotomatisasi sebagian tahapan tersebut sehingga eksperimen pemodelan dapat dilakukan secara lebih sistematis, tetapi hasil otomatis tetap perlu dievaluasi berdasarkan kualitas data, tujuan analisis, dan konteks penggunaan (Hutter, Kotthoff, dan Vanschoren, 2019). Oleh sebab itu, penerapan intelligent data mining tidak cukup hanya dengan menjalankan algoritma; prosesnya mencakup pemilihan metode, pengujian kinerja, interpretasi hasil, serta

pertimbangan terhadap kebutuhan organisasi. Bab ini membahas Ensemble Learning, Bagging, dan Boosting; Random Forest, Gradient Boosting, dan XGBoost; Artificial Neural Network dan Deep Learning; Explainable AI dan Interpretasi Model; serta AutoML dan Perkembangan Intelligent Data Mining.

C. Ensemble Learning, Bagging, dan Boosting

Pembahasan algoritma pada bab-bab sebelumnya memperkenalkan model yang dilatih untuk mengenali pola dari kumpulan data. Pada kasus yang lebih rumit, satu model dapat menghasilkan prediksi yang kurang stabil atau gagal menangkap beragam pola sekaligus. *Ensemble learning* mengatasi keterbatasan tersebut dengan menggabungkan keluaran beberapa model. Kualitas gabungan bergantung pada kinerja model penyusunnya dan pada sejauh mana kesalahan antarmodel berbeda satu sama lain. Pendekatan ini digunakan dalam tugas klasifikasi maupun prediksi, tetapi penambahan jumlah model saja tidak menjamin peningkatan kinerja (Unknown, 2011; Unknown, 2019).

Dua strategi yang umum adalah *bagging* dan *boosting*. Pada *bagging* (*bootstrap aggregating*), beberapa kumpulan data pelatihan dibentuk melalui pengambilan sampel dengan pengembalian dari data asal. Model sejenis dilatih secara terpisah pada tiap sampel, kemudian prediksinya

digabungkan. Untuk klasifikasi, penggabungan dapat menggunakan suara terbanyak; untuk prediksi nilai kontinu, keluaran dapat dirata-ratakan. Variasi sampel membantu mengurangi ketergantungan hasil pada satu komposisi data pelatihan. *Boosting* membangun model secara berurutan: model berikutnya difokuskan pada kesalahan atau pola yang belum tertangani oleh model sebelumnya. Strategi ini dapat menghasilkan batas keputusan yang lebih baik, tetapi pelatihan bertahapnya membuat pengaturan dan evaluasi menjadi penting (Unknown, 2011; Unknown, 2019).

Perbedaan strategi tersebut terlihat pada cara model belajar dari data. Dalam *bagging*, model dapat dilatih secara paralel karena masing-masing menggunakan sampel yang dibentuk dari data asal. Dalam *boosting*, pelatihan model berikutnya bergantung pada keluaran model sebelumnya. Karena itu, *bagging* sering dipertimbangkan ketika variasi prediksi antarsampel menjadi perhatian, sedangkan *boosting* dapat dipilih ketika model perlu secara bertahap memperbaiki kesalahan. Keduanya tetap harus dibandingkan menggunakan pembagian data dan metrik evaluasi yang sesuai dengan tujuan analisis, sebagaimana dibahas pada Bab 5 dan Bab 6. Penerapan model ensemble pada deteksi anomali kode kontrak Ethereum menunjukkan bahwa gabungan model juga dapat digunakan untuk mengenali pola berisiko dalam data keamanan digital (Hisham, 2025).

Sebagai ilustrasi dalam konteks Indonesia, sebuah koperasi di Yogyakarta dapat menggunakan *bagging* untuk mengklasifikasikan pengajuan pinjaman sebagai berisiko atau tidak berisiko berdasarkan riwayat pembayaran dan karakteristik pengajuan. Tabel berikut menunjukkan lima keluaran model pada satu pengajuan; nilai “berisiko” memperoleh suara terbanyak. Contoh ini bersifat ilustratif dan tidak merepresentasikan data koperasi tertentu.

Tabel 13.1 Suara Model pada Klasifikasi Pengajuan Pinjaman

Model	Prediksi
Model 1	Berisiko
Model 2	Tidak berisiko
Model 3	Berisiko
Model 4	Berisiko
Model 5	Tidak berisiko

Hasil tabel memperlihatkan bagaimana penggabungan suara dapat menghasilkan keputusan yang lebih konsisten daripada mengandalkan satu model. Namun, suara terbanyak tidak menghapus kemungkinan bias pada data pelatihan atau konsekuensi kesalahan klasifikasi terhadap pemohon. Sebelum digunakan, pengelola perlu memeriksa kinerja model pada data pengujian dan memastikan bahwa keputusan dapat ditinjau oleh petugas.

Dalam penerapan, pemilihan antara *bagging* dan *boosting* perlu mempertimbangkan ukuran dan kualitas data, sumber daya komputasi, sasaran prediksi, serta kebutuhan organisasi untuk menjelaskan hasil. Jika kelas

yang menjadi perhatian jarang muncul, misalnya transaksi penyalahgunaan yang jumlahnya jauh lebih sedikit daripada transaksi normal, evaluasi perlu memperhatikan kemampuan model mengenali kelas tersebut, bukan hanya ketepatan keseluruhan. Teknik penyeimbangan kelas dan metrik klasifikasi telah dibahas pada bab-bab terdahulu; pada model ensemble, keduanya tetap menjadi bagian dari proses pengujian, bukan pengganti pemilihan algoritma yang sesuai (Unknown, 2019; Hisham, 2025).

D. Random Forest, Gradient Boosting, dan XGBoost

Random Forest mengembangkan gagasan *bagging* dengan melatih banyak pohon keputusan pada sampel bootstrap dan membatasi fitur yang dipertimbangkan saat pemisahan simpul. Pengacakan sampel dan fitur mendorong keragaman antarpohon, sedangkan gabungan prediksi menghasilkan klasifikasi melalui suara terbanyak atau prediksi numerik melalui perataan. Metode ini dapat menangani hubungan nonlinier dan interaksi fitur tanpa mengharuskan analis menentukan bentuk hubungan terlebih dahulu. Untuk menggunakannya secara tepat, jumlah pohon, kedalaman, dan jumlah fitur yang dipertimbangkan perlu disesuaikan melalui validasi; pohon yang makin kompleks tidak otomatis memberi hasil yang lebih baik pada data baru (Unknown, 2011; Unknown, 2019).

Gradient Boosting membangun pohon secara bertahap. Setiap pohon baru memperbaiki kesalahan prediksi kumpulan pohon yang telah dibangun, dengan memperbarui model menurut arah yang mengurangi fungsi kerugian. *XGBoost* merupakan implementasi *gradient boosting* yang menyediakan mekanisme pengaturan kompleksitas dan dukungan komputasi yang dirancang untuk pemodelan berskala besar. Dalam praktik, parameter seperti laju pembelajaran, kedalaman pohon, jumlah iterasi, serta proporsi sampel yang digunakan memengaruhi keseimbangan antara kecocokan terhadap data pelatihan dan kemampuan menggeneralisasi. Karena istilah dan pilihan parameter dapat berbeda antarperangkat lunak, pengguna perlu membaca dokumentasi algoritma yang digunakan dan menguji konfigurasi secara sistematis (Unknown, 2019; Unknown, 2011).

Ketiga algoritma dapat diterapkan pada permasalahan yang berbeda. *Random Forest* sering menjadi titik awal yang berguna ketika dibutuhkan model pohon gabungan yang relatif mudah diterapkan. *Gradient Boosting* dan *XGBoost* dapat dipertimbangkan ketika pemodelan bertahap memberi peningkatan pada data validasi, dengan konsekuensi bahwa pengaturan parameter dan waktu pelatihan perlu diperhatikan. Perbandingan harus menggunakan pembagian data yang sama dan metrik yang berkaitan dengan tujuan analisis. Religia (2021) membahas

optimasi algoritma pada *Random Forest* untuk klasifikasi data pemasaran bank; konteks tersebut memperlihatkan bahwa pilihan pengaturan model merupakan bagian dari penerapan, bukan keputusan yang dapat dilepaskan dari karakteristik data dan tujuan klasifikasi.

Misalnya, tim analitik perusahaan ritel di Bandung dapat membandingkan ketiga metode untuk memprediksi kemungkinan pelanggan menanggapi penawaran produk. Fitur yang digunakan dapat mencakup frekuensi transaksi dan lama sejak pembelian terakhir, sementara label berasal dari riwayat respons yang telah disiapkan melalui prapemrosesan data. Tim dapat melatih model dengan kumpulan pelatihan yang sama, memilih parameter berdasarkan data validasi, lalu melaporkan hasil pada data pengujian yang tidak digunakan selama pemilihan. Jika tujuan organisasi adalah mengurangi pelanggan potensial yang terlewat, recall kelas respons perlu menjadi salah satu ukuran evaluasi, disertai precision agar dampak promosi yang salah sasaran tetap terlihat (Religia, 2021; Unknown, 2011).

Pada data yang memiliki kelas tidak seimbang, akurasi tunggal dapat menutupi kegagalan model mengenali kelas minoritas. Contohnya, model deteksi transaksi berisiko di sebuah bank Indonesia perlu dievaluasi dengan mempertimbangkan konsekuensi transaksi yang tidak terdeteksi dan peringatan palsu yang diterima petugas.

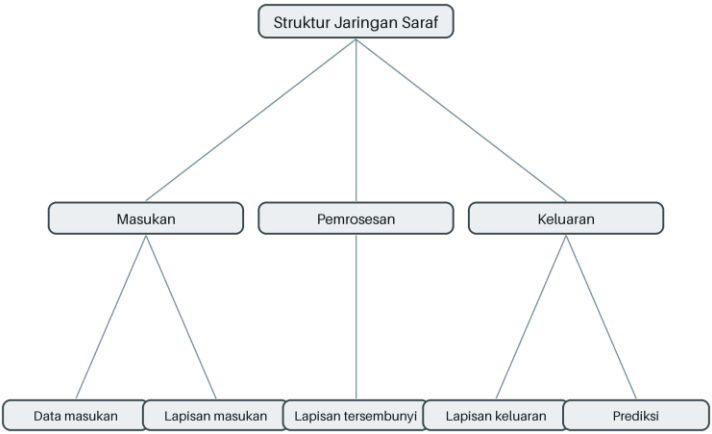
Pemilihan ambang keputusan dapat mengubah keseimbangan antara kedua jenis kesalahan tersebut. Model dengan nilai evaluasi terbaik belum tentu menjadi pilihan operasional apabila kebutuhan waktu respons, biaya pemeriksaan, atau kemampuan staf dalam menindaklanjuti peringatan tidak terpenuhi. Karena itu, evaluasi model pohon gabungan perlu mengaitkan metrik prediksi dengan alur kerja organisasi (Dua, 2016; Unknown, 2019).

E. Artificial Neural Network dan Deep Learning

Artificial Neural Network (ANN) merupakan model pembelajaran mesin yang tersusun atas unit komputasi yang terhubung dan diatur dalam lapisan. Lapisan masukan menerima nilai fitur, lapisan tersembunyi mengolah representasi antara, dan lapisan keluaran menghasilkan nilai prediksi atau peluang kelas. Selama pelatihan, parameter jaringan disesuaikan untuk mengurangi kesalahan prediksi. Bentuk jaringan, fungsi aktivasi, fungsi kerugian, dan metode pembaruan parameter perlu disesuaikan dengan jenis tugas, skala data, serta sumber daya komputasi. Pembahasan model jaringan saraf untuk prediksi kasus gizi buruk di Nusa Tenggara Barat menunjukkan penggunaan pendekatan neural network dalam konteks sistem informasi dan data layanan publik di Indonesia (Kurniati, 2021).

Deep learning merujuk pada penggunaan jaringan dengan beberapa lapisan tersembunyi untuk mempelajari representasi data secara bertingkat. Pendekatan ini dapat diterapkan pada data yang memiliki pola kompleks, tetapi pelatihan biasanya memerlukan perhatian terhadap jumlah contoh, kualitas label, biaya komputasi, dan risiko model menghafal data pelatihan. Untuk data tabular berukuran terbatas, model yang lebih sederhana dapat menjadi pembanding yang relevan. Oleh sebab itu, penggunaan jaringan dalam tidak seharusnya ditentukan hanya karena kompleksitasnya; pemilihan perlu didasarkan pada hasil evaluasi terhadap tujuan analisis dan karakteristik kumpulan data (Unknown, 2011; Nurlaela, 2025).

Gambar berikut menyajikan struktur konseptual jaringan saraf yang dapat digunakan untuk memahami perpindahan informasi dari masukan menuju keluaran. Representasi ini membantu menghubungkan susunan lapisan dengan peran fitur dan keluaran model dalam tugas klasifikasi atau prediksi.



Gambar 13.1 Struktur Jaringan Saraf untuk Prediksi. Sumber: Diolah penulis

Susunan lapisan pada gambar menekankan bahwa prediksi dihasilkan melalui transformasi bertahap atas masukan, bukan melalui satu aturan tunggal yang langsung terlihat. Jumlah lapisan dan unit tersembunyi harus ditentukan melalui pelatihan serta evaluasi, bukan diasumsikan semakin banyak selalu semakin baik.

Penerapan di Indonesia dapat berupa prediksi kebutuhan layanan kesehatan kabupaten berdasarkan data historis yang telah dianonimkan, dengan fitur agregat seperti periode, wilayah, dan indikator layanan. Tim perlu memisahkan data pelatihan, validasi, dan pengujian menurut waktu atau kelompok yang tepat agar evaluasi tidak mengambil informasi masa depan. Jika sumber data

berupa citra atau sinyal, rancangan jaringan dapat berbeda dari jaringan untuk tabel; jika jumlah data berlabel terbatas, pendekatan transfer learning dapat dipertimbangkan dan diuji sebagai pilihan, bukan dianggap selalu unggul. Kerangka penambangan data berbasis jaringan saraf dalam Nurlaela (2025) memberi konteks bagi penggunaan deep neural network dan eksplorasi transfer learning untuk prediksi serta klasifikasi, sedangkan Kurniati (2021) memberikan contoh penggunaan neural network pada kasus informasi kesehatan di Indonesia.

Keputusan menggunakan ANN atau deep learning juga perlu mempertimbangkan cara hasil akan digunakan. Dalam sistem pendukung keputusan, keluaran probabilitas dapat membantu petugas memprioritaskan pemeriksaan, tetapi tidak dengan sendirinya menjadi keputusan final. Pengguna perlu memantau perubahan distribusi data dan meninjau kesalahan pada kelompok yang relevan, terutama bila prediksi berpengaruh pada layanan publik. Langkah tersebut melengkapi evaluasi teknis dengan pemeriksaan kesesuaian model terhadap proses organisasi, sebagaimana menjadi bagian dari siklus pengembangan model yang dibahas pada Bab 3 (Nurlaela, 2025; Unknown, 2011).

F. Explainable AI dan Interpretasi Model

Explainable AI (XAI) mencakup pendekatan untuk membantu manusia memahami alasan model menghasilkan

prediksi tertentu. Kebutuhan penjelasan berbeda menurut pengguna dan keputusan yang didukung. Pengembang dapat memerlukan informasi untuk mendiagnosis kesalahan, sedangkan manajer membutuhkan ringkasan faktor yang memengaruhi keputusan operasional. Interpretasi model sebaiknya membedakan penjelasan global, yang merangkum perilaku model secara umum, dari penjelasan lokal, yang menjelaskan satu prediksi. Penjelasan merupakan dukungan bagi evaluasi model, bukan bukti bahwa prediksi pasti benar atau bahwa hubungan yang ditemukan bersifat kausal (Kononenko, 2007).

Beberapa teknik dapat diterapkan sesuai pertanyaan analisis. Kepentingan fitur global merangkum fitur yang banyak digunakan atau berpengaruh terhadap keluaran model, sementara metode lokal dapat menunjukkan fitur yang mendorong prediksi pada satu kasus. Kurva ketergantungan parsial dapat membantu melihat perubahan rata-rata keluaran ketika suatu fitur diubah, dengan keterbatasan apabila fitur saling berkorelasi. Untuk model pohon, kepentingan fitur yang dihitung dari struktur pohon perlu ditafsirkan hati-hati karena bisa memihak fitur dengan karakteristik tertentu. Rekayasa fitur juga memengaruhi makna penjelasan: bila beberapa kolom digabung menjadi indikator, arti kontribusi indikator

tersebut perlu diterangkan berdasarkan cara pembentukannya (Liu, 2018; Kononenko, 2007).

Dalam konteks Indonesia, sebuah bank yang menggunakan model untuk menandai pengajuan kredit berisiko dapat menyajikan alasan prediksi kepada petugas, misalnya rasio kewajiban terhadap pendapatan dan riwayat keterlambatan. Petugas dapat memeriksa apakah alasan tersebut selaras dengan data pengajuan dan aturan proses yang berlaku. Jika model menempatkan kode wilayah sebagai fitur dominan, hal itu perlu ditelusuri: kode tersebut mungkin bertindak sebagai proksi bagi variabel lain atau mencerminkan ketimpangan pada data. Penjelasan semacam ini membantu menentukan kasus yang perlu ditinjau, tetapi tidak menggantikan pemeriksaan kualitas data maupun validasi kinerja (Dua, 2016; Liu, 2018).

Penerapan XAI dapat dimulai dengan menetapkan keputusan yang perlu dijelaskan, pihak yang akan menggunakan penjelasan, dan tingkat rinciannya. Selanjutnya, analis memilih teknik yang sesuai dengan model, menghitung penjelasan pada data yang tepat, lalu memeriksa kestabilannya. Sebagai contoh, jika perubahan kecil pada satu fitur menyebabkan penjelasan lokal berubah drastis, pengguna perlu mempertimbangkan apakah keluaran tersebut cukup konsisten untuk mendukung proses keputusan. Penjelasan juga sebaiknya dibandingkan dengan pengetahuan proses bisnis dan ditinjau oleh ahli

domain. Kerangka interpretasi dalam pembelajaran mesin menekankan pentingnya mengaitkan keluaran model dengan representasi fitur serta konteks prediksinya (Kononenko, 2007; Liu, 2018).

Penjelasan yang mudah dibaca tetap memiliki batas. Grafik kepentingan fitur dapat menyederhanakan interaksi yang rumit, dan penjelasan lokal tidak selalu mencerminkan perilaku model pada kumpulan data lain. Pengguna juga dapat menyamakan kontribusi prediktif dengan sebab-akibat; padahal, model yang dilatih dari data observasional tidak dengan sendirinya membuktikan bahwa perubahan suatu fitur menyebabkan perubahan hasil. Karena itu, hasil XAI perlu disampaikan bersama keterbatasan data, lingkup penggunaan, dan ukuran evaluasi model. Pada sistem keamanan, penjelasan dapat membantu analis memprioritaskan pemeriksaan peringatan, namun keputusan tindak lanjut tetap membutuhkan validasi terhadap bukti dan konteks insiden (Dua, 2016; Kononenko, 2007).

G. AutoML dan Perkembangan Intelligent Data Mining

AutoML merujuk pada penggunaan perangkat dan prosedur otomatis untuk membantu sebagian tahapan pemodelan, seperti pemilihan algoritma, prapemrosesan, pencarian parameter, dan perbandingan hasil. Tujuannya

adalah membuat eksperimen lebih sistematis serta mengurangi pekerjaan konfigurasi yang berulang. Otomasi tidak menghilangkan kebutuhan untuk memahami masalah, memeriksa kualitas data, atau menentukan ukuran keberhasilan. Hasil tetap bergantung pada kumpulan data masukan, rancangan eksperimen, ruang pilihan algoritma, dan metrik yang ditetapkan pengguna (Unknown, 2011; Unknown, 2019).

Penerapan AutoML dapat dimulai dengan menerjemahkan kebutuhan organisasi menjadi sasaran pemodelan yang terukur. Sebuah koperasi di Surabaya, misalnya, dapat membandingkan beberapa model untuk memperkirakan kemungkinan keterlambatan pembayaran. Tim perlu menetapkan label yang tepat, memilih metrik yang mencerminkan dampak kesalahan, lalu memastikan bahwa data yang sama digunakan untuk membandingkan konfigurasi. Jika AutoML menguji prapemrosesan dan algoritma sekaligus, catatan eksperimen perlu mencakup transformasi yang dipilih dan parameter terbaik, agar hasil dapat ditinjau serta diulang. Prosedur pemodelan praktis dan penggunaan perangkat pembelajaran mesin menyediakan landasan untuk memahami eksperimen semacam ini (Unknown, 2011).

Salah satu risiko penggunaan AutoML adalah kebocoran informasi ketika transformasi atau pemilihan fitur dilakukan sebelum pembagian data. Risiko lain muncul saat terlalu

banyak konfigurasi diuji pada data validasi yang sama; pilihan akhir dapat terlalu menyesuaikan diri terhadap data tersebut. Untuk mengurangi masalah, prapemrosesan yang mempelajari parameter dari data harus dipasang dalam alur pelatihan yang sesuai, dan evaluasi akhir dilakukan pada data pengujian yang tidak dipakai untuk memilih model. Selain kinerja, tim perlu mempertimbangkan durasi pencarian, kebutuhan komputasi, kemudahan interpretasi, serta kemampuan menerapkan model di lingkungan organisasi (Unknown, 2019; Kononenko, 2007).

AutoML juga perlu dipahami sebagai bagian dari *intelligent data mining*, bukan pengganti pertimbangan manusia. Sistem dapat menghasilkan kandidat model dan ringkasan evaluasi, sementara analis memeriksa apakah data pelatihan mewakili keadaan operasional dan apakah keluaran model sesuai dengan tujuan. Dalam deteksi transaksi mencurigakan di perusahaan finansial Indonesia, misalnya, pencarian otomatis dapat membantu membandingkan model, tetapi tim keamanan tetap perlu menentukan biaya pemeriksaan peringatan dan prosedur eskalasi. Jika data berubah karena pola transaksi baru, model terpilih perlu dievaluasi kembali sebelum keputusan otomatis diteruskan kepada pengguna. Kerangka pembelajaran mesin dan penambangan data menempatkan pemilihan metode serta evaluasi dalam konteks masalah,

bukan sebagai prosedur mekanis yang berdiri sendiri (Unknown, 2011; Unknown, 2019).

Pemilihan AutoML sebaiknya mengikuti kapasitas dan kebutuhan organisasi. Untuk tim kecil dengan sumber daya terbatas, otomasi dapat mempercepat penyaringan kandidat; untuk penggunaan berdampak tinggi, hasilnya tetap perlu disertai dokumentasi, pemeriksaan penjelasan, dan persetujuan pihak yang bertanggung jawab. Organisasi juga perlu menghindari pemilihan model hanya karena satu skor tertinggi apabila selisih kinerjanya kecil atau biaya penerapannya besar. Dengan demikian, AutoML bermanfaat sebagai sarana eksperimen yang terstruktur, sedangkan keputusan akhir memerlukan penilaian atas kinerja, interpretabilitas, sumber daya, dan kecocokan dengan proses kerja (Unknown, 2011; Unknown, 2019; Kononenko, 2007).

H. Rangkuman

Ensemble learning menggabungkan prediksi beberapa model untuk menangani keterbatasan model tunggal, tetapi manfaatnya bergantung pada kinerja dan keragaman kesalahan model penyusun, bukan semata-mata jumlah model. Bagging membentuk sampel bootstrap melalui pengambilan data dengan pengembalian, melatih model secara terpisah, lalu menggabungkan hasilnya dengan suara terbanyak untuk klasifikasi atau perataan untuk prediksi

numerik. Karena model dapat dilatih paralel, pendekatan ini berguna ketika variasi hasil akibat komposisi data menjadi perhatian. Sebaliknya, boosting melatih model secara berurutan; setiap model berikutnya berupaya memperbaiki kesalahan atau pola yang belum tertangani. Strategi bertahap ini berpotensi memperbaiki prediksi, tetapi menuntut pengaturan dan evaluasi yang cermat.

Random Forest menerapkan prinsip bagging pada banyak pohon keputusan, sekaligus mengacak fitur yang dipertimbangkan saat pemisahan simpul. Keragaman sampel dan fitur membantu menghasilkan prediksi gabungan serta menangani hubungan nonlinier dan interaksi fitur. Gradient Boosting membangun pohon secara bertahap dengan mengurangi fungsi kerugian, sedangkan XGBoost merupakan implementasi yang menyediakan pengaturan kompleksitas dan dukungan komputasi untuk pemodelan berskala besar. Laju pembelajaran, kedalaman pohon, jumlah iterasi, dan proporsi sampel memengaruhi keseimbangan antara kecocokan pada data pelatihan dan kemampuan generalisasi.

Dalam praktik Sistem Informasi, pemilihan algoritma harus disesuaikan dengan sasaran analisis, mutu dan ukuran data, sumber daya komputasi, serta kebutuhan organisasi untuk memahami dan menindaklanjuti hasil. Model perlu dibandingkan menggunakan pembagian data yang sama; data pengujian tidak boleh digunakan saat memilih

parameter. Metrik pun perlu mencerminkan konsekuensi kesalahan: pada kelas minoritas, akurasi dapat menyamarkan kegagalan deteksi, sehingga recall dan precision perlu dipertimbangkan bersama. Ambang keputusan mengubah keseimbangan antara kasus yang terlewat dan peringatan palsu. Karena itu, keputusan operasional juga perlu mempertimbangkan biaya pemeriksaan, waktu respons, kapasitas staf, potensi bias, dan mekanisme peninjauan manusia. Ensemble dapat membantu mengidentifikasi risiko atau mendukung keputusan organisasi, tetapi keluaran model tetap harus diuji dan diterapkan secara bertanggung jawab.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana keragaman kesalahan antarmodel memengaruhi manfaat ensemble learning. Dalam jawaban, bandingkan kondisi ketika penambahan model berpotensi memperbaiki prediksi dengan kondisi ketika jumlah model bertambah tetapi kualitas gabungan tidak meningkat.
2. Bandingkan bagging dan boosting dari sisi pembentukan data atau urutan pelatihan, cara menggabungkan hasil, peluang pemrosesan paralel, serta jenis masalah yang dapat mendorong pemilihannya. Rumuskan pula satu kondisi ketika salah satu pendekatan lebih sesuai untuk organisasi.

3. Evaluasi alasan Random Forest dapat menjadi titik awal yang berguna untuk klasifikasi data organisasi. Telaah peran sampel bootstrap, pengacakan fitur, dan pengaturan kompleksitas pohon, lalu jelaskan bagaimana validasi dapat mencegah pemilihan model yang hanya cocok pada data pelatihan.
4. Telaah bagaimana tim analitik seharusnya membandingkan Random Forest, Gradient Boosting, dan XGBoost untuk tujuan prediksi organisasi. Rumuskan rancangan evaluasi yang menjaga pemisahan data pelatihan, validasi, dan pengujian, serta mengaitkan pemilihan parameter dan metrik dengan sasaran operasional.
5. Argumentasikan mengapa akurasi saja dapat menyesatkan dalam deteksi transaksi berisiko yang jarang terjadi. Analisis peran recall, precision, ambang keputusan, biaya peringatan palsu, kasus yang terlewat, dan kapasitas petugas dalam menentukan kelayakan model untuk digunakan.

b. Pilihan Berganda

1. Manakah alasan utama bahwa penambahan jumlah model dalam ensemble tidak otomatis meningkatkan kualitas prediksi?
    a. Model ensemble hanya dapat digunakan untuk klasifikasi.
    b. Penggabungan selalu menghilangkan kesalahan setiap model.
    c. Manfaat gabungan bergantung pada kinerja dan keragaman kesalahan model penyusunnya.

d. Model yang dilatih pada data berbeda tidak dapat digabungkan.
e. Jumlah model menentukan metrik evaluasi yang digunakan.

2. Bagaimana sampel pelatihan untuk model-model dalam bagging umumnya dibentuk?
   a. Data dibagi menjadi kelompok berurutan tanpa pengembalian.
   b. Sampel dibentuk melalui pengambilan data dengan pengembalian dari data asal.
   c. Setiap model hanya dilatih menggunakan data yang salah diprediksi model sebelumnya.
   d. Data pengujian ditambahkan ke data pelatihan secara bergantian.
   e. Fitur dihapus sampai tersisa satu prediktor untuk setiap model.

3. Ciri yang membedakan proses pelatihan boosting dari bagging adalah bahwa boosting:
   a. melatih semua model secara paralel pada sampel yang sama.
   b. menggabungkan keluaran hanya dengan perataan.
   c. selalu menggunakan pohon dengan kedalaman yang sama.
   d. melatih model secara berurutan untuk memperbaiki kesalahan yang belum tertangani.
   e. membatasi setiap model pada satu fitur.

4. Bagaimana Random Forest mendorong keragaman antarpohon?
    a. Menggunakan sampel bootstrap dan membatasi fitur yang dipertimbangkan pada pemisahan simpul.
    b. Melatih satu pohon yang sangat dalam pada seluruh data.
    c. Menggunakan data pengujian untuk menentukan pemisahan setiap pohon.
    d. Mengurutkan pohon berdasarkan waktu pelatihan.
    e. Menghindari penggabungan prediksi dari pohon-pohon.
5. Mekanisme utama Gradient Boosting dalam membangun model adalah:
    a. melatih pohon secara independen, lalu memilih pohon dengan akurasi tertinggi.
    b. mengacak label data agar kesalahan model saling meniadakan.
    c. mengurangi jumlah fitur setelah setiap pohon dilatih.
    d. menggabungkan prediksi dengan suara terbanyak tanpa memperbarui model.
    e. menambahkan pohon secara bertahap untuk mengurangi fungsi kerugian.
6. Dalam perbandingan algoritma yang adil, penggunaan data pengujian yang tepat adalah:

a. memakai data pengujian untuk memilih parameter terbaik sebelum pelatihan.
b. menggunakan data pengujian berulang kali untuk memperbaiki model.
c. menyimpan data pengujian dari proses pemilihan parameter dan melaporkannya setelah model ditetapkan.
d. mengganti data validasi dengan data pengujian setelah setiap iterasi.
e. mencampurkan data pengujian ke dalam data pelatihan agar ukuran sampel meningkat.

7. Pada masalah klasifikasi dengan kelas berisiko yang jarang, mengapa akurasi tunggal dapat menyesatkan?
   a. Akurasi hanya dapat dihitung untuk model berbasis pohon.
   b. Prediksi mayoritas dapat menghasilkan akurasi tinggi meskipun kelas minoritas gagal dikenali.
   c. Akurasi selalu bernilai lebih rendah daripada recall.
   d. Akurasi tidak dapat digunakan jika data memiliki lebih dari dua fitur.
   e. Akurasi secara langsung mengukur biaya pemeriksaan petugas.
8. Apa dampak utama perubahan ambang keputusan dalam deteksi transaksi berisiko?
   a. Mengubah jumlah fitur yang digunakan Random Forest.

b. Menentukan apakah data perlu diambil dengan pengembalian.
c. Menghilangkan kebutuhan untuk mengevaluasi kelas minoritas.
d. Mengubah keseimbangan antara transaksi berisiko yang terlewat dan peringatan palsu.
e. Menjamin bahwa model tidak mengalami bias.

9. Pernyataan yang tepat tentang XGBoost adalah:
   a. XGBoost merupakan implementasi Gradient Boosting yang menyediakan pengaturan kompleksitas dan dukungan komputasi untuk pemodelan berskala besar.
   b. XGBoost adalah metode bagging yang melatih semua pohon secara independen.
   c. XGBoost hanya dapat digunakan untuk memprediksi nilai kontinu.
   d. XGBoost menghindari penggunaan parameter model.
   e. XGBoost selalu menghasilkan kinerja terbaik tanpa validasi.

10. Kapan bagging dapat menjadi pilihan yang masuk akal dibandingkan boosting?
    a. Ketika model berikutnya harus berfokus pada kesalahan model sebelumnya.
    b. Ketika pelatihan harus bergantung pada keluaran model sebelumnya.
    c. Ketika organisasi hanya memiliki satu observasi pelatihan.

d. Ketika semua prediksi harus berasal dari satu model tunggal.
e. Ketika variasi prediksi akibat komposisi sampel menjadi perhatian dan model dapat dilatih secara terpisah.

c. Soal Project / Studi Kasus

1. Sebuah koperasi simpan pinjam di Yogyakarta ingin menggunakan model untuk mengelompokkan pengajuan pinjaman berdasarkan riwayat pembayaran dan karakteristik pemohon. Tim analitik mengusulkan bagging dengan suara terbanyak, sedangkan pengelola mempertimbangkan boosting karena ingin mengurangi kesalahan klasifikasi. Data historis memiliki proporsi pengajuan berisiko yang lebih kecil daripada pengajuan lainnya. Petugas juga harus dapat meninjau rekomendasi sebelum keputusan akhir dibuat. Analisis kebutuhan koperasi, bandingkan kedua pendekatan, lalu rumuskan rancangan validasi dan metrik yang sesuai. Jelaskan pula cara menangani kelas minoritas dan memasukkan peninjauan manusia tanpa menganggap prediksi model sebagai keputusan final.
2. Perusahaan ritel di Bandung hendak memprediksi pelanggan yang akan merespons penawaran produk. Tim memiliki frekuensi transaksi, waktu sejak pembelian terakhir, dan riwayat respons. Mereka membandingkan Random Forest, Gradient Boosting, dan XGBoost, tetapi kampanye yang terlalu luas menambah biaya, sedangkan pelanggan potensial yang terlewat mengurangi peluang penjualan. Seorang

analis mengusulkan memilih model dengan akurasi tertinggi pada seluruh data yang tersedia. Telaah kelemahan usulan tersebut dan rumuskan prosedur perbandingan yang memisahkan data pelatihan, validasi, dan pengujian. Tentukan metrik serta pertimbangan operasional untuk memilih model dan ambang keputusan.

BAB 14
IMPLEMENTASI DAN PENGEMBANGAN PROYEK DATA MINING TERINTEGRASI

A. Tujuan Pembelajaran

Setelah menyelesaikan pembahasan pada bab ini, mahasiswa diharapkan mampu:

1. Menerapkan tahapan perancangan proyek data mining end-to-end, mulai dari perumusan masalah dan penyiapan data hingga evaluasi hasil.
2. Mengimplementasikan teknik prapemrosesan data, pelatihan model, dan evaluasi kinerja menggunakan Python sesuai karakteristik permasalahan.
3. Mengintegrasikan model data mining dengan sistem IoT untuk mengolah data yang dihasilkan perangkat dan mendukung kebutuhan analitik.
4. Menggunakan teknik analitik data mining untuk mengidentifikasi pola dan potensi ancaman pada sistem informasi.
5. Menentukan dan menerapkan langkah perlindungan etika, privasi, pengurangan bias, serta keamanan data dalam deployment dan monitoring model.

B. Pendahuluan

Proyek data mining dalam organisasi kerap menghadapi kesenjangan antara keberhasilan membangun model dan kemampuan menggunakan model tersebut secara andal

dalam proses operasional. Model dengan akurasi tinggi belum tentu menghasilkan manfaat apabila tujuan analisis tidak terhubung dengan kebutuhan bisnis, data tidak mewakili kondisi nyata, atau hasil prediksi tidak dapat diintegrasikan dengan sistem yang digunakan organisasi. Kerangka CRISP-DM menempatkan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment sebagai tahapan yang saling berkaitan, sehingga pengembangan solusi perlu dirancang sebagai proses menyeluruh, bukan sekadar pemilihan algoritma (Chapman et al., 2000). Bagi mahasiswa Sistem Informasi, pemahaman tersebut diperlukan karena proyek analitik berada pada persilangan antara teknologi, data, proses organisasi, dan pengambilan keputusan. Oleh sebab itu, penerapan data mining perlu mempertimbangkan kecocokan solusi teknis dengan konteks penggunaannya.

Penerjemahan rancangan ke dalam implementasi menuntut keterampilan mengolah data dan membangun model secara terstruktur. Python menyediakan ekosistem pustaka yang mendukung tahapan tersebut, tetapi penggunaan pustaka saja tidak menjamin hasil yang dapat direproduksi atau dipelihara. Dalam lingkungan produksi, perubahan sumber data, ketergantungan antarkomponen, dan perbedaan antara data pelatihan dengan data operasional dapat menimbulkan persoalan teknis yang berkelanjutan; Sculley et al. (2015) menyebut akumulasi

persoalan semacam ini sebagai utang teknis dalam sistem pembelajaran mesin. Tantangan menjadi lebih beragam ketika model terhubung dengan sistem IoT, yang menghasilkan data berkelanjutan dari perangkat dan lingkungan fisik (Gubbi et al., 2013). Integrasi tersebut mengharuskan perancang memperhatikan aliran data, keterbatasan sumber daya, latensi, serta keandalan komunikasi, agar keluaran model dapat dimanfaatkan tepat waktu dalam proses organisasi.

Penerapan analitik pada keamanan sistem informasi juga memerlukan pertimbangan yang melampaui kinerja prediksi. Pola akses atau aktivitas jaringan dapat membantu mengidentifikasi perilaku yang menyimpang, tetapi hasil analisis tetap perlu ditafsirkan dalam konteks operasional dan ditindaklanjuti melalui prosedur keamanan yang sesuai. Selain itu, pengumpulan serta pemanfaatan data dapat menimbulkan risiko terhadap privasi, memperkuat bias yang telah ada, atau membuka celah keamanan baru. Kerangka manajemen risiko kecerdasan buatan menekankan perlunya mengelola risiko sepanjang siklus hidup sistem, termasuk ketika sistem diterapkan dan dipantau (NIST, 2023). Dengan demikian, implementasi yang bertanggung jawab menggabungkan validasi teknis dengan perlindungan data, pemantauan perubahan kinerja, dan evaluasi dampak terhadap pihak yang terdampak. Bab ini membahas perancangan end-to-end proyek data mining,

implementasi data mining dan machine learning dengan Python, integrasi model dengan sistem IoT, analitik untuk keamanan sistem informasi, serta etika, privasi, bias, keamanan data, deployment, dan monitoring model.

C. Perancangan End-to-End Data Mining Project

Bab-bab sebelumnya telah membahas tahapan KDD dan CRISP-DM, prapemrosesan data, pemilihan algoritma, serta evaluasi model. Pada tahap perancangan proyek, pengetahuan tersebut diterapkan sebagai satu rangkaian kerja yang menghubungkan kebutuhan organisasi dengan keluaran analitik dan tindakan operasional. Proyek tidak dimulai dari pemilihan algoritma, melainkan dari rumusan masalah yang dapat diterjemahkan menjadi keputusan yang hendak didukung. Sebagai contoh, sebuah koperasi di Yogyakarta dapat merancang proyek untuk memperkirakan keterlambatan pembayaran agar petugas dapat memprioritaskan tindak lanjut, bukan sekadar menghasilkan skor prediksi. Pemilihan metode perlu mengikuti karakteristik masalah dan data, sebagaimana ditekankan dalam pembahasan praktis pembelajaran mesin dan penambangan data (Kononenko, 2007; Unknown, 2011).

Perancangan end-to-end menetapkan tujuan, batas cakupan, peran pemangku kepentingan, sumber data,

ukuran keberhasilan, serta cara hasil model digunakan. Ukuran keberhasilan teknis, seperti F1-score, tidak selalu sama dengan ukuran manfaat operasional, seperti berkurangnya waktu pemeriksaan atau meningkatnya ketepatan prioritas. Karena itu, kriteria penerimaan perlu disepakati sebelum pelatihan model. Rencana juga perlu mencakup kebutuhan sumber daya, jadwal, risiko kualitas data, mekanisme persetujuan akses, dan penanggung jawab setelah penerapan. Pengalaman pengembangan sistem monitoring proyek berbasis web menunjukkan nilai pemantauan dan rekapitulasi bagi keteraturan alur kerja; prinsip tersebut dapat diterapkan untuk mencatat status, hambatan, dan keputusan pada proyek analitik (Wahyuningtyas, 2026).

Tabel berikut menggambarkan contoh rencana proyek untuk memperkirakan keterlambatan pembayaran pada koperasi. Setiap tahap menghasilkan keluaran yang dapat diperiksa oleh tim sebelum pekerjaan berlanjut.

Tabel 14.1 Rencana Kerja Proyek Prediksi Keterlambatan Pembayaran

Tahap	Aktivitas	Keluaran	Penanggung jawab	Kriteria pemeriksaan
Perumusan	Tetapkan keputusan dan populasi sasaran	Pernyataan masalah	Pemilik proses	Target dapat ditindaklanjuti

Pemahaman data	Inventarisasi riwayat pembayaran	Profil data	Analis data	Cakupan dan mutu terdokumentasi
Persiapan	Bersihkan data dan susun fitur	Kumpulan data analitik	Analis data	Tidak ada kebocoran target
Pemodelan	Latih kandidat model	Model terpilih	Tim pemodelan	Kinerja memenuhi ambang
Evaluasi	Uji kinerja dan dampak proses	Laporan evaluasi	Pemilik proses dan analis	Hasil sesuai konteks penggunaan
Penerapan	Hubungkan prediksi ke alur kerja	Layanan prediksi	Tim sistem informasi	Hasil dapat ditelusuri

Tabel tersebut menegaskan bahwa pemodelan hanya satu bagian dari proyek; keluaran tiap tahap harus menjadi masukan yang dapat ditinjau pada tahap berikutnya. Kriteria pemeriksaan membantu tim mendeteksi kesenjangan antara kebutuhan operasional dan hasil teknis sebelum model digunakan.

Dalam penerapan di Indonesia, koperasi perlu memeriksa apakah data historis mencerminkan perubahan kebijakan, variasi pola pembayaran, dan perbedaan karakteristik anggota. Jika anggota tertentu memiliki riwayat transaksi lebih pendek, model dapat menghasilkan prediksi yang kurang stabil untuk kelompok tersebut. Tim dapat menguji hasil berdasarkan periode dan kelompok

yang relevan, lalu mendiskusikan konsekuensinya dengan petugas yang akan memakai rekomendasi. Pendekatan bertahap ini sejalan dengan penggunaan hasil penambangan data sebagai dukungan keputusan, bukan sebagai pengganti penilaian organisasi (Unknown, 2011).

Sebelum proyek ditutup, dokumentasikan definisi label, asal data, keputusan prapemrosesan, versi kode, konfigurasi model, hasil evaluasi, serta batas penggunaan. Dokumentasi memungkinkan tim lain memahami alasan suatu prediksi dihasilkan dan mengulang eksperimen ketika data atau kebutuhan berubah. Proyek yang dirancang menyeluruh juga menetapkan sejak awal siapa yang meninjau keluaran, bagaimana kesalahan dilaporkan, dan kapan model perlu diperbarui. Dengan demikian, keberhasilan tidak hanya ditentukan oleh metrik pada kumpulan data uji, tetapi juga oleh keterpaduan hasil analitik dengan proses organisasi yang hendak diperbaiki (Kononenko, 2007; Wahyuningtyas, 2026).

D. Implementasi Data Mining dan Machine Learning dengan Python

Perancangan proyek menjadi nyata ketika kumpulan data dipersiapkan, model dilatih, dan hasil evaluasinya dapat direproduksi. Python menyediakan lingkungan yang mendukung tahap-tahap tersebut melalui pustaka pengolahan data dan pembelajaran mesin. Akan tetapi,

implementasi yang baik bukan sekadar menjalankan algoritma: pemisahan data, pemilihan fitur, transformasi, dan evaluasi harus mengikuti rancangan masalah. Pada kasus prediksi keterlambatan pembayaran koperasi di Yogyakarta, misalnya, kolom yang baru diketahui setelah jatuh tempo tidak boleh digunakan sebagai fitur karena akan membocorkan informasi target. Pemilihan fitur dan transformasinya perlu disesuaikan dengan makna data, bukan hanya kemudahan komputasi (Liu, 2018; Unknown, 2011).

Contoh berikut memperlihatkan alur klasifikasi yang dapat dijalankan setelah data tabular tersedia. Pipeline menggabungkan imputasi nilai numerik, standardisasi, dan regresi logistik, sehingga transformasi dipelajari hanya dari data pelatihan. Pembagian data menggunakan stratifikasi agar proporsi kelas lebih terjaga pada kedua bagian. Kolom `terlambat` menjadi target, sedangkan fitur harus dipilih berdasarkan informasi yang tersedia pada saat prediksi.

python

```
import pandas as pd
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.linear_model import
LogisticRegression
from sklearn.metrics import classification_report
from sklearn.model_selection import
train_test_split
```

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

data = pd.read_csv("pembayaran_koperasi.csv")

fitur = ["jumlah_pinjaman", "pendapatan_bulanan",
"jumlah_tanggungan"]
X = data[fitur]
y = data["terlambat"]

X_train, X_test, y_train, y_test =
train_test_split(
    X, y, test_size=0.2, random_state=42,
stratify=y
)

pra = Pipeline([
    ("imputasi",
SimpleImputer(strategy="median")),
    ("skala", StandardScaler())
])

model = Pipeline([
    ("pra", pra),
    ("klasifikasi",
LogisticRegression(max_iter=1000))
])

model.fit(X_train, y_train)
prediksi = model.predict(X_test)
```

```
print(classification_report(y_test, prediksi,
zero_division=0))
```

Pada tahap penerapan, evaluasi perlu mempertimbangkan biaya kesalahan. Apabila kasus yang terlambat tetapi tidak terdeteksi berisiko tinggi, recall kelas tersebut perlu ditinjau bersama precision, F1-score, dan confusion matrix. Ambang keputusan tidak selalu harus mengikuti nilai bawaan model; perubahan ambang sebaiknya didasarkan pada konsekuensi operasional dan kapasitas petugas menindaklanjuti peringatan. Pengetahuan tentang pemilihan metrik dan pembagian data yang dibahas pada bab-bab sebelumnya digunakan di sini untuk menafsirkan keluaran, bukan sekadar mencetak skor (Unknown, 2011; Kononenko, 2007).

Hasil program perlu disertai catatan tentang versi Python dan pustaka, sumber serta tanggal ekstraksi data, definisi label, fitur yang digunakan, dan nilai `random_state`. *Penyimpanan pipeline secara utuh mengurangi risiko perbedaan antara prapemrosesan saat pelatihan dan saat prediksi. Dalam praktik, data koperasi dapat mengalami perubahan format atau satuan, sehingga pemeriksaan skema masukan perlu dilakukan sebelum prediksi diterima. Implementasi Python juga dapat mendukung metode yang lebih kompleks; penerapan jaringan saraf untuk prediksi kasus gizi buruk di Nusa Tenggara Barat, misalnya,*

memperlihatkan pemanfaatan pemodelan untuk suatu persoalan kontekstual, tetapi keluaran tetap perlu dinilai sesuai mutu dan tujuan data yang digunakan (Kurniati, 2021).

Bahasa Python juga dimanfaatkan dalam analisis lain, termasuk analisis jalur, sehingga keterampilan menyusun data dan menjalankan prosedur analitik dapat mendukung beragam kebutuhan sistem informasi (Kaestria, 2023). Untuk proyek yang berkembang, kode sebaiknya dipisahkan menjadi fungsi atau modul, disertai pengujian sederhana terhadap masukan dan keluaran, serta disimpan dalam pengelolaan versi. Dengan praktik tersebut, perubahan fitur atau algoritma dapat ditinjau dan direproduksi. Model yang berhasil dilatih belum otomatis siap dipakai; perlu dipastikan pula bahwa kinerja pada data uji sesuai konteks, proses prediksi dapat diulang, dan pengguna memahami batas interpretasi hasilnya (Liu, 2018; Unknown, 2011).

E. Integrasi Model Data Mining dengan Sistem IoT

Integrasi model dengan Internet untuk Segala (IoT) menghubungkan pengukuran perangkat dengan analisis dan tindakan. Aliran umumnya mencakup sensor, perangkat pengendali, jaringan komunikasi, layanan penyimpanan atau pemrosesan, model, dan antarmuka pengguna. Setiap bagian menghasilkan kebutuhan teknis yang berbeda:

perangkat mengukur kondisi, jaringan mengirimkan catatan, sedangkan model mengubah fitur masukan menjadi prediksi atau peringatan. Karena itu, rancangan perlu menentukan format pesan, identitas perangkat, satuan pengukuran, cap waktu, frekuensi pengiriman, dan tindakan ketika koneksi terputus. Implementasi pemantauan kualitas udara berbasis sensor dan IoT menunjukkan penggunaan sensor DHT11, MQ-135, ESP8266, serta platform penerima data untuk pemantauan waktu nyata (Alfirdaus, 2024).

Model dapat ditempatkan di perangkat, gerbang lokal, atau layanan komputasi jarak jauh. Pemrosesan di perangkat atau gerbang dapat mengurangi ketergantungan pada koneksi dan mempercepat respons, tetapi sumber daya komputasinya terbatas. Pemrosesan pada layanan jarak jauh memudahkan pembaruan model dan pengelolaan data, tetapi menambah kebutuhan komunikasi serta perlindungan data. Pilihan lokasi harus mengikuti toleransi latensi, kapasitas perangkat, ketersediaan jaringan, dan konsekuensi keputusan. Penerapan prediksi kadar CO_2 dengan sensor DHT22 dan MQ135 yang terhubung ke ESP32 memperlihatkan integrasi pengukuran sensor dan model pembelajaran mesin dalam konteks IoT (Agung Suburdjati, 2026).

Misalnya, pengelola ruang kelas di Bandung dapat menggunakan pembacaan suhu, kelembapan, dan kualitas

udara untuk mengidentifikasi kondisi yang memerlukan pemeriksaan. Sebelum model digunakan, satuan, rentang pengukuran, interval pencatatan, dan waktu pada tiap sensor harus diseragamkan. Pembacaan yang hilang atau terlambat perlu ditandai, bukan diam-diam diperlakukan sebagai nilai normal. Jika model dilatih pada rentang sensor tertentu, pergantian jenis sensor dapat mengubah distribusi masukan sehingga keluaran lama tidak lagi sebanding. Pencatatan identitas perangkat dan versi model memungkinkan petugas menelusuri sumber perbedaan hasil (Alfirdaus, 2024; Agung Suburdjati, 2026).

Integrasi bukan hanya pengiriman data ke model; keluaran perlu disampaikan dalam bentuk yang dapat dipakai. Pada sistem pemantauan kateter berbasis IoT, misalnya, informasi kondisi perangkat ditujukan untuk membantu tenaga kesehatan memantau kapasitas dan mencegah luapan. Kasus tersebut menunjukkan bahwa rancangan keluaran perlu mempertimbangkan pengguna dan tindakan yang mungkin dilakukan, bukan hanya nilai sensor (Banjarnahor, 2022). Demikian pula, sistem deteksi tingkat dehidrasi berbasis IoT dan Android menghubungkan masukan perangkat dengan informasi yang disajikan melalui aplikasi; konteks penggunaan menuntut penyampaian hasil secara jelas dan tepat (Jumawal, 2026).

Untuk menghindari keputusan otomatis yang tidak sesuai, perancang dapat menetapkan ambang peringatan,

aturan eskalasi, dan prosedur pemeriksaan manusia. Peringatan berulang perlu dikendalikan agar pengguna tidak mengalami kelelahan alarm, sedangkan kegagalan sensor harus dibedakan dari anomali lingkungan. Catatan mentah, fitur yang dikirim ke model, keluaran, dan tindakan pengguna sebaiknya memiliki cap waktu yang selaras. Informasi tersebut berguna untuk menilai apakah kesalahan berasal dari sensor, komunikasi, pemrosesan, atau model. Dengan demikian, pengujian integrasi mencakup keseluruhan rantai, dari pembacaan sensor sampai keputusan operasional, bukan hanya ketepatan prediksi pada kumpulan data uji (Alfirdaus, 2024; Banjarnahor, 2022).

F. Implementasi Analitik untuk Keamanan Sistem Informasi

Analitik keamanan menerapkan penambangan data pada catatan aktivitas sistem, log autentikasi, lalu lintas jaringan, dan sumber telemetri lain untuk membantu menemukan pola yang memerlukan pemeriksaan. Dalam proyek operasional, tujuan perlu dirumuskan sebagai tindakan yang dapat dilakukan, misalnya menandai percobaan masuk yang menyimpang agar analis memeriksa akun terkait. Data keamanan sering memuat ketidakseimbangan kelas: aktivitas berbahaya dapat jauh lebih jarang daripada aktivitas normal. Oleh sebab itu, akurasi keseluruhan saja tidak cukup untuk menilai

kegunaan model; tim perlu memeriksa jenis kesalahan dan beban investigasi yang dihasilkan. Pembahasan penambangan data dalam keamanan siber menempatkan metode analitik sebagai salah satu perangkat untuk memahami serta menangani pola ancaman (Dua, 2016).

Implementasi dimulai dengan menggabungkan sumber data yang relevan secara aman dan menyelaraskan cap waktu, identitas aset, serta format kejadian. Fitur dapat mencakup jumlah percobaan autentikasi dalam jendela waktu, variasi alamat sumber, atau perubahan pola akses. Label insiden perlu diperiksa bersama tim keamanan karena laporan insiden yang terlambat atau tidak lengkap dapat menyebabkan contoh pelatihan yang keliru. Jika label terbatas, model dapat digunakan untuk mengurutkan kejadian berdasarkan prioritas pemeriksaan, dengan catatan bahwa skor bukan bukti tunggal terjadinya serangan. Penelitian tentang penerapan sistem deteksi intrusi berbantuan kecerdasan buatan pada big data menekankan kebutuhan untuk menangani volume dan pergerakan data keamanan yang besar (Amanda, 2025).

Sebagai contoh di Indonesia, sebuah perguruan tinggi di Surabaya dapat menerapkan analitik pada log autentikasi layanan akademik untuk membantu mengidentifikasi pola akses yang tidak biasa. Model dapat memberi skor lebih tinggi pada rangkaian percobaan login yang cepat dari sumber yang tidak lazim, tetapi petugas tetap perlu

memeriksa apakah aktivitas tersebut berasal dari pengguna sah yang lupa kata sandi, pengujian sistem terjadwal, atau penggunaan akun yang telah disalahgunakan. Penjelasan fitur yang berkontribusi terhadap skor membantu analis menentukan langkah pemeriksaan berikutnya. Rancangan seperti ini mengikuti orientasi keamanan siber yang menggabungkan pemrosesan data dengan pengetahuan tentang konteks ancaman (Dua, 2016).

Analitik dapat disusun sebagai lapisan pendukung sistem deteksi intrusi yang sudah ada. Aturan sederhana berguna untuk kejadian yang diketahui dan harus segera ditangani, sedangkan model dapat membantu mengurutkan pola yang lebih beragam. Integrasi keduanya perlu menghindari penghapusan sinyal mentah: catatan sumber, versi aturan atau model, skor, dan keputusan analis harus tetap dapat ditelusuri. Penelitian tentang deteksi intrusi berbantuan kecerdasan buatan dalam big data menyoroti penerapan analitik untuk membantu menghadapi ancaman yang sulit ditangani dengan cara konvensional, tetapi penerapan tetap membutuhkan evaluasi sesuai karakteristik lingkungan (Amanda, 2025).

Prosedur respons perlu menjelaskan siapa yang menerima peringatan, bagaimana bukti tambahan dikumpulkan, dan kondisi yang memerlukan eskalasi. Model sebaiknya tidak langsung memblokir akun atau perangkat tanpa mempertimbangkan dampak layanan dan jalur

pemulihan, kecuali kebijakan organisasi memang mengizinkannya. Sistem pertukaran data proyek juga memerlukan perlindungan agar informasi yang digunakan untuk analisis tidak menimbulkan kebocoran baru; kajian keamanan data pada pertukaran informasi proyek menekankan kebutuhan melindungi kerahasiaan dan integritas informasi (Kennedy, 2025). Dengan menggabungkan pemeringkatan kejadian, telaah analis, dan pencatatan tindakan, organisasi dapat menjadikan analitik sebagai bagian dari proses keamanan yang dapat dievaluasi, bukan sekadar mekanisme yang menghasilkan alarm.

G. Etika, Privasi, Bias, Keamanan Data, Deployment, dan Monitoring Model

Penerapan model mengubah data menjadi keluaran yang dapat memengaruhi individu, layanan, atau keputusan organisasi. Oleh karena itu, pemeriksaan teknis perlu disertai pertimbangan mengenai tujuan penggunaan, pihak yang terdampak, akses terhadap data, dan peluang untuk mengoreksi keputusan. Pada proyek analitik layanan kesehatan di puskesmas Indonesia, misalnya, data pasien hanya semestinya dipakai sesuai tujuan yang jelas dan akses petugas perlu dibatasi menurut perannya. Pembahasan mengenai pengelolaan informasi kesehatan menyoroti bahwa mutu layanan dan efisiensi sistem informasi berjalan berdampingan dengan tantangan pengelolaan data pasien (Faisal, 2025). Perlindungan kerahasiaan dokumen digital

juga memerlukan prosedur penyimpanan dan akses yang terkelola (Hutama, 2025).

Bias dapat muncul ketika kumpulan data tidak mewakili kelompok yang akan menerima layanan, label historis mencerminkan keputusan lama yang tidak adil, atau fitur menjadi proksi bagi atribut sensitif. Pemeriksaan praktis dapat dilakukan dengan membandingkan kesalahan dan kinerja model pada kelompok yang relevan, lalu meninjau apakah perbedaan tersebut menimbulkan dampak yang tidak dapat dibenarkan. Hasil pemeriksaan tidak seharusnya ditafsirkan sebagai bukti otomatis bahwa model adil atau tidak adil; konteks, tujuan, dan konsekuensi keputusan tetap perlu dikaji. Rekayasa fitur juga perlu memperhatikan asal dan makna setiap variabel agar hubungan yang dipelajari model tidak disalahartikan sebagai hubungan sebab-akibat (Liu, 2018).

Keamanan data perlu diterapkan sepanjang pengumpulan, penyimpanan, pemrosesan, dan pertukaran. Praktik yang dapat digunakan mencakup pembatasan hak akses, pencatatan aktivitas, pemisahan lingkungan pengembangan dan produksi, serta perlindungan data dalam penyimpanan dan pengiriman. Untuk data sensitif yang diproses pada infrastruktur bersama, pendekatan komputasi rahasia berupaya memberi perlindungan terhadap data yang sedang digunakan, meskipun rancangan dan overhead kinerjanya perlu dipertimbangkan (Arman,

2026). Sistem pertukaran pesan dapat menggunakan pendekatan kriptografi untuk melindungi komunikasi, tetapi pengelolaan kunci dan konfigurasi tetap menjadi bagian dari keamanan, bukan persoalan yang selesai hanya dengan memilih algoritma enkripsi (Mulyadi, 2026).

Deployment mencakup cara model disajikan kepada pengguna atau sistem lain, versi artefak yang dijalankan, serta mekanisme pengembalian ke versi sebelumnya jika ditemukan masalah. Sebelum peluncuran penuh, model dapat diuji pada sebagian alur kerja atau berjalan dalam mode bayangan, yaitu menghasilkan keluaran tanpa langsung mengubah keputusan. Setelah diluncurkan, monitoring perlu memeriksa mutu masukan, perubahan distribusi fitur, tingkat kesalahan berdasarkan label yang tersedia, dan pola penggunaan keluaran. Perubahan karakteristik data operasional dapat membuat kinerja model menurun meskipun kode tidak berubah; karena itu, monitoring menghubungkan perilaku produksi dengan rencana evaluasi dan pembaruan model (Unknown, 2011; Unknown, 2026).

Dashboard monitoring dapat membantu tim membedakan masalah kualitas data, perubahan pola, dan gangguan layanan. Gambar berikut menampilkan rancangan antarmuka untuk meninjau indikator model serta menelusuri peringatan yang memerlukan pemeriksaan.



Gambar 14.1 Antarmuka Monitoring Model. Sumber: Diolah penulis

Panel pada antarmuka tersebut mengarahkan perhatian dari satu skor menuju beberapa sinyal operasional, termasuk perubahan data dan status layanan. Tim tetap perlu menetapkan siapa yang meninjau peringatan, batas tindakan yang diizinkan, dan prosedur eskalasi ketika model atau sumber data bermasalah.### Monitoring juga perlu memasukkan mekanisme umpan balik untuk mencatat koreksi pengguna serta hasil pemeriksaan kasus, sehingga evaluasi berikutnya dapat menggunakan bukti operasional yang lebih lengkap. Perubahan atau penghentian model harus didokumentasikan, terutama apabila keluaran memengaruhi akses layanan atau keputusan terhadap individu. Pendekatan keamanan yang mencakup perlindungan data saat diproses maupun saat

dipertukarkan melengkapi tata kelola penerapan model (Arman, 2026; Hutama, 2025).

H. Rangkuman

Keberhasilan proyek data mining terintegrasi bergantung pada keterhubungan antara kebutuhan organisasi, pengolahan data, pemodelan, evaluasi, dan tindakan operasional. Proyek sebaiknya dimulai dengan merumuskan keputusan yang perlu didukung, bukan dengan memilih algoritma. Pada kasus koperasi, misalnya, tujuan prediksi keterlambatan pembayaran adalah membantu petugas menentukan prioritas tindak lanjut. Karena itu, tim perlu menyepakati cakupan, pemangku kepentingan, sumber data, tanggung jawab, sumber daya, risiko, jadwal, kriteria penerimaan, serta cara hasil digunakan. Metrik teknis seperti F1-score perlu dibedakan dari manfaat operasional, seperti ketepatan prioritas atau pengurangan waktu pemeriksaan.

Rencana kerja membagi proyek menjadi tahap perumusan, pemahaman data, persiapan, pemodelan, evaluasi, dan penerapan. Setiap tahap menghasilkan keluaran yang dapat diperiksa sebelum pekerjaan berlanjut. Pemeriksaan mutu data, kesesuaian fitur dengan waktu prediksi, serta pengujian hasil pada periode dan kelompok anggota yang relevan membantu menemukan keterbatasan model. Prediksi harus mendukung pertimbangan organisasi,

bukan menggantikan keputusan petugas. Dokumentasi definisi label, asal data, prapemrosesan, versi kode, konfigurasi model, hasil evaluasi, dan batas penggunaan mendukung keterlacakan serta pengulangan eksperimen.

Implementasi Python menerjemahkan rancangan tersebut ke dalam proses yang dapat dijalankan dan direproduksi. Pemisahan data latih dan uji, stratifikasi kelas, imputasi, standardisasi, dan pemilihan fitur perlu mengikuti karakteristik masalah. Pipeline membantu memastikan transformasi dipelajari dari data pelatihan dan diterapkan secara konsisten saat prediksi. Fitur yang baru diketahui setelah kejadian target harus dikeluarkan untuk mencegah kebocoran informasi. Kinerja dinilai dengan metrik yang sesuai konsekuensi kesalahan; recall, precision, F1-score, dan confusion matrix dapat ditinjau bersama. Ambang keputusan perlu mempertimbangkan biaya kesalahan dan kapasitas petugas menindaklanjuti peringatan. Dengan demikian, praktik Sistem Informasi yang baik memadukan mutu teknis, tata kelola, konteks penggunaan, dan kesiapan alur kerja agar keluaran analitik dapat dipercaya serta memberi manfaat nyata.

I. Latihan Mahasiswa

a. Essay

1. Analisis bagaimana perumusan keputusan operasional memengaruhi pemilihan data, fitur, metrik evaluasi,

dan bentuk penerapan model untuk memprediksi keterlambatan pembayaran anggota koperasi. Jelaskan pula risiko yang muncul jika proyek dimulai dengan memilih algoritma sebelum masalah dan pengguna hasilnya ditetapkan.

2. Bandingkan ukuran keberhasilan teknis seperti F1-score dengan ukuran manfaat operasional seperti ketepatan prioritas tindak lanjut. Rumuskan cara tim koperasi menyepakati kriteria penerimaan yang menghubungkan kedua jenis ukuran tersebut sebelum model dilatih dan diterapkan.
3. Evaluasi penggunaan pipeline Python yang menggabungkan imputasi, standardisasi, dan regresi logistik. Telaah bagaimana pemisahan data, stratifikasi, pemilihan fitur berdasarkan waktu prediksi, dan penyimpanan pipeline memengaruhi validitas evaluasi serta konsistensi prediksi saat diterapkan.
4. Analisis konsekuensi penggunaan ambang klasifikasi bawaan ketika kesalahan tidak mendeteksi anggota yang akan terlambat memiliki biaya tinggi. Rumuskan pendekatan evaluasi yang mempertimbangkan recall, precision, F1-score, confusion matrix, dan kapasitas petugas dalam menindaklanjuti peringatan.
5. Argumentasikan mengapa dokumentasi proyek perlu mencakup definisi label, asal data, keputusan prapemrosesan, versi kode, konfigurasi model, hasil evaluasi, dan batas penggunaan. Telaah bagaimana dokumentasi tersebut membantu keterlacakan, pengulangan eksperimen, dan peninjauan model setelah kondisi organisasi berubah.

b. Pilihan Berganda

1. Sebuah koperasi ingin memprediksi keterlambatan pembayaran agar petugas dapat menentukan prioritas tindak lanjut. Langkah awal perancangan proyek yang paling tepat adalah...
    a. Memilih algoritma klasifikasi dengan akurasi tertinggi
    b. Menetapkan keputusan yang hendak didukung dan populasi sasaran
    c. Mengubah semua variabel menjadi skala yang sama
    d. Menentukan jumlah lapisan jaringan saraf
    e. Menyusun laporan hasil sebelum menginventarisasi data
2. Tim memperoleh F1-score tinggi, tetapi waktu pemeriksaan petugas tidak berkurang dan prioritas tindak lanjut tetap tidak tepat. Kesimpulan yang paling sesuai adalah...
    a. Model pasti mengalami kebocoran target
    b. Data uji perlu dihapus dan diganti data latih
    c. Kinerja teknis belum membuktikan tercapainya manfaat operasional
    d. F1-score harus selalu menggantikan ukuran operasional
    e. Model perlu diterapkan tanpa evaluasi tambahan

3. Dalam memprediksi keterlambatan pembayaran pada awal periode, fitur yang paling berisiko menimbulkan kebocoran target adalah...
    a. Jumlah pinjaman saat pengajuan
    b. Pendapatan bulanan yang dicatat sebelum jatuh tempo
    c. Jumlah tanggungan yang tercatat saat pendaftaran
    d. Status pembayaran yang baru diketahui setelah melewati jatuh tempo
    e. Lama keanggotaan yang dihitung pada tanggal pengajuan
4. Apa manfaat utama menggabungkan imputasi, standardisasi, dan klasifikasi dalam satu pipeline yang dilatih pada data latih?
    a. Menghilangkan kebutuhan untuk memisahkan data latih dan data uji
    b. Menjamin semua kelas memiliki jumlah anggota yang sama
    c. Memastikan transformasi dipelajari dari data latih dan diterapkan konsisten
    d. Menghapus kebutuhan untuk menentukan fitur
    e. Menjamin model selalu menghasilkan prediksi yang benar
5. Mengapa pembagian data dengan stratifikasi digunakan dalam contoh klasifikasi?
    a. Agar proporsi kelas lebih terjaga pada data latih dan data uji

b. Agar seluruh data masuk ke dalam data uji
c. Agar nilai yang hilang otomatis menjadi nol
d. Agar semua fitur memiliki kontribusi yang sama
e. Agar model tidak perlu dievaluasi dengan metrik

6. Jika kasus keterlambatan yang tidak terdeteksi membawa risiko tinggi, metrik mana yang perlu mendapat perhatian khusus?
   a. Akurasi saja
   b. Waktu komputasi saja
   c. Jumlah fitur yang digunakan
   d. Recall untuk kelas terlambat
   e. Nilai random*state*
7. Tim hendak mengubah ambang keputusan model. Pertimbangan paling tepat adalah...
   a. Memilih ambang yang menghasilkan jumlah prediksi positif paling banyak
   b. Menyesuaikan ambang dengan konsekuensi kesalahan dan kapasitas tindak lanjut
   c. Menggunakan ambang yang sama untuk semua konteks tanpa evaluasi
   d. Mengubah ambang sampai semua prediksi menjadi kelas terlambat
   e. Menetapkan ambang hanya berdasarkan banyaknya fitur
8. Anggota koperasi dengan riwayat transaksi lebih pendek mungkin memperoleh prediksi yang kurang stabil. Tindakan evaluasi yang paling tepat adalah...

a. Menghapus seluruh anggota dengan riwayat pendek tanpa pemeriksaan
b. Menganggap hasil agregat berlaku sama bagi semua anggota
c. Menguji kinerja berdasarkan kelompok relevan dan membahas konsekuensinya dengan petugas
d. Mengganti semua nilai yang hilang dengan nol
e. Menggunakan data pelatihan sebagai satu-satunya dasar evaluasi

9. Dokumentasi definisi label, asal data, prapemrosesan, versi kode, dan konfigurasi model terutama mendukung...
    a. Penghapusan kebutuhan untuk menguji model
    b. Keterlacakan dan pengulangan eksperimen
    c. Peningkatan jumlah observasi secara otomatis
    d. Penggantian tanggung jawab pemilik proses
    e. Penghapusan seluruh risiko perubahan data
10. Sebelum layanan prediksi menerima masukan baru, pemeriksaan skema data diperlukan terutama untuk...
    a. Memastikan format dan satuan masukan sesuai dengan yang diharapkan
    b. Mengganti proses evaluasi model
    c. Menghilangkan kebutuhan pencatatan versi pustaka
    d. Menentukan populasi sasaran tanpa masukan pemilik proses

e. Menjamin perubahan kebijakan tidak memengaruhi prediksi

c. Soal Project / Studi Kasus

1. Koperasi di Yogyakarta akan menerapkan model prediksi keterlambatan pembayaran. Tim menemukan bahwa sebagian anggota memiliki riwayat transaksi panjang, sedangkan anggota baru hanya memiliki sedikit catatan. Data juga mencakup status pembayaran setelah jatuh tempo, tetapi petugas ingin menerima peringatan sebelum tanggal tersebut. Manajer mengusulkan penggunaan akurasi sebagai satu-satunya ukuran keberhasilan dan meminta model langsung dihubungkan ke sistem layanan. Petugas lapangan khawatir jumlah peringatan akan melebihi kapasitas tindak lanjut mereka. Rumuskan rancangan evaluasi dan penerapan bertahap, termasuk pemilihan fitur, pengujian kelompok, metrik, kriteria penerimaan, dan peran petugas dalam penggunaan hasil.
2. Sebuah tim Sistem Informasi membangun layanan prediksi menggunakan Python dari data pembayaran koperasi. Kode awal mengimputasi dan menstandardisasi seluruh data sebelum membagi data menjadi latih dan uji. Model menunjukkan hasil evaluasi yang tinggi, tetapi format ekstraksi data berubah ketika layanan akan dipasang. Tim juga belum mencatat definisi label, tanggal ekstraksi, versi pustaka, atau konfigurasi model. Pimpinan meminta layanan diluncurkan segera karena hasil uji tampak menjanjikan. Telaah risiko teknis dan tata kelola proyek tersebut, lalu susun langkah koreksi sebelum

penerapan, rencana pemeriksaan masukan, dokumentasi, dan mekanisme pemantauan setelah peluncuran.

DAFTAR PUSTAKA

Agung Suburdjati, B., Sukoco, L., & Cedric, X. (2026). Analisis Prediksi Kadar CO_2 (PPM) Berdasarkan Suhu Dan Kelembapan Menggunakan Sensor DHT22 Dan MQ135 Dengan Perbandingan Model Machine Learning Berbasis IoT Pada ESP32. *Jurnal Sistem Informasi Galuh*, *4*(1), 99-106. https://doi.org/10.25157/jsig.v4i1.5726

Al Riyami, N., & Al Hinai, A. (2026). Fast Pattern Discovery in Big Data Mining Using Incremental Mining for Social Media Analysis. *PatternIQ Mining*, *3*(1), 66. https://doi.org/10.70023/piqm314

Alfirdaus, M. H., & Martiano, M. (2024). Perancangan dan Implementasi Sistem Monitoring Kualitas Udara Berbasis IoT. *Jurnal Sistem Informasi dan Teknologi Jaringan*, *5*(1). https://doi.org/10.63703/sisfotekjar.v5i1.28

Amanda, D. P., & Absharina, E. D. (2025). Implementasi ai-powered intrusion detection systems untuk mendeteksi ancaman keamanan pada big data. *Simtek : jurnal sistem informasi dan teknik komputer*, *10*(1), 29-33. https://doi.org/10.51876/simtek.v10i1.1381

Angga Wijaya, M., & Meinawati (2026). Penerapan data mining menggunakan algoritma apriori untuk menentukan menu makanan bergizi gratis. *Jurnal Sistem Informasi Galuh*, *4*(2), 78-85. https://doi.org/10.25157/jsig.v4i2.6092

Arman, M., Wijaya, N., ., M., & ., I. (2026). Confidential Virtual Machines Untuk Keamanan Data Sensitif Desain

Arsitektur, Overhead Kinerja, Dan Strategi Penempatan Beban Kerja. *JuSiTik : Jurnal Sistem dan Teknologi Informasi Komunikasi*, *9*(2), 71-83. https://doi.org/10.32524/jusitik.v9i2.1999

Ayang Prabu, P., Sidiq, M., & Maulana, H. (2026). Penerapan data mining untuk clustering potensi kesehatan masyarakat desa menggunakan algoritma k-means. *Jurnal Mahasiswa Sistem Informasi Galuh*, *2*(2), 121-132. https://doi.org/10.25157/jmsig.v2i2.5839

B., S. S., & C., T. (2026). Privacy-Preserving Federated Analytics Framework for Secure OLAP-Driven Data Mining in Enterprise Data Warehouses. *International Journal of Data Mining Techniques and Applications*, *15*(1), 1. https://doi.org/10.20894/ijdmta.102.015.001.001

Banjarnahor, W. S. A. (2022). Rancang Bangun Sistem Monitoring Kateter Pasien Berbasis Iot (Sismoniterin) Pada Rumah Sakit Mitra Medika Medan. *Jurnal Multimedia dan Teknologi Informasi (Jatilima)*, *4*(01), 25-37. https://doi.org/10.54209/jatilima.v4i01.320

Blázquez-García, A. (2022). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, *54*(3). https://doi.org/10.1145/3444690

Budiawati, R., Rohayati, T., & Dwicahya Supriatman, R. (2026). Penerapan data mining untuk prediksi penjualan produk terlaris menggunakan metode k-nearest neighbor (knn) (studi kasus: Toko grosir rizki). *Jurnal Mahasiswa Sistem Informasi Galuh*, *2*(2), 166-171. https://doi.org/10.25157/jmsig.v2i2.5983

Dua, S., & Du, X. (2016). *Data Mining and Machine Learning in Cybersecurity*. Auerbach Publications. https://doi.org/10.1201/b10867

Faisal, L. (2025). Analisis sistematis literatur tentang optimalisasi sistem informasi kesehatan untuk meningkatkan efisiensi dan kualitas layanan di puskesmas. *Simtek : jurnal sistem informasi dan teknik komputer*, *10*(1), 63-69. https://doi.org/10.51876/simtek.v10i1.1482

Fernando, J. G., & Jain, D. S. K. (2022). Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines. *TechnoareteTransactions on Intelligent Data Mining and Knowledge Discovery*, *2*(3). https://doi.org/10.36647/ttidmkd/02.03.a003

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Hartono, B., Lusiana, V., & Al Amin, I. H. (2025). Perbandingan Proses Klasterisasi Data Menggunakan K-Means Clustering dan Agglomerative Hierarchical Clustering. *JURIKOM (Jurnal Riset Komputer)*, *12*(4), 628-635. https://doi.org/10.30865/jurikom.v12i4.8766

Hassan, M., & El-Shennawy, Y. (2025). Pattern Mining in Smart Grid Energy Data Using Enhanced Binary Firefly Optimization. *PatternIQ Mining*, 13. https://doi.org/10.70023/sahd/25020302

Hisham, S. (2025). An Extensive Mining of Ethereum Contract Code by Adapting Relevant Feature Generators to Optimize The Performance of Anomaly Detection Using an Ensemble Model. *Asia Pacific Journal of Information Technology and Multimedia*, *14*(1), 258-294. https://doi.org/10.17576/apjitm-2025-1401-15

Huffaker, R., Bittelli, M., & Rosa, R. (2018). *Data Preprocessing*. Oxford University Press. https://doi.org/10.1093/oso/9780198782933.003.0006

Hussin, M. (2022). Privacy based classification model of public data by utilizing two-steps validation approach. *Asia Pacific Journal of Information Technology and Multimedia*, *11*(2), 123-132. https://doi.org/10.17576/apjitm-2022-1102-10

Hutama, D. J., & Ramadhan, Y. F. (2025). Implementasi konsep cyber notary dan peran notaris dalam perlindungan data pribadi untuk menjamin kerahasiaan. *Jurnal Hukum Kenotariatan Otentik's*, *7*(1), 1-19. https://doi.org/10.35814/otentik.v7i1.7881

Imron, I., Fridianto, T., & Hari Nugroho, D. (2023). Identifikasi Bahaya Keamanan Siber pada Perizinan Bahan Nuklir Online di Indonesia. *Jurnal Pengawasan Tenaga Nuklir*, *3*(2), 14-20. https://doi.org/10.53862/jupeten.v3i2.003

Iqbal, I., Hidayat, N., Gevano, D. P., & Ilahi, A. P. R. (2025). Segmentasi Pelanggan Menggunakan K-Means Clustering Berdasarkan Data Kepribadian dan Pola

Konsumsi. *Jurnal Teknik Informatika (Jutif)*, *6*(5), 3914-3924. https://doi.org/10.52436/1.jutif.2025.6.5.5140

Jain, V. (2023). An Overview on Data Mining and Data Fusion. *Indian Journal of Data Mining*, *3*(1), 1-5. https://doi.org/10.54105/ijdm.a1624.053123

Jain, D. S. K., & Fernando, J. G. (2022). Cybersecurity using Data Mining Techniques. *TechnoareteTransactions on Intelligent Data Mining and Knowledge Discovery*, *2*(4). https://doi.org/10.36647/ttidmkd/02.04.a003

Jumawal, J., Mahpuz, M., Sadali, M., & Wasil, M. (2026). Implementasi Algoritma K-Nearest Neighbor pada Sistem Deteksi Tingkat Dehidrasi Berdasarkan Warna Urin Berbasis IoT dan Android. *Infotek : Jurnal Informatika dan Teknologi*, *9*(2), 732-739. https://doi.org/10.29408/jit.v9i2.35273

Kaestria, R., & Himmah, E. F. (2023). Implementasi Bahasa Pemrograman Python untuk Path analysis. *Jurnal Komputasi*, *11*(2), 105-117. https://doi.org/10.23960/komputasi.v11i2.6634

Kahar, A.M. (2021). Child Marriage Data Classification in Maluku Utara Using Kernel Regression and Support Vector Machine. *Barekeng*, *15*(3), 401-408. https://doi.org/10.30598/barekengvol15iss3pp401-408

Kennedy, C. M., & Doni, R. (2025). Implementasi Sistem Keamanan Data untuk Proyek Konstruksi Menggunakan Metode Algoritma Rivest Cipher 4 (RC4) Berbasis Android. *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi*

Akuntansi, *5*(2), 334-343. https://doi.org/10.46880/tamika.vol5no2.pp334-343

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining*. Woodhead Publishing Limited. https://doi.org/10.1533/9780857099440

KRYVENCHUK, Y., & KHANAS, M. Y. (2022). Algorithm of data mining and processing of related data in social networks. *Herald of Khmelnytskyi National University. Technical sciences*, *311*(4), 115-118. https://doi.org/10.31891/2307-5732-2022-311-4-115-118

Kurniati, I., Dewi, C. S., & Juniantika, R. (2021). Penerapan data mining dengan algoritma neural network pada sistem informasi prediksi kasus balita gizi buruk di provinsi nusa tenggara barat. *JRIS: JURNAL REKAYASA INFORMASI SWADHARMA*, *1*(1), 20-27. https://doi.org/10.56486/jris.vol1no1.61

Liu, H. (2018). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press. https://doi.org/10.1201/9781315181080

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). *Big Data Preprocessing*. Springer International Publishing. https://doi.org/10.1007/978-3-030-39105-8

Mulyadi, B. H., innuddin, M., Asroni, O., Azwar, M., & Latif, K. A. (2026). Impelementasi algoritma aes dan rsa untuk keamanan pada aplikasi web chat team sync. *Jurnal Manajemen Informatika dan Sistem Informasi*, *9*(1), 20-28. https://doi.org/10.36595/misi.v9i1.1807

Nurlaela, L., Suhanda, Y., Sopian, A., Dewi, C. S., & Syahrial, R. (2025). Pengembangan framework data mining berbasis deep neural network dengan eksplorasi teknik transfer learning untuk prediksi dan klasifikasi data. *JRIS : Jurnal Rekayasa Informasi Swadharma*, *5*(1), 132-141. https://doi.org/10.56486/jris.vol5no1.723

Priandini, B., Asfi, M., & Magdalena, L. (2025). IMPLEMENTASI K-MEANS RFM DAN HOLT-WINTERS EXPONENTIAL SMOOTHING ADDITIVE DALAM SISTEM BUSINESS INTELLIGENCE UNTUK STRATEGI PENGELOLAAN PELANGGAN PADA PERUSAHAAN TRANSPORTASI. Pembuatan Dashboard BI Segmentasi pelanggan dan peramalan Jumlah pelanggan menggunakan Tools Tableau menggunakan metode Kmeans RFM dan Holtwinters Exponential Smoothing. *METHODIKA: Jurnal Teknik Informatika dan Sistem Informasi*, *11*(2), 7-16. https://doi.org/10.46880/mtk.v11i2.4511

Putri, D. A., & Absharina, E. D. (2025). Eksplorasi penerapan teknologi big data dalam mendorong inovasi kesehatan di era digital. *Simtek : jurnal sistem informasi dan teknik komputer*, *10*(1), 19-22. https://doi.org/10.51876/simtek.v10i1.1382

Religia, Y. (2021). Comparative Analysis of Optimization Algorithms in Random Forest for Classification of Bank Marketing Data. *Jurnal Resti*, *5*(1), 187-192. https://doi.org/10.29207/resti.v5i1.2813

Setiawan, S. (2025). *Pengolahan data dan analisa statistik dengan spss*. UIN Gus Dur Press. https://doi.org/10.28918/ugp.19

Solimun, S., Fernandes, A. A. R., Nurjannah, N., Erwinda, E. G., Hardianti, R., & Arini, L. H. Y. (2023). *Metodologi Penelitian: Variabel Mining Berbasis Big Data dalam Pemodelan Sistem untuk Mengungkap Research Novelty*. Universitas Brawijaya Press. https://doi.org/10.11594/ubpress9786232967052

Unknown (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier. https://doi.org/10.1016/c2009-0-19715-5

Unknown (2019). *Introduction to Algorithms for Data Mining and Machine Learning*. Elsevier. https://doi.org/10.1016/c2018-0-02034-4

Unknown (2012). *Data Mining*. Elsevier. https://doi.org/10.1016/c2009-0-61819-5

Unknown (2026). *Machine Learning and Data Mining Annual Volume 2026 [Working Title]*. IntechOpen. https://doi.org/10.5772/intechopen.1010763

Wahyuningtyas, D., Permatasari, H., & Srirahayu, A. (2026). Pengembangan sistem monitoring dan rekapitulasi proyek berbasis web untuk optimasi workflow di digimizu digital management. *METHODIKA: Jurnal Teknik Informatika dan Sistem Informasi*, *12*(2), 31-38. https://doi.org/10.46880/mtk.v12i2.5628

Wulandari, S. a., Kuswara, H., & Palasara, N. (2021). Analisis penerapan data mining pada penjualan kerupuk rambak menggunakan metode naïve bayes classifer untuk optimasi strategi pemasaran. *Jurnal SITECH : Sistem Informasi dan Teknologi*, *3*(2), 83-94. https://doi.org/10.24176/sitech.v3i2.5444

Yuyun (2021). Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter. *Jurnal Resti*, *5*(4), 820-826. https://doi.org/10.29207/resti.v5i4.3146

LAMPIRAN: KUNCI JAWABAN

Kunci Jawaban BAB 1

a. Essay

6. Metode harus mengikuti kebutuhan keputusan, bukan popularitas algoritma. Perkiraan permintaan memerlukan regresi dan keluaran berupa nilai prediksi untuk persediaan; segmentasi pelanggan memerlukan klasterisasi dan keluaran berupa label kelompok untuk penyesuaian layanan. Keduanya berbeda dalam sasaran, data, dan tindakan yang didukung.
7. KDD mencakup pemilihan dan persiapan data, penambangan data, serta penilaian dan interpretasi; penambangan data adalah tahap analitisnya. Data yang tidak sesuai dapat menghasilkan pola yang menyesatkan. Tim perlu memeriksa relevansi dan mutu data, menilai pola dalam konteks, melibatkan pemangku kepentingan, dan meninjau ulang tahap sebelumnya bila diperlukan.
8. Pola komputasional belum tentu berguna tanpa kaitan dengan keputusan dan konteks. Pada catatan pasien, tenaga kesehatan perlu menilai makna kelompok dan keterbatasan data. Hasil dapat mendukung perencanaan program, tetapi tidak otomatis menjadi keputusan klinis. Dampak penggunaan dan kemungkinan salah tafsir perlu dipertimbangkan.
9. Penambangan data berfokus menemukan pengetahuan untuk kebutuhan tertentu; pembelajaran

mesin menyediakan pendekatan pembentukan model; kecerdasan buatan mencakup perancangan sistem yang melaksanakan tugas cerdas. Dalam klasifikasi transaksi, metode pembelajaran mesin dapat membentuk model sebagai bagian dari analisis penambangan data dalam sistem yang lebih luas.

10. Jika tujuannya memperkirakan kategori sasaran pemasaran dan tersedia label, klasifikasi dapat dipilih. Jika ingin menemukan kelompok pembeli tanpa label, klasterisasi lebih sesuai. Keluaran dan tindak lanjutnya berbeda. Sebelum menetapkan metode, pelaku usaha perlu memperjelas keputusan yang dituju, menilai mutu serta kelengkapan data, dan memastikan hasil dapat ditafsirkan.

b. Pilihan Ganda

11. B
12. D
13. A
14. E
15. C
16. B
17. D
18. A
19. E
20. C

c. Studi Kasus / Proyek

21. Pisahkan tujuan peramalan permintaan dari segmentasi pelanggan. Gunakan regresi untuk memperkirakan kebutuhan produk per periode dan klasterisasi bila ingin membentuk kelompok pelanggan tanpa label. Periksa konsistensi pencatatan antarwaktu dan cabang, kelengkapan data, serta keterwakilan periode. Nilai prediksi melalui pengujian yang sesuai dan nilai kelompok berdasarkan keterpakaian untuk promosi. Uji hasil bersama manajer sebelum menggunakannya untuk persediaan atau layanan.
22. Rumuskan kebutuhan sebagai dukungan perencanaan program, bukan penentuan diagnosis otomatis. Tinjau kelengkapan, konsistensi, dan keterwakilan data antar-fasilitas sebelum klasterisasi. Nilai kelompok bersama tenaga kesehatan dan pengelola program; periksa apakah pola relevan serta dapat ditindaklanjuti. Gunakan hasil sebagai bahan pertimbangan perencanaan, sertakan batas interpretasi, dan jangan menetapkan prioritas pasien hanya dari label kelompok tanpa penilaian profesional serta informasi kontekstual tambahan.

Kunci Jawaban BAB 2

a. Essay

23. Unit pelanggan memerlukan satu objek per pelanggan dan aturan agregasi transaksi, misalnya frekuensi, total, atau rata-rata belanja dalam periode yang sama. Tetapkan periode, cara menangani transaksi batal, dan atribut yang diringkas. Tanpa aturan konsisten,

pelanggan dengan transaksi lebih banyak dapat terhitung berulang dan mendominasi pola.

24. Label layak digunakan jika berasal dari hasil pemeriksaan historis yang dapat diverifikasi dan kriteria “perlu diperiksa” diterapkan konsisten. Catat waktu penetapan label. Keluarkan atribut yang baru diketahui setelah pemeriksaan karena dapat membocorkan sasaran dan membuat hasil analisis tampak lebih baik daripada kemampuan prediksi sebenarnya.
25. Nominal diringkas dengan frekuensi atau modus; ordinal dapat diringkas dengan median urutan; numerik dengan ukuran seperti median atau rentang sesuai konteks. Waktu dianalisis menurut urutan dan selang, sedangkan teks ditelaah isi atau temanya. Operasi yang tidak sesuai menghasilkan kesimpulan tanpa makna.
26. Data terstruktur dapat dieksplorasi melalui skema tabel; data semi-terstruktur memerlukan pembacaan penanda dan aturan organisasinya; data tidak terstruktur memerlukan ekstraksi informasi yang relevan. Integrasi perlu mempertahankan asal, konteks, dan makna data serta mencatat transformasi yang dilakukan sebelum membandingkan hasilnya.
27. Definisi, satuan, rentang, dan makna nilai kosong menjaga agar ringkasan serta visualisasi tidak salah ditafsirkan. Jika nilai kosong dianggap nol, misalnya, rata-rata suhu dapat tampak lebih rendah dan memicu keputusan keliru. Dokumentasi memungkinkan analis membedakan kondisi nyata dari masalah pencatatan.

b. Pilihan Ganda

1. B
2. D
3. A
4. C
5. B
6. A
7. D
8. E
9. C
10. B

c. Studi Kasus / Proyek

1. Tetapkan transaksi sebagai unit analisis dan verifikasi definisi serta konsistensi kriteria label antar cabang. Periksa kapan status pemeriksaan dicatat; gunakan atribut yang tersedia sebelum keputusan pemeriksaan dibuat. Keluarkan “hasil tindak lanjut” dari prediktor karena muncul setelah pemeriksaan dan berisiko membocorkan sasaran. Dokumentasikan kualitas, nilai kosong, dan perbedaan pencatatan. Sarankan uji coba terbatas serta peninjauan manusia sebelum hasil dipakai untuk memprioritaskan pemeriksaan.
2. Tetapkan identitas ruang dan waktu sebagai dasar pengaitan sumber, lalu periksa skema berkas semi-terstruktur dan metode ekstraksi catatan teks. Konversikan suhu ke satuan yang disepakati dengan

menyimpan nilai serta satuan asal; verifikasi zona waktu sebelum membandingkan waktu. Bedakan nilai kosong karena alat rusak, tidak berlaku, atau tidak tercatat. Ringkas suhu secara numerik dan telaah catatan sebagai teks. Tandai keterbatasan data dan validasi hasil sebelum menetapkan prioritas pemeriksaan.

Kunci Jawaban BAB 3

a. Essay

1. KDD mencakup keseluruhan proses dari pemilihan data hingga pemanfaatan temuan, sedangkan penambangan data merupakan tahap penerapan teknik untuk menemukan pola atau membangun model. Hasil pengelompokan perlu diuji kestabilannya, ditafsirkan berdasarkan konteks koperasi, dan dinilai kegunaannya sebelum menjadi dasar tindakan.
2. Jika model tidak mendukung pengaturan ventilasi, tim perlu meninjau kembali pengguna, keputusan, atau kriteria keberhasilan pada pemahaman bisnis. Jika data sensor tidak memadai, tim dapat kembali ke pemahaman dan persiapan data. Iterasi memastikan model menanggapi kebutuhan operasional serta kondisi data yang tersedia.
3. Artefak dapat berupa pernyataan masalah dan kriteria keberhasilan, inventaris sumber serta kualitas data, catatan transformasi dan dataset analitik, dokumentasi model dan eksperimen, hasil perbandingan dengan kriteria, serta rencana deployment. Artefak tersebut

memungkinkan pihak terkait meninjau keputusan dan keterkaitan hasil dengan tujuan proyek.

4. Kinerja statistik menunjukkan mutu model menurut ukuran analitik, tetapi tidak memastikan kegunaan organisasi. Tim perlu memeriksa kestabilan kelompok, karakteristik anggota, keterjelasan interpretasi, kesesuaian dengan pertanyaan awal, dan kemungkinan tindakan. Jika sulit dimaknai, tinjau fitur atau representasi data sebelum menggunakan hasil.
5. Dokumentasi merekam alasan pemilihan data dan transformasi, sehingga keputusan dapat diperiksa saat temuan tidak masuk akal. Catatan interpretasi dan batasan membantu pengguna menilai kekuatan bukti. Dengan demikian, tim dapat melacak kesalahan, mengulang tahap yang bermasalah, dan menghindari rekomendasi yang melampaui konteks data.

b. Pilihan Ganda

1. B.
2. B.
3. C.
4. C.
5. C.
6. A.
7. B.
8. D.
9. E.

10. C.

c. Studi Kasus / Proyek

1. Periksa duplikasi, konsistensi identitas, cakupan periode, dan keterwakilan anggota; dokumentasikan dasar pemilihan rentang waktu. Tinjau prapemrosesan serta transformasi fitur, lalu jalankan kembali pengelompokan dan evaluasi kestabilannya. Tafsirkan ciri setiap kelompok bersama pengurus. Tunda rekomendasi sampai hasil menjawab tujuan promosi, dapat dijelaskan, memiliki batasan yang tercatat, dan mendukung tindakan yang layak.
2. Awali dengan menetapkan pengguna, batas operasional aman, dan kriteria keberhasilan. Inventarisasi kelengkapan sensor, tetapkan penanganan data hilang, lalu dokumentasikan persiapan data dan eksperimen model. Evaluasi bukan hanya akurasi, tetapi juga kejelasan penjelasan dan keamanan keputusan ventilasi. Sebelum deployment, rencanakan integrasi, pemantauan, penanganan kegagalan sensor, serta mekanisme peninjauan oleh pengelola fasilitas.

Kunci Jawaban BAB 4

a. Essay

1. Periksa jumlah serta pola kehilangan menurut unit, periode, dan atribut; pastikan tekanan darah relevan bagi analisis dan telusuri proses pencatatannya. Pilih imputasi yang sesuai atau pertahankan penanda

kehilangan. Penghapusan baris perlu dibatasi karena dapat mengurangi data dan menimbulkan bias.

2. Rata-rata sesuai untuk distribusi relatif simetris tanpa pengaruh ekstrem, sedangkan median lebih tahan terhadap kemencengan. Kategori "tidak diketahui" dapat dipakai jika ketiadaan informasi bermakna. Pilihan harus sesuai tipe atribut dan tujuan analisis karena imputasi tunggal dapat mengurangi variasi.
3. Periksa ringkasan statistik dan visualisasi, lalu bandingkan temuan dengan batas domain, sumber pencatatan, serta konteks waktu. Koreksi atau keluarkan nilai hanya setelah kesalahan terverifikasi. Pertahankan pencilan yang mewakili kejadian sah dan dokumentasikan keputusan pembersihan.
4. Gunakan identitas utama, nomor referensi, waktu, alamat, jumlah, dan status untuk pencocokan. Bedakan salinan dari pembatalan atau transaksi sah yang dibuat ulang. Tandai pasangan dengan kecocokan tidak pasti untuk pemeriksaan manual, bukan langsung menghapusnya berdasarkan kemiripan nama atau waktu.
5. Pilih penyelarasan skala numerik sesuai kebutuhan algoritma, encoding sesuai sifat kategori, dan transformasi sesuai satuan atau distribusi. Pastikan representasi baru tidak menimbulkan makna yang keliru. Nilai pilihan berdasarkan tujuan analisis dan pertahankan informasi yang relevan bagi interpretasi.

b. Pilihan Ganda

1. C

2. A
3. D
4. B
5. D
6. A
7. B
8. C
9. A
10. B

c. Studi Kasus / Proyek

1. Kelompokkan kehilangan tekanan darah menurut unit dan periode, lalu periksa proses pencatatan serta kondisi alat. Jangan mengimputasi otomatis dengan rata-rata seluruh pasien; pilih imputasi yang sesuai atau pertahankan penanda. Keluarkan nomor telepon hanya jika tidak diperlukan. Selidiki nilai ekstrem berdasarkan konteks klinis. Standarkan tanggal menurut aturan sumber, catat seluruh perubahan, dan bandingkan distribusi sebelum serta sesudah prapemrosesan.
2. Cocokkan pesanan dengan nomor referensi, waktu, jumlah, status, dan atribut pelanggan; tinjau manual pasangan yang meragukan. Pertahankan catatan pembatalan dan pembuatan ulang dengan status yang tepat. Validasi harga ekstrem melalui katalog dan satuan sebelum menandainya sebagai kesalahan. Encoding kategori harus sesuai kebutuhan algoritma,

sedangkan penyelarasan skala numerik dipilih berdasarkan model dan tujuan prediksi stok serta penjualan.

Kunci Jawaban BAB 5

a. Essay

1. Perkiraan stok menghasilkan nilai numerik dan memerlukan data berlabel historis, sehingga sesuai untuk pembelajaran terawasi serta dievaluasi berdasarkan ketepatan estimasi. Segmentasi mencari struktur tanpa kategori acuan, sehingga sesuai untuk pembelajaran tak terawasi dan memerlukan penilaian makna serta kegunaan kelompok.
2. Pembelajaran terawasi memerlukan label yang memadai dan konsisten; tak terawasi dapat mencari kelompok tanpa label, tetapi kelompok harus ditafsirkan; semi-terawasi menggabungkan sebagian tiket berlabel dengan tiket tanpa label. Tim perlu memeriksa mutu, konsistensi, dan keterwakilan label serta tujuan dan konsekuensi kategorisasi.
3. Label tidak konsisten dapat membuat model mempelajari perbedaan kebijakan unit, bukan pola risiko yang relevan. Samakan definisi operasional, tinjau contoh yang berbeda penilaian, periksa kelengkapan data, dan dokumentasikan sumber label. Setelah itu, evaluasi apakah label mendukung keputusan yang dituju.
4. Model harus menghasilkan keluaran sebelum keputusan dibuat, sesuai sumber daya komputasi dan

pola pembaruan. Untuk pemeriksaan transaksi harian, keterlambatan hasil mengurangi kegunaan meskipun prediksi akurat. Pertimbangkan integrasi alur kerja, beban petugas, serta konsekuensi transaksi normal yang keliru ditandai.

5. Rumusan masalah menetapkan objek transaksi, keluaran berupa prioritas pemeriksaan, data dan label yang tersedia atau perlu dikumpulkan, serta keputusan penutupan laporan yang akan didukung. Unsur tersebut mengarahkan pemilihan paradigma, fitur, ukuran evaluasi, waktu penyediaan hasil, dan penanganan konsekuensi kesalahan.

b. Pilihan Ganda

1. B
2. D
3. A
4. E
5. C
6. B
7. D
8. A
9. C
10. E

c. Studi Kasus / Proyek

1. Tetapkan tujuan sebagai pemeringkatan tiket untuk peninjauan, bukan keputusan otomatis atas status

mendesak. Seragamkan definisi kategori dan audit kesepakatan petugas sebelum memakai label. Periksa apakah 8 persen tiket berlabel mewakili fakultas, periode, dan jenis kasus; jika memadai, uji pendekatan semi-terawasi secara hati-hati. Evaluasi kesalahan per kategori dan fakultas. Sajikan prioritas serta alasan yang dapat ditinjau petugas, lalu pantau kinerja dan akses dasbor dalam alur layanan.

2. Rumuskan keluaran sebagai daftar transaksi untuk pemeriksaan sebelum laporan ditutup. Karena label belum tersedia, mulai dengan pencarian pola tak terawasi atau penanda awal berbasis penyimpangan, tanpa menganggap hasilnya sebagai bukti pelanggaran. Gunakan tinjauan harian untuk membangun label dan menguji konsistensi; pertimbangkan semi-terawasi setelah label cukup mewakili. Ukur temuan relevan dan beban salah tandai. Terapkan peninjauan manusia, ambang yang disepakati, serta pemrosesan yang selesai sebelum penutupan.

Kunci Jawaban BAB 6

a. Essay

1. Definisikan label dan waktu penetapannya secara konsisten; gunakan hanya fitur yang tersedia saat transaksi dinilai. Pisahkan data sebelum mempelajari parameter prapemrosesan, gunakan validasi untuk pemilihan model, dan simpan data uji untuk evaluasi akhir. Langkah ini mengurangi kebocoran dan menghasilkan estimasi yang lebih realistis.

2. Decision Tree menyediakan jalur aturan yang lebih langsung untuk ditinjau, tetapi pohon kompleks dapat peka terhadap perubahan data. Random Forest menggabungkan banyak pohon sehingga prediksi dapat lebih stabil, tetapi alasan prediksi gabungan lebih sulit ditelusuri. Pilihan bergantung pada kebutuhan penjelasan, karakteristik data, dan tujuan penggunaan.
3. Akurasi dapat menyamarkan jenis kesalahan yang biaya bisnisnya berbeda. Evaluasi perlu memeriksa kesalahan mengenali pemohon berisiko dan kesalahan menandai pemohon layak. Ambang keputusan disesuaikan dengan konsekuensi dan kapasitas pemeriksaan; prediksi risiko dapat digunakan sebagai sinyal bagi analis, bukan otomatis menjadi penolakan.
4. Entropi mengukur ketidakmurnian simpul, sedangkan information gain membantu memilih pemisahan yang menurunkannya. Kedalaman, jumlah minimum data per simpul, atau pemangkasan dapat membatasi kompleksitas. Bandingkan konfigurasi menggunakan validasi, kemudian ukur kinerja akhir pada data uji yang tidak terlibat dalam pemilihan model.
5. Naïve Bayes layak dipertimbangkan ketika perhitungan efisien diperlukan atau data pelatihan terbatas. Asumsi independensi bersyarat menyederhanakan penghitungan peluang, tetapi dapat kurang sesuai ketika fitur saling berkaitan. Periksa karakteristik fitur dan kinerja pada data validasi serta uji sebelum menerapkannya dalam proses operasional.

b. Pilihan Ganda

1. C
2. B
3. D
4. B
5. E
6. C
7. A
8. D
9. B
10. E

c. Studi Kasus / Proyek

1. Tetapkan label risiko berdasarkan hasil pembayaran dan tentukan batas waktu pengamatan. Keluarkan status tunggakan yang baru tercatat setelah pengajuan karena menimbulkan kebocoran; gunakan fitur yang tersedia saat pengajuan. Pisahkan data sebelum prapemrosesan dan pertahankan proporsi kelas. Bandingkan Decision Tree dan Random Forest melalui validasi, lalu uji pada data terpisah. Ukur kesalahan sesuai biayanya dan gunakan tanda risiko untuk prioritas pemeriksaan analis, bukan penolakan otomatis.
2. Definisikan kelas berhenti secara operasional dan tetapkan jangka waktu prediksi. Singkirkan data promosi yang dicatat setelah pelanggan berhenti serta batasi fitur pada informasi yang tersedia saat prediksi. Bandingkan pohon yang dapat ditelusuri dengan

Random Forest yang berpotensi lebih stabil menggunakan validasi dan data uji terpisah. Nilai kesalahan berdasarkan dampak program retensi; gunakan hasil sebagai prioritas tindak lanjut, bukan keputusan otomatis, dan pantau perubahan kinerja dari waktu ke waktu.

Kunci Jawaban BAB 7

a. Essay

1. Regresi memperkirakan keluaran numerik, misalnya durasi layanan; klasifikasi menetapkan kategori, misalnya risiko keterlambatan. Keduanya memerlukan sasaran yang jelas dan evaluasi sesuai tujuan. Akurasi prediksi menunjukkan kemampuan memperkirakan, bukan bukti bahwa perubahan fitur menyebabkan perubahan respons.
2. Intersep 20 adalah prediksi saat promosi nol dalam model, sedangkan koefisien 3 menyatakan kenaikan prediksi penjualan Rp3 juta per tambahan Rp1 juta promosi. Pola itu hanya ilustrasi, bukan bukti kausal. Keputusan produksi memerlukan evaluasi data baru dan pertimbangan musim serta saluran penjualan.
3. Regresi sederhana memakai satu fitur, sedangkan regresi berganda mempertimbangkan beberapa fitur dengan fitur lain dianggap tetap. Korelasi kuat dapat menyulitkan interpretasi koefisien; fitur yang tidak tercatat dapat mengaburkan hubungan. Fitur juga harus tersedia saat prediksi dibuat agar model berguna secara operasional.

4. Data pengujian mengukur kemampuan prediksi pada observasi baru, sedangkan residual membantu mengenali pola ketidaksesuaian. Kelengkungan dapat mendorong pengujian transformasi fitur; kesalahan yang membesar perlu diperiksa lebih lanjut. Bandingkan alternatif pada data pengujian dan pilih model berdasarkan kinerja serta kegunaannya.
5. Catat prediksi dan hasil aktual, lalu pantau kesalahan serta perubahan konteks pasar dan mutu data. Tinjau ulang fitur dan kinerja model ketika pola operasional berubah. Organisasi sebaiknya menyesuaikan atau mengganti model bila hasilnya tidak lagi mendukung keputusan secara andal.

b. Pilihan Ganda

1. B
2. C
3. D
4. B
5. A
6. C
7. D
8. B
9. C
10. D

c. Studi Kasus / Proyek

1. Tetapkan target berupa penjualan minggu berikutnya dan gunakan hanya fitur yang telah tersedia saat prediksi, misalnya promosi terencana serta persediaan awal. Jangan gunakan jumlah produk tersedia yang baru dicatat pada akhir periode target. Data satu musim membatasi generalisasi; tambahkan periode dan konteks musiman. Jangan tafsirkan koefisien sebagai sebab-akibat. Uji pada data yang tidak dipakai untuk pelatihan, bandingkan prediksi dengan hasil aktual, dan pantau kinerja lintas musim.
2. Garis lurus mungkin tidak memadai bila hubungan melengkung dan kesalahan membesar. Periksa konsistensi catatan mesin, jam kerja, volume produksi, dan konsumsi listrik; benahi definisi serta nilai yang hilang. Uji suku kuadrat atau representasi fitur lain, lalu bandingkan kinerja pada data pengujian. Sebelum menganggarkan energi, laporkan ketidakpastian dan batas penerapan. Catat hasil aktual setelah penggunaan serta pantau perubahan mesin dan proses.

Kunci Jawaban BAB 8

a. Essay

1. Rentang transaksi yang besar dapat mendominasi jarak dan membuat frekuensi kurang berpengaruh. Standardisasi atau normalisasi fitur dapat menyeimbangkan kontribusinya. Analis kemudian membandingkan profil dan metrik kelompok sebelum serta sesudah prapemrosesan, lalu menilai relevansi hasil terhadap tujuan segmentasi pelanggan.

2. Metode partisi sesuai jika jumlah kelompok diketahui dan kelompok relatif kompak; metode hierarkis membantu meninjau susunan bertingkat; metode kepadatan mengenali wilayah padat serta noise. Jika jumlah kelompok belum diketahui dan pencilan perlu dibedakan, pendekatan kepadatan layak dipertimbangkan, dengan evaluasi kesesuaian parameter dan konteks.
3. Centroid awal yang berbeda dapat menghasilkan penempatan serta pembaruan centroid berbeda, sehingga nilai fungsi objektif dan keanggotaan kelompok berubah. Analis dapat menjalankan K-Means beberapa kali dengan inisialisasi berbeda, membandingkan nilai objektif dan struktur hasil, lalu memeriksa stabilitas serta interpretasi kelompok.
4. Ukuran teknis tidak menjamin manfaat operasional. Analis perlu menelaah profil fitur setiap kelompok, mencocokkannya dengan tujuan analisis, dan menilai apakah organisasi dapat melakukan tindakan berbeda berdasarkan temuan tersebut. Kelompok yang tidak mendukung keputusan yang relevan mungkin tidak berguna meskipun terpisah secara matematis.
5. RFM menggabungkan kebaruan, frekuensi, dan nilai transaksi untuk membedakan pola penggunaan layanan. Profil kelompok dapat mengarahkan prioritas retensi atau penawaran, tetapi interpretasi perlu diperiksa terhadap operasi perusahaan dan kualitas fitur. Nama segmen tidak boleh diperlakukan sebagai label yang diberikan otomatis oleh algoritma.

b. Pilihan Ganda

1. B
2. C
3. D
4. C
5. B
6. A
7. C
8. B
9. A
10. C

c. Studi Kasus / Proyek

1. Periksa kualitas data dan nilai ekstrem, lalu normalisasi atau standardisasi fitur RFM agar rentang belanja tidak mendominasi. Karena empat segmen diminta, K-Means dapat diuji dengan inisialisasi berulang dan hasilnya dibandingkan. Telaah kekompakan, keterpisahan, stabilitas, serta profil fitur tiap kelompok. Periksa apakah akun bernilai tinggi tetapi jarang bertransaksi membentuk pola yang konsisten. Beri nama segmen setelah interpretasi kontekstual, kemudian usulkan tindak lanjut yang sesuai dan dapat dievaluasi.
2. Periksa kelengkapan, satuan, skala, dan kemungkinan kesalahan sensor pada fitur suhu serta getaran. Metode berbasis kepadatan dapat dipertimbangkan karena jumlah kelompok belum diketahui dan titik terpencil perlu dikenali sebagai noise; metode hierarkis

dapat membantu meninjau struktur kelompok bertingkat. Uji parameter dan bandingkan hasil, lalu telaah profil kelompok bersama petugas fasilitas. Bedakan anomali operasi dari gangguan sensor sebelum menjadikannya dasar pemeliharaan atau tindakan terhadap mesin.

Kunci Jawaban BAB 9

a. Essay

1. Unit nota menangkap barang yang dibeli bersama dalam satu pembelian, sedangkan unit pelanggan menggabungkan perilaku pembelian lintas nota. Penggabungan dapat menaikkan kemunculan bersama dan mengubah support serta aturan. Periksa kelengkapan ID, duplikasi, konsistensi kode item, dan periode penggabungan sebelum membandingkan hasil.
2. Support pasangan adalah 0,60, confidence 0,75, dan lift 0,9375. Pasangan cukup sering muncul, tetapi keterkaitannya tidak melampaui frekuensi dasar susu. Pengelola dapat menguji promosi secara terukur, tetapi tidak boleh menyatakan roti menyebabkan pembelian susu atau menjadikan confidence sebagai satu-satunya dasar keputusan.
3. Minimum support menyaring itemset menurut frekuensi seluruh transaksi, sedangkan confidence menilai proporsi konsekuen pada transaksi yang memuat anteseden. Ambang support terlalu tinggi menghilangkan pola berguna; terlalu rendah

menghasilkan banyak kandidat. Confidence tinggi dapat menyesatkan tanpa lift dan konteks.

4. Asosiasi hanya menunjukkan kemunculan bersama, yang dapat dipengaruhi popularitas, promosi lain, atau karakteristik pelanggan. Uji kausalitas dengan eksperimen terkontrol, misalnya membandingkan kelompok yang menerima promosi dengan kelompok pembanding, sambil mengendalikan periode dan faktor lain serta menetapkan ukuran keberhasilan.
5. Bentuk transaksi yang konsisten, bersihkan duplikasi dan kode item, lalu hitung support item tunggal. Gunakan sifat anti-monotonik untuk membentuk dan memangkas kandidat bertahap. Bentuk aturan dari itemset frequent, nilai confidence dan lift, lalu pertimbangkan kebutuhan gizi serta kebijakan menu sebelum merekomendasikan kombinasi.

b. Pilihan Ganda

1. B
2. C
3. C
4. C
5. B
6. B
7. B
8. B
9. B

10. C

c. Studi Kasus / Proyek

1. Rekonstruksi transaksi dari nomor nota dan tangani catatan tanpa ID melalui prosedur yang dapat diaudit; hapus duplikasi setelah verifikasi. Gunakan nota sebagai unit untuk analisis keranjang dan pisahkan pembelian grosir bila perlu. Hitung support untuk menyaring pola, lalu baca confidence bersama lift. Uji promosi dalam periode terbatas dan laporkan asosiasi sebagai dasar eksplorasi, bukan bukti sebab-akibat.
2. Tetapkan unit yang sesuai dengan keputusan: nota untuk paket pembelian atau mahasiswa-hari untuk perilaku harian, dengan penanganan pembelian berulang yang konsisten. Uji beberapa ambang support agar pola menu kurang populer tidak otomatis hilang. Gunakan lift untuk menilai keterkaitan relatif terhadap popularitas menu, lalu validasi paket melalui uji terbatas. Tinjau gizi, porsi, harga, dan kebijakan kampus bersama ahli terkait.

Kunci Jawaban BAB 10

a. Essay

1. Frekuensi sampling perlu mengikuti kecepatan perubahan suhu dan kebutuhan respons aplikasi. Peringatan cepat memerlukan interval yang cukup rapat, tetapi analisis tren gangguan mungkin menggunakan agregasi. Pertimbangkan kapasitas jaringan, penyimpanan, perangkat, dan nilai informasi

setiap pengukuran; uji rancangan sebelum diterapkan luas.

2. Tampilan kondisi terutama membutuhkan data aktual dan penyajian yang mudah dibaca. Peringatan memerlukan ambang, validasi, dan jalur respons yang cepat. Prediksi gangguan memerlukan data historis berkonteks, fitur relevan, dan evaluasi model. Perbedaan tujuan menentukan kebutuhan penyimpanan, analitik, serta lokasi pemrosesan.
3. Tanpa identitas, waktu, satuan, dan lokasi, nilai sulit dikaitkan dengan perangkat serta konteks yang benar. Penggabungan sensor dapat mencampur skala atau waktu, pola menjadi keliru, dan gangguan sukar dilacak. Model berisiko mempelajari hubungan yang tidak sahih sehingga hasil analisis kurang dapat dipercaya.
4. Periksa rentang berdasarkan karakteristik sensor, lalu tandai data hilang, duplikat, urutan waktu janggal, dan pola pengiriman berubah. Simpan rekaman meragukan untuk pemeriksaan konteks dan kondisi perangkat. Keluarkan data dari pelatihan hanya jika terbukti tidak layak, bukan sekadar berbeda dari pola umum.
5. Pemetaan menunjukkan titik asal dan jalur data sehingga pemeriksaan dapat dilakukan berurutan. Jika nilai tidak sampai, periksa sensor dan konektivitas; jika diterima tetapi tidak tampil, telusuri platform, analitik, dan aplikasi. Pemetaan yang sama membantu menentukan tempat pemrosesan dan tanggung jawab operasional.

b. Pilihan Ganda

1. C
2. B
3. A
4. C
5. B
6. C
7. A
8. A
9. B
10. B

c. Studi Kasus / Proyek

1. Tetapkan sampling berdasarkan kebutuhan peringatan suhu dan perubahan kondisi mesin; evaluasi pengurangan frekuensi atau agregasi untuk mengendalikan beban. Seragamkan satuan dan catat identitas, waktu, serta status perangkat. Validasi rentang, data hilang, dan jeda pengiriman; tandai data meragukan. Gunakan analitik peringatan untuk ambang operasional dan data historis berkualitas untuk prediksi gangguan. Pantau kapasitas jaringan dan penyimpanan.
2. Periksa kalibrasi serta konfigurasi satuan pada perangkat, sinkronisasi waktu, dan status konektivitas. Pada platform, validasi identitas, satuan, duplikasi, rentang, dan kelengkapan waktu; tandai jeda tanpa

menganggapnya perubahan kualitas udara. Verifikasi transformasi serta penyajian dasbor. Dokumentasikan aturan pembersihan dan kelayakan data, lalu analisis hanya data yang konteksnya memadai. Sampaikan keterbatasan cakupan sebelum menetapkan prioritas lokasi.

Kunci Jawaban BAB 11

a. Essay

1. Gabungkan catatan berdasarkan waktu dan identitas akun atau perangkat setelah menyelaraskan zona waktu serta format. Log autentikasi menunjukkan akses, log aplikasi memperlihatkan tindakan setelah masuk, dan lalu lintas memberi konteks komunikasi. Periksa kemungkinan identitas bersama, perubahan alamat jaringan, dan aktivitas sah sebelum menyimpulkan keterkaitan.
2. Klasifikasi terawasi sesuai ketika tersedia label tepercaya dan tujuan kategorisasi jelas, tetapi rentan terhadap label tidak lengkap dan ketimpangan kelas. Pengelompokan membantu menemukan kemiripan tanpa label awal, tetapi kelompok tidak otomatis bermakna sebagai ancaman. Pilihan bergantung pada tujuan, mutu data, dan kemampuan analis menafsirkan hasil.
3. Fitur tersebut dapat menandai lonjakan autentikasi atau perubahan komunikasi. Jendela pendek menonjolkan perubahan cepat, sedangkan jendela panjang menunjukkan kebiasaan dan tren. Interpretasi

harus mempertimbangkan jadwal layanan, pemeliharaan, dan penggunaan sah. Indikator merupakan petunjuk pemeriksaan, bukan bukti mandiri terjadinya serangan.

4. Gunakan skor untuk mengurutkan pemeriksaan, lalu periksa bukti lintas-sumber dan konteks aset. Catat alasan keputusan serta ketidakpastiannya. Pilih tindakan bertahap, seperti verifikasi pengguna atau pembatasan sementara, sesuai risiko. Hasil pemeriksaan dapat digunakan untuk memperbaiki label dan aturan analisis.
5. Label yang tidak lengkap dan ketimpangan kelas dapat membuat model melewatkan ancaman atau menghasilkan prioritas yang kurang berguna. Periksa konsistensi, kelengkapan, dan asal data; tinjau fitur dengan pengetahuan sistem; kemudian gunakan temuan analis untuk memperbarui label atau aturan. Pantau perubahan lingkungan dan kegunaan keluaran.

b. Pilihan Ganda

1. B.
2. C.
3. C.
4. A.
5. B.
6. B.
7. A.
8. C.

9. B.
10. B.

c. Studi Kasus / Proyek

1. Selaraskan zona waktu dan telusuri kemungkinan penggunaan bersama alamat sumber. Bandingkan kegagalan masuk, variasi akun, tujuan jaringan, serta volume keluar dengan jadwal ujian dan pemeliharaan. Korelasikan log portal dan jaringan sebelum menilai dugaan insiden. Verifikasi aktivitas pengguna atau perangkat; jangan menetapkan serangan hanya dari lonjakan. Jika diperlukan, lakukan verifikasi akun atau pembatasan sementara yang terarah, lalu dokumentasikan bukti, keputusan, dan ketidakpastian.
2. Tinjau kelengkapan label dan dampak ketimpangan kelas, lalu periksa perubahan aplikasi serta pola akses jarak jauh sebagai kemungkinan penyebab pergeseran data. Urutkan peringatan dengan skor, tetapi verifikasi bukti lintas-sumber dan dahulukan pemeriksaan berisiko tinggi. Hindari pemblokiran otomatis menyeluruh tanpa validasi; gunakan tindakan bertahap sesuai dampak layanan. Catat keputusan analis dan gunakan hasilnya untuk memperbaiki label, aturan, serta evaluasi model.

Kunci Jawaban BAB 12

a. Essay

1. Anomali titik menyimpang pada satu pengamatan; anomali kontekstual bergantung pada kondisi,

misalnya waktu atau kebutuhan suhu ruang; anomali kolektif muncul dari pola beberapa pengamatan. Unit analisis dapat berupa pembacaan, pembacaan dalam konteks operasi, atau rangkaian waktu, sehingga fitur dan tindak lanjutnya berbeda.

2. Statistik mengukur ketidaklaziman berdasarkan distribusi, tetapi sensitif terhadap asumsi distribusi dan nilai ekstrem. Jarak membandingkan kedekatan titik sehingga skala fitur perlu diperhatikan. Kepadatan menilai kelangkaan lokal dan dapat menangkap perbedaan kepadatan kelompok, tetapi tetap bergantung pada representasi data dan parameter yang sesuai.
3. Skor baku dapat menyesatkan saat distribusi, tren, atau musim berubah. Periksa nilai hilang, pencilan pencatatan, satuan, dan skala; kelompokkan menurut kondisi operasi dan waktu; pertimbangkan ukuran robust; lalu verifikasi alarm dengan kondisi ruangan, status pendingin, serta kualitas sensor.
4. LOF membandingkan kepadatan titik dengan lingkungan lokalnya, sedangkan Isolation Forest menilai kemudahan titik dipisahkan melalui pemisahan fitur bertahap. LOF dapat membantu saat anomali bersifat lokal; Isolation Forest dapat dipertimbangkan untuk pemisahan titik. Keduanya memberi skor, bukan diagnosis, sehingga konteks dan data sensor perlu diperiksa.
5. Evaluasi perlu mengukur alarm palsu dan keterlambatan deteksi, memeriksa kualitas data serta kestabilan skor, dan menilai kapasitas petugas

merespons notifikasi. Ambang ditetapkan sesuai konsekuensi operasional. Alarm harus dapat diverifikasi dengan konteks waktu, perangkat, dan proses sebelum tindakan, sehingga kinerja model dinilai berdasarkan manfaat dan risiko organisasi.

b. Pilihan Ganda

1. B
2. D
3. A
4. C
5. E
6. B
7. D
8. A
9. C
10. E

c. Studi Kasus / Proyek

1. Bedakan lonjakan singkat terkait pembukaan pintu dari pola berulang atau berkepanjangan dan dugaan sensor macet. Catat jadwal pemindahan barang, kondisi operasi, identitas ruang, serta status sensor; tangani nilai hilang dan validasi pencatatan. Uji pendekatan kontekstual dan metode statistik atau lokal setelah mempertimbangkan pola data. Tampilkan skor sebagai peringatan untuk pemeriksaan teknisi, bukan diagnosis

otomatis. Evaluasi alarm palsu dan waktu respons sebelum mengaktifkan notifikasi otomatis.

2. Standardisasi fitur numerik sebelum menguji metode jarak; bandingkan skor statistik, jarak, kepadatan, LOF, atau Isolation Forest sesuai struktur data. Gunakan label insiden yang terbatas untuk validasi dan tinjau kasus yang ditandai bersama petugas. Ukur alarm palsu, keterlambatan, serta beban tindak lanjut. Terapkan eskalasi bertahap: pemeriksaan telemetri dan konteks, verifikasi petugas, lalu tindakan yang disetujui. Hindari pemblokiran perangkat hanya berdasarkan skor model.

Kunci Jawaban BAB 13

a. Essay

1. Ensemble bermanfaat jika model penyusunnya cukup baik dan kesalahannya tidak sepenuhnya sejalan, sehingga penggabungan dapat menyeimbangkan prediksi. Model tambahan dengan kesalahan serupa tidak banyak menambah informasi. Karena itu, jumlah model perlu dinilai bersama keragaman, kinerja validasi, dan tujuan prediksi.
2. Bagging melatih model secara terpisah pada sampel bootstrap, sehingga pemrosesan dapat paralel dan hasil digabungkan melalui suara atau perataan. Boosting menyusun model berurutan untuk memperbaiki kesalahan sebelumnya. Bagging sesuai saat variasi prediksi menjadi perhatian; boosting layak

diuji saat perbaikan bertahap memberi hasil validasi lebih baik.

3. Random Forest menggabungkan banyak pohon yang dilatih pada sampel bootstrap dengan fitur acak pada pemisahan simpul. Mekanisme ini mendorong keragaman dan membantu menangkap hubungan nonlinier. Validasi terhadap kedalaman, jumlah pohon, dan fitur membantu memilih konfigurasi yang menggeneralisasi, bukan sekadar cocok pada data pelatihan.
4. Gunakan pembagian data yang sama untuk ketiga algoritma; latih pada data pelatihan, pilih parameter memakai validasi, dan evaluasi sekali pada data pengujian yang terpisah. Tuning mencakup kompleksitas serta jumlah iterasi. Pilih metrik yang sejalan dengan sasaran organisasi dan pertimbangkan biaya, waktu pelatihan, serta tindak lanjut operasional.
5. Pada kelas berisiko yang jarang, model dapat memperoleh akurasi tinggi dengan mengutamakan kelas mayoritas tetapi melewatkan banyak risiko. Recall mengukur deteksi kasus berisiko, sedangkan precision mencerminkan ketepatan peringatan. Ambang keputusan dan kapasitas petugas harus menyeimbangkan kasus terlewat dengan biaya pemeriksaan peringatan palsu.

b. Pilihan Ganda

1. C.
2. B.

3. D.
4. A.
5. E.
6. C.
7. B.
8. D.
9. A.
10. E.

c. Studi Kasus / Proyek

1. Bandingkan bagging dan boosting menggunakan pembagian data yang sama. Atur parameter pada data validasi dan pertahankan data pengujian untuk evaluasi akhir. Karena kelas berisiko minoritas, laporkan recall dan precision selain akurasi; pertimbangkan penyeimbangan kelas tanpa menggantikan evaluasi. Pilih ambang berdasarkan biaya kasus terlewat dan kapasitas pemeriksaan. Gunakan keluaran model sebagai rekomendasi yang dapat ditinjau petugas, serta periksa potensi bias dan dokumentasikan alasan keputusan.
2. Tolak pemilihan berdasarkan akurasi pada seluruh data karena dapat menyebabkan kebocoran informasi dan menyamarkan pelanggan respons yang terlewat. Latih setiap algoritma pada data pelatihan, pilih parameter serta ambang melalui validasi, lalu bandingkan sekali pada data pengujian. Gunakan recall dan precision kelas respons, disertai perkiraan biaya promosi dan

kapasitas tindak lanjut. Pertimbangkan waktu pelatihan, stabilitas, dan kebutuhan operasional; pilih konfigurasi yang menyeimbangkan peluang respons dan kampanye salah sasaran.

Kunci Jawaban BAB 14

a. Essay

1. Rumusan keputusan menentukan kapan prediksi dibutuhkan, siapa sasarannya, dan tindakan yang tersedia. Hal tersebut membatasi fitur pada informasi yang telah tersedia saat prediksi, memandu pemilihan metrik sesuai biaya kesalahan, serta menentukan bentuk layanan dan pengguna. Memilih algoritma lebih dahulu berisiko menghasilkan model yang tidak relevan atau tidak dapat ditindaklanjuti.
2. F1-score menggambarkan keseimbangan precision dan recall, sedangkan ukuran operasional menilai perubahan proses. Tim perlu menetapkan ambang teknis sekaligus sasaran manfaat, seperti ketepatan prioritas dan kemampuan petugas menindaklanjuti. Kriteria penerimaan sebaiknya diuji pada data yang relevan dan disepakati pemilik proses sebelum pelatihan.
3. Pipeline menjaga urutan transformasi dan mencegah transformasi dipelajari dari data uji. Stratifikasi membantu mempertahankan proporsi kelas, sedangkan fitur harus tersedia pada waktu prediksi. Penyimpanan pipeline secara utuh mendukung konsistensi antara pelatihan dan layanan, tetapi tetap

memerlukan evaluasi dan pemeriksaan skema masukan.

4. Ambang bawaan dapat menghasilkan terlalu banyak kasus yang tidak terdeteksi atau terlalu banyak peringatan. Tim perlu meninjau confusion matrix, recall, precision, dan F1-score pada beberapa ambang, lalu memperhitungkan biaya kesalahan serta kapasitas petugas. Ambang terpilih diuji sebelum digunakan dalam alur kerja.
5. Dokumentasi menjelaskan asal dan makna data, cara pemrosesan, konfigurasi, hasil, serta batas pemakaian model. Catatan tersebut memungkinkan tim menelusuri alasan prediksi, mengulang eksperimen, dan membandingkan versi. Ketika data atau proses berubah, dokumentasi membantu menentukan apakah model masih layak digunakan atau perlu diperbarui.

b. Pilihan Ganda

1. B
2. C
3. D
4. C
5. A
6. D
7. B
8. C
9. B

10. A

c. Studi Kasus / Proyek

1. Keluarkan status pembayaran yang baru diketahui setelah jatuh tempo dan gunakan hanya fitur yang tersedia pada waktu peringatan. Evaluasi kinerja agregat serta kelompok anggota dengan riwayat berbeda, lalu tinjau recall, precision, F1-score, dan dampak operasional. Tetapkan ambang berdasarkan risiko dan kapasitas petugas. Mulai dengan uji terbatas; petugas meninjau rekomendasi sebelum tindakan, sementara pemilik proses menyepakati kriteria penerimaan.
2. Ulangi eksperimen dengan membagi data terlebih dahulu, lalu melatih imputasi dan standardisasi hanya melalui pipeline pada data latih. Evaluasi kembali pada data uji yang terpisah. Periksa skema, tipe, dan satuan masukan sebelum prediksi diterima. Catat definisi label, asal dan tanggal data, versi pustaka, kode, serta konfigurasi. Luncurkan bertahap dengan pemantauan kinerja, kesalahan, perubahan data, dan mekanisme pelaporan serta peninjauan berkala.

GLOSARIUM

Algoritma: Rangkaian langkah komputasional yang digunakan untuk mengolah data atau menemukan pola. Dalam penambangan data, algoritma dipilih sesuai pertanyaan analisis dan bentuk keluaran yang dibutuhkan.

Data mining: Tahap analitis yang menerapkan metode komputasional untuk menemukan pola dari data. Dalam KDD, data mining merupakan bagian dari proses yang lebih luas, bukan keseluruhan proses penemuan pengetahuan.

Deteksi anomali: Pendekatan untuk mengenali data atau perilaku yang menyimpang dari pola umum. Hasilnya dapat digunakan sebagai penanda untuk pemeriksaan lebih lanjut.

KDD (Knowledge Discovery in Databases): Keseluruhan proses memperoleh pengetahuan yang berguna dari data, mulai dari pemilihan dan penyiapan data hingga penilaian serta penafsiran pola. Data mining merupakan salah satu tahap analitis di dalamnya.

Klasifikasi: Pendekatan analisis untuk menempatkan objek ke dalam kategori yang telah ditentukan. Keluarannya dapat berupa kategori, seperti tingkat risiko suatu transaksi.

Klasterisasi: Pendekatan untuk membentuk kelompok objek berdasarkan kemiripan karakteristik tanpa harus menetapkan kategori terlebih dahulu. Hasil

pengelompokan dapat membantu memahami segmen atau kelompok dalam data.

Model: Representasi pola atau hubungan yang dibentuk melalui analisis data. Model dapat menghasilkan keluaran seperti kategori atau nilai prediksi untuk mendukung keputusan.

Naïve Bayes: Metode klasifikasi yang menggunakan prinsip probabilitas untuk memperkirakan kategori suatu objek berdasarkan karakteristiknya. Dalam cuplikan ini, metode tersebut dicontohkan untuk analisis penjualan kerupuk rambak.

Penambangan data: Proses komputasional untuk menemukan pola, keteraturan, atau hubungan yang berguna dalam kumpulan data. Hasilnya perlu ditafsirkan sesuai persoalan dan konteks penggunaannya.

Pola asosiasi: Hubungan keterkaitan yang ditemukan antara unsur-unsur dalam data, misalnya produk yang kerap dibeli bersama. Pola ini dapat dimanfaatkan untuk mendukung penataan promosi.

Regresi: Pendekatan analisis untuk memperkirakan nilai, seperti kebutuhan pada periode mendatang. Hasilnya dapat digunakan dalam perencanaan persediaan.

Segmentasi pelanggan: Pengelompokan pelanggan berdasarkan karakteristik atau pola tertentu. Hasil segmentasi dapat digunakan untuk menyesuaikan layanan.

Analisis terawasi: Pendekatan analisis yang menggunakan label yang telah diketahui sebagai sasaran pembelajaran.

Atribut: Karakteristik yang dicatat pada suatu objek dalam kumpulan data.

Atribut kategoris: Atribut yang nilainya menyatakan kelompok atau kategori, bukan besaran numerik.

Atribut nominal: Atribut kategoris yang nilainya tidak memiliki urutan.

Atribut numerik: Atribut yang nilainya berupa angka, baik diskret maupun kontinu.

Atribut ordinal: Atribut kategoris dengan nilai yang memiliki urutan, tetapi jarak antartingkatnya belum tentu setara.

Data teks: Data dalam bentuk bahasa, seperti isi aduan, yang membawa informasi dan tidak dapat diperlakukan begitu saja sebagai bilangan.

Data waktu: Data yang mencatat waktu dan memiliki urutan kronologis.

Kumpulan data: Representasi terpilih dari suatu fenomena yang disusun untuk menjawab pertanyaan analitis.

Label: Atribut sasaran yang nilainya telah diketahui pada data tertentu dan digunakan dalam analisis terawasi.

Objek: Unit pengamatan yang direpresentasikan dalam kumpulan data.

Prapemrosesan data: Tahap menyiapkan atau mengubah data sebelum analisis dan pemodelan dengan tetap mempertahankan maknanya.

Skala interval: Skala pengukuran dengan jarak antarnilai yang bermakna, tetapi tanpa titik nol mutlak.

Skala rasio: Skala pengukuran dengan jarak bermakna dan titik nol yang memungkinkan perbandingan rasio.

Unit analisis: Jenis unit yang ditetapkan sebagai objek pengamatan untuk menjawab pertanyaan analitis.

CRISP-DM: Metodologi untuk mengorganisasikan proyek data mining ke dalam fase-fase kerja, mulai dari pemahaman bisnis hingga deployment. Fase-fasenya dapat ditinjau kembali sesuai kebutuhan proyek.

Data mining: Penerapan teknik komputasional untuk menemukan pola atau membentuk model dari data. Kegiatan ini merupakan bagian dari proses KDD yang lebih luas.

Deployment: Fase penerapan hasil proyek data mining agar dapat digunakan dalam konteks operasional.

Fitur: Representasi atau karakteristik data yang digunakan dalam analisis atau pemodelan. Fitur dapat dibentuk melalui transformasi data, misalnya dengan merangkum transaksi.

Interpretasi dan evaluasi: Tahap untuk memeriksa makna, kualitas, dan kegunaan temuan dalam konteks masalah. Hasil pemeriksaan dapat mendorong peninjauan kembali tahap sebelumnya.

Klasterisasi: Teknik pengelompokan data berdasarkan kemiripan karakteristik. Kelompok yang dihasilkan perlu ditafsirkan dan dinilai kestabilannya sebelum digunakan.

Knowledge Discovery in Databases (KDD): Proses sistematis untuk menemukan pengetahuan yang valid, baru, berguna, dan dapat dipahami melalui pengolahan serta penafsiran data.

Pemahaman bisnis: Fase CRISP-DM untuk mengenali kebutuhan organisasi dan tujuan penggunaan hasil analisis.

Pemahaman data: Fase CRISP-DM untuk menelaah data yang tersedia, termasuk kesesuaiannya dengan kebutuhan proyek.

Pemilihan data: Tahap penetapan sumber dan bagian data yang relevan dengan pertanyaan analitik.

Penambangan data: Tahap penggunaan algoritma untuk menemukan pola atau membangun model dari data yang telah disiapkan.

Prapemrosesan: Tahap pengolahan awal untuk mengurangi gangguan pada data, seperti nilai yang tidak konsisten atau tidak lengkap.

Transformasi: Tahap penyesuaian representasi data agar sesuai dengan kebutuhan analisis, termasuk pembentukan fitur turunan.

Atribut kategorikal: Atribut yang nilainya berupa kategori atau label, bukan ukuran numerik. Nilai hilangnya

dapat ditangani dengan modus atau kategori khusus jika sesuai dengan maknanya.

Atribut numerik: Atribut yang nilainya dinyatakan sebagai angka dan dapat dianalisis secara kuantitatif. Penanganan nilai hilangnya dapat mempertimbangkan distribusi serta keberadaan nilai ekstrem.

Data cleaning: Tahap prapemrosesan untuk memperbaiki atau menandai nilai tidak valid, format yang tidak seragam, dan informasi yang tidak lengkap. Keputusan pembersihan perlu mempertimbangkan konteks pencatatan dan tujuan analisis.

Deret waktu: Data yang dicatat secara berurutan menurut waktu. Urutan dan konteks temporal perlu dipertahankan saat data dibersihkan agar pola penting tidak hilang.

Duplikasi: Pencatatan suatu kejadian atau entitas lebih dari sekali. Penentuannya memerlukan aturan identitas karena kemiripan beberapa atribut belum tentu menunjukkan catatan ganda.

Imputasi: Penggantian nilai hilang dengan estimasi tertentu, misalnya median untuk atribut numerik atau modus untuk atribut kategorikal. Metode yang digunakan perlu dicatat agar dampaknya dapat dievaluasi.

Inkonsistensi data: Ketidakseragaman atau ketidaksesuaian dalam data, seperti perbedaan format atau ejaan. Pemeriksaannya membantu memastikan nilai dan catatan memiliki makna yang selaras.

Interpolasi: Perkiraan nilai di antara pengamatan berdasarkan nilai di sekitarnya. Penggunaannya pada data berurutan menurut waktu bergantung pada kewajaran asumsi kesinambungan pengukuran.

Median: Nilai tengah setelah data diurutkan. Ukuran ini lebih tahan terhadap kemencengan distribusi dan nilai ekstrem dibandingkan rata-rata.

Modus: Nilai atau kategori yang paling sering muncul dalam kumpulan data. Modus dapat digunakan untuk mengisi nilai hilang pada atribut kategorikal.

Noise: Variasi atau kesalahan pada data yang tidak merepresentasikan sinyal yang hendak dianalisis. Noise dapat berasal dari gangguan alat, pembulatan, kesalahan input, atau perubahan kondisi pencatatan.

Outlier: Pengamatan yang tampak jauh dari sebagian besar data menurut ukuran tertentu. Nilai tersebut dapat berupa kesalahan atau kejadian nyata, sehingga perlu diperiksa sebelum dihapus.

Rentang antarkuartil: Selisih antara kuartil atas dan kuartil bawah dalam distribusi data. Ukuran ini dapat membantu menandai nilai yang jauh dari bagian tengah distribusi.

Nilai hilang: Nilai yang tidak tercatat atau tidak tersedia untuk suatu atribut. Penyebab dan polanya perlu diperiksa sebelum memilih penghapusan, imputasi, atau penandaan.

Algoritma: Prosedur komputasional yang digunakan untuk mengolah data atau membangun model guna menyelesaikan tugas analitis tertentu.

Evaluasi: Penilaian terhadap kinerja atau kegunaan model berdasarkan ukuran yang sesuai dengan tujuan analisis.

Fitur: Atribut atau variabel masukan yang digunakan model untuk mengenali pola dan menghasilkan keluaran.

Klasterisasi: Proses mengelompokkan observasi berdasarkan kemiripan fitur tanpa menggunakan label sebagai jawaban acuan.

Kumpulan data: Sekumpulan observasi beserta atributnya yang digunakan dalam analisis atau pemodelan.

Label: Informasi yang menunjukkan keluaran atau kategori yang diharapkan untuk suatu observasi.

Model: Representasi pola yang dipelajari dari data dan digunakan untuk menghasilkan keluaran pada kasus baru.

Paradigma pembelajaran: Kerangka yang menggambarkan jenis pengalaman atau data yang digunakan model untuk belajar.

Pembelajaran mesin: Pendekatan komputasional yang memungkinkan model menyesuaikan perilakunya berdasarkan pengalaman berupa data.

Pembelajaran semi-terawasi: Pendekatan yang memanfaatkan sejumlah kecil contoh berlabel bersama data yang belum berlabel.

Pembelajaran tak terawasi: Pendekatan yang mencari struktur atau pola dalam data tanpa menggunakan label sebagai acuan.

Pembelajaran terawasi: Pendekatan yang mempelajari hubungan antara fitur masukan dan label yang tersedia.

Prediksi: Keluaran model berupa perkiraan untuk suatu nilai atau keadaan, termasuk pada kasus yang belum diamati.

Rekayasa fitur: Proses menyiapkan atau membentuk atribut data agar lebih sesuai untuk digunakan dalam pemodelan.

Ambang keputusan: Batas nilai yang digunakan untuk mengubah keluaran model menjadi keputusan kelas. Penetapannya perlu mempertimbangkan konsekuensi tiap jenis kesalahan.

Data pelatihan: Bagian data berisi fitur dan label kelas yang digunakan algoritma untuk mempelajari hubungan dan membentuk model.

Data pengujian: Bagian data yang disisihkan untuk mengukur kinerja akhir model setelah pengembangan, tanpa digunakan dalam proses pelatihan.

Data validasi: Bagian data yang digunakan untuk memilih konfigurasi model selama pengembangan, sebelum kinerja akhir diukur pada data pengujian.

Decision Tree: Model klasifikasi berbentuk pohon yang menetapkan kelas melalui pengujian kondisi pada fitur secara bertahap.

Entropi: Ukuran ketidakmurnian simpul berdasarkan proporsi kelas di dalamnya. Nilai ini dapat digunakan untuk menilai kualitas pemisahan data.

Fitur: Karakteristik atau variabel yang digunakan model untuk membedakan kelas objek.

Information gain: Ukuran penurunan entropi yang dihasilkan oleh pemisahan data berdasarkan suatu fitur.

Kebocoran informasi: Keadaan ketika informasi yang seharusnya tidak tersedia saat prediksi masuk ke proses pelatihan atau evaluasi, sehingga penilaian kinerja model menjadi tidak mencerminkan penggunaan sebenarnya.

Kelas: Kategori sasaran yang ditetapkan untuk objek, seperti fraud atau bukan fraud.

Klasifikasi: Tugas pembelajaran terawasi yang menggunakan fitur untuk menetapkan objek ke dalam kategori atau kelas.

Model klasifikasi: Hasil pembelajaran algoritma dari data berlabel yang digunakan untuk menetapkan kelas bagi objek baru.

Pembagian berstrata: Cara membagi data dengan menjaga proporsi kelas pada setiap bagian, terutama ketika distribusi kelas tidak seimbang.

Prapemrosesan data: Persiapan dan transformasi data sebelum digunakan dalam pelatihan model, termasuk penanganan data hilang dan penentuan skala.

Random Forest: Model klasifikasi yang menggabungkan banyak pohon yang dilatih dengan variasi sampel dan fitur, lalu menggabungkan keluarannya.

Fitur: Variabel yang digunakan sebagai masukan model untuk memperkirakan nilai respons. Fitur dapat berasal dari data mentah atau dibentuk melalui rekayasa fitur.

Intersep: Nilai prediksi respons ketika seluruh prediktor dalam model bernilai nol. Dalam regresi linear sederhana, intersep dinyatakan dengan β_0.

Koefisien: Parameter model yang menggambarkan hubungan antara prediktor dan respons. Dalam regresi linear, koefisien menunjukkan perubahan nilai prediksi yang berkaitan dengan perubahan pada prediktor.

Pembelajaran terawasi: Pendekatan pembelajaran mesin yang menggunakan contoh berlabel, yaitu data yang telah memiliki nilai target, untuk mempelajari hubungan antara fitur dan respons.

Predictive analytics: Rangkaian kegiatan yang menggunakan data, metode statistik, dan model komputasional untuk menghasilkan perkiraan bagi pengambilan keputusan.

Regresi: Pendekatan pemodelan untuk mengestimasi hubungan antara satu atau lebih variabel prediktor dan variabel respons numerik, lalu memperkirakan respons pada observasi baru.

Regresi linear: Metode regresi yang memodelkan respons sebagai kombinasi linear dari satu atau lebih fitur.

Regresi linear berganda: Regresi linear yang menggunakan beberapa fitur secara simultan untuk memperkirakan respons.

Residual: Selisih antara nilai respons aktual dan nilai respons yang diprediksi oleh model.

Respons: Nilai target yang ingin diperkirakan oleh model regresi.

Variabel prediktor: Variabel masukan yang digunakan model untuk memperkirakan nilai respons.

Variabel respons: Variabel numerik yang menjadi sasaran estimasi dalam analisis regresi.

Centroid: Titik pusat yang mewakili suatu kelompok dalam K-Means, dihitung dari rata-rata fitur objek anggotanya.

Data kategorikal: Data yang nilainya menyatakan kategori, sehingga kemiripannya perlu dinilai dengan ukuran yang sesuai dengan bentuk representasinya.

Data numerik: Data yang dinyatakan dalam nilai angka dan dapat dibandingkan menggunakan ukuran jarak.

Fitur: Karakteristik atau atribut yang digunakan untuk merepresentasikan objek dalam ruang data.

Hierarkis: Pendekatan klasterisasi yang membangun kelompok dalam susunan bertingkat.

Jarak Euclidean: Ukuran jarak garis lurus antara dua titik dalam ruang fitur, yang peka terhadap perbedaan skala fitur.

Jarak Manhattan: Ukuran jarak antara dua titik yang dihitung dari jumlah selisih absolut pada setiap dimensinya.

K-Means: Algoritma berbasis partisi yang mengelompokkan objek ke dalam sejumlah kelompok dengan menempatkannya pada centroid terdekat.

Klasterisasi: Pendekatan pembelajaran tak terawasi untuk mengelompokkan objek berdasarkan kemiripan tanpa label kelas yang telah ditentukan.

Noise: Titik terpencil yang tidak tergabung secara jelas dalam wilayah kelompok, terutama pada pendekatan berbasis kepadatan.

Normalisasi: Penskalaan nilai fitur agar rentangnya sebanding dan tidak mendominasi perhitungan jarak.

Pembelajaran tak terawasi: Pendekatan pemodelan yang menemukan pola atau struktur pada data tanpa menggunakan label kelas yang telah ditentukan.

Ruang fitur: Representasi data tempat setiap objek digambarkan sebagai titik berdasarkan nilai fitur-fiturnya.

Segmentasi: Pembagian objek ke dalam kelompok yang memiliki karakteristik atau pola tertentu untuk mendukung analisis.

Standardisasi: Transformasi skala fitur agar nilainya dapat dibandingkan secara lebih seimbang dalam pemodelan.

Ukuran jarak: Metode untuk mengukur perbedaan atau kedekatan antara objek berdasarkan representasi fiturnya.

Anteseden: Kelompok item di sisi kiri aturan asosiasi, yang menjadi kondisi bagi kemunculan konsekuen.

Apriori: Algoritma untuk menemukan frequent pattern dengan memanfaatkan sifat bahwa setiap subset dari itemset frequent juga frequent.

Association rule: Aturan yang menyatakan keterkaitan bersyarat antara anteseden dan konsekuen tanpa menyiratkan hubungan sebab-akibat.

Confidence: Proporsi transaksi yang memuat anteseden dan konsekuen di antara seluruh transaksi yang memuat anteseden.

Frequent pattern: Pola item yang muncul dalam data dengan frekuensi atau proporsi sekurang-kurangnya sebesar ambang yang ditentukan.

Item: Elemen yang dicatat dalam suatu transaksi dan dianalisis sebagai bagian dari pola.

Itemset: Gabungan satu atau lebih item dalam kumpulan data transaksi.

Konsekuen: Kelompok item di sisi kanan aturan asosiasi yang dikaitkan dengan kemunculan anteseden.

Lift: Ukuran yang membandingkan confidence suatu aturan dengan support konsekuennya untuk menilai kekuatan asosiasi relatif terhadap frekuensi dasar.

Minimum support: Ambang frekuensi atau proporsi transaksi yang harus dipenuhi itemset agar dianggap frequent.

Support: Proporsi seluruh transaksi yang memuat itemset tertentu.

Transaksi: Satuan catatan yang memuat kumpulan item, seperti satu nota penjualan atau satu kunjungan, sesuai definisi unit analisis.

Akuisisi data sensor: Proses mengambil pengukuran dari sensor, mencatat identitas dan waktu pengukuran, lalu mengirimkan data ke media penerimaan.

Aktuator: Komponen yang menerima perintah untuk menghasilkan respons atau tindakan pada lingkungan fisik.

Arsitektur IoT: Susunan lapisan dan komponen yang menghubungkan perangkat fisik, jaringan komunikasi, platform pemrosesan, dan aplikasi.

Data sensor: Data digital yang dihasilkan dari pengukuran kondisi fisik oleh sensor.

Frekuensi sampling: Tingkat keseringan pengambilan pengukuran sensor dalam suatu rentang waktu.

Gerbang IoT: Komponen perantara yang menerima atau meneruskan data antara perangkat IoT dan platform.

Internet untuk Segala (IoT): Sistem yang menghubungkan objek fisik melalui jaringan agar data dapat dikumpulkan, diproses, dan dimanfaatkan oleh aplikasi.

Konektivitas: Kemampuan perangkat dan komponen sistem untuk saling terhubung dan mengirimkan data melalui jaringan.

Penanda waktu: Informasi yang mencatat waktu suatu pengukuran dilakukan, sehingga data dapat diurutkan atau digabungkan.

Prapemrosesan data: Tahap menyiapkan data sebelum analisis, termasuk pemeriksaan dan penanganan mutu data.

Sensor: Perangkat yang mengukur kondisi fisik dan mengubah hasil pengukuran menjadi data digital.

Validasi data: Pemeriksaan untuk memastikan data memenuhi kriteria mutu dan konsistensi sebelum digunakan.

Aliran komunikasi: Ringkasan pertukaran data antarperangkat dalam jaringan, yang dapat mencakup jumlah koneksi, durasi, dan volume data.

Analisis log: Pemeriksaan catatan peristiwa yang dihasilkan sistem operasi, aplikasi, perangkat keamanan, atau layanan autentikasi untuk memahami aktivitas sistem.

Autentikasi: Proses memverifikasi identitas pengguna atau entitas sebelum memberikan akses ke sistem.

Data endpoint: Catatan aktivitas pada perangkat akhir, seperti komputer atau perangkat lain yang terhubung ke sistem.

Firewall: Perangkat atau sistem keamanan yang mengatur lalu lintas jaringan berdasarkan aturan tertentu.

Fitur: Atribut atau ukuran yang dibentuk dari data mentah untuk menggambarkan karakteristik aktivitas dalam analisis atau pemodelan.

Jendela waktu: Interval waktu yang digunakan untuk merangkum atau mengukur aktivitas, seperti jumlah kegagalan masuk.

Klasifikasi: Proses menempatkan data atau kejadian ke dalam kategori yang telah ditentukan berdasarkan karakteristiknya.

Log aplikasi: Catatan peristiwa yang merekam aktivitas atau tindakan dalam suatu aplikasi.

Lalu lintas jaringan: Komunikasi data antarperangkat yang dapat diamati sebagai paket atau ringkasan aliran komunikasi.

Penambangan data: Penerapan teknik analisis terhadap kumpulan data untuk menemukan pola, hubungan, atau penyimpangan yang berguna.

Prapemrosesan data: Tahap menyiapkan data agar lebih konsisten untuk dianalisis, misalnya dengan menormalkan format dan menangani nilai kosong.

Rekayasa fitur: Proses membentuk atau memilih atribut dari data mentah agar karakteristik yang relevan dapat digunakan dalam analisis.

Anomali: Pola yang berbeda dari perilaku yang diharapkan dalam konteks tertentu. Penyimpangan tersebut dapat menandakan kesalahan, kejadian langka yang sah, atau gangguan.

Anomali kolektif: Pola tidak biasa yang terbentuk dari sekumpulan pengamatan, meskipun setiap nilai secara individual tampak normal.

Anomali kontekstual: Penyimpangan yang hanya dianggap abnormal dalam kondisi tertentu, seperti suhu yang tidak wajar untuk ruang penyimpanan dingin.

Anomali titik: Satu pengamatan yang menyimpang dari pengamatan lain, misalnya pembacaan suhu yang jauh di luar rentang operasi.

Data runtun waktu: Data yang pengamatannya tersusun menurut urutan waktu. Analisisnya perlu memperhatikan urutan dan konteks, bukan hanya besar-kecilnya nilai.

Deteksi anomali: Proses mengenali atau memberi skor pada pengamatan yang menyimpang dari pola perilaku normal untuk ditindaklanjuti.

Jarak Euclidean: Ukuran jarak antara titik-titik berdasarkan perbedaan nilai pada fitur numerik. Perbedaan skala fitur dapat memengaruhi hasil perhitungannya.

Kepadatan lingkungan lokal: Tingkat kepadatan pengamatan di sekitar suatu titik dibandingkan dengan lingkungan sekitarnya; kepadatan yang tidak biasa dapat menjadi petunjuk anomali.

Outlier: Objek atau pengamatan yang terletak jauh dari sebagian besar data menurut ukuran tertentu. Nilai ekstrem tidak selalu merupakan anomali operasional.

Pembelajaran tak terawasi: Pendekatan pembelajaran yang mencari struktur atau pola dalam data tanpa menggunakan label kejadian.

Pencilan: Pengamatan yang menyimpang dari kumpulan data, termasuk yang muncul akibat kesalahan pencatatan.

Simpangan baku: Ukuran sebaran nilai terhadap rata-rata yang digunakan, antara lain, untuk menghitung skor baku.

Skor penyimpangan: Nilai yang menunjukkan tingkat ketidaklaziman suatu pengamatan dan dapat digunakan untuk memeringkat kasus yang perlu diperiksa.

Standardisasi: Penyesuaian skala fitur agar perbedaan satuan atau rentang nilainya tidak mendominasi perhitungan jarak.

Bagging: Strategi ensemble yang melatih beberapa model secara terpisah pada sampel bootstrap dari data asal, lalu menggabungkan prediksinya. Penggabungan dapat dilakukan melalui suara terbanyak atau perataan keluaran.

Boosting: Strategi ensemble yang melatih model secara berurutan, dengan setiap model berikutnya berfokus pada kesalahan atau pola yang belum tertangani oleh model sebelumnya.

Ensemble learning: Pendekatan yang menggabungkan keluaran beberapa model untuk menghasilkan prediksi. Kinerjanya dipengaruhi oleh kemampuan tiap model dan perbedaan kesalahan antarmodel.

Fungsi kerugian: Ukuran kesalahan prediksi yang digunakan untuk mengarahkan pembaruan model. Dalam gradient boosting, model diperbarui ke arah yang mengurangi nilai fungsi ini.

Gradient Boosting: Metode boosting yang membangun pohon secara bertahap dengan memperbaiki

kesalahan prediksi kumpulan pohon yang telah dibangun.

Laju pembelajaran: Parameter yang mengatur besarnya kontribusi setiap pembaruan model dalam proses pembelajaran bertahap.

Random Forest: Metode ensemble yang melatih banyak pohon keputusan pada sampel bootstrap dan membatasi fitur yang dipertimbangkan pada setiap pemisahan simpul. Prediksi antarpohon digabungkan melalui suara terbanyak atau perataan.

Sampel bootstrap: Kumpulan data pelatihan yang dibentuk dengan mengambil sampel dari data asal menggunakan pengembalian.

Suara terbanyak: Cara menggabungkan prediksi model klasifikasi dengan memilih kelas yang paling banyak dipilih oleh model-model penyusun.

Validasi: Proses menguji konfigurasi model pada data yang tidak digunakan untuk melatihnya, guna membantu menilai kinerja dan menyesuaikan parameter.

XGBoost: Implementasi gradient boosting yang menyediakan pengaturan kompleksitas model dan dukungan komputasi untuk pemodelan berskala besar.

CRISP-DM: Kerangka kerja untuk mengatur tahapan proyek penambangan data, dari pemahaman masalah hingga penerapan hasil.

F1-score: Ukuran kinerja klasifikasi yang merangkum keseimbangan antara precision dan recall.

Fitur: Variabel masukan yang digunakan model untuk menghasilkan prediksi. Pemilihannya harus mempertimbangkan makna data dan waktu ketersediaan informasi.

Imputasi: Pengisian nilai yang hilang dalam data, misalnya dengan nilai tertentu yang dipelajari dari data pelatihan.

KDD: Proses penemuan pengetahuan dari data yang mencakup tahapan pengolahan hingga interpretasi hasil.

Kebocoran target: Kondisi ketika informasi yang seharusnya tidak tersedia saat prediksi masuk ke dalam fitur, sehingga evaluasi model dapat menjadi tidak realistis.

Machine learning: Pendekatan komputasi yang memungkinkan model mempelajari pola dari data untuk mendukung tugas seperti klasifikasi atau prediksi.

Pembelajaran mesin: Proses membangun model yang mempelajari pola dari data dan menggunakannya untuk menghasilkan keluaran.

Pipeline: Rangkaian langkah pemrosesan dan pemodelan yang dijalankan secara terpadu, sehingga transformasi dapat dipelajari dari data pelatihan.

Prapemrosesan data: Tahap menyiapkan data sebelum pemodelan, termasuk pembersihan dan transformasi.

Regresi logistik: Algoritma klasifikasi yang memperkirakan peluang suatu observasi termasuk ke dalam kelas tertentu.

Standardisasi: Transformasi fitur numerik agar berada pada skala yang sebanding.

Stratifikasi: Cara membagi data dengan menjaga proporsi kelas agar relatif serupa pada setiap bagian data.