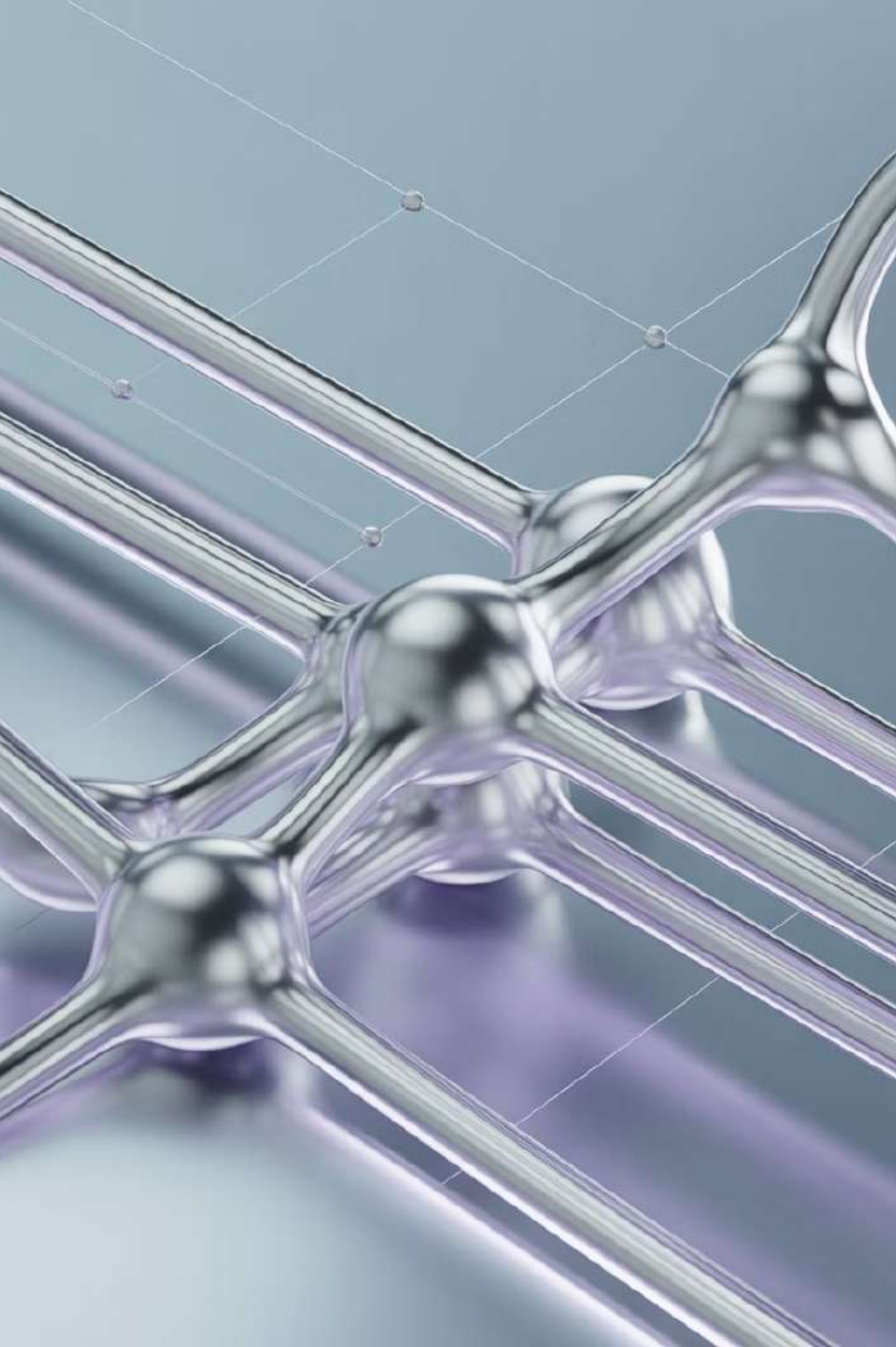




Data Mining untuk Business Intelligence

Memahami metode-metode data mining: Teknik asosiasi, klasifikasi, dan clustering untuk mengoptimalkan keputusan bisnis berbasis data

Tujuan Pembelajaran

1

Memahami Konsep Data Mining

Menjelaskan definisi, tujuan, dan peran data mining dalam konteks Business Intelligence serta bagaimana data mining membantu organisasi mengekstrak pengetahuan dari data besar

2

Menguasai Teknik Asosiasi

Mengidentifikasi pola hubungan antar item dalam dataset menggunakan algoritma seperti Apriori dan FP-Growth untuk analisis keranjang belanja

3

Menerapkan Metode Klasifikasi

Membangun model prediktif untuk mengkategorikan data ke dalam kelas-kelas tertentu menggunakan algoritma Decision Tree, Naive Bayes, dan lainnya

4

Mengimplementasikan Clustering

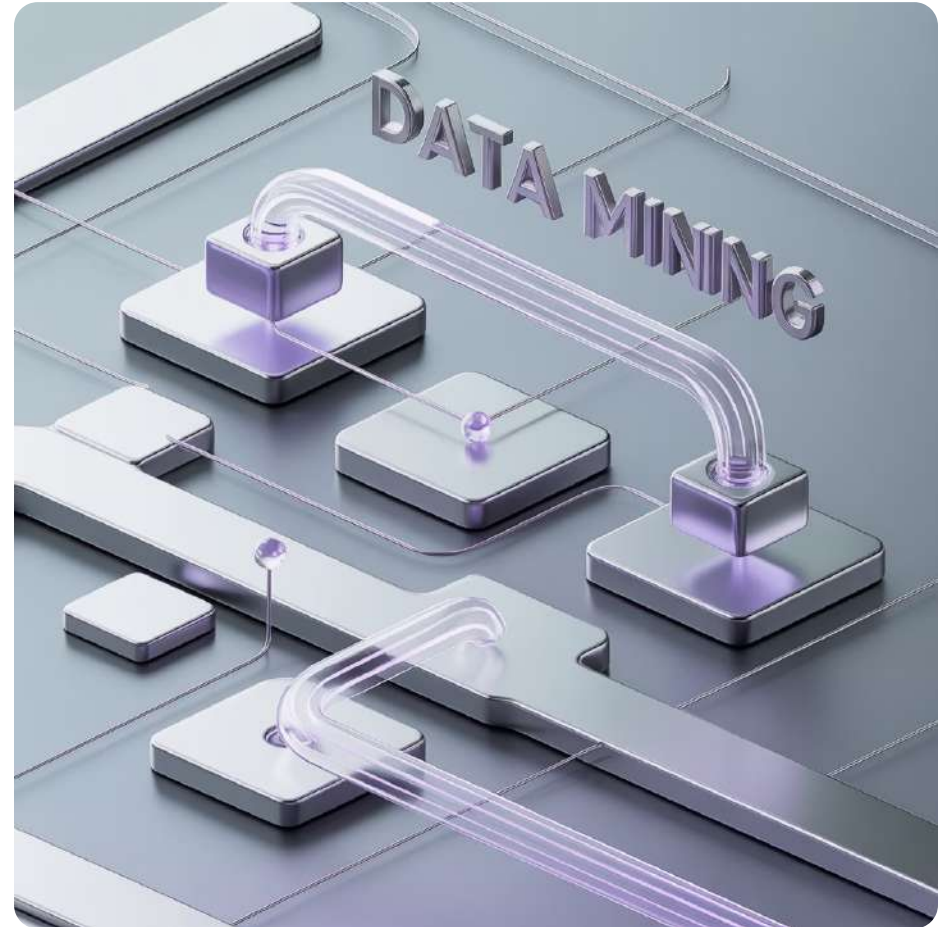
Mengelompokkan data berdasarkan kesamaan karakteristik tanpa label yang telah ditentukan sebelumnya menggunakan K-Means dan metode hierarchical clustering

Apa Itu Data Mining?

Data mining adalah proses eksplorasi dan analisis data dalam jumlah besar untuk menemukan pola, korelasi, dan insight yang bermakna. Proses ini menggunakan metode statistik, machine learning, dan kecerdasan buatan untuk mengubah data mentah menjadi informasi yang dapat ditindaklanjuti.

Dalam konteks Business Intelligence, data mining memungkinkan organisasi untuk membuat keputusan strategis berdasarkan bukti empiris, memprediksi tren masa depan, mengidentifikasi peluang bisnis baru, dan mengoptimalkan operasi.

Data mining berbeda dari query database tradisional karena fokusnya pada penemuan pola tersembunyi yang tidak diketahui sebelumnya, bukan hanya mengambil informasi yang sudah ada.



Proses Knowledge Discovery in Database (KDD)



Data Selection

Memilih dan mengumpulkan data yang relevan dari berbagai sumber untuk analisis



Preprocessing

Membersihkan data dari noise, missing values, dan inkonsistensi



Transformation

Mengubah data ke format yang sesuai untuk proses mining



Data Mining

Menerapkan algoritma untuk menemukan pola dan pengetahuan



Evaluation

Mengevaluasi dan menginterpretasi hasil temuan untuk pengambilan keputusan

Tiga Teknik Utama Data Mining

Asosiasi

Menemukan hubungan "jika-maka" antar item dalam dataset

- Market basket analysis
- Cross-selling strategy
- Product recommendation

Klasifikasi

Memprediksi kategori atau kelas dari data baru berdasarkan pola historis

- Customer segmentation
- Credit risk assessment
- Fraud detection

Clustering

Mengelompokkan data berdasarkan kesamaan tanpa label yang telah ditentukan

- Customer profiling
- Anomaly detection
- Image segmentation

Teknik Asosiasi: Market Basket Analysis

Teknik asosiasi mengidentifikasi hubungan antar item yang sering muncul bersamaan dalam transaksi. Metode ini sangat berguna untuk memahami perilaku pembelian pelanggan dan mengoptimalkan strategi penjualan.



Support

Frekuensi kemunculan itemset dalam seluruh transaksi. Contoh: 30% transaksi mengandung {Roti, Susu}



Confidence

Probabilitas item Y dibeli jika item X dibeli. Contoh: 80% pembeli roti juga membeli susu



Lift

Kekuatan asosiasi dibanding kejadian acak. Lift > 1 menunjukkan asosiasi positif yang kuat



Algoritma Apriori untuk Asosiasi

Cara Kerja Algoritma

1. Identifikasi frequent itemsets dengan support minimum
2. Generate candidate itemsets dari frequent itemsets
3. Prune candidates yang tidak memenuhi minimum support
4. Ulangi hingga tidak ada candidate baru
5. Generate association rules dengan confidence minimum

📌 **Prinsip Apriori:** Jika suatu itemset jarang muncul, maka semua supersetnya juga akan jarang muncul. Prinsip ini mengurangi ruang pencarian secara signifikan.

Contoh Aturan Asosiasi

{Roti, Mentega} → {Susu}

Support: 25%

Confidence: 75%

Lift: 2.5

{Laptop} → {Mouse, Mousepad}

Support: 40%

Confidence: 85%

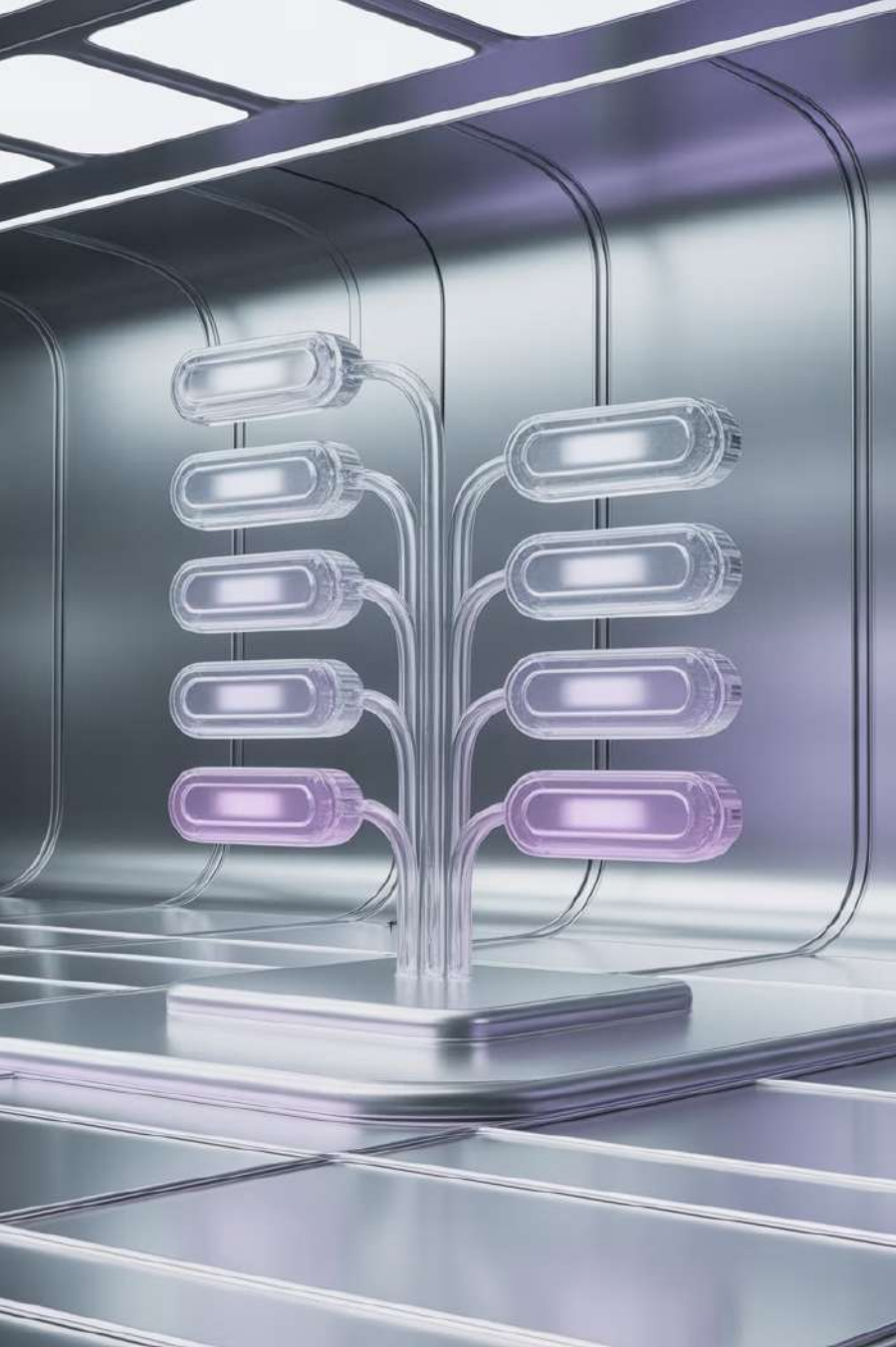
Lift: 3.2

{Popok Bayi} → {Bir}

Support: 15%

Confidence: 60%

Lift: 1.8



Teknik Klasifikasi: Memprediksi Kategori

Klasifikasi adalah teknik supervised learning yang membangun model prediktif berdasarkan data training yang sudah dilabeli. Model ini kemudian digunakan untuk memprediksi kelas dari data baru yang belum diketahui kategorinya.

01

Training Phase

Model belajar dari historical data yang sudah memiliki label kelas untuk memahami pola dan karakteristik setiap kategori

02

Validation

Model diuji dengan validation set untuk mengukur akurasi dan menghindari overfitting pada data training

03

Prediction

Model yang sudah divalidasi digunakan untuk memprediksi kelas dari data baru dalam aplikasi nyata

Algoritma Klasifikasi Populer



Decision Tree

Membuat pohon keputusan berdasarkan atribut yang paling informatif. Mudah diinterpretasi dan visualisasi, namun rentan terhadap overfitting jika tidak dipruning dengan baik.



Naive Bayes

Menggunakan teorema Bayes dengan asumsi independensi antar atribut. Sangat efisien untuk dataset besar dan efektif untuk text classification seperti spam filtering.



Neural Network

Meniru cara kerja otak manusia dengan layers of neurons. Sangat powerful untuk pola kompleks tetapi membutuhkan data training yang besar dan computational resources tinggi.



Support Vector Machine

Mencari hyperplane optimal yang memisahkan kelas dengan margin maksimum. Efektif untuk high-dimensional data dan robust terhadap outliers.



K-Nearest Neighbors

Mengklasifikasi berdasarkan mayoritas kelas dari K tetangga terdekat. Sederhana dan tidak memerlukan training phase, namun komputasi intensive saat prediksi.



Random Forest

Ensemble method yang mengkombinasikan banyak decision trees. Lebih akurat dan robust dibanding single tree karena mengurangi variance melalui averaging.

Evaluasi Model Klasifikasi

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Metrik Evaluasi

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Proporsi prediksi yang benar dari total data

Precision

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

Proporsi prediksi positif yang benar

Recall (Sensitivity)

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

Proporsi actual positif yang terdeteksi

F1-Score

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

Harmonic mean dari precision dan recall

Teknik Clustering: Pengelompokan Data

Clustering adalah teknik unsupervised learning yang mengelompokkan data berdasarkan kesamaan karakteristik tanpa memerlukan label yang telah ditentukan sebelumnya. Teknik ini sangat berguna untuk exploratory data analysis dan menemukan struktur tersembunyi dalam data.



Customer Segmentation

Mengelompokkan pelanggan berdasarkan demografi, perilaku pembelian, dan preferensi untuk strategi marketing yang lebih targeted



Anomaly Detection

Mengidentifikasi outliers atau data yang tidak normal, berguna untuk fraud detection dan quality control



Image Segmentation

Mempartisi gambar menjadi region-region berdasarkan warna, texture, atau karakteristik visual lainnya



Document Clustering

Mengelompokkan dokumen atau artikel berdasarkan topik dan konten untuk sistem rekomendasi dan search engines

Algoritma K-Means Clustering

Langkah-Langkah K-Means



Menentukan Jumlah Cluster Optimal

- **Elbow Method:** Plot SSE (Sum of Squared Errors) vs K, cari "siku" di grafik
- **Silhouette Score:** Ukur seberapa mirip object dengan clusternya vs cluster lain
- **Gap Statistic:** Bandingkan total within-cluster variation dengan ekspektasi null reference
- **Domain Knowledge:** Pertimbangkan context bisnis dan interpretability

Kesimpulan: Kekuatan Data Mining untuk BI



Data mining mengubah data menjadi actionable insights yang mendorong keputusan bisnis yang lebih cerdas. Dengan menguasai ketiga teknik ini, organisasi dapat mengoptimalkan operasi, meningkatkan customer satisfaction, dan mencapai competitive advantage di era digital.

"Data mining adalah seni mengubah data mentah menjadi emas informasi yang bernilai bagi organisasi."