UNIVERSITAS PAHLAWAN

MODUL AJAR
DATA MAINING

Disusun Oleh:
Kasini S.Kom,M.kom

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah SWT atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan modul pembelajaran Data Mining ini. Modul ini disusun dengan tujuan untuk menjadi panduan bagi mahasiswa dalam memahami konsep-konsep fundamental dalam mata kuliah Data Mining. Diharapkan modul ini dapat mempermudah proses belajar dan mengajar di lingkungan akademik.

Pada kesempatan ini, penulis ingin mengucapkan terima kasih yang tulus kepada semua pihak yang telah memberikan dukungan, baik secara moril maupun materil, terutama kepada rekan-rekan dosen dan semua yang terlibat yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari sepenuhnya bahwa modul ini masih jauh dari kata sempurna. Oleh karena itu, segala bentuk kritik dan saran yang membangun akan penulis terima dengan lapang dada demi perbaikan di masa mendatang.

Akhir kata, besar harapan penulis semoga modul ini dapat memberikan manfaat yang nyata dan menjadi sumber belajar yang mudah dipahami bagi para mahasiswa serta pembaca yang budiman.

Bangkinang, 6 Agustus 2019

Kasini S.Kom., M.Kom

DAFTAR ISI

BAB I

PENGANTAR DATA MINING

1.1 Definisi dan Latar Belakang Data Mining

Seiring dengan pesatnya perkembangan teknologi informasi, data yang dihasilkan dan disimpan dalam berbagai basis data tumbuh secara eksponensial. Basis data ini dapat diibaratkan sebagai sebuah gudang data raksasa. Tumpukan data yang sangat besar ini memunculkan sebuah pertanyaan penting: "pengetahuan atau informasi berharga apa yang dapat kita peroleh dari tumpukan data tersebut?".Untuk menjawab tantangan ini, lahirlah Data Mining.

Secara sederhana, Data Mining adalah proses mengekstrak atau "menggali" pengetahuan yang bermanfaat dari sekumpulan data bervolume besar, di mana pengetahuan tersebut sebelumnya tidak diketahui secara manual. Menurut Tacbir, Data Mining didefinisikan sebagai "proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan

machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari database yang besar". Tujuan utamanya adalah menemukan pola-pola tersembunyi yang valid dan dapat ditindaklanjuti.

1.2 Evolusi dan Sejarah Data Mining

Meskipun istilah "Data Mining" populer dalam beberapa dekade terakhir, akarnya berasal dari bidang ilmu yang sudah matang lebih dulu. Data Mining bukanlah sebuah disiplin ilmu yang berdiri sendiri, melainkan sebuah integrasi dari berbagai teknik dan pendekatan.

Disiplin ilmu yang menjadi fondasi utama Data Mining antara lain:

- Statistik
- Kecerdasan Buatan (*Artificial Intelligence*)
- *Machine Learning*
- Teknologi Basis Data dan *Data Warehouse*
- Pengenalan Pola (*Pattern Recognition*)
- *Neural Network*

1.3 Perbedaan Data Mining dengan Statistika

Walaupun Data Mining banyak memanfaatkan teknik statistik, keduanya memiliki pendekatan yang berbeda. Analisis statistik tradisional biasanya bersifat **verifikatif**, di mana analis memulai dengan sebuah hipotesis lalu menguji kebenarannya menggunakan data.

Sebaliknya, Data Mining sering kali menggunakan pendekatan berbasis penemuan (**discovery-based**). Dalam pendekatan ini, algoritma digunakan untuk menjelajahi data secara otomatis

untuk menemukan relasi-relasi kunci dan pola-pola yang sebelumnya tidak terduga, tanpa perlu diawali oleh sebuah hipotesis.

1.4 Manfaat dan Aplikasi di Berbagai Industri

Dengan kemampuannya menemukan pola tersembunyi, Data Mining memberikan manfaat signifikan di banyak bidang seperti manajemen bisnis, pendidikan, dan kesehatan. Dua kemampuan utamanya adalah:

- Mengotomatisasi Prediksi Tren Bisnis: Data Mining dapat menganalisis data historis untuk memprediksi tren di masa depan.
- Mengotomatisasi Penemuan Pola Tersembunyi: Algoritma Data Mining dapat "menyapu" basis data untuk mengidentifikasi pola-pola yang tidak diketahui sebelumnya hanya dalam satu proses.

Salah satu contoh penerapan yang paling terkenal adalah analisis keranjang belanja (*market basket analysis*) di industri ritel. Dengan menganalisis data transaksi, sebuah supermarket dapat menemukan produk-produk yang sering dibeli bersamaan oleh pelanggan, misalnya roti dan susu. Pengetahuan ini dapat digunakan untuk mengatur tata letak produk atau merancang strategi promosi yang lebih efektif.

BAB II

PROSES *KNOWLEDGE DISCOVERY IN DATABASES* (KDD)

2.1 Definisi Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) adalah keseluruhan proses terorganisir untuk menemukan pengetahuan yang berguna dari sekumpulan data bervolume besar. Ini adalah sebuah kerangka kerja yang sistematis untuk mengidentifikasi pola-pola yang valid, bermanfaat, dan mudah dipahami dari dalam data yang besar dan kompleks.Hubungan antara KDD dan Data Mining

Seringkali istilah KDD dan Data Mining digunakan secara bergantian, namun keduanya memiliki makna yang berbeda. KDD merujuk pada keseluruhan proses penemuan pengetahuan, mulai dari pengumpulan data mentah hingga menjadi informasi yang bisa digunakan.Sementara itu, Data Mining adalah salah satu langkah inti di dalam proses KDD. Langkah inilah yang berfokus pada penerapan algoritma cerdas untuk menjelajahi data, mengembangkan model, dan menemukan pola-pola yang sebelumnya tidak diketahui. Jadi, Data Mining adalah jantung dari proses KDD, tetapi bukan keseluruhan proses itu sendiri.

2.2. Tahapan Lengkap dalam Proses KDD

Proses KDD merupakan sebuah alur kerja yang terdiri dari beberapa tahapan berurutan. Proses ini mengubah data mentah menjadi sebuah pengetahuan yang siap pakai.

Berikut adalah tahapan-tahapannya:

1. **Seleksi Data (*Data Selection*)** Pada tahap ini, data yang relevan dengan tujuan analisis dipilih dan diambil dari kumpulan data yang lebih besar (database). Tidak semua data yang ada akan digunakan, hanya data target yang akan diproses lebih lanjut.
2. **Pembersihan & Pra-pemrosesan (*Data Cleaning & Preprocessing*)** Data yang telah dipilih kemudian dibersihkan. Proses ini mencakup penghapusan *noise* (data yang mengganggu), penanganan data yang tidak konsisten, serta penggabungan data dari berbagai sumber (*data integration*).
3. **Transformasi Data (*Data Transformation*)** Data yang sudah bersih kemudian diubah atau digabung ke dalam format yang sesuai untuk diproses oleh algoritma Data Mining.
4. **Proses Mining (*Data Mining*)** Ini adalah tahap inti di mana berbagai metode dan algoritma cerdas diterapkan untuk mengekstrak pola-pola atau aturan yang bermakna dan tersembunyi dari dalam data.
5. **Evaluasi & Interpretasi Pola (*Pattern Evaluation & Interpretation*)** Tidak semua pola yang ditemukan bermanfaat. Pada tahap ini, pola-pola menarik yang telah ditemukan diidentifikasi dan dievaluasi untuk memastikan validitas dan kegunaannya. Pola yang sudah dievaluasi kemudian ditafsirkan menjadi pengetahuan atau informasi baru yang dapat dipahami oleh manusia.

BAB III

DATA PREPROCESSING

3.1 Pentingnya Persiapan Data

Data yang ada di dunia nyata seringkali "kotor", artinya data tersebut bisa jadi tidak lengkap, mengandung *error* (*noise*), dan tidak konsisten. Proses Data Mining sangat sensitif terhadap kualitas data yang digunakan. Jika data yang dimasukkan berkualitas buruk, maka hasil yang didapatkan juga tidak akan akurat dan tidak dapat diandalkan. Prinsip "sampah masuk, sampah keluar" (*garbage in, garbage out*) sangat berlaku di sini.

Oleh karena itu, Data Preprocessing atau pra-pemrosesan data adalah sebuah tahap fundamental dan krusial dalam rangkaian proses KDD. Tujuannya adalah untuk membersihkan dan mempersiapkan data mentah agar menjadi data berkualitas tinggi yang siap untuk dianalisis.

3.3 Teknik Data Cleaning

Data Cleaning (pembersihan data) adalah proses untuk menghilangkan *noise* dan data yang tidak konsisten atau tidak relevan dari kumpulan data. Beberapa tugas utama dalam tahap ini meliputi:

- Menangani Data yang Hilang (*Missing Values*): Mengisi nilai yang kosong, misalnya dengan nilai rata-rata atau nilai yang paling sering muncul.
- Menghaluskan *Noisy Data*: Mengidentifikasi dan memperbaiki data yang mengandung kesalahan atau nilai-nilai pencilan (*outlier*) yang ekstrem.
- Mengatasi Inkonsistensi**:** Memperbaiki perbedaan dalam penulisan data, misalnya antara "JKT", "Jakarta", dan "DKI Jakarta" yang seharusnya merujuk pada hal yang sama.

3.4 Teknik Data Integration

Data Integration adalah proses penggabungan data dari berbagai sumber yang berbeda (misalnya dari beberapa database atau file) ke dalam satu penyimpanan data yang terpadu, seperti *data warehouse*. Tahap ini penting ketika informasi yang dibutuhkan untuk analisis tersebar di beberapa lokasi. Tantangan dalam integrasi data termasuk menyelaraskan skema data yang berbeda dan mengatasi redundansi data.

3.5 Teknik Data Transformation

Data Transformation adalah proses mengubah data ke dalam format yang sesuai untuk diproses oleh algoritma Data Mining. Beberapa bentuk transformasi yang umum dilakukan antara lain:

- Normalisasi: Menyesuaikan rentang nilai dari atribut-atribut data ke dalam skala yang sama (misalnya, 0 hingga 1) untuk mencegah atribut dengan rentang nilai besar mendominasi proses analisis.
- Agregasi: Menggabungkan dan meringkas data. Contohnya, data penjualan harian dapat diagregasi menjadi data penjualan bulanan atau tahunan.
- Inisialisasi Data**:** Mengubah data yang bersifat non-numerik (kategorikal) menjadi format numerik. Sebagai contoh, dalam analisis data mahasiswa, data nominal seperti "Jurusan" (misalnya, 'IT', 'Akuntansi') dan "Kota Asal" ('Jakarta', 'Bandung') harus diubah terlebih dahulu ke dalam bentuk angka (misalnya, 1, 2, 3) agar dapat diolah oleh algoritma seperti K-Means Clustering.

BAB IV

TEKNIK KLASIFIKASI

4.1 Konsep Dasar Klasifikasi

Klasifikasi adalah salah satu teknik yang paling umum diterapkan dalam Data Mining. Tujuannya adalah untuk menetapkan atau memprediksi label kelas dari suatu objek berdasarkan atribut-atribut yang dimilikinya. Label kelas ini bersifat kategorikal atau diskrit (misalnya: "Lulus" atau "Tidak Lulus", "Setuju" atau "Tidak Setuju"). Karena menggunakan data yang label kelasnya sudah diketahui untuk membangun model, klasifikasi tergolong dalam metode *supervised learning*.

4.2 Proses Membangun Model Klasifikasi

Proses klasifikasi umumnya terdiri dari dua tahap utama:

1. Tahap Belajar (*Learning*): Pada tahap ini, sebuah model klasifikasi dibangun dengan menganalisis sekumpulan data latih (*training data*). Data latih ini berisi objek-objek beserta label kelasnya yang sudah benar. Algoritma akan "belajar" untuk menemukan pola yang membedakan satu kelas dengan kelas lainnya.
2. Tahap Klasifikasi (*Classification*)**:** Setelah model dibangun, kinerjanya diuji menggunakan data uji (*testing data*) untuk memperkirakan akurasi aturan klasifikasi yang telah dibuat. Jika akurasi model dianggap cukup baik, model tersebut dapat diterapkan untuk mengklasifikasikan data baru yang label kelasnya belum diketahui.

Pengenalan Algoritma: Decision Tree

Decision Tree (Pohon Keputusan) adalah salah satu metode klasifikasi yang paling populer karena model yang dihasilkannya mudah untuk diinterpretasi oleh manusia. *Decision Tree* adalah sebuah struktur yang menyerupai *flowchart* atau pohon, di mana:

- Setiap **simpul internal** (*internal node*) menandakan sebuah tes pada suatu atribut.
- Setiap **cabang** (*branch*) merepresentasikan hasil dari tes tersebut.
- Setiap **simpul daun** (*leaf node*) merepresentasikan label kelas atau prediksi akhir.Alur pada

Decision tree ditelusuri dari simpul akar (paling atas) hingga ke salah satu simpul daun, yang akan memberikan prediksi kelas untuk data yang diuji. Model pohon keputusan ini sangat mudah untuk diubah menjadi aturan klasifikasi berbentuk "JIKA ... MAKA ...".

BAB V

TEKNIK CLUSTERING

5.1 Konsep Dasar Clustering

Clustering atau klasterisasi adalah proses pengorganisasian atau pengelompokan objek-objek data ke dalam beberapa grup atau *cluster*. Tujuannya adalah agar objek-objek yang berada dalam satu *cluster* memiliki tingkat kemiripan yang tinggi, sementara objek-objek yang berada di *cluster* yang berbeda memiliki tingkat kemiripan yang rendah. Perbedaan Clustering dengan Klasifikasi Perbedaan utama antara *clustering* dengan klasifikasi terletak pada ketersediaan label kelas pada data latih.

- Klasifikasi adalah metode *supervised learning*. Prosesnya memerlukan data latih yang sudah memiliki label kelas yang benar untuk membangun model.
- Clustering adalah metode *unsupervised learning* (klasifikasi tanpa arahan). Prosesnya tidak memerlukan data latih berlabel. Algoritma akan secara otomatis menemukan struktur atau pola pengelompokan alami yang ada di dalam data berdasarkan karakteristiknya.

5.2 Tipe-tipe Metode Clustering

Metode clustering secara umum dapat dibagi menjadi dua kategori utama:

1. ***Hierarchical Clustering*** Metode ini mengelompokkan data secara bertingkat, sehingga membentuk sebuah hierarki atau struktur pohon yang disebut dendrogram. Prosesnya dimulai dengan mengelompokkan dua atau lebih objek yang paling mirip, kemudian proses dilanjutkan ke objek lain hingga semua objek membentuk satu *cluster* besar.
2. ***Partitional (Non-Hierarchical) Clustering*** Berbeda dengan metode hierarki, metode ini langsung mempartisi atau membagi data ke dalam sejumlah *cluster* yang telah ditentukan di awal (misalnya, 3 *cluster*, 5 *cluster*, dll.). Proses pengelompokan dilakukan tanpa melalui tahapan hierarki. Metode yang paling populer dalam kategori ini adalah K-Means Clustering.

5.3 Pengukuran Jarak/Kemiripan

Dalam *clustering*, kemiripan antar objek data biasanya diukur berdasarkan jarak. Semakin dekat jarak antara dua objek, maka semakin mirip keduanya. Salah satu formula yang paling umum digunakan untuk menghitung jarak adalah *Euclidean Distance*.

BAB VI

STUDI KASUS MENDALAM: ALGORITMA K-MEANS

6.1 Cara Kerja Algoritma K-Means

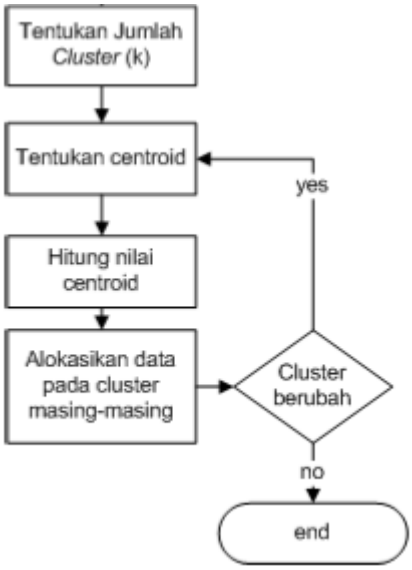
K-Means adalah salah satu metode *clustering* non-hirarki yang berusaha untuk mempartisi sekumpulan objek data ke dalam **K** buah *cluster* (di mana K ditentukan di awal). Pengelompokan ini dilakukan berdasarkan karakteristik dari objek-objek tersebut, sehingga objek yang memiliki karakteristik sama akan dikelompokkan ke dalam satu *cluster* yang sama.

Langkah-langkah Algoritma K-Means Proses algoritma K-Means bersifat iteratif (berulang) dengan langkah-langkah sebagai berikut:

1. Pilih secara acak **K** buah data sebagai pusat *cluster* (*centroid*) awal.
2. Hitung jarak setiap data ke masing-masing *centroid* awal tersebut.
3. Kelompokkan setiap data ke dalam *cluster* berdasarkan jarak terdekatnya ke *centroid*.
4. Hitung kembali posisi *centroid* baru dengan cara mengambil nilai rata-rata dari seluruh data yang ada di dalam *cluster* tersebut.
5. Ulangi langkah 2 hingga 4 sampai posisi *centroid* tidak lagi berubah (konvergen), yang menandakan bahwa proses *clustering* telah selesai.

6.2 Rumus Jarak Euclidean

Langkah-langkah melakukan clustering dengan metode K-Means adalah :



1. Tentukan jumlah nilai k sebagai jumlah cluster.

2. Alokasikan data kedalam kelompok secara random.

3. Hitung pusat cluster (centroid) menggunakan mean untuk masing-masing cluster dengan persamaan Euclidean yaitu sebagai berikut :

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

Dimana:

$D_{(i,j)=}$ jarak data ke *i* ke pusat *cluster j*

$X_{ki=}$ Data ke *i* pada atribut data ke *k*

$X_{kj=}$ Titik pusat ke *j* pada atribut ke *k*

Dimana : D(i,j) = Jarak data ke i ke pusat cluster j Xki = Data ke i pada atribut data ke k Xki = Titik pusat ke j pada atribut ke k

4. Alokasikan data berdasarkan jarak terdekat antara data dengan centroidnya.

5. Kembali kelangkah sebelumnya, jika ternyata masih ada data yang berpindah cluster atau jika nilai centroid diatas nilai ambang, atau jika nilai pada fungsi objektif yang digunakan masih diatas ambang.

Pada penerapan metode k-means cluster analysis, data yang bisa diolah dalam perhitungan adalah data numerik yang berbentuk angka. Sedangkan data selain angka juga bisa diterapkan tetapi terlebih dahulu harus dilakukan pengkodean untuk mempermudah perhitungan jarak/kesamaan karakteristik yang dimiliki dari setiap objek. Setiap objek dihitung kedekatan jaraknya berdasarkan karakter yang dimiliki dengan pusat cluster yang sudah ditentukan sebelumnya, jarak terkecil antara objek dengan masing-masing cluster merupakan anggota cluster yang terdekat. Setelah jumlah cluster ditentukan, selanjutnya dipilih sebanyak 3 objek secara acak sesuai jumlah cluster yang dibentuk sebagai pusat cluster awal untuk dihitung jarak kedekatannya terhadap semua objek yang ada.

2.3 Implementasi *RapidMiner*

Tahapan berikutnya yang akan dilakukan dalam penelitian ini adalah melakukan implementasi. Pada tahap implementasi ini penulis akan menghitung data karakter siswa yang telah diperoleh menggunakan aplikasi *RapidMiner* dengan perhitungan algoritma *K-Means.*

2.4 Hasil

Pada tahapan hasil penulis akan mengetahui hasil perhitungan algoritma *K-Means* menggunakan *excel* dan aplikasi *RapidMiner*. Dari hasil ini didapatkan kesimpulan dan saran.

Contoh Penerapan (Studi Kasus Mahasiswa)

Misalkan kita ingin mengelompokkan data mahasiswa menjadi 3 *cluster* (K=3). Data mahasiswa memiliki atribut non-numerik seperti "Jurusan" dan "Kota Asal".

- Transformasi Data: Atribut non-numerik tersebut harus diubah menjadi angka terlebih dahulu (inisialisasi).
- Penentuan Pusat Awal: Tiga data dipilih secara acak sebagai pusat *cluster* awal (C1, C2, C3).
- Perhitungan Jarak: Jarak dari setiap data mahasiswa ke tiga pusat *cluster* dihitung menggunakan rumus Euclidean. Contohnya, untuk data mahasiswa pertama, dihitung jaraknya ke C1, C2, dan C3.
    - Jarak ke C1: 5,390
    - Jarak ke C2: 13,000
    - Jarak ke C3: 13,038
- **Pengelompokan:** Karena jarak terdekat adalah ke C1 (5,390), maka data mahasiswa pertama dimasukkan ke dalam *Cluster* 1. Proses ini dilakukan untuk semua data.
- **Iterasi:** Setelah semua data dikelompokkan, *centroid* baru dihitung. Proses ini diulangi hingga tidak ada lagi data yang berpindah *cluster*.

Interpretasi Hasil Cluster

Setelah proses iterasi selesai, setiap *cluster* akan memiliki anggota dengan karakteristik yang cenderung homogen. Kita kemudian dapat menganalisis dan menafsirkan setiap *cluster*. Sebagai contoh, dari studi kasus tersebut dapat disimpulkan bahwa:

- Mahasiswa di **Cluster 1** didominasi oleh jurusan IT dan Marketing yang berasal dari Jakarta dan Jawa Barat.
- Mahasiswa di **Cluster 3** didominasi oleh jurusan Public Relation dan Akuntansi yang berasal dari Sulawesi, Jawa Timur, dan Sumatera Selatan.

Informasi atau pengetahuan seperti inilah yang menjadi hasil akhir dari proses Data Mining.

BAB VII

TEKNIK ATURAN ASOSIASI (*ASSOCIATION RULES*)

7.1 Konsep Dasar Aturan Asosiasi

Teknik Aturan Asosiasi adalah metode untuk menemukan hubungan atau asosiasi yang menarik antar item dalam sebuah kumpulan data. Teknik ini paling sering diilustrasikan dengan Analisis Keranjang Belanja (*Market Basket Analysis*), yang bertujuan untuk mengenali perilaku pembelian pelanggan di supermarket. Tujuannya adalah untuk menjawab pertanyaan seperti: "Jika seorang pelanggan membeli produk A, seberapa besar kemungkinannya ia juga akan membeli produk B?".

Hasil dari analisis ini adalah sekumpulan aturan asosiasi. Contoh aturan yang terkenal adalah "Jika seorang pelanggan membeli bir, maka ia juga akan membeli popok". Aturan semacam ini memberikan wawasan berharga yang dapat digunakan untuk pengambilan keputusan bisnis.

7.2 Metrik Penting: Support, Confidence, dan Lift

Untuk mengukur kekuatan dan signifikansi sebuah aturan, digunakan tiga metrik utama:

1. *Support* (Dukungan) *Support* adalah ukuran popularitas sebuah item atau kombinasi item. Metrik ini menunjukkan persentase dari total transaksi yang memuat item tersebut.
    - Contoh: Jika ada 1.000 transaksi dan 100 di antaranya memuat {Roti, Susu}, maka *support* untuk itemset {Roti, Susu} adalah 10%. Aturan dengan *support* yang rendah mungkin hanya terjadi secara kebetulan.
2. *Confidence* (Kepercayaan) *Confidence* adalah ukuran kekuatan hubungan antar item dalam sebuah aturan. Metrik ini mengukur probabilitas kondisional dari kemunculan item Y jika item X sudah ada dalam transaksi.
    - Contoh: Jika ada 150 transaksi yang memuat Roti, dan 100 dari transaksi tersebut juga memuat Susu, maka *confidence* dari aturan {Roti} => {Susu} adalah (100/150) = 66.7%. Ini berarti, 66.7% pelanggan yang membeli Roti juga membeli Susu.
3. *Lift* (Daya Angkat) *Lift* adalah metrik yang mengukur seberapa besar kemungkinan item Y dibeli jika item X dibeli, sambil mengontrol popularitas item Y itu sendiri. Nilai *lift* > 1 menunjukkan bahwa kedua item tersebut memang memiliki kemungkinan untuk dibeli bersamaan (asosiasi positif).
    - Contoh: Jika *lift* dari aturan {Roti} => {Susu} adalah 3, ini berarti pelanggan yang membeli Roti memiliki kemungkinan 3 kali lebih besar untuk membeli Susu dibandingkan pelanggan secara umum.

7.3 Struktur Aturan dan Algoritma Apriori

Sebuah aturan asosiasi biasanya ditulis dalam bentuk **X => Y**, di mana X adalah *antecedent* (pendahulu) dan Y adalah *consequent* (akibat).

Untuk menemukan aturan-aturan ini dari jutaan data transaksi, digunakan algoritma khusus. Salah satu yang paling terkenal adalah Algoritma Apriori. Prinsip utama Apriori adalah: "Jika sebuah *itemset* (kombinasi item) sering muncul, maka semua *subset* (bagian dari kombinasi itu) juga pasti sering muncul." Prinsip ini digunakan untuk memangkas pencarian dan membuat prosesnya menjadi lebih efisien.

BAB VIII

EVALUASI MODEL DATA MINING

8.1 Pentingnya Evaluasi Model

Membangun sebuah model Data Mining hanyalah setengah dari pekerjaan. Tahap **evaluasi** sangat krusial untuk memastikan bahwa model yang dihasilkan tidak hanya akurat pada data yang digunakan untuk melatihnya, tetapi juga dapat berkinerja baik pada data baru yang belum pernah dilihat sebelumnya. Tanpa evaluasi, kita berisiko menggunakan model yang salah dan membuat keputusan yang keliru. Proses ini bertujuan untuk mengidentifikasi pola-pola yang benar-benar menarik dan signifikan, bukan hanya sekadar kebetulan statistik.

8.2 Teknik Evaluasi untuk Klasifikasi

- Metrik Kinerja dari *Confusion Matrix*
    - *Accuracy* (Akurasi): (TP + TN) / Total. Mengukur persentase prediksi yang benar secara keseluruhan. Metrik ini bisa menyesatkan pada dataset yang tidak seimbang.
    - *Precision* (Presisi): TP / (TP + FP). Mengukur tingkat keakuratan dari prediksi Positif. Berguna jika biaya kesalahan *False Positive* tinggi.
    - *Recall* (Sensitivitas): TP / (TP + FN). Mengukur kemampuan model untuk menemukan semua data yang benar-benar Positif. Berguna jika biaya kesalahan *False Negative* tinggi.

8.3 Teknik Evaluasi untuk Clustering

Evaluasi *clustering* lebih kompleks karena tidak ada "jawaban benar" atau label kelas sebagai pembanding. Evaluasi biasanya berfokus pada kualitas geometris dari *cluster* yang terbentuk.

- Kohesi dan Separasi: *Cluster* yang baik harus memiliki:
    - Kohesi Tinggi: Jarak antar anggota di dalam satu *cluster* sangat dekat (anggota sangat mirip).
    - Separasi Tinggi: Jarak antar *cluster* yang berbeda sangat jauh (*cluster* yang satu sangat berbeda dari *cluster* yang lain).
- *Elbow Method* Ini adalah teknik visual yang sering digunakan untuk membantu menentukan jumlah *cluster* (K) yang optimal dalam algoritma K-Means. Caranya adalah dengan menjalankan K-Means dengan jumlah K yang berbeda-beda dan memplot tingkat *error* (biasanya *Sum of Squared Errors*). Titik "siku" (*elbow*) pada grafik menunjukkan titik di mana penambahan jumlah *cluster* tidak lagi memberikan penurunan *error* yang signifikan, yang sering dianggap sebagai nilai K yang optimal.

BAB IX

TOOLS DAN TREN MASA DEPAN DATA MINING

9.1 Pengenalan Tools Populer

Untuk menerapkan Data Mining, praktisi dan akademisi sering menggunakan perangkat lunak yang menyediakan berbagai algoritma siap pakai.

- Tools Berbasis Antarmuka Grafis (GUI):
    - Weka: Aplikasi *open-source* berbasis Java yang sangat populer di lingkungan akademik. Weka menyediakan koleksi lengkap algoritma *machine learning* dan alat visualisasi yang mudah digunakan untuk pemula.
    - RapidMiner & Orange: Platform yang menggunakan pendekatan alur kerja visual. Pengguna dapat merancang proses Data Mining dengan menyusun dan menyambungkan blok-blok operator, membuatnya sangat intuitif.
- Tools Berbasis Kode (Programming):
    - Python: Telah menjadi bahasa standar de-facto di industri *data science*. Kekuatannya terletak pada ekosistem *library*-nya yang sangat kaya, seperti:
        - Pandas: Untuk manipulasi dan analisis data.
        - Scikit-learn: Untuk implementasi berbagai algoritma klasik.
        - TensorFlow & PyTorch: Untuk pengembangan model *Deep Learning*.
        - **Matplotlib & Seaborn:** Untuk visualisasi data.

9.2 Tren Masa Depan Data Mining

Bidang Data Mining terus berevolusi dengan cepat. Beberapa tren yang akan mendominasi di masa depan antara lain:

- ***Big Data***: Analisis data dengan volume, kecepatan, dan variasi yang ekstrem. Ini mendorong pengembangan algoritma yang lebih efisien dan dapat berjalan pada sistem terdistribusi (seperti Apache Spark).
- ***Deep Learning***: Penggunaan *Artificial Neural Network* (Jaringan Saraf Tiruan) yang sangat dalam untuk secara otomatis mengekstrak pola-pola kompleks dari data tidak terstruktur seperti gambar, suara, dan teks.
- ***Explainable AI* (XAI):** Ada tuntutan yang semakin besar agar model AI tidak lagi menjadi "kotak hitam" (*black box*). XAI adalah bidang yang berfokus pada pengembangan teknik untuk memahami dan menjelaskan mengapa sebuah model membuat keputusan tertentu, yang sangat krusial untuk transparansi dan akuntabilitas.
- ***Automated ML* (AutoML):** Platform dan teknik yang bertujuan untuk mengotomatiskan proses pemilihan data, rekayasa fitur, serta pemilihan dan penyetelan model, sehingga mempercepat siklus pengembangan Data Mining.

DAFTAR PUSTAKA

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze student's performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63–69.

Begum, S. H. (2013). Data mining tools and trends: An overview. *International Journal of Emerging Research in Management & Technology*, *2*(6).

Ediyanto, Novitasari Mara, M., & Satyahadewi, N. (2013). Pengklasifikasian karakteristik dengan metode K-means cluster analysis. *Buletin Ilmiah Matematika, Statistika dan Terapannya (BILMASTER), 2*(2), 133-138.

Kaparang, D. R., & Sediyono, E. (2013). Penentuan alih fungsi lahan marginal menjadi lahan pangan berbasis algoritma K-means di wilayah Kabupaten Boyolali. *Jurnal JdC*, *2*(2), 18-24.

Nugraha, D. D. C., Naimah, Z., Fahmi, M., & Setiani, N. (2014). Klasterisasi judul buku dengan menggunakan metode K-Means. *Prosiding Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, G-2.

Ong, J. O. (2013). Implementasi algoritma K-means clustering untuk menentukan strategi marketing President University. *Jurnal Ilmiah Teknik Industri*, *12*(1), 13–20.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of K-Means clustering algorithm for prediction of student's academic performance. *International Journal of Computer Science and Information Security*, *7*(1), 292-295.

Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan data mining untuk evaluasi kinerja akademik mahasiswa menggunakan algoritma naive bayes classifier. *Jurnal EECCIS, 7*(1), 59-64.

Seddawy, A. B. E., Khedr, A., & Sultan, T. (2012). Adapted framework for data mining technique to improve decision support system in an uncertain situation. *International Journal of Data Mining & Knowledge Management Process*, *2*(5), 1-10.

Sudirman, & Ani, N. (2012). Implementasi teknik data mining dengan algoritma K-means clustering dan fungsi kernel polynominal untuk klasterisasi objek data. *Prosiding Seminar Nasional Efisiensi Energi untuk Peningkatan Daya Saing Industri Manufaktur & Otomotif Nasional*, B-50.

Yedla, M., Pathakota, S. R., & Srinivasa, T. M. (2010). Enhancing K-means clustering algorithm with improved initial center. *International Journal of Computer Science and Information Technologies*, *1*(2), 121-125.