BUKU AJAR
DATA MINING

Tim Penulis :
Prastyadi Wibawa Rahayu, S.Kom., M.Kom
I Gede Iwan Sudipa, S.Kom., M.Cs
Suryani, S.Kom., M.T
Arie Surachman, M.Kom
Achmad Ridwan, S.Kom., M.Kom
I Gede Mahendra Darmawiguna, S.Kom., M.Sc
Ir. Muh. Nurtanzis Sutoyo, S.Kom., M.Cs., IPP
Drs. Isnandar Slamet, M.Sc., Ph.D
Sitti Harlina, SE., M.Kom
I Made Dendi Maysanjaya, S.Pd., M.Eng

SONPEDIA.COM
PT. Sonpedia Publishing Indonesia

BUKU AJAR
DATA MINING

Tim Penulis :
Prastyadi Wibawa Rahayu, S.Kom., M.Kom
I Gede Iwan Sudipa, S.Kom., M.Cs
Suryani, S.Kom., M.T
Arie Surachman, M.Kom
Achmad Ridwan, S.Kom., M.Kom
I Gede Mahendra Darmawiguna, S.Kom., M.Sc
Ir. Muh. Nurtanzis Sutoyo, S.Kom., M.Cs., IPP
Drs. Isnandar Slamet, M.Sc., Ph.D
Sitti Harlina, SE., M.Kom
I Made Dendi Maysanjaya, S.Pd., M.Eng

Penerbit

SONPEDIA.COM
PT. Sonpedia Publishing Indonesia

BUKU AJAR
DATA MINING

Tim Penulis :
Prastyadi Wibawa Rahayu, S.Kom., M.Kom
I Gede Iwan Sudipa, S.Kom., M.Cs
Suryani, S.Kom., M.T
Arie Surachman, M.Kom
Achmad Ridwan, S.Kom., M.Kom
I Gede Mahendra Darmawiguna, S.Kom., M.Sc
Ir. Muh. Nurtanzis Sutoyo, S.Kom., M.Cs., IPP
Drs. Isnandar Slamet, M.Sc., Ph.D
Sitti Harlina, SE., M.Kom
I Made Dendi Maysanjaya, S.Pd., M.Eng

ISBN : 978-623-8483-96-9 (PDF)

Editor :
Efitra
Penyunting :
Ida Kumala Sari
Desain sampul dan Tata Letak :
Yayan Agusdi

Penerbit :
PT. Sonpedia Publishing Indonesia
Redaksi :
Jl. Kenali Jaya No 166 Kota Jambi 36129 Telp. +6282177858344
Email : sonpediapublishing@gmail.com
Website : www.buku.sonpedia.com

Anggota IKAPI : 006/JBI/2023

Cetakan Pertama, Januari 2024

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa yang telah memberikan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan buku ini dengan baik. Buku ini berjudul ***"BUKU AJAR DATA MINING".*** Tidak lupa kami ucapkan terima kasih bagi semua pihak yang telah membantu dalam penulisan dan penerbitan buku ini.

Buku ini disusun sebagai buku panduan komprehensif yang menjelajahi kompleksitas dan mendalamnya tentang ilmu sistem informasi. Buku ini dapat digunakan oleh pendidik dalam melaksanakan kegiatan pembelajaran di bidang ilmu sistem informasi dan diberbagai bidang Ilmu terkait lainnya. Buku ini dapat digunakan sebagai panduan dan referensi mengajar mata kuliah data mining dan menyesuaikan dengan Rencana Pembelajaran Semester tingkat Perguruan Tinggi masing-masing.

Secara garis besar, buku ajar ini pembahasannya mulai dari data mining and knowledge discovery process, data understanding, knowledge representation, data preprocessing. Buku ini juga membahas materi penting lainnya seperti data mining roles, classification and prediction, cluster analysis, association rules. Selain itu materi mengenai Text Mining dan Feature extraction and selection Method dibahas secara mendalam. Buku ajar ini disusun secara sistematis, ditulis dengan bahasa yang jelas dan mudah dipahami, dan dapat digunakan dalam kegiatan pembelajaran.

Buku ini mungkin masih terdapat kekurangan dan kelemahan. Oleh karena itu, saran dan kritik para pemerhati sungguh penulis harapkan. Semoga buku ajar ini memberikan manfaat dan menambah khasanah ilmu pengetahuan dalam pembalajaran.

Bandung, Januari 2024
Tim Penulis

DAFTAR ISI

KATA PENGANTAR ..... ***ii***
DAFTAR ISI ..... ***iii***
KEGIATAN BELAJAR 1 *DATA MINING AND KNOWLEDGE DISCOVERY PROCESS* ***1***
DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PENGERTIAN DATA MINING 2
B. TUJUAN DATA MINING 3
C. MANFAAT DATA MINING 4
D. PENERAPAN DATA MINING 5
E. TEKNIK-TEKNIK DATA MINING 7
F. KNOWLEDGE DISCOVERY PROCESS 8
G. RANGKUMAN 9
H. TES FORMATIF 10
I. LATIHAN 11
KEGIATAN BELAJAR 2 *DATA UNDERSTANDING DALAM DATA MINING* ***12***
DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PENGANTAR DATA UNDERSTANDING DALAM DATA MINING 13
B. FUNGSI DATA UNDERSTANDING DALAM DATA MINING 16
C. MANFAAT DATA UNDERSTANDING 17
D. KONSEP DASAR DATA UNDERSTANDING 19
E. KELEBIHAN DAN KEKURANGAN DATA UNDERSTANDING 21
F. RANGKUMAN 23
G. STUDI KASUS 24

H. TEST FORMATIF 27

I. LATIHAN 29

KEGIATAN BELAJAR 3 *KNOWLEDGE REPRESENTATION* *30*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN

A. PENGERTIAN KNOWLEDGE REPRESENTATION 31

B. FUNGSI KNOWLEDGE REPRESENTATION 31

C. DATA VISUALIZATION 32

D. RANGKUMAN 42

E. TES FORMATIF 42

F. LATIHAN 43

KEGIATAN BELAJAR 4 *DATA PREPROCESSING* *44*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN

A. DATA PREPROCESSING 46

B. Langkah-langkah Data Preprocessing 49

C. Tujuan Data Preprocessing 64

D. Hasil Data Preprocessing 66

E. Dokumentasi Data Preprocessing 68

F. RANGKUMAN 70

G. TES FORMATIF 72

H. LATIHAN 73

KEGIATAN BELAJAR 5 *DATA MINING ROLES (Peran Data Mining)* *76*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN

A. PENGERTIAN DATA MINING 77

B. PERAN DATA MINING DALAM KLASIFIKASI (CLASSIFICATION) 79

C. PERAN DATA MINING DALAM ASOSIASI (ASSOCIATION) 81
D. PERAN DATA MINING DALAM KLASTERING (CLUSTERING) 83
E. PERAN DATA MINING DALAM ESTIMASI (ESTIMATION) 85
F. PERAN DATA MINING DALAM PREDIKSI (FORECASTING) 87
G. RANGKUMAN 89
H. TES FORMATIF 90
I. LATIHAN 90

KEGIATAN BELAJAR 6 *PREDICTION AND CLASSIFICATION* *91*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PREDICTIVE ANALYTICS 92
B. METODE PREDIKSI 93
C. METODE KLASIFIKASI 97
D. EVALUASI METODE PREDIKSI DAN KLASIFIKASI 101
E. RANGKUMAN 107
F. TES FORMATIF 108
G. LATIHAN 109

KEGIATAN BELAJAR 7 *CLUSTERING ANALYSIS* *110*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. KONSEP DASAR CLUSTERING 111
B. JENIS-JENIS CLUSTERING 112
C. PREPROCESSING DATA 115
D. ALGORITMA CLUSTERING 117
E. METRIK KESAMAAN JARAK 118
F. EVALUASI CLUSTERING 120
G. TANTANGAN DAN SOLUSI DALAM CLUSTERING 122

H. KASUS 124
I. RANGKUMAN 130
J. TES FORMATIF 132
K. LATIHAN 134

KEGIATAN BELAJAR 8 *TEORI DASAR ASSOCIATION RULES* *135*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PENDAHULUAN 136
B. KONSEP ASOSIASI 139
C. ALGORITMA APRIORI 142
D. ALGORITMA *FP-Growth* 152
E. PENERAPAN PADA R 156
F. RANGKUMAN 158
G. TES FORMATIF 159
H. LATIHAN 160

KEGIATAN BELAJAR 9 *TEXT DATA MINING* *162*

DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PENGERTIAN TEXT MINING 163
B. MANFAAT TEXT MINING 165
C. TAHAPAN PROSES TEXT MINING 166
D. CARA KERJA TEXT MINING 167
E. TEKNIK DALAM TEXT MINING 168
F. RANGKUMAN 170
G. TES FORMATIF 171
H. LATIHAN 171

KEGIATAN BELAJAR 10 *METODE EKSTRAKSI*
DAN SELEKSI FITUR ***172***
DESKRIPSI, KOMPETENSI DAN PETA KONSEP PEMBELAJARAN
A. PENGERTIAN EKSTRAKSI DAN SELEKSI FITUR 173
B. METODE EKSTRAKSI FITUR 175
C. METODE SELEKSI FITUR 181
D. RANGKUMAN 184
E. TES FORMATIF 185
F. LATIHAN 185
DAFTAR PUSTAKA ***186***
TENTANG PENULIS ***202***

KEGIATAN BELAJAR I

DATA MINING AND KNOWLEDGE DISCOVERY PROCESS

DESKRIPSI PEMBELAJARAN

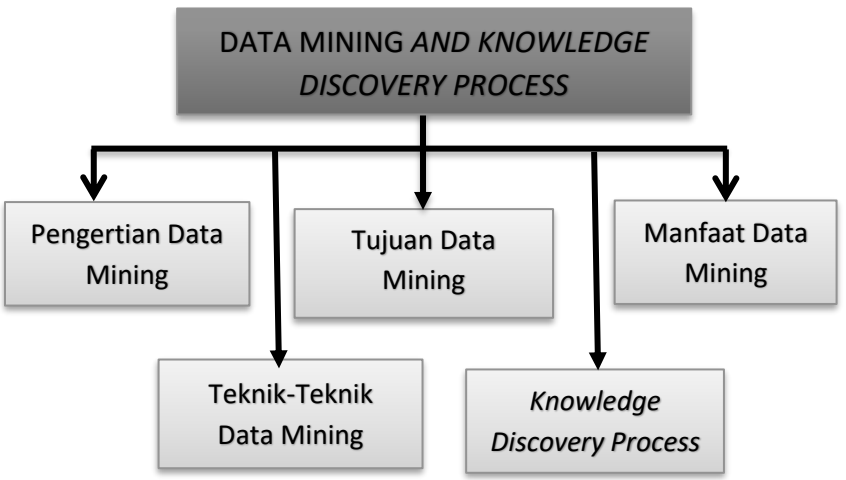
Pada bab ini mahasiswa mempelajari pengenalan dan konsep data mining and *knowledge discovery process*. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk mempelajari berbagai jenis teknik-teknik data mining.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

1. Mampu menguraikan definisi data mining
2. Mempu menjelaskan tujuan, manfaat, penerpan dan teknik-teknik data mining
3. Mampu menjelaskan *knowledge discovery process*

PETA KONSEP PEMBELAJARAN



A. PENGERTIAN DATA MINING

Data mining adalah suatu proses ekstraksi pengetahuan atau informasi yang berharga dari suatu set data yang besar dan kompleks. Tujuan utama dari data mining adalah mengidentifikasi pola, hubungan, atau informasi yang mungkin tidak terlihat secara langsung dalam data, sehingga dapat memberikan wawasan yang lebih dalam dan bernilai.

Proses data mining melibatkan penggunaan berbagai teknik statistik, matematis, dan kecerdasan buatan untuk menganalisis data dengan cara yang sistematis dan otomatis. Hasil dari data mining dapat digunakan untuk mendukung pengambilan keputusan, mengidentifikasi tren pasar, meningkatkan efisiensi operasional, atau merumuskan strategi bisnis. Berikut adalah beberapa definisi Data Mining menurut para ahli:

1. Menurut Han dan Kamber: "Data mining adalah proses penemuan pola atau informasi yang berguna dari basis data besar dengan menggunakan metode termasuk teknik statistik, matematika, dan kecerdasan buatan."
2. Menurut Berry dan Linoff: "Data mining adalah proses penemuan pola berharga atau informasi pengetahuan baru dalam database besar dengan menggunakan algoritma pencarian atau algoritma pembelajaran mesin."
3. Menurut Fayyad, Piatetsky-Shapiro, dan Smyth: "Data mining adalah proses ekstraksi pengetahuan yang berharga, pola, atau informasi tersembunyi dari sejumlah besar data.
4. Margaret H. Dunham: "Data mining adalah proses otomatisasi dalam penemuan pola yang bermakna, kecocokan, korelasi, dan tren yang tersembunyi di dalam basis data."
5. Usama Fayyad: "Data mining adalah proses penggalian pengetahuan yang dimaksudkan untuk menemukan pola-pola yang bermanfaat, pengetahuan baru, dan informasi yang tersembunyi dari data yang tersimpan di basis data."

B. TUJUAN DATA MINING

Tujuan utama dari data mining adalah mengidentifikasi pola, hubungan, atau pengetahuan yang berharga dan tersembunyi dalam suatu set data besar atau kompleks. Proses data mining bertujuan untuk menggali wawasan yang tidak dapat ditemukan secara langsung melalui pengamatan sederhana terhadap data. Beberapa tujuan khusus dari data mining meliputi:

1. Pencarian Pola atau Hubungan
   Mengidentifikasi pola atau hubungan yang mungkin tidak terlihat secara langsung dalam data.
2. Prediksi dan Klasifikasi
   Membangun model yang dapat memprediksi nilai yang belum diketahui atau mengklasifikasikan data ke dalam kategori tertentu.
3. Segmentasi Pelanggan
   Membagi pelanggan atau data ke dalam segmen atau kelompok berdasarkan karakteristik tertentu, seperti perilaku pembelian atau preferensi.
4. Analisis Asosiasi
   Mengidentifikasi hubungan atau asosiasi antara variabel-variabel dalam data, seperti hubungan antara produk yang sering dibeli bersama-sama.
5. Deteksi Anomali
   Mendeteksi pola yang tidak biasa atau anomali dalam data, yang dapat mengindikasikan situasi atau kejadian yang tidak normal atau patut dicurigai.
6. Optimasi Proses Bisnis
   Meningkatkan efisiensi dan produktivitas dengan mengidentifikasi area-area di mana proses bisnis dapat dioptimalkan.
7. Penentuan Profil Konsumen
   Membangun profil konsumen berdasarkan data pembelian, preferensi, dan perilaku konsumen untuk mendukung strategi pemasaran yang lebih efektif.

8. Penentuan Tren Pasar
   Mengidentifikasi tren dan pola perilaku pasar yang dapat digunakan untuk merumuskan strategi bisnis dan pemasaran.
9. Analisis Risiko
   Menganalisis risiko dalam keputusan bisnis, seperti risiko kredit dalam industri keuangan.
10. Penemuan Pengetahuan Baru
    Menemukan informasi atau pengetahuan baru yang dapat memberikan wawasan atau keuntungan kompetitif.

C. MANFAAT DATA MINING

Data mining memiliki banyak manfaat dan dapat memberikan nilai tambah yang signifikan untuk berbagai industri dan organisasi. Berikut adalah beberapa manfaat utama dari data mining:

1. Pengambilan Keputusan yang Lebih Baik
   Data mining membantu organisasi membuat keputusan yang lebih informasional dan terarah. Dengan menganalisis data historis, model prediktif, dan pola tersembunyi, pengambilan keputusan dapat didukung oleh wawasan yang lebih mendalam.
2. Penemuan Pola dan Tren
   Data mining memungkinkan identifikasi pola dan tren yang tidak terlihat secara langsung oleh manusia. Ini dapat membantu organisasi dalam merespons perubahan pasar, mengidentifikasi peluang baru, atau menanggapi tantangan bisnis.
3. Prediksi Masa Depan
   Dengan menggunakan teknik prediktif, data mining dapat membantu organisasi memprediksi kejadian masa depan berdasarkan data historis. Contohnya termasuk prediksi penjualan, perilaku pelanggan, atau peristiwa risiko.

4. Segmentasi Pelanggan
   Melalui pengelompokan atau segmentasi pelanggan, data mining memungkinkan organisasi untuk memahami preferensi dan perilaku pelanggan secara lebih baik. Ini dapat digunakan untuk pengembangan produk, penargetan pemasaran, dan personalisasi layanan.
5. Efisiensi Operasional
   Data mining membantu mengidentifikasi area di mana efisiensi operasional dapat ditingkatkan. Analisis ini dapat membantu organisasi mengoptimalkan proses bisnis, mengurangi biaya, dan meningkatkan produktivitas.
6. Pengelolaan Risiko
   Dengan mendeteksi pola anomali atau melalui model risiko, data mining dapat membantu organisasi dalam pengelolaan risiko. Ini berlaku untuk industri keuangan (deteksi penipuan), kesehatan (prediksi penyakit), dan lainnya.
7. Pemasaran yang Lebih Efektif
   Melalui analisis data pelanggan, preferensi, dan perilaku pembelian, data mining dapat membantu perusahaan menyusun strategi pemasaran yang lebih efektif dan berfokus.
8. Peningkatan Layanan Pelanggan
   Dengan memahami lebih baik kebutuhan dan preferensi pelanggan, organisasi dapat meningkatkan layanan pelanggan mereka, memberikan pengalaman yang lebih baik, dan meningkatkan retensi pelanggan.

D. PENERAPAN DATA MINING

Data mining memiliki berbagai penerapan di berbagai industri dan bidang. Berikut adalah beberapa contoh penerapan data mining:

1. Pemasaran dan Analisis Pelanggan
   Analisis data pelanggan dapat membantu perusahaan mengidentifikasi pola pembelian, preferensi pelanggan, dan perilaku konsumen. Ini memungkinkan perusahaan untuk

merancang kampanye pemasaran yang lebih efektif dan menyediakan layanan yang lebih personal.

2. Keuangan dan Perbankan
   Data mining digunakan untuk mendeteksi penipuan kredit, analisis risiko kredit, prediksi perilaku pasar keuangan, dan identifikasi tren dalam transaksi keuangan. Ini membantu lembaga keuangan dalam pengambilan keputusan yang lebih baik dan pengelolaan risiko.
3. Kesehatan dan Perawatan Medis
   Data mining dapat membantu dalam analisis data kesehatan untuk prediksi penyakit, identifikasi pola penyebaran penyakit, dan pengembangan model diagnostik. Ini juga dapat digunakan untuk personalisasi perawatan pasien dan manajemen data klinis.
4. Manufaktur dan Produksi
   Dalam industri manufaktur, data mining dapat digunakan untuk meningkatkan efisiensi operasional, merencanakan rantai pasokan, dan memperbaiki kualitas produk. Analisis data dapat membantu mengoptimalkan proses produksi dan mengurangi biaya.
5. Sumber Daya Manusia
   Data mining dapat digunakan dalam manajemen sumber daya manusia untuk rekrutmen yang lebih efektif, retensi karyawan, dan analisis kinerja karyawan. Ini membantu organisasi dalam pengelolaan sumber daya manusia secara lebih efisien.
6. Telekomunikasi
   Dalam industri telekomunikasi, data mining digunakan untuk analisis pelanggan, manajemen kapasitas jaringan, dan deteksi kecurangan. Ini membantu penyedia layanan telekomunikasi dalam meningkatkan kualitas layanan dan kepuasan pelanggan.
7. Pendidikan
   Data mining dapat membantu institusi pendidikan dalam meningkatkan efektivitas pengajaran dan membantu pengambilan keputusan berbasis data. Ini dapat digunakan

untuk memahami pola belajar siswa, memprediksi keberhasilan akademis, dan mengidentifikasi area yang memerlukan perhatian khusus

8. Riset Ilmiah
   Dalam penelitian ilmiah, data mining digunakan untuk mengekstrak pola-pola dari data eksperimental dan analisis besar-besaran. Ini membantu peneliti dalam membuat hipotesis baru atau mengidentifikasi tren yang tidak terlihat secara langsung.
9. Pemerintahan dan Keamanan
   Pemerintah dapat menggunakan data mining untuk analisis keamanan publik, deteksi aktivitas kriminal, dan prediksi tren keamanan. Ini juga dapat membantu dalam manajemen data pemerintah untuk pengambilan keputusan yang lebih baik.
10. *E-commerce*
    Perusahaan *e-commerce* menggunakan data mining untuk rekomendasi produk yang lebih akurat kepada pelanggan, analisis pola pembelian, dan manajemen rantai pasokan. Ini dapat meningkatkan pengalaman belanja *online* dan kepuasan pelanggan.

E. TEKNIK-TEKNIK DATA MINING

Ada beberapa teknik data mining yang digunakan untuk mengekstrak pola dan pengetahuan dari data. Berikut adalah beberapa teknik utama:

1. Klasifikasi (*Classification*):
   Teknik ini digunakan untuk mengelompokkan data ke dalam kelas atau kategori tertentu berdasarkan karakteristik atau atribut tertentu. Contohnya termasuk pohon keputusan, jaringan saraf tiruan, dan algoritma klasifikasi lainnya.
2. Regresi (*Regression*)
   Digunakan untuk memodelkan hubungan antara variabel dependen dan independen. Regresi digunakan untuk

memprediksi nilai kontinu berdasarkan hubungan linier atau non-linier antara variabel-variabel tersebut.

3. Klastering (*Clustering*):
   Klastering digunakan untuk mengelompokkan data ke dalam kelompok atau klaster berdasarkan kesamaan fitur atau karakteristik tertentu. Algoritma k-means dan hierarchical clustering adalah contoh teknik klastering.
4. Asosiasi (*Association*)
   Teknik ini digunakan untuk menemukan hubungan atau asosiasi antara item atau variabel dalam data. Algoritma terkenal seperti Apriori dan Eclat digunakan untuk menemukan aturan asosiasi dalam data transaksional.

F. KNOWLEDGE DISCOVERY PROCESS

Knowledge Discovery in Databases (KDD) adalah proses secara umum yang melibatkan beberapa tahapan atau langkah. Proses ini dikenal sebagai *Knowledge Discovery Process* atau Data Mining *Process*. Tahapan-tahapan tersebut mencakup:

1. Seleksi dan Pemahaman Data (*Selection and Preprocessing*)
   Tahap ini melibatkan pemilihan data yang relevan untuk analisis dan pemahaman awal terhadap data. Hal ini juga melibatkan pembersihan data, pengelolaan nilai yang hilang, dan preproses data untuk mempersiapkannya bagi tahapan analisis berikutnya.
2. Pemahaman Data (Data *Cleaning and Understanding*)
   Data yang telah dipilih dianalisis lebih lanjut untuk pemahaman yang lebih mendalam. Ini melibatkan pemahaman statistik sederhana, visualisasi data, dan analisis deskriptif untuk mengidentifikasi pola awal atau anomali yang mungkin perlu ditangani.
3. Transformasi Data (Data *Transformation*)
   Proses ini melibatkan konversi atau transformasi data agar sesuai dengan kebutuhan analisis. Ini bisa mencakup

normalisasi data, konversi format, atau pembentukan fitur baru yang lebih relevan untuk analisis.

4. Pembentukan Model atau Pemodelan (Data Mining)
   Ini adalah tahap inti dari proses KDD. Pada tahap ini, teknik-teknik data mining seperti klasifikasi, klastering, regresi, atau asosiasi diterapkan untuk mengekstraksi pola-pola dari data. Model prediktif atau deskriptif dibangun selama tahap ini.
5. Evaluasi Model (Model *Evaluation*)
   Setelah model dibangun, tahap evaluasi dilakukan untuk menilai sejauh mana model tersebut efektif dan dapat diandalkan. Evaluasi dapat melibatkan pengujian model menggunakan data yang tidak terlihat sebelumnya atau dengan menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan sebagainya.
6. Penginterpretasian dan Visualisasi Hasil (*Interpretation* and *Visualization*)
   Hasil dari analisis data mining diinterpretasikan dalam konteks bisnis atau ilmiah. Visualisasi sering digunakan untuk membantu pemahaman dan komunikasi hasil dengan pemangku kepentingan yang mungkin tidak memiliki latar belakang analisis data yang mendalam.
7. Penggunaan Pengetahuan yang Ditemukan (*Knowledge Utilization*)
   Pengetahuan yang ditemukan atau pola yang diidentifikasi selama proses KDD digunakan untuk mendukung pengambilan keputusan atau tindakan yang lebih baik. Implementasi solusi atau perubahan berdasarkan temuan dapat melibatkan penggunaan teknologi informasi atau perubahan prosedur bisnis.

G. RANGKUMAN

Berdasarkan uraian di atas di mulai dari pengertian data mining adalah suatu proses ekstraksi pengetahuan atau informasi yang

berharga dari suatu set data yang besar dan kompleks, Tujuan utama dari data mining adalah mengidentifikasi pola, hubungan, atau pengetahuan yang berharga dan tersembunyi dalam suatu set data besar atau kompleks. Proses data mining bertujuan untuk menggali wawasan yang tidak dapat ditemukan secara langsung melalui pengamatan sederhana terhadap data. Manfaat utama dari data mining adalah pengambilan keputusan yang lebih baik, penemuan pola dan tren, prediksi masa depan, segmentasi pelanggan, efisiensi operasional, pengelolaan risiko, pemasaran yang lebih efektif dan peningkatan layanan pelanggan. Penerapan data mining dalam berbagai industri dan bidang ada beberapa contoh yaitu pemasaran dan analisis pelanggan, keuangan dan perbankan, Kesehatan dan perawatan medis, manufaktur dan produksi, sumber daya manusia, telekomunikasi dan pendidikan. Teknik – teknik data mining utama terdiri dari klasifikasi, regresi, klastering, asosiasi. Tahapan *knowledge discovery process* dimulai dari seleksi dan pemahaman data, pemahaman data, transformasi data, pembentukan model atau pemodelan, evaluasi model, penginterpretasian dan visualisasi hasil dan penggunaan pengetahuan yang ditemukan.

H. TES FORMATIF

1. Tahapan *knowledge discovery process*, kecuali ?
   a. Seleksi dan Pemahaman Data Python
   b. Transformasi Data
   c. Penggunaan Pengetahuan yang Ditemukan
   d. Analisis data
2. Teknik-teknik utama data mining, kecuali ?
   a. Klasifikasi
   b. Klastering
   c. *Regresi.*
   d. Regresi Linear

I. LATIHAN

Berikan beberapa contoh penerapan data mining yang digunakan dalam kehidupan sehari-hari untuk menunjang aktifitas kerja dan bisnis lainya sebagainya!

KEGIATAN BELAJAR 2

DATA UNDERSTANDING DALAM DATA MINING

DESKRIPSI PEMBELAJARAN

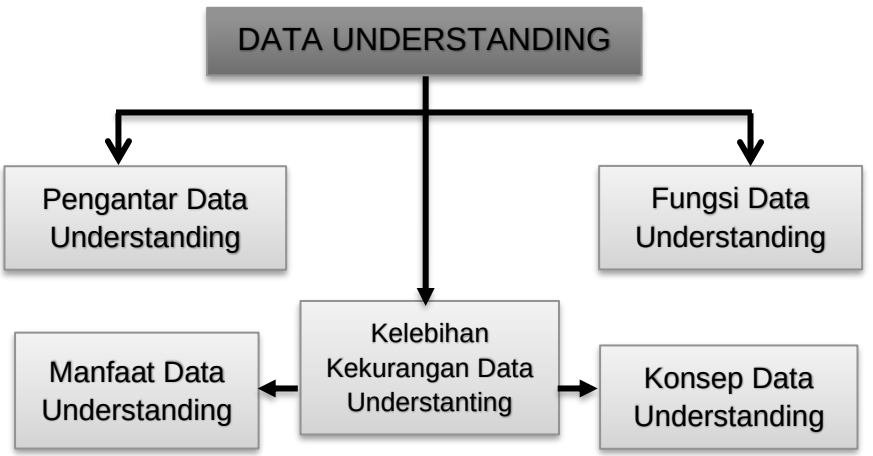
Pada bab ini mahasiswa mempelajari pengenalan dan konsep Media Pembelajaran. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk modal dasar mempelajari data *understanding* dalam *data mining.*

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

1. Mampu menguraikan pengantar data *understanding* dalam data *mining*.
2. Mampu menjelaskan fungsi dan manfaat *data understanding*.
3. Mampu menjelaskan konsep dasar data *understanding*

PETA KONSEP PEMBELAJARAN



A. PENGANTAR DATA UNDERSTANDING DALAM DATA MINING

Data *mining* adalah proses ekstraksi pengetahuan yang berguna, pola, dan informasi yang tersembunyi dari dataset besar. Tujuan utamanya adalah untuk mendapatkan pemahaman yang lebih dalam tentang data dan membuat keputusan yang informasional. Data mining melibatkan penerapan teknik dan algoritma dari berbagai disiplin ilmu seperti statistika, pembelajaran mesin, dan kecerdasan buatan.

Data *understanding* merupakan langkah kritis dalam proses analisis data yang bertujuan untuk mendapatkan wawasan mendalam tentang karakteristik dan struktur *dataset*. Proses ini dimulai dengan pengumpulan data dari berbagai sumber untuk kemudian menjalankan analisis eksploratif. Dalam tahap ini, analis data memeriksa kualitas data, mengidentifikasi nilai yang hilang atau *outlier*, dan merinci distribusi variabel untuk memahami pola yang mungkin ada. Melalui visualisasi data, seperti grafik atau diagram, para analis dapat mengenali hubungan dan korelasi antar variabel, membentuk dasar untuk pemilihan model dan metode analisis selanjutnya (Anjelita et al., 2020; Sudipa, Sarasvananda, et al., 2023). Selain itu, pemahaman data juga melibatkan integrasi pengetahuan domain untuk memastikan interpretasi hasil data sesuai dengan konteks bisnis. Proses ini bukan hanya satu kali, tetapi bersifat iteratif, memungkinkan para analis untuk terus memperdalam pemahaman mereka seiring perjalanan proyek analisis data. Pemahaman data yang kokoh menjadi dasar yang diperlukan sebelum melangkah ke tahap berikutnya dalam analisis data, seperti pemodelan atau pengembangan strategi.

Di era ini yang didominasi oleh ledakan data, pemahaman yang mendalam terhadap informasi menjadi keharusan mutlak. Bab pengantar ini bertujuan untuk membuka pintu gerbang ke

kompleksitas dan keberagaman data, membawa pembaca melintasi lorong-lorong labirin informasi untuk memahami esensi dari proses pemahaman data dalam konteks data *mining*.

Pemahaman data bukan sekadar langkah awal yang biasa, melainkan fondasi yang memastikan keberhasilan seluruh proses data mining. Sub bab awal membahas bagaimana evolusi data telah memunculkan tantangan baru, dan peran utama pemahaman data dalam mengatasi kompleksitas tersebut. Kita akan mengeksplorasi konsep dasar pemahaman data, menyoroti perbedaan antara pemahaman data dan analisis data, serta bagaimana kedua elemen ini saling melengkapi (Hariyono et al., 2023; Muhammad Wali et al., 2023; Sudipa et al., 2022).

Manfaat pemahaman data menjadi jelas di mana kita akan meneliti betapa pentingnya pemahaman data dalam mengidentifikasi pola yang tersembunyi, menghubungkan informasi yang tampaknya terpisah, dan menyaring keberagaman data untuk menemukan inti yang bernilai. Dengan menguraikan aspek-aspek penting ini, pembaca akan mendapatkan gambaran menyeluruh tentang bagaimana pemahaman data berfungsi sebagai fondasi yang kuat dalam data mining (Asana et al., 2022; Kwintiana et al., 2023; Risqi Ananda et al., 2023). Dengan membaca buku ini, diharapkan pembaca akan mampu mengimplementasikan pemahaman data secara efektif, membuka pintu menuju penemuan wawasan berharga yang terkandung dalam setiap titik data yang ada.

Dalam perspektif fundamental, pemahaman data menjadi kunci utama untuk menggali potensi maksimal dari setiap dataset. Dalam bab pengantar ini, penulis akan membahas peran integral pemahaman data dalam proses data mining. Dinamika cepat evolusi data dalam beberapa tahun terakhir menunjukkan bahwa hanya melalui pemahaman yang mendalam kita dapat mengatasi kompleksitas dan potensi kehilangan informasi yang signifikan.

Dengan demikian, tujuan saya sebagai penulis adalah membimbing pembaca melalui langkah-langkah esensial dalam pemahaman data, mulai dari tahap pengumpulan hingga eksplorasi, dengan maksud untuk membentuk landasan yang kokoh bagi pembaca dalam mengeksplorasi potensi yang terkandung dalam dataset mereka.

Konsep dasar pemahaman data, dengan penekanan pada perbedaan antara pemahaman data dan analisis data. Pembaca akan diperkenalkan pada esensi dari pemahaman data dan signifikansinya dalam perjalanan data mining. Pada tahap ini, kita akan mendiskusikan bahwa pemahaman data tidak hanya merupakan langkah awal, melainkan suatu sikap holistik terhadap data yang harus diterapkan dalam setiap tahap analisis.

Langkah selanjutnya dalam eksplorasi ini adalah memahami manfaat konkret dari pemahaman data dalam konteks data mining. Pembaca akan mendapatkan pemahaman tentang bagaimana pemahaman data memberikan fondasi yang kuat untuk mengidentifikasi pola dan hubungan dalam dataset, sambil menyaring informasi yang relevan dan kritis(Arifin, Djumat, et al., 2023; Hayadi et al., 2021; Prahendratno et al., 2023). Kami akan membahas kontribusi pemahaman data dalam mengidentifikasi informasi yang berpotensi membentuk dasar keputusan yang lebih solid.

Dengan memahami pentingnya pemahaman data, pembaca akan diakui untuk melihat setiap aspek data sebagai potensi sumber pengetahuan yang dapat digali. Melalui eksplorasi konsep dan manfaat ini, diharapkan pembaca akan memperoleh pemahaman yang lebih mendalam tentang bagaimana memanfaatkan pemahaman data dalam mencapai keberhasilan dalam proses data mining(Radhitya & Sudipa, 2020; Sudipa, Riana, et al., 2023; Suryawan et al., 2023; Wiguna et al., 2023).

B. FUNGSI DATA UNDERSTANDING DALAM DATA MINING

Data *Understanding* berguna dalam proses Data *Mining*, yang merupakan rangkaian kegiatan untuk menggali pengetahuan yang berharga dari data. Berikut adalah beberapa fungsi penting dari Data *Understanding* dalam konteks Data *Mining* (Wahyuddin et al., 2023):

1. **Identifikasi Pola dan Hubungan :** Data *Understanding* memungkinkan identifikasi pola dan hubungan yang tersembunyi dalam *dataset*. Melalui analisis yang mendalam, kita dapat mengenali struktur atau tren yang mungkin tidak terlihat secara langsung. Dengan demikian, pemahaman data menjadi landasan untuk mengidentifikasi pola analisis data (Dewantara & Giovanni, 2023; Dewi et al., 2021; Saputra et al., 2022; Sudipa, Asana, et al., 2023).
2. **Penentuan Variabel Penting :** Dalam proses Data *Mining*, tidak semua variabel memiliki dampak yang sama terhadap hasil. Data *Understanding* membantu dalam menentukan variabel yang paling relevan dan berpengaruh, sehingga memudahkan fokus pada aspek-aspek kritis yang perlu dianalisis lebih lanjut.
3. **Optimasi Proses Data Mining :** Dengan memahami karakteristik data, kita dapat mengoptimalkan proses Data Mining. Pemahaman tentang struktur dan distribusi data membantu pemilihan model yang sesuai dan parameter yang optimal, meningkatkan performa keseluruhan analisis.
4. **Pembersihan Data :** Sebelum menerapkan teknik Data *Mining*, dataset harus bersih dan bebas dari kesalahan. Pemahaman data memungkinkan identifikasi dan penanganan data yang hilang, *outlier*, atau kesalahan lainnya(Damanik et al., 2022; Simanjuntak et al., 2023). Pembersihan data yang efektif membantu memastikan keandalan hasil analisis.
5. **Perancangan Model yang Efektif :** Pemahaman data menjadi dasar untuk merancang model prediktif atau klasifikasi yang efektif. Dengan mengetahui keunikan dataset,

kita dapat memilih algoritma dan metode yang paling cocok untuk mencapai tujuan analisis yang diinginkan.

6. **Pemilihan Fitur :** Seiring dengan penentuan variabel penting, Data Understanding membantu dalam pemilihan fitur (*feature selection*). Proses ini memilih subset fitur yang paling informatif dan relevan, mengurangi dimensi data dan meningkatkan efisiensi analisis.
7. **Evaluasi Kualitas Data :** Data *Understanding* memungkinkan evaluasi kualitas data secara menyeluruh. Pemahaman tentang distribusi, keberagaman, dan karakteristik lainnya membantu menentukan apakah data memiliki keandalan yang cukup untuk digunakan dalam analisis yang mendalam(N. P. M. E. Putri et al., 2024; R. M. A. Putri et al., 2024).
8. **Pemilihan Model dan Metode Analisis :** Dengan memahami data, kita dapat membimbing pemilihan model dan metode analisis yang paling sesuai dengan sifat dataset. Ini membantu dalam menghindari pemilihan model yang tidak tepat dan meningkatkan interpretabilitas hasil analisis.

Dengan memanfaatkan Data *Understanding* secara efektif, praktisi Data Mining dapat meningkatkan kualitas dan validitas analisis, membimbing pengambilan keputusan yang lebih baik (Afifuddin & Hakim, 2023; Mahendra, Hariyono, et al., 2023; Mahendra, Wardoyo, et al., 2023; Sudipa, Wardoyo, et al., 2023), dan meraih wawasan yang bernilai dari kerumitan data yang terkandung.

C. MANFAAT DATA UNDERSTANDING

Pemahaman data memainkan peran penting dalam proses data mining, memberikan beberapa manfaat yang berkontribusi pada keberhasilan proyek data mining (Agarwal, 2014; Safitri & Bella, 2022). Berikut adalah beberapa keuntungan utama dari pemahaman data dalam data *mining*:

1. **Identifikasi Pola dan Tren :** Dengan memahami data secara menyeluruh, analis dapat mengidentifikasi pola tersembunyi, tren, dan hubungan dalam dataset. Pengetahuan ini penting untuk membuat keputusan dan prediksi yang berbasis informasi.
2. **Seleksi Fitur :** Memahami signifikansi fitur-fitur yang berbeda membantu dalam memilih variabel-variabel yang paling relevan untuk analisis. Proses ini krusial untuk membangun model yang akurat dan efisien, karena fitur yang tidak relevan atau berlebihan dapat dikecualikan.
3. **Optimasi Kinerja Model :** Pemahaman data yang solid memungkinkan pemilihan algoritma dan model yang sesuai dengan karakteristik dataset. Optimasi ini meningkatkan kinerja keseluruhan model data *mining*.
4. **Penanganan Data yang Hilang :** Pemahaman data membantu mengidentifikasi nilai yang hilang dan pencilan, memungkinkan implementasi strategi yang sesuai untuk menanganinya. Hal ini memastikan integritas data dan mencegah dampak negatif terhadap hasil analisis.
5. **Peningkatan Kualitas Data :** Pemahaman data memfasilitasi identifikasi masalah kualitas data, memungkinkan kegiatan pembersihan dan pra-pemrosesan data. Peningkatan kualitas data menghasilkan hasil analisis yang lebih dapat diandalkan dan akurat pada langkah-langkah analisis berikutnya.
6. **Meningkatkn Interpretabilitas :** Memahami konteks dan makna variabel-variabel data berkontribusi pada interpretabilitas hasil data *mining*. Hal ini penting agar pemangku kepentingan dapat mempercayai dan bertindak berdasarkan wawasan yang dihasilkan oleh model data *mining*.
7. **Mengurangi Overfitting :** Dengan pemahaman yang jelas tentang distribusi data, analis dapat membuat keputusan yang lebih informasional tentang kompleksitas model. Ini membantu menghindari overfitting, di mana model berperforma baik pada

data pelatihan tetapi gagal generalisasi pada data baru yang belum pernah dilihat sebelumnya.

8. **Alokasi Sumber Daya yang Efisien :** Mengetahui karakteristik data memungkinkan alokasi sumber daya komputasi dan waktu yang lebih efisien. Hal ini terutama penting dalam proyek data mining berskala besar di mana pemrosesan yang efisien sangat krusial.
9. **Peningkatan Integrasi Pengetahuan Domain :** Pemahaman data memfasilitasi kolaborasi dengan ahli domain. Dengan menggabungkan wawasan data dengan pengetahuan domain, hasil data mining menjadi lebih relevan dan dapat diimplementasikan dalam aplikasi dunia nyata.
10. **Mengurangi Bias :** Memahami data membantu mengidentifikasi bias potensial yang mungkin ada dalam dataset. Kesadaran ini memungkinkan implementasi strategi untuk mengurangi bias dan memastikan keadilan dalam hasil analisis.

D. KONSEP DASAR DATA UNDERSTANDING

Konsep dasar Data *Understanding* adalah fondasi penting dalam proses analisis data, khususnya dalam konteks Data *Mining*. Berikut adalah beberapa elemen kunci yang membentuk konsep dasar Data *Understanding*:

1. **Pengumpulan Data :** Konsep dasar pertama dari Data *Understanding* adalah pengumpulan data. Ini melibatkan proses mengakuisisi data dari berbagai sumber yang relevan dengan tujuan analisis yang diinginkan. Pengumpulan data harus memperhitungkan keakuratan, kelengkapan, dan representativitas dataset.
2. **Pemahaman Terhadap Variabel** : Penting untuk memahami setiap variabel dalam dataset, termasuk jenis variabelnya (kategorikal, numerik, dan lainnya.), unit pengukuran, dan interpretasinya. Pemahaman mendalam terhadap variabel

membantu dalam menentukan implikasi dan hubungan antar variabel.

3. **Eksplorasi Statistik Deskriptif :** Analisis statistik deskriptif merupakan konsep dasar untuk memahami karakteristik utama dari data. Rangkuman statistik seperti mean, median, dan deviasi standar digunakan untuk memberikan gambaran umum tentang distribusi dan variasi data.
4. **Visualisasi Data :** Konsep dasar Data *Understanding* mencakup penggunaan visualisasi data. Grafik dan plot dapat membantu memperjelas pola, tren, dan anomali dalam data, sehingga mempermudah interpretasi data oleh praktisi analisis.
5. **Identifikasi dan Penanganan Outlier :** Pengidentifikasian dan penanganan outlier (poin data yang berbeda secara signifikan dari mayoritas) adalah konsep dasar yang penting. Outlier dapat mempengaruhi hasil analisis, dan oleh karena itu, pemahaman tentang cara menangani atau melaporkannya menjadi esensial.
6. **Pemahaman Distribusi Data :** Pemahaman tentang distribusi data membantu dalam mengidentifikasi apakah data terdistribusi normal atau memiliki karakteristik distribusi lainnya. Hal ini dapat memengaruhi pilihan model statistik atau algoritma machine learning yang tepat.
7. **Pemahaman Hubungan Antar Variabel :** Konsep dasar Data Understanding melibatkan analisis korelasi dan hubungan antar variabel. Mengidentifikasi apakah variabel-variabel tersebut saling berhubungan atau tidak(Arifin et al., 2018; Arifin, Prajayanti, et al., 2023; Rony et al., 2023) dapat membantu dalam menentukan arah analisis yang lebih mendalam.
8. **Pemilihan dan Penanganan Missing Value :** Ketika data memiliki nilai yang hilang, penting untuk memahami cara menangani missing value. Ini melibatkan pemilihan metode imputasi yang tepat atau pertimbangan terhadap penghapusan data yang hilang.

9. **Penentuan Frekuensi dan Keberagaman Kategori :** Jika data mencakup variabel kategorikal, pemahaman tentang frekuensi dan keberagaman kategori dalam setiap variabel kategorikal adalah konsep dasar yang diperlukan. Ini dapat memberikan wawasan tentang representasi dan variasi dalam data.
10. **Pemahaman Domain dan Konteks Bisnis :** Pemahaman tentang domain dan konteks bisnis merupakan unsur kritis dalam konsep dasar Data *Understanding*. Hal ini memungkinkan praktisi untuk mengaitkan temuan analisis dengan kebutuhan dan tujuan bisnis secara lebih baik.

Dengan memahami konsep dasar Data *Understanding*, praktisi data dapat memastikan bahwa data yang digunakan dalam analisis memiliki kualitas yang memadai, dan interpretasi hasil analisis menjadi lebih kontekstual dan bermakna.

E. KELEBIHAN DAN KEKURANGAN DATA UNDERSTANDING

❖ **Kelebihan Data Understanding**

1. **Informasi yang Akurat :** Pemahaman data membantu memastikan bahwa informasi yang diperoleh dari dataset merupakan representasi yang akurat dan dapat diandalkan.
2. **Optmimasi Model :** Mengidentifikasi karakteristik data memungkinkan pemilihan model dan algoritma yang paling cocok untuk mencapai kinerja optimal.
3. **Reduksi Risiko :** Dengan memahami data, risiko kesalahan interpretasi atau pemodelan yang buruk dapat dikurangi, meningkatkan keandalan hasil analisis.
4. **Efisiensi Pengolahan Data :** Memahami struktur data membantu dalam pengolahan data yang lebih efisien, mempercepat langkah-langkah analisis selanjutnya.

5. **Pemilihan Fitur yang Tepat :** Memahami data memungkinkan seleksi fitur yang tepat, menghindarkan penggunaan variabel yang tidak relevan dan dapat meningkatkan performa model.
6. **Peningkatan Interpretabilitas :** Interpretabilitas hasil analisis meningkat karena pemahaman yang baik tentang data memungkinkan penjelasan yang lebih baik kepada pemangku kepentingan.
7. **Peningkatan Kualitas Model :** Pemahaman data memberikan dasar untuk pra-pemrosesan data yang efektif, yang pada gilirannya dapat meningkatkan kualitas model.

❖ **Kekurangan Data Understanding**

1. **Kompleksitas Proses :** Pemahaman data merupakan tahap yang kompleks dan membutuhkan waktu. Proses ini bisa menjadi tantangan, terutama dalam proyek dengan volume data besar.
2. **Keterbatasan Sumber Daya :** Memahami data dengan baik dapat memerlukan sumber daya komputasi dan manusia yang signifikan, terutama dalam hal pemrosesan dan interpretasi.
3. **Keterbatasan Informasi :** Beberapa dataset mungkin tidak memberikan informasi yang cukup untuk pemahaman yang mendalam, terutama jika data tersebut kurang representatif atau tidak lengkap.
4. **Keterbatasan Waktu :** Dalam situasi di mana waktu sangat terbatas, pemahaman data yang mendalam mungkin sulit dicapai, dan keputusan harus diambil dengan cepat.
5. **Perubahan Sifat Data :** Data dapat berubah seiring waktu, dan pemahaman yang dibangun pada awal proyek mungkin perlu diperbarui untuk mencerminkan perubahan tersebut.
6. **Kompleksitas Penanganan Anomali :** Identifikasi dan penanganan anomali dalam data seringkali memerlukan pemahaman yang mendalam dan bisa menjadi tugas yang rumit.

7. **Tantangan Integrasi dengan Domain :** Memahami data seringkali memerlukan kolaborasi dengan ahli domain, dan terkadang, kesulitan dapat muncul dalam mengintegrasikan pemahaman teknis dengan pengetahuan domain.

Penting untuk diingat bahwa sementara pemahaman data memiliki kelebihan dan kekurangan, tahap ini tetap menjadi langkah penting untuk memastikan bahwa proses data mining berjalan dengan baik dan hasilnya dapat diandalkan.

F. RANGKUMAN

Pemahaman data dalam data *mining* memainkan peran krusial sebagai fondasi bagi kesuksesan analisis data. Proses ini melibatkan pengumpulan data, analisis eksploratif, dan visualisasi untuk mengidentifikasi pola, hubungan, serta outliers dalam dataset. Tidak hanya sebagai langkah awal, pemahaman data menjadi fondasi iteratif yang memastikan keberhasilan seluruh proses analisis. Fungsinya mencakup identifikasi pola tersembunyi, penentuan variabel penting, optimasi proses data mining, pembersihan data, perancangan model yang efektif, pemilihan fitur, evaluasi kualitas data, dan membimbing pemilihan model.

Manfaat dari pemahaman data sangat signifikan dalam mengidentifikasi pola, memilih fitur yang relevan, dan meningkatkan kualitas model data mining. Selain itu, konsep dasar pemahaman data mencakup pengumpulan data, pemahaman variabel, eksplorasi statistik deskriptif, visualisasi data, identifikasi outlier, pemahaman distribusi data, hubungan antar variabel, pemilihan dan penanganan nilai yang hilang, penentuan frekuensi kategori, serta pemahaman domain dan konteks bisnis.

Kelebihan pemahaman data melibatkan informasi yang akurat, optimasi model, reduksi risiko, efisiensi pengolahan data, pemilihan fitur yang tepat, peningkatan interpretabilitas, dan peningkatan kualitas model. Meskipun demikian, terdapat kekurangan seperti kompleksitas proses, keterbatasan sumber daya, keterbatasan informasi, keterbatasan waktu, perubahan sifat data, kompleksitas penanganan anomali, dan tantangan integrasi dengan domain. Namun, pemahaman data tetap menjadi langkah penting untuk memastikan hasil analisis data mining yang dapat diandalkan.

G. STUDI KASUS

Studi Kasus: Analisis Pemasaran Produk *E-commerce*
Sebuah perusahaan e-commerce yang beroperasi secara global ingin meningkatkan efektivitas kampanye pemasaran mereka. Mereka memiliki dataset besar yang mencakup informasi tentang transaksi, preferensi pelanggan, dan hasil kampanye pemasaran sebelumnya. Perusahaan ini menginginkan pemahaman data yang mendalam untuk mengidentifikasi pola-pola yang dapat meningkatkan target pemasaran dan mengoptimalkan strategi penjualan.

❖ **Langkah-Langkah Data *Understanding*:**

1. Pengumpulan Data : Mengumpulkan data transaksi, preferensi pelanggan, dan data kampanye pemasaran dari sumber-sumber internal dan eksternal.
2. Pemahaman Terhadap Variabel : Menilai jenis variabel yang ada dalam dataset, seperti variabel numerik (jumlah transaksi, total belanja), variabel kategorikal (kategori produk, metode pembayaran), dan variabel waktu (tanggal transaksi, waktu respons kampanye).
3. Eksplorasi Statistik Deskriptif : Menggunakan statistik deskriptif untuk merangkum karakteristik utama dari data,

termasuk mean, median, dan deviasi standar untuk variabel numerik.

4. Visualisasi Data : Membuat grafik dan diagram untuk memvisualisasikan pola-pola seperti tren penjualan, preferensi produk, dan efektivitas kampanye.
5. Identifikasi dan Penanganan Outlier : Mengidentifikasi nilai yang di luar norma dalam jumlah transaksi atau total belanja, dan memutuskan apakah nilai tersebut memerlukan penanganan khusus.
6. Pemahaman Distribusi Data : Menganalisis distribusi data untuk memahami apakah data penjualan terdistribusi normal atau memiliki karakteristik distribusi lainnya.
7. Pemahaman Hubungan Antar Variabel : Menganalisis korelasi antara variabel-variabel, seperti apakah produk tertentu lebih sering dibeli bersamaan atau apakah metode pembayaran mempengaruhi jumlah transaksi.
8. Pemilihan dan Penanganan Missing Value : Memeriksa apakah ada nilai yang hilang dalam dataset dan menentukan apakah perlu dilakukan imputasi atau penghapusan.
9. Penentuan Frekuensi dan Keberagaman Kategori : Menentukan frekuensi kategori produk atau metode pembayaran dan memastikan keberagaman dalam representasi kategori tersebut.
10. Pemahaman Domain dan Konteks Bisnis : Berkolaborasi dengan tim pemasaran untuk memahami konteks bisnis, seperti apakah ada preferensi geografis pelanggan atau tren musiman tertentu.

❖ **Jawaban Terhadap Temuan Pemahaman Data:**

1. Identifikasi Pola dan Hubungan : Mengidentifikasi bahwa produk X sering dibeli bersamaan dengan produk Y, yang dapat digunakan untuk mengoptimalkan penempatan produk atau menawarkan bundel diskon.
2. Penentuan Variabel Penting : Menemukan bahwa jenis produk dan metode pembayaran adalah variabel yang sangat

mempengaruhi total belanja, memungkinkan penyesuaian strategi pemasaran.
3. Optimasi Proses Data Mining : Memilih model dan algoritma yang paling sesuai berdasarkan karakteristik dataset, meningkatkan efisiensi analisis.
4. Pembersihan Data : Menangani nilai yang hilang dan outlier untuk memastikan hasil analisis lebih dapat diandalkan.
5. Perancangan Model yang Efektif : Membangun model prediktif untuk memprediksi preferensi pelanggan atau respons terhadap kampanye pemasaran.
6. Pemilihan Fitur : Memilih fitur yang paling relevan untuk meningkatkan akurasi model dan mengurangi dimensi data.
7. Evaluasi Kualitas Data : Mengevaluasi kualitas data untuk memastikan keandalan hasil analisis.
8. Membimbing Pemilihan Model dan Metode Analisis : Memberikan panduan pada pemilihan model yang sesuai dengan sifat dataset dan tujuan analisis.

❖ **Manfaat dari Data Understanding :**

1. Mengidentifikasi pola pembelian pelanggan.
2. Menyesuaikan strategi pemasaran berdasarkan preferensi produk.
3. Meningkatkan efektivitas kampanye dengan mengoptimalkan waktu dan metode pemasaran.
4. Menyaring informasi kritis untuk meningkatkan target pemasaran.

❖ **Kesimpulan :**

Dengan langkah-langkah pemahaman data yang mendalam, perusahaan e-*commerce* dapat meningkatkan strategi pemasaran mereka, meningkatkan kepuasan pelanggan, dan mencapai keberhasilan dalam dunia yang kompetitif. Pemahaman data menjadi pondasi untuk menggali wawasan berharga dari dataset besar, membimbing keputusan yang lebih baik, dan meningkatkan efektivitas operasional.

H. TEST FORMATIF

Berikut 10 latihan soal formatif desain pembelajaran

1. Apa yang dimaksud dengan *data mining?*
   a. Proses ekstraksi informasi dari dataset kecil
   b. Proses analisis data untuk mendapatkan pemahaman yang lebih dalam
   c. Proses pembersihan data dari outlier
   d. Proses pembuatan dataset besar
2. Mengapa *data understanding* menjadi langkah kritis dalam proses analisis data?
   a. Karena hanya dilakukan sekali pada awal proyek
   b. Karena melibatkan integrasi pengetahuan domain
   c. Karena hanya fokus pada pemilihan model
   d. Karena tidak berhubungan dengan pemahaman data mining
3. Apa fungsi utama dari *data understanding* dalam *data mining*?
   a. Menentukan warna grafik yang digunakan
   b. Menentukan variabel yang paling penting
   c. Menjalankan analisis eksploratif sekali saja
   d. Menghitung nilai mean dari dataset
4. Manfaat apa yang didapatkan dari pemahaman data dalam mengidentifikasi pola dan hubungan?
   a. Menentukan variabel penting
   b. Memilih fitur yang relevan
   c. Mengoptimalkan kinerja model
   d. Menentukan frekuensi kategori
5. Apa yang menjadi konsep dasar pertama dari *data understanding*?
   a. Pemahaman distribusi data
   b. Pemahaman domain dan konteks bisnis
   c. Pengumpulan data
   d. Eksplorasi statistik deskriptif

6. Kelebihan apa yang diperoleh dari pemahaman data dalam proses *data mining?*
    a. Keterbatasan informasi
    b. Reduksi risiko kesalahan interpretasi
    c. Keterbatasan sumber daya
7. Apa yang termasuk dalam fungsi *data understanding?*
    a. Menentukan kelebihan dan kekurangan
    b. Mengidentifikasi pola dan hubungan
    c. Menghapus semua outlier
    d. Menentukan warna grafik yang digunakan
8. Mengapa pemahaman data memainkan peran penting dalam mengurangi bias?
    a. Karena memilih variabel yang tidak relevan
    b. Karena mengidentifikasi pola tersembunyi
    c. Karena mengurangi keberagaman kategori
    d. Karena mengenali bias potensial dalam dataset
9. Apa kelemahan yang mungkin muncul dalam proses *data understanding* ?
    a. Peningkatan kualitas model
    b. Keterbatasan waktu
    c. Identifikasi dan penanganan outlier yang mudah
    d. Perubahan sifat data yang terdapat dalam setiap proyek
10. Apa yang dilibatkan dalam konsep dasar *data understanding* untuk pemahaman terhadap variabel?
    a. Penentuan variabel yang paling penting
    b. Penanganan nilai yang hilang
    c. Evaluasi kualitas data
    d. Pemahaman domain dan konteks bisnis

I. LATIHAN

1. Jelaskan peran data understanding dalam proses data mining dan mengapa dianggap sebagai langkah krusial.
2. Diskusikan, bagaimana visualisasi data dapat membantu analis data dalam memahami pola dan hubungan antar variabel. Berikan contoh penggunaan visualisasi data yang efektif.
3. Apa saja fungsi utama Data Understanding dalam konteks Data Mining? Berikan penjelasan mendalam mengenai salah satu fungsi tersebut dan bagaimana dampaknya pada kualitas analisis.
4. Jelaskan konsep dasar Data Understanding, termasuk elemen-elemen kunci seperti pengumpulan data, eksplorasi statistik deskriptif, dan pemahaman terhadap variabel. Mengapa konsep-konsep ini dianggap penting dalam analisis data?
5. Sebutkan dan jelaskan dua kelebihan dan dua kekurangan utama dari Data Understanding dalam proses data mining. Bagaimana praktisi data dapat mengatasi atau memitigasi kekurangan tersebut?

KEGIATAN BELAJAR 3

KNOWLEDGE REPRESENTATION

DESKRIPSI PEMBELAJARAN

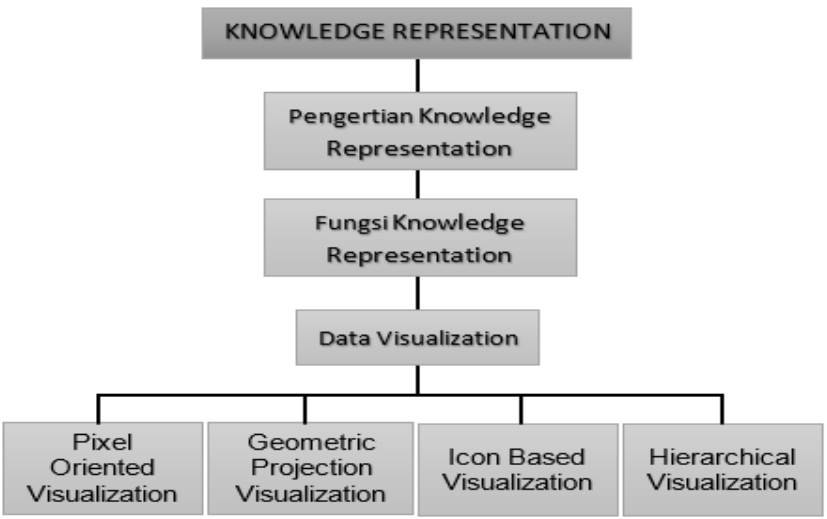
Pada bab ini mahasiswa mempelajari representasi atau penyajian pengetahuan dalam data mining. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk memvisualisasikan hasil data mining dalam berbagai bentuk atau teknik representasi pengetahuan.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

1. Mampu menguraikan definisi knowledge representation.
2. Mempu menjelaskan berbagai teknik knowledge representation

PETA KONSEP PEMBELAJARAN



A. PENGERTIAN KNOWLEDGE REPRESENTATION

Knowledge representation merupakan salah satu tahapan proses pada *knowledge discovery*. Pada tahapan ini, data yang ada di dalam *knowledge based* akan ditampilkan atau diperlihatkan berkaitan dengan metode yang digunakan kepada pengguna yang melakukan *knowledge discovery* tersebut (Wahyuddin et al., 2023).

Knowledge representation dapat diartikan sebagai visualisasi dan penyajian pengetahuan dari proses yang telah dilakukan untuk membantu pengguna dalam memahami pengetahuan hasil *data mining* (Purwati et al., 2023).

B. FUNGSI KNOWLEDGE REPRESENTATION

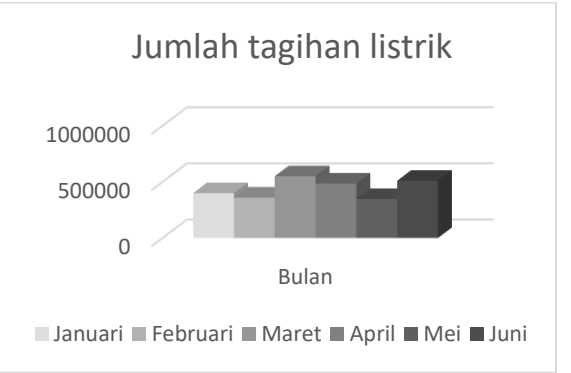
Visualisasi data dapat membantu pengguna untuk memahami data dengan lebih baik, mengungkap pola-pola yang tidak terlihat dalam tabel atau angka, dan memudahkan dalam pengambilan keputusan (Nurirwan Saputra, n.d.).

Knowledge representation dapat divisualisasikan dalam bentuk diagram pohon, diagram batang, tabel, *rules graphs, charts* atau bagan, diagram garis, diagram lingkaran, matriks dan lain sebagainya.

Teknik ini dapat membantu pengguna untuk memahami hubungan antara variabel, melihat perubahan dalam data, dan mengidentifikasi pola atau anomali dalam data.

Contoh: *Histogram*
Histogram memberikan representasi distribusi nilai suatu atribut, yang terdiri dari satu set persegi panjang, yang mencerminkan jumlah atau frekuensi kelas yang ada dalam data yang diberikan.

Contoh *Histogram* tagihan listrik yang dihasilkan selama 12 bulan, seperti terlihat pada gambar 3.1 berikut:

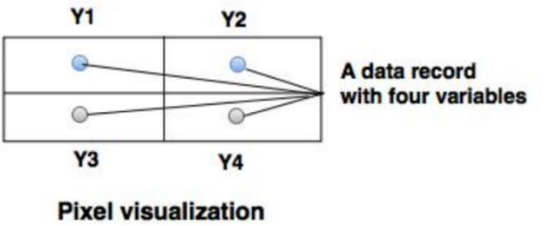


Gambar 3.1: Contoh *Histogram*

C. DATA VISUALIZATION

Visualisasi data adalah penyajian atau representasi data dalam format gambar atau grafis (Jeovano, 2020). Dengan menggunakan teknik visualisasi data, pola dalam data ditandai dengan mudah. Beberapa teknik visualisasi data penting antara lain:

1. Teknik visualisasi berorientasi *pixel (Pixel Oriented Visualization Technique)*
   Teknik visualisasi berorientasi *pixel* yaitu memetakan tiap-tiap data *value* ke titik tertentu dan setiap data *value* memiliki satu atribut dalam jendela terpisah seperti pada gambar 3.2 berikut:



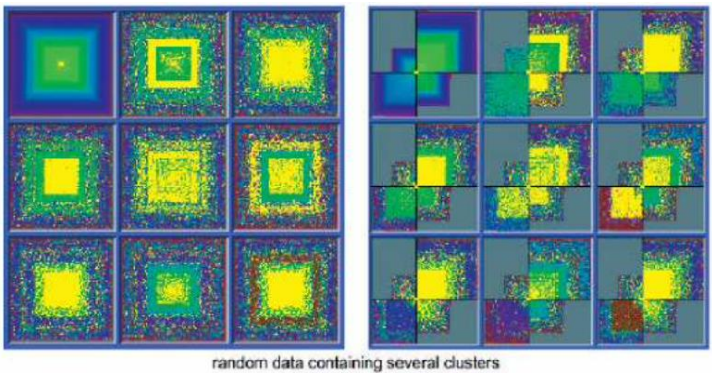
Gambar 3.2: Visualisasi *Pixel*

Visualisasi *pixel* daapat memaksimalkan jumlah informasi yang disajikan pada satu waktu tanpa tumpang tindih.

Tuple dengan variabel 'm' memiliki *pixel* berwarna 'm' yang berbeda untuk mewakili setiap variabel dan setiap variabel memiliki *sub* jendela.

Pemetaan warna *pixel* ditentukan berdasarkan karakteristik data dan tugas visualisasi.

Secara umum teknik ini menggunakan satu titik setiap satu data value, maka teknik ini dapat mengalokasikan visualisasi data dalam ukuran yang sangat besar, yaitu memungkinkan untuk menampilkan lebih dari 1.000.000 data *value*. Atribut yang berbeda ditampilkan pada *sub* jendela yang berbeda dan *range* nilai data yang mungkin dipetakan pada titik sesuai warna tertentu seperti pada gambar 3.3 berikut (Mulyana et al., 2009):



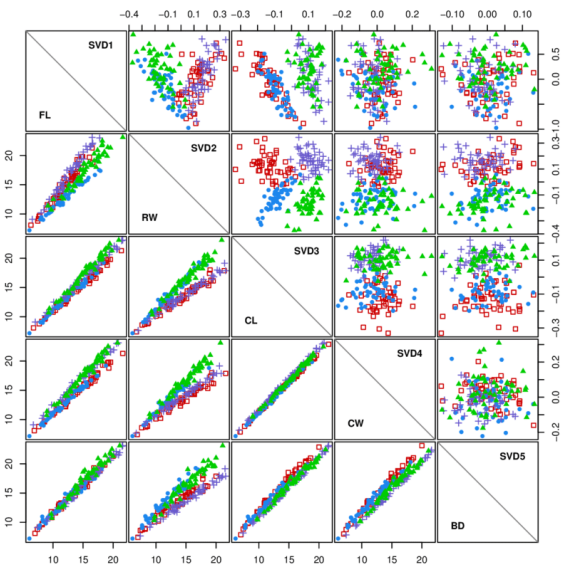
Gambar 3.3: Visualisasi berbasis *pixel* menggunakan spiral(kiri) dan sumbu(kanan)

2. Teknik visualisasi proyeksi geometris (*Geometric Projection Visualization Technique*)
   Teknik yang digunakan untuk mencari transformasi geometri adalah:

a. Matrks plot sebar (*Scatter Plot Matrix*)

Scatter Plot Matrix adalah teknik terkenal untuk analisis visual data berdimensi tinggi yang terdiri dari plot sebar dari semua kemungkinan pasangan variabel dalam kumpulan data. Saja salah satu kekurangan dari *Scatter Plot Matrix* adalah biasanya tidak semua tampilan berpotensi relevan dengan tugas analisis atau pengguna tertentu.

Matriks plot sebar menunjukkan semua plot sebar berpasangan dari atribut pada satu tampilan dengan beberapa plot sebar dalam format matriks (W. B. Wang et al., 2016). Contoh matriks plot sebar dapat dilihat pada gambar 3.4 berikut:



Gambar 3.4: Matriks plot sebar
(Scrucca & Raftery, 2015)

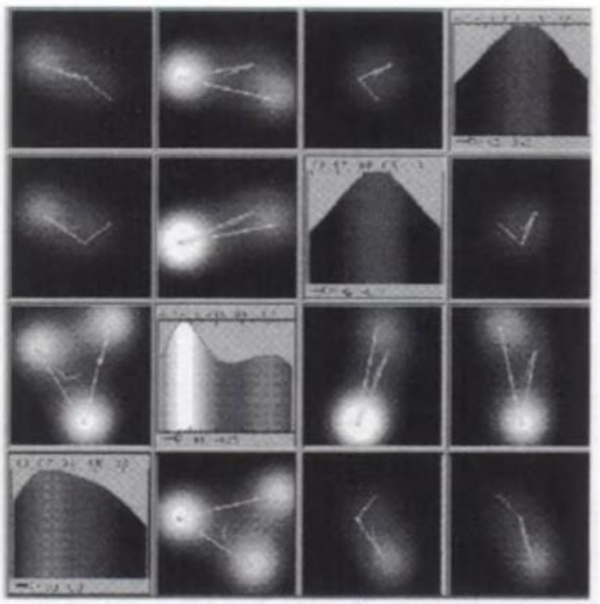
Pada contoh matriks plot sebar pada gambar 3.4 menampilkan data Kepiting. Panel bawah menampilkan plot sebar untuk pasangan variabel dalam skala asli,

sedangkan panel atas menunjukkan fitur yang diperoleh dengan menerapkan transformasi SVD berskala.

b. Hyper slice

Hyper slice adalah perluasan dari matriks plot sebar yang mewakili multidimensi dan berfungsi sebagai matriks irisan dua dimensi ortogonal.

Hyper Slice adalah panel matriks di mana “irisan” fungsi *multivariate* ditampilkan pada titik fokus tertentu (c1, c2, c3...cn) yang memungkinkan seseorang menelusuri fungsi *multivariate* secara interaktif (Hoffman & Grinstein, 1997). Contoh *hyper slice* dapat dilihat pada gambar 3.5 berikut:



Gambar 3.5: Contoh *Hyper Slice*
(GRINSTEIN, 2002)

c. Koordinat Paralel (*Parallel coordinates*)

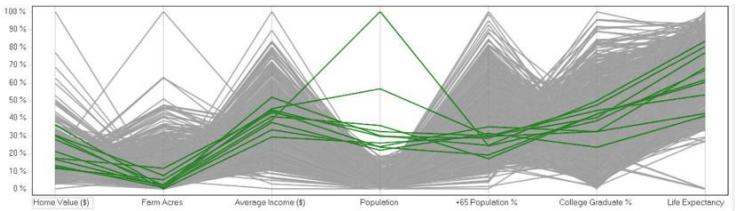
Koordinat Paralel merupakan teknik visualisasi yang digunakan untuk memplot elemen data individual di berbagai ukuran kinerja. Masing-masing ukuran

berhubungan dengan sumbu vertikal dan setiap elemen data ditampilkan sebagai serangkaian titik yang terhubung di berbagai ukuran/sumbu.

Garis vertikal sejajar yang dipisahkan sumbu atau sebuah titik dalam koordinat *Cartesian* yang terhubung dengan *poly line* dalam koordinat paralel.

Koordinat Paralel merupakan teknik visualisasi yang banyak digunakan untuk data *multivariate* dan geometri dimensi tinggi (Heinrich & Weiskopf, 2013).

Visualisasi koordinat paralel terkenal sebagai representasi yang sulit dipahami dan hanya digunakan oleh ahli (Siirtola & Räihä, 2006). Contoh visualisasi koordinat paralel dapat dilihat pada gambar 3.6 berikut:

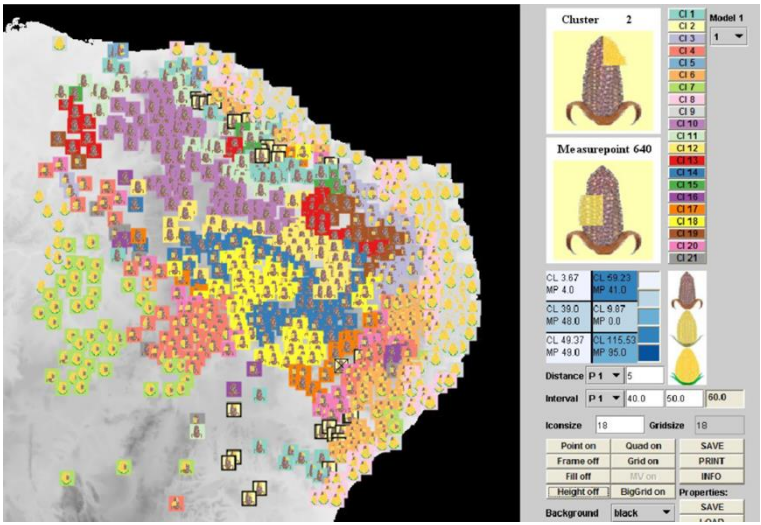


Gambar 3.6: Contoh koordinat paralel
(Few, 2006)

Sebagian besar garis pada koordinat paralel digunakan untuk menyandikan data deret waktu. Perubahan sepanjang waktu dari satu nilai ke nilai berikutnya ditunjukkan oleh kemiringan garis yang naik dan turun. Garis-garis dalam tampilan koordinat paralel tidak menunjukkan perubahan, akan tetapi satu garis dalam grafik koordinat paralel menghubungkan serangkaian nilai masing-masing terkait dengan variabel berbeda yang mengukur berbagai aspek dari suatu hal, seperti orang, produk, atau negara (Few, 2006).

3. Teknik Visualisasi Berbasis Ikon (*Icon Based Visualization Technique*)

   Teknik Visualisasi Berbasis Ikon memetakan variabel data ke atribut visual geometris (misalnya, bentuk, ukuran, orientasi) dan non geometris (misalnya, warna dan tekstur) (Nocke et al., 2005). Contoh visualisasi berbasis ikon dapat dilihat pada gambar 3.7 berikut:



Gambar 3.7: Contoh visualisasi berbasis ikon (Nocke et al., 2005)

4. Teknik Visualisasi Hierarki (*Hierarchical Visualization Technique*)

   Teknik Visualisasi Hierarki adalah teknik visualisasi data menggunakan partisi hierarki ke dalam *sub* ruang.

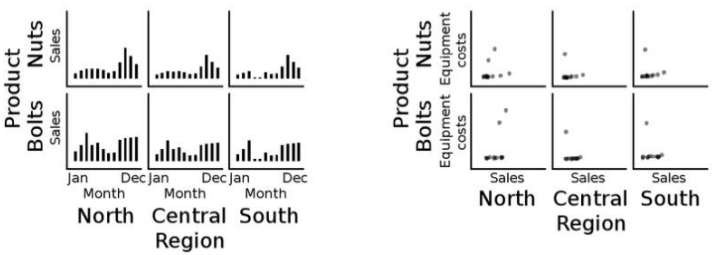
   Dalam mengeksplorasi hubungan hierarki secara visual, dapat menerapkan beberapa metode visualisasi seperti diagram *node link* dan peta pohon (H. Wang et al., 2021). Untuk diagram *node link*, *node* mewakili objek data dan garis penghubung antar *node* mewakili hubungan hierarki. Tata letak diagram *node link* dapat dibagi dua jenis yaitu:

a. Tata letak ortogonal
b. Tata letak radial

Untuk metode berbasis area, peta pohon, hierarki dikodekan dengan penyertaan dan nilai divisualisasikan berdasarkan area. Jumlah nilai dari *node* anak menentukan nilai induknya dalam hierarki. Beberapa metode yang dapat digunakan antara lain:

a. Dimensional Stacking

Dimensional stacking memungkinkan lebih dari satu variabel kategori dipetakan ke sumbu spasial yang sama, dan digunakan dalam visualisasi *database* (Im et al., 2013). *Dimensional stacking* dapat diartikan melakukan partisi ruang atribut n dimensi dalam *sub* ruang 2-D, yang 'ditumpuk' satu sama lain. Contoh *dimensional stacking* dapat dilihat pada gambar 3.8 berikut:

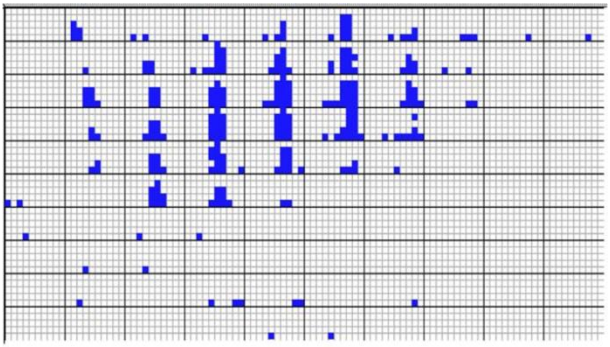


Gambar 3.8: Contoh *Dimensional Stacking* Produk (Im et al., 2013)

Contoh *dimensional stacking* pada gambar 3.8 merupakan *dimensional stacking* data mur dan baut. Kiri: Produk dan Penjualan dipetakan ke sumbu vertikal, Wilayah dan Bulan dipetakan ke sumbu horizontal. Kanan: Biaya Produk dan Peralatan dipetakan ke vertikal, Wilayah dan Penjualan ke horizontal.

Contoh lain visualisasi data tambang minyak dengan garis bujur dan lintang yang dipetakan ke sumbu x, y dan kadar

bijih terluar, serta kedalaman yang dipetakan ke sumbu x, y *axes* seperti pada gambar 3.9 berikut (Han et al., 2006):

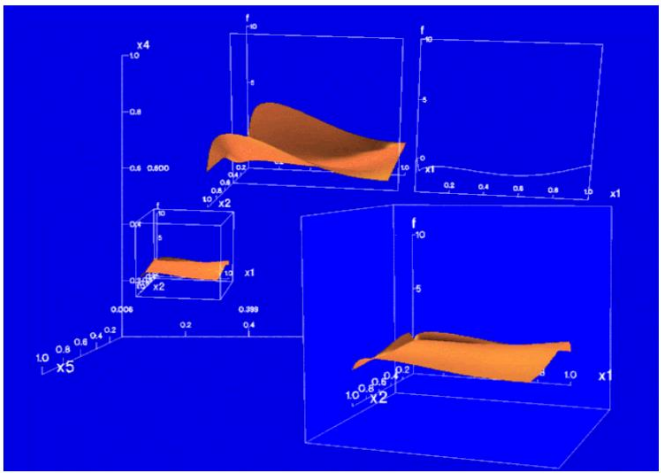


Gambar 3.9: Contoh *Dimensional Stacking* Tambang Minyak

b. Worlds Within Worlds

Tetapkan fungsi dan dua parameter terpenting ke *world* terdalam kemudian perbaiki semua parameter lainnya pada nilai konstan gambar lainnya (memilih 1 atau 2 atau 3 dimensi world sebagai sumbu) (Han et al., 2006).

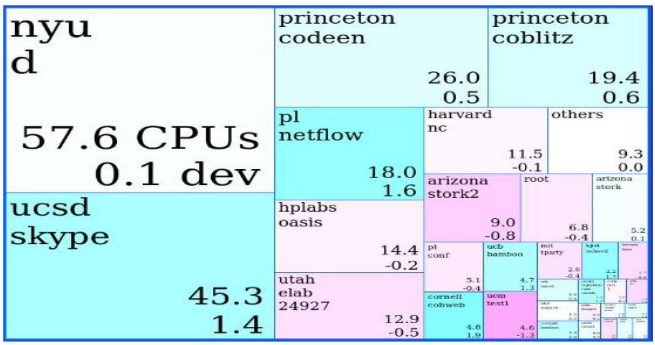
Software yang dapat digunakan yaitu N vision dan Visual Otomatis. Contoh worlds-within-worlds dapat dilihat pada gambar 3.10 berikut:



Gambar 3.10: Contoh *worlds Within Worlds* (Card & Mackinlay, 1997)

c. *Tree* map

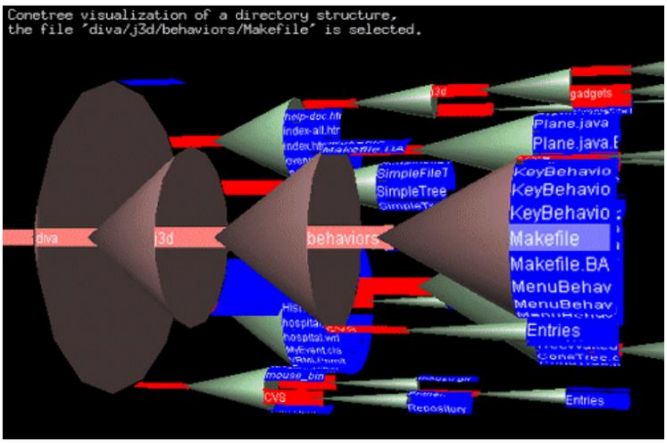
Tree map adalah metode pengisian layar yang menggunakan partisi hierarkis layar menjadi beberapa wilayah bergantung pada nilai atribut. Dimensi x dan y layar dipartisi secara bergantian sesuai dengan nilai atribut (kelas) (Kopanakis, 2003). Contoh tree map dapat dilihat pada gambar 3.11 berikut:



Gambar 3.11: Contoh Treemap beban CPU (Schnorr et al., n.d.-a)

d. *3D Cone Trees*

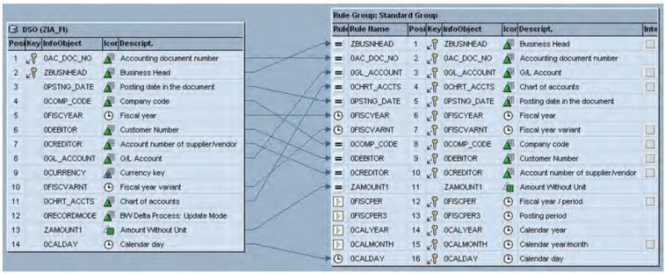
3D Cone trees adalah teknik visualisasi pohon kerucut 3D yang bekerja dengan baik hingga sekitar seribu *node*. Pertama-tama buatlah pohon lingkaran 2D yang menyusun simpul-simpulnya dalam lingkaran konsentris dan berpusat pada simpul akar. Pada 3D-Cone *trees* tidak dapat menghindari tumpang tindih saat diproyeksikan ke 2D (Pleuss et al., 2011). Contoh 3D cone trees dapat dilihat pada gambar 3.12 berikut:



Gambar 3.12: Contoh 3D-Cone *Trees* sistem *file* (Schnorr et al., n.d.-b)

e. InfoCube

Teknik visualisasi 3-D dimana informasi hierarki ditampilkan sebagai kubus *semi transparan* bertumpuk. Kubus terluar berhubungan dengan data tingkat atas, sedangkan subnode atau data tingkat bawah direpresentasikan sebagai kubus kecil di dalam kubus terluar, dan seterusnya (Sastry et al., 2013) , contoh:



Gambar 3.13: Contoh *InfoCube*

D. RANGKUMAN

Berdasarkan uraian di atas *knowledge representation* adalah visualisasi dan penyajian pengetahuan dari proses yang telah dilakukan untuk membantu pengguna dalam memahami pengetahuan hasil *data mining*. Fungsi *knowledge representation* yaitu dapat membantu pengguna untuk memahami data dengan lebih baik, memahami hubungan antara variabel, melihat perubahan dalam data, dan mengidentifikasi pola atau anomali dalam data, serta mengungkap pola-pola yang tidak terlihat dalam bentuk tabel atau angka, diagram pohon, diagram batang, tabel, *rules graphs, charts* atau bagan, diagram garis, diagram lingkaran dan matriks. Visualisasi data adalah representasi data dalam format gambar atau grafis dengan menggunakan teknik visualisasi data antar lain *Pixel Oriented Visualization Technique*, *Geometric Projection Visualization Technique, Icon Based Visualization Technique* ataupun *Hierarchical Visualization Technique.*

E. TES FORMATIF

1. Fungsi *knowledge representation* adalah membantu pengguna dalam hal:
   a. Pemahaman data yang lebih baik
   b. Memahami hubungan antara variabel
   c. Mengetahui perubahan dalam data

d. Mengidentifikasi dan mengungkap pola atau anomali dalam data
e. Benar semua

2. Teknik visualisasi mana yang tidak termasuk teknik visualisasi data mining?
   a. Pixel Oriented Visualization Technique
   b. Icon Based Visualization Technique
   c. Geometric Projection Visualization Technique
   d. Hierarchical Visualization Technique
   e. Benar semua

F. LATIHAN

Buatlah salah satu contoh visualisasi *data mining* menggunakan salah satu teknik visualisasi data, jelaskan!

KEGIATAN BELAJAR 4

DATA PREPROCESSING

DESKRIPSI PEMBELAJARAN

Pada bab ini, mahasiswa akan diajak untuk memahami esensi dan konsep dasar teoritis dari proses Data Preprocessing. Pembelajaran ini bertujuan memberikan mahasiswa wawasan yang kuat dan pemahaman mendalam terkait prinsip-prinsip dasar yang melandasi tahapan-tahapan penting dalam menyiapkan data untuk analisis atau penggunaan dalam model pembelajaran mesin.

Keberhasilan memahami Data Preprocessing diharapkan akan memberikan mahasiswa modal dasar yang kokoh, memberikan landasan yang mendukung untuk mengeksplorasi dan menguasai konsep-konsep bahasa pemrograman lebih lanjut dalam konteks pengolahan data.

KOMPETENSI PEMBELAJARAN

Setelah menyelesaikan perkuliahan ini, diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan sebagai berikut :

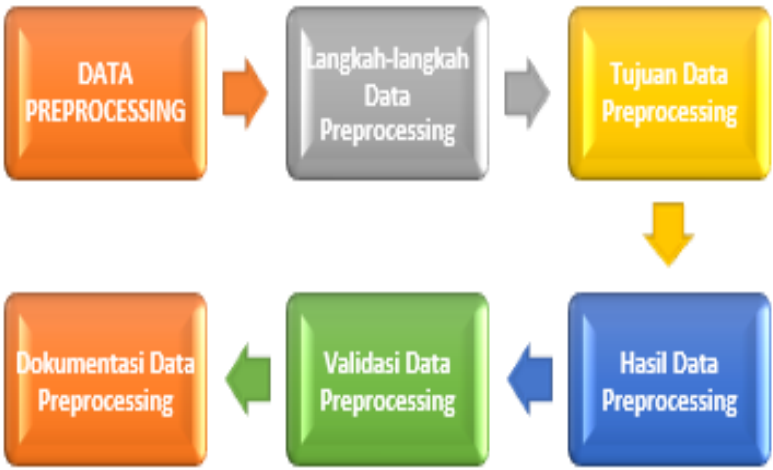
1. **Mampu Menguraikan Definisi Data Preprocessing**
   Mahasiswa diharapkan dapat merinci konsep dasar Data Preprocessing dan memahami langkah-langkah esensial dalam mempersiapkan data untuk analisis atau pemodelan. Ini melibatkan pemahaman mendalam terkait dengan membersihkan, mentransformasi, dan mengorganisir data secara sistematis.
2. **Mampu Menjelaskan Fungsi dan Manfaat Data Preprocessing**
   Setelah mengikuti pembelajaran, mahasiswa diharapkan mampu menjelaskan secara komprehensif mengenai fungsi

dan manfaat Data Preprocessing dalam konteks analisis data. Ini mencakup pemahaman bagaimana proses ini dapat meningkatkan kualitas data, meminimalkan kesalahan, dan mengoptimalkan performa model pembelajaran mesin.

3. **Mampu Menjelaskan Tingkatan, Struktur, Jenis-Jenis Data Preprocessing**
   Mahasiswa diharapkan dapat memberikan penjelasan terinci mengenai tingkatan Data Preprocessing, struktur dasar dari proses ini, dan berbagai jenis metode yang dapat diterapkan. Ini mencakup pemahaman mengenai pembersihan data, transformasi, encoding, dan langkah-langkah lainnya yang terlibat dalam merawat kualitas data.

Dengan mencapai kompetensi-kompetensi ini, diharapkan mahasiswa dan mahasiswi dapat memahami peran krusial Data Preprocessing dalam menghasilkan data yang berkualitas tinggi untuk mendukung analisis atau pengembangan model pembelajaran mesin.

PETA KONSEP PEMBELAJARAN



A. DATA PREPROCESSING

1. Definisi Data Preprocessing

Data preprocessing, atau pra-pemrosesan data, adalah serangkaian langkah atau tahapan yang dilakukan pada data mentah sebelum data tersebut digunakan untuk analisis lebih lanjut atau pengembangan model. Tujuan utama dari data preprocessing adalah untuk meningkatkan kualitas data, memastikan keakuratan hasil analisis, dan mengatasi masalah atau kekurangan yang mungkin muncul dalam data mentah.

Berikut adalah beberapa komponen utama dari definisi data preprocessing :

a. Pembersihan Data (*Data Cleaning*)
   - Identifikasi dan Penanganan Outliers
     Mendeteksi dan mengatasi data yang di luar pola umum, yang dapat mempengaruhi hasil analisis secara negatif.
   - Penanganan Duplikasi
     Mengidentifikasi dan menghapus data duplikat yang dapat menghasilkan hasil yang tidak akurat.

b. Pengisian Nilai yang Hilang (*Missing Values Handling*)
   - Imputasi Nilai
     Melakukan estimasi atau pengisian nilai yang hilang menggunakan metode tertentu, seperti nilai rata-rata atau median.

c. Integrasi Data (*Data Integration*)
   - Penggabungan Data
     Menggabungkan data dari berbagai sumber untuk membuat kumpulan data yang lebih lengkap dan bermakna.
   - Penanganan Perbedaan Format
     Menangani perbedaan dalam format, skema, atau struktur data.

d. Transformasi Data (*Data Transformation*)
   - Standarisasi dan Normalisasi

Menyesuaikan skala dan bentuk distribusi data agar lebih konsisten.

- Encoding Variabel Kategorikal
  Mengubah variabel kategorikal menjadi bentuk yang dapat diproses oleh algoritma, seperti menggunakan one-hot encoding.

e. Seleksi Fitur (*Feature Selection*)
   Memilih subset fitur yang paling relevan dan signifikan untuk analisis atau pemodelan.
f. Manajemen Data Tidak Seimbang (*Handling Imbalanced Data*)
   - Oversampling dan Undersampling
     Menangani masalah ketidakseimbangan kelas dengan meningkatkan atau mengurangi jumlah sampel pada kelas tertentu.
   - Pengelolaan Kesalahan atau Inkonsistensi (*Error Handling*)
     Penanganan Kesalahan Data: Mendeteksi dan memperbaiki kesalahan atau inkonsistensi dalam data yang dapat memengaruhi keakuratan analisis.
g. Optimasi Kinerja (*Performance Optimization*)
   - Pemrosesan Paralel
     Menggunakan teknik pemrosesan paralel untuk meningkatkan efisiensi dan kecepatan pemrosesan data.

Data preprocessing adalah tahap kritis dalam siklus analisis data atau pengembangan model. Dengan memastikan data bersih, terstruktur, dan relevan, analisis atau model yang dihasilkan akan lebih dapat diandalkan dan memberikan wawasan yang lebih berharga.

2. Pentingnya Data Preprocessing dalam Analisis Data

Data preprocessing memiliki peran yang sangat penting dalam analisis data, dan keberhasilan analisis atau pengembangan model seringkali sangat bergantung pada kualitas data yang diproses. Berikut adalah beberapa alasan mengapa data preprocessing sangat penting :

a. Meningkatkan Kualitas Data
   Membersihkan data dari duplikat, outlier, dan nilai yang hilang membantu meningkatkan kualitas dan integritas data. Data yang bersih dan terstruktur meminimalkan risiko menghasilkan hasil yang bias atau tidak akurat.
b. Memastikan Keakuratan Analisis
   Data preprocessing membantu memastikan bahwa data yang digunakan dalam analisis adalah representatif dan akurat. Tanpa preprocessing, analisis dapat terpengaruh oleh kesalahan atau ketidakpastian yang mungkin muncul dari data yang tidak bersih atau tidak terstruktur.
c. Mengatasi Missing Values
   Pengisian nilai yang hilang memastikan bahwa tidak ada informasi yang hilang atau tidak lengkap, yang dapat mengarah pada kesimpulan yang keliru atau keputusan yang tidak akurat.
d. Integrasi Data dari Sumber yang Berbeda
   Dalam kasus penggabungan data dari sumber yang berbeda, preprocessing membantu menyatukan data dengan format yang berbeda atau skema yang berbeda, memastikan konsistensi dan integritas data.
e. Standarisasi dan Normalisasi
   Standarisasi dan normalisasi data membantu menghindari bias yang mungkin muncul akibat perbedaan skala atau unit pengukuran dalam variabel-variabel yang digunakan dalam analisis.
f. Transformasi Data untuk Analisis yang Optimal
   Transformasi data, seperti encoding variabel kategorikal atau pengurangan dimensi, membantu mempersiapkan data agar sesuai dengan kebutuhan analisis atau model tertentu.
g. Seleksi Fitur yang Relevan
   Memilih fitur yang paling relevan mengurangi kompleksitas model dan meningkatkan interpretabilitas hasil. Hal ini juga dapat menghindari overfitting dan meningkatkan kinerja model.
h. Manajemen Data Tidak Seim bång

Mengatasi masalah data tidak seimbang membantu memastikan bahwa analisis klasifikasi atau model pembelajaran mesin tidak didominasi oleh satu kelas tertentu, yang dapat menghasilkan model yang tidak seimbang dan tidak akurat.

i. Pemrosesan Paralel dan Optimasi Kinerja
   Dalam situasi di mana kinerja waktu adalah faktor kunci, data preprocessing dapat mencakup pemrosesan paralel untuk meningkatkan efisiensi dan kecepatan pemrosesan data.
j. Peningkatan Validitas dan Reliabilitas
   Dengan memastikan bahwa data telah melalui proses preprocessing yang tepat, hasil analisis atau model yang dihasilkan lebih valid dan dapat diandalkan.

Dengan melakukan data preprocessing dengan baik, analisis data menjadi lebih efektif dan hasil yang dihasilkan menjadi lebih dapat dipercaya. Ini merupakan langkah esensial dalam siklus analisis data yang menyediakan fondasi yang kuat untuk pengambilan keputusan yang tepat.

B. Langkah-langkah Data Preprocessing

1. Pengumpulan Data

Proses pengumpulan data adalah langkah awal dalam siklus analisis data, tetapi sebelum data dapat digunakan untuk analisis lebih lanjut, seringkali diperlukan langkah-langkah pra-pemrosesan data atau Data Preprocessing. Data preprocessing adalah serangkaian langkah yang bertujuan untuk membersihkan, mengorganisir, dan mengubah data mentah menjadi bentuk yang lebih sesuai untuk analisis.

2. Pembersihan Data (*Handling Missing Values*, *Outliers*, *Anomalies*)

Pembersihan data (*data cleaning*) adalah tahap kritis dalam proses analisis data yang melibatkan identifikasi, penanganan, dan/atau penghapusan nilai yang hilang (*missing values*), pencilan

(*outliers*), dan anomali dalam dataset. Tujuan utama dari pembersihan data adalah memastikan bahwa data yang digunakan untuk analisis atau pemodelan adalah akurat, konsisten, dan dapat diandalkan. Berikut adalah penjelasan secara detail mengenai penanganan nilai yang hilang, pencilan, dan anomali :

a. *Handling Missing Values* (Penanganan Nilai yang Hilang)

Identifikasi Nilai yang Hilang

- Inspeksi Awal Data
  Periksa dataset untuk mengidentifikasi apakah ada nilai yang hilang.
- Statistik Deskriptif
  Gunakan metode statistik deskriptif seperti mean, median, dan perhitungan jumlah nilai yang hilang untuk mengevaluasi sejauh mana masalah nilai yang hilang.

Strategi Penanganan Nilai yang Hilang

- Hapus Baris atau Kolom
  Hapus baris atau kolom yang memiliki nilai yang hilang jika jumlahnya signifikan dan tidak dapat diimputasi.
- Imputasi
  Isi nilai yang hilang dengan nilai yang dapat diestimasi berdasarkan statistik (mean, median), model prediktif, atau metode lainnya.

b. Handling Outliers (Penanganan Pencilan)

Identifikasi Pencilan

- Visualisasi Data
  Gunakan grafik seperti box plot atau scatter plot untuk mengidentifikasi pencilan.
- Metode Statistik
  Gunakan metode statistik seperti Interquartile Range (IQR) untuk mengidentifikasi nilai-nilai yang di luar jangkauan normal.

Strategi Penanganan Pencilan

- Transformasi Data
  Gunakan transformasi data seperti log-transform untuk mengurangi dampak pencilan.
- Penyesuaian Nilai
  Periksa dan, jika diperlukan, sesuaikan nilai pencilan agar sesuai dengan rentang data yang realistis.
- Hapus atau Isolasi
  Hapus atau isolasi nilai pencilan jika mempengaruhi analisis secara signifikan.

c. Handling Anomalies (Penanganan Anomali)

Identifikasi Anomali

- Analisis Domain
  Gunakan pengetahuan domain untuk mengidentifikasi anomali yang mungkin tidak dapat dijelaskan oleh aturan umum.
- Algoritma Deteksi Anomali
  Terapkan algoritma deteksi anomali seperti isolation forest, k-nearest neighbors, atau DBSCAN.

Strategi Penanganan Anomali

- Verifikasi
  Verifikasi apakah anomali tersebut benar-benar mencerminkan kesalahan atau kejadian langka yang perlu dipertahankan.
- Isolasi atau Hapus
  Isolasi anomali jika mungkin atau hapus jika dapat dikonfirmasi sebagai kesalahan.

Langkah Tambahan

- Dokumentasi
  Selalu dokumentasikan langkah-langkah pembersihan data yang diambil, termasuk alasan di balik setiap keputusan.
- Uji Validitas

Uji ulang data setelah pembersihan untuk memastikan bahwa data tetap konsisten dan sesuai dengan ekspektasi.

- Iterasi

  Pembersihan data mungkin melibatkan iterasi berulang seiring berlanjutnya analisis data atau model pembangunan.

  Pembersihan data merupakan bagian penting dari siklus analisis data dan merupakan aspek kunci untuk memastikan keandalan dan validitas hasil analisis.

3. Transformasi Data (*Normalisasi, Transformasi Logaritma*, dll.)

Transformasi data adalah suatu teknik yang digunakan dalam analisis data untuk mengubah nilai-nilai variabel agar memenuhi asumsi atau persyaratan tertentu.

Transformasi ini dapat membantu meningkatkan distribusi data, mengurangi heteroskedastisitas, atau mempermudah interpretasi. Beberapa transformasi data yang umum digunakan melibatkan normalisasi, transformasi logaritma, dan lainnya.

a. Normalisasi

Normalisasi bertujuan untuk mengubah nilai-nilai variabel sehingga memiliki rentang atau skala yang seragam. Ini membantu mencegah variabel dengan rentang nilai yang besar mendominasi perhitungan atau analisis.

Rumus Normalisasi Min-Max :

- $Xnormalized = \frac{X - X\ min}{X\ \max - X\ min}$

- Keuntungan : Menjamin bahwa semua variabel memiliki skala yang seragam, sehingga tidak ada variabel yang mendominasi yang lain.

b. Transformasi Logaritma

Digunakan ketika data memiliki distribusi yang tidak normal atau memiliki skewness. Transformasi logaritma dapat

membantu mengurangi efek skewness dan membuat distribusi data lebih simetris.

Rumus Transformasi Logaritma :

- $Y = \log(X)$
- Keuntungan : Mengurangi variabilitas data dan menghasilkan distribusi yang lebih simetris, yang berguna dalam analisis statistik seperti regresi.

c. Transformasi Akar Kuadrat

Sama seperti transformasi logaritma, transformasi akar kuadrat dapat digunakan untuk mengurangi efek skewness dan membuat distribusi data lebih simetris.

Rumus Transformasi Akar Kuadrat :

- $Y = \sqrt{X}$
- Keuntungan : Berguna ketika data memiliki distribusi yang condong ke kanan dan perlu diubah agar lebih simetris.

d. Box-Cox Transformation

Box-Cox adalah transformasi parametrik yang dapat digunakan untuk mengatasi variasi heteroskedastisitas dan membuat distribusi data lebih normal. Namun, hanya dapat digunakan pada data yang bernilai positif.

Rumus Box-Cox :

- $Y\ (\lambda) = \frac{Y^{\lambda}-2}{\lambda}$
  $untuk\ \lambda \neq 0\ ; dan \log(Y), untuk\ \lambda = 0$
- Keuntungan : Mengoptimalkan nilai λ untuk menghasilkan transformasi terbaik.

e. Z-Score Transformation

Mengubah nilai-nilai variabel menjadi skor-z, yang menyatakan seberapa jauh suatu nilai dari rata-rata dalam satuan deviasi standar.

Rumus Z-Score :

- $Z = \frac{X-\mu}{\sigma}$

- Keuntungan : Memungkinkan perbandingan relatif antara nilai-nilai variabel yang awalnya memiliki satuan yang berbeda.

Setiap transformasi data harus diterapkan dengan pertimbangan matang tergantung pada sifat data dan tujuan analisis. Beberapa metode mungkin lebih sesuai untuk situasi tertentu, dan pemilihan transformasi yang tepat dapat meningkatkan kualitas analisis statistik.

4. Penggabungan Data

Penggabungan data, atau sering disebut juga dengan istilah "merging" atau "joining" data, adalah proses menggabungkan dua set data atau lebih berdasarkan suatu kriteria tertentu. Tujuan utama dari penggabungan data adalah untuk menggabungkan informasi dari dua atau lebih sumber data yang berbeda agar dapat dianalisis bersama-sama. Proses ini umumnya dilakukan dalam analisis data dan pengolahan data di berbagai aplikasi, seperti database, spreadsheet, atau bahkan dalam lingkungan pemrograman.

Berikut adalah beberapa konsep dasar dan langkah-langkah yang terlibut dalam penggabungan data :

a. Jenis Penggabungan Data

- Inner Join
  Menggabungkan hanya baris yang memiliki nilai kunci yang cocok di kedua set data.
- Left (Outer) Join
  Menyertakan semua baris dari set data kiri dan baris yang cocok dari set data kanan.
- Right (Outer) Join
  Kebalikan dari Left Join; menyertakan semua baris dari set data kanan dan baris yang cocok dari set data kiri.
- Full (Outer) Join
  Menyertakan semua baris dari kedua set data, meskipun ada nilai kunci yang tidak cocok.

b. Kunci Gabungan (Join Key)

- Ini adalah kolom atau kolom-kolom di setiap set data yang digunakan sebagai acuan untuk menggabungkan data.
- Kunci ini harus memiliki nilai yang sama atau setidaknya dapat diubah (diubah ke format yang sama) agar data dapat digabungkan dengan benar.

c. Langkah-langkah Penggabungan Data

- Identifikasi Kunci Gabungan
  Tentukan kunci gabungan yang sesuai antara kedua set data.
- Pilih Jenis Penggabungan
  Pilih jenis penggabungan yang sesuai dengan kebutuhan analisis.
- Lakukan Penggabungan
  Terapkan operasi penggabungan menggunakan perintah atau fungsi yang sesuai dengan alat atau platform yang digunakan (SQL, Pandas di Python, dll.).
- Penanganan Data yang Tidak Cocok
  Tentukan cara menangani data yang tidak memiliki nilai kunci yang cocok di kedua set data (misalnya, mengganti dengan nilai tertentu atau mengabaikannya).

d. Contoh Penggabungan Data dalam SQL

```
SELECT *
FROM tabel1
INNER JOIN tabel2 ON tabel1.kunci = tabel2.kunci;
```

e. Contoh Penggabungan Data dalam Python (menggunakan Pandas)

```
import pandas as pd
df1 = pd.DataFrame({'kunci': [1, 2, 3], 'data1': ['A', 'B', 'C']})
df2 = pd.DataFrame({'kunci': [2, 3, 4], 'data2': ['X', 'Y', 'Z']})
merged_data = pd.merge(df1, df2, on='kunci', how='inner')
```

f. Penanganan Duplikasi Kunci

Pastikan untuk menangani dengan baik situasi di mana terdapat duplikasi nilai kunci di satu atau kedua set data.

g. Pemantauan Kinerja

Pada dataset yang sangat besar, perlu mempertimbangkan efisiensi penggabungan data untuk meminimalkan waktu eksekusi.

Penting untuk diingat bahwa penggabungan data harus dilakukan dengan hati-hati dan dipahami dengan baik untuk menghindari hasil yang tidak diinginkan atau analisis yang tidak akurat. Setiap jenis penggabungan memiliki penggunaan dan situasi yang sesuai, tergantung pada kebutuhan analisis spesifik.

5. Encoding Data (*One-Hot Encoding, Label Encoding*)

Encoding data adalah proses mengonversi data dari suatu bentuk ke bentuk lain agar dapat diolah oleh model atau algoritma tertentu. Dalam konteks machine learning, encoding data sering digunakan untuk mengubah variabel kategori atau teks menjadi bentuk numerik agar dapat dimengerti oleh algoritma pembelajaran mesin.

Dua metode encoding data yang umum digunakan adalah One-Hot Encoding dan Label Encoding.

a. One-Hot Encoding

One-Hot Encoding digunakan untuk mengubah variabel kategori menjadi vektor biner (0 dan 1), dengan setiap kategori direpresentasikan sebagai vektor di mana hanya satu elemen bernilai 1 (hot), dan yang lainnya bernilai 0 (cold).

Contoh :

Misalkan kita memiliki variabel kategori "Warna" dengan kategori "Merah", "Biru", dan "Hijau". One-Hot Encoding akan menghasilkan vektor sebagai berikut: Merah: [1, 0, 0] Biru: [0, 1, 0] Hijau: [0, 0, 1]

Dengan menggunakan library Python, seperti scikit-learn, proses One-Hot Encoding dapat dilakukan dengan mudah :

```
from sklearn.preprocessing import OneHotEncoder
import pandas as pd
# Contoh data
data = {'Warna': ['Merah', 'Biru', 'Hijau']}
df = pd.DataFrame(data)

# Membuat objek OneHotEncoder
encoder = OneHotEncoder()

# Melakukan One-Hot Encoding dan menggabungkan dengan DataFrame
encoded_data = pd.DataFrame(encoder.fit_transform(df[['Warna']]).toarray(),
columns=encoder.get_feature_names_out(['Warna']))
df_encoded = pd.concat([df, encoded_data], axis=1)
```

b. Label Encoding

Label Encoding digunakan untuk mengonversi setiap nilai dalam suatu kolom menjadi nilai numerik unik. Setiap nilai kategori diberikan label berupa bilangan bulat secara berurutan.

Contoh :

Jika kita menggunakan Label Encoding pada variabel "Warna", maka kita akan mendapatkan: Merah: 1,
Biru: 2, Hijau: 3

Contoh implementasi Label Encoding dengan library scikit-learn :

```
from sklearn.preprocessing import LabelEncoder

# Contoh data
data = {'Warna': ['Merah', 'Biru', 'Hijau']}
df = pd.DataFrame(data)

# Membuat objek LabelEncoder
encoder = LabelEncoder()

# Melakukan Label Encoding
df['Warna_Encoded'] = encoder.fit_transform(df['Warna'])
```

Perhatian :

1. One-Hot Encoding lebih cocok digunakan ketika tidak ada hubungan ordinal antar nilai kategori, sedangkan Label Encoding lebih cocok digunakan jika ada hubungan ordinal antar nilai kategori.
2. One-Hot Encoding dapat menyebabkan peningkatan jumlah fitur dalam dataset, yang dapat mempengaruhi kinerja model jika dataset sangat besar. Label Encoding menghasilkan dataset dengan jumlah fitur yang tetap.

6. Pembagian Data (*Training Set, Testing Set*)

Pembagian data menjadi set pelatihan (training set) dan set pengujian (testing set) adalah langkah kritis dalam proses pembangunan model dalam pembelajaran mesin. Tujuan utamanya adalah untuk mengukur sejauh mana model yang dibangun mampu melakukan generalisasi pada data yang tidak pernah dilihat sebelumnya. Berikut adalah penjelasan secara detail mengenai kedua set tersebut :

a. Set Pelatihan (*Training Set*)

- Set pelatihan adalah bagian dari dataset yang digunakan untuk melatih model mesin. Model mempelajari pola dan relasi dari data ini.
- Biasanya, set pelatihan merupakan mayoritas dari keseluruhan dataset, mungkin sekitar 70-80% dari total data. Proses Pelatihan: Model dibangun dengan memasukkan set pelatihan ke dalamnya, dan parameter-model diperbarui iteratif melalui algoritma pembelajaran.
- Menciptakan model yang dapat memahami pola dan hubungan yang terkandung dalam data pelatihan sehingga dapat membuat prediksi yang baik pada data baru.

b. Set Pengujian (*Testing Set*)

- Set pengujian adalah bagian dari dataset yang tidak digunakan selama proses pelatihan. Ini berfungsi sebagai pengukur kinerja model pada data yang belum pernah dilihat sebelumnya.

- Biasanya, set pengujian merupakan sekitar 20-30% dari keseluruhan dataset.
- Setelah model dilatih, set pengujian digunakan untuk mengukur seberapa baik model tersebut dapat menggeneralisasi pola pada data baru.
- Mengidentifikasi sejauh mana model dapat memberikan prediksi yang akurat dan dapat diandalkan pada data yang tidak terlibat dalam proses pelatihan.

c. Langkah-langkah Pembagian Data

- Randomization

  Data umumnya diacak sebelum pembagian untuk memastikan bahwa setiap subset mencerminkan variasi data yang merata.
- Stratifikasi

  Jika ada kelas atau grup yang signifikan dalam data, stratifikasi dapat digunakan untuk memastikan distribusi yang seimbang antara set pelatihan dan pengujian.
- Pemisahan acak atau berurutan

  Data dapat dipisahkan secara acak atau berurutan. Pemisahan acak lebih umum dan dapat membantu mencegah bias.

d. Pentingnya Validasi

Set Validasi kadang-kadang, data dapat dibagi menjadi tiga bagian: pelatihan, validasi, dan pengujian. Set validasi digunakan selama pelatihan untuk menyesuaikan parameter-model dan mencegah overfitting.

e. Evaluasi Model

- Metrik Kinerja

  Setelah model dilatih dan diuji, metrik kinerja seperti akurasi, presisi, recall, dan F1-score dapat digunakan untuk mengukur kinerja model.
- Confusion Matrix

  Berguna untuk mengevaluasi sejauh mana model dapat membedakan kelas-kelas yang berbeda.

f. Catatan Tambahan

Pembaruan Model: Jika model tidak memberikan kinerja yang baik pada set pengujian, mungkin diperlukan revisi model atau proses feature engineering untuk meningkatkan generalisasi.

Pembagian yang baik antara set pelatihan dan pengujian sangat penting untuk menghasilkan model yang memiliki kemampuan generalisasi yang baik pada data yang belum pernah dilihat sebelumnya.

7. Standarisasi

Standarisasi adalah proses pengembangan dan penerapan standar untuk mencapai tujuan tertentu dalam berbagai bidang, seperti industri, teknologi, atau layanan. Standar adalah dokumen yang memuat kriteria, pedoman, atau spesifikasi teknis yang digunakan sebagai referensi umum untuk memastikan konsistensi, interoperabilitas, keamanan, dan kualitas produk atau layanan.

Berikut adalah beberapa aspek terperinci mengenai standarisasi :

a. Tujuan Standarisasi

- Membantu memastikan bahwa produk atau layanan yang dihasilkan memiliki kualitas yang konsisten.
- Memungkinkan produk dari berbagai produsen atau penyedia layanan dapat bekerja bersama atau saling terhubung.
- Menetapkan pedoman untuk memastikan tingkat keamanan yang memadai dalam penggunaan produk atau layanan.
- Membantu meningkatkan efisiensi proses produksi atau penyediaan layanan dengan menetapkan metode terbaik.
- Memberikan kerangka kerja untuk inovasi dengan memberikan dasar yang jelas untuk pengembangan produk atau layanan baru.

b. Proses Standarisasi

- Menentukan kebutuhan dan masalah yang perlu dipecahkan melalui standar.
- Melibatkan penyusunan dokumen standar, yang dapat melibatkan banyak pihak termasuk ahli teknis, produsen, pengguna, dan regulator.
- Mencapai persetujuan dan konsensus dari semua pihak yang terlibat melalui proses konsultasi dan pemilihan.
- Menerapkan standar tersebut dalam proses produksi, produk, atau layanan.

c. Jenis Standar

- Menetapkan spesifikasi teknis untuk produk tertentu.
- Menyediakan pedoman untuk proses produksi atau layanan.
- Berkaitan dengan interoperabilitas dan integrasi sistem.
- Menetapkan prinsip-prinsip untuk manajemen organisasi.

d. Organisasi Standarisasi

Ada beberapa organisasi standarisasi teknologi yang memiliki peran penting dalam mengembangkan dan menetapkan standar di berbagai bidang. Beberapa di antaranya termasuk :

1. ISO (*International Organization for Standardization*)
   - ISO adalah organisasi internasional yang mengembangkan dan menerbitkan standar internasional untuk berbagai industri, termasuk teknologi.
   - Berperan dalam menyediakan standar yang diakui secara global untuk memastikan kualitas, keamanan, dan interoperabilitas produk dan layanan.
2. IEEE (*Institute of Electrical and Electronics Engineers*)
   - IEEE adalah organisasi profesional yang fokus pada pengembangan standar di bidang teknologi informasi, elektronika, dan listrik.
   - Menetapkan standar untuk perangkat keras, perangkat lunak, jaringan, dan protokol komunikasi.

3. W3C (*World Wide Web Consortium*)
    - W3C adalah organisasi internasional yang bertanggung jawab untuk mengembangkan standar web.
    - Menetapkan spesifikasi dan pedoman untuk teknologi web, seperti HTML, CSS, dan protokol web.
4. ITU (*International Telecommunication Union*)
    - ITU adalah badan khusus Perserikatan Bangsa-Bangsa yang bertanggung jawab untuk mengembangkan standar di bidang telekomunikasi dan teknologi informasi.
    - Menetapkan standar untuk teknologi telekomunikasi, termasuk jaringan seluler dan protokol komunikasi.
5. IETF (*Internet Engineering Task Force*)
    - IETF adalah sebuah kelompok kerja sukarela yang berkumpul untuk mengembangkan dan memperbaiki standar internet.
    - Menetapkan standar untuk protokol internet, seperti TCP/IP, DNS, dan HTTP.
6. ETSI (*European Telecommunications Standards Institute*)
    - ETSI adalah organisasi standarisasi Eropa yang fokus pada teknologi telekomunikasi dan informasi.
    - Menetapkan standar untuk berbagai teknologi, termasuk 5G, IoT, dan jaringan seluler.
7. NIST (*National Institute of Standards and Technology*)
    - NIST adalah lembaga Amerika Serikat yang mengembangkan standar di berbagai sektor, termasuk teknologi.
    - Menetapkan standar keamanan, pengukuran, dan teknologi informasi.
8. IEC (*International Electrotechnical Commission*)
    - IEC adalah organisasi internasional yang fokus pada standar teknologi listrik, elektronika, dan terkait.
    - Menetapkan standar untuk perangkat elektronik, peralatan listrik, dan teknologi terkait.

9. OASIS (*Organization for the Advancement of Structured Information Standards*)
    - OASIS adalah konsorsium industri yang mengembangkan standar terbuka untuk teknologi informasi dan layanan web.
    - Menetapkan standar untuk format data terbuka, seperti XML dan Web Services.

Setiap organisasi standarisasi memiliki fokusnya sendiri dan berkontribusi pada perkembangan teknologi dengan menyediakan pedoman yang umumnya diterima secara global. Kolaborasi antara organisasi ini membantu memastikan interoperabilitas dan keberlanjutan teknologi di seluruh dunia.

e. Manfaat Standarisasi

- Meningkatkan keamanan dan kualitas produk atau layanan.
- Membangun kepercayaan konsumen dengan memberikan standar sebagai acuan kualitas.
- Meningkatkan efisiensi produksi dan penggunaan sumber daya.
- Mempermudah akses ke pasar global dengan mengadopsi standar internasional.

f. Penerapan Standar

- Proses penilaian oleh pihak ketiga untuk memastikan bahwa suatu produk atau sistem memenuhi standar tertentu.
- Melibatkan pelatihan tenaga kerja untuk memahami dan mengimplementasikan standar.
- Melibatkan peninjauan dan pembaruan standar secara berkala sesuai dengan perkembangan teknologi dan kebutuhan industri.

g. Tantangan dalam Standarisasi

- Standar harus mampu beradaptasi dengan cepat terhadap perubahan teknologi.
- Mengatasi perbedaan budaya, regulasi, dan kebutuhan di berbagai pasar global.

- Memastikan partisipasi yang luas dari berbagai pihak untuk mencapai konsensus.

Standarisasi adalah elemen kunci dalam membangun fondasi yang kuat untuk keberlanjutan dan kemajuan dalam berbagai industri dan sektor.

C. Tujuan Data Preprocessing

Tujuan dari data preprocessing melibatkan beberapa aspek yang dapat meningkatkan kualitas data, memastikan konsistensi dan kebersihan data, serta memungkinkan model pembelajaran mesin untuk berkonvergensi lebih cepat.

Adapun tujuannya sebagai berikut :

1. Meningkatkan Kualitas Data

a. Penanganan *Missing Values*
   - Identifikasi dan penanganan nilai-nilai yang hilang agar tidak merugikan analisis atau model.
   - Penggantian nilai yang hilang dengan metode seperti interpolasi, mean, median, atau menggunakan model prediktif.

b. Penanganan *Outliers*
   - Identifikasi dan penanganan data ekstrem yang dapat memengaruhi hasil analisis atau model secara negatif.
   - Penggunaan teknik seperti trimming (memotong) atau transformasi data untuk menangani outliers.

c. Normalisasi Data
   Menstandarisasi skala data agar memiliki rentang yang seragam, sehingga model tidak terpengaruh oleh perbedaan skala antar fitur.

d. Encoding Variabel Kategorikal
   Mengubah variabel kategorikal menjadi bentuk yang dapat dimengerti oleh algoritma pembelajaran mesin, seperti one-hot encoding atau label encoding.

2. Memastikan Konsistensi dan Kebersihan Data

a. Deteksi dan Penanganan Duplikasi
   Mengidentifikasi dan menghapus data duplikat untuk memastikan ketepatan dan kebersihan analisis.
b. Standarisasi Format
   Menyesuaikan format data agar konsisten, seperti mengubah format tanggal atau satuan ke dalam format yang seragam.
c. Penanganan Kesalahan Tertulis (*Typos*)
   Mendeteksi dan memperbaiki kesalahan penulisan atau ketidaksesuaian format data.
d. Validasi Data
   Memastikan bahwa data memenuhi aturan dan batasan yang telah ditetapkan, sehingga dapat diandalkan dan valid.

3. Memungkinkan Model Pembelajaran Mesin untuk Berkonvergensi Lebih Cepat

a. Reduksi Dimensi
   Menggunakan teknik reduksi dimensi seperti Principal Component Analysis (PCA) untuk mengurangi kompleksitas data dan mempercepat konvergensi model.
b. Pemilihan Fitur (*Feature Selection*)
   Memilih subset fitur yang paling relevan untuk meningkatkan efisiensi dan kecepatan konvergensi model.
c. Pengelompokan (*Clustering*)
   Menggunakan teknik clustering untuk mengelompokkan data menjadi subset yang lebih teratur, mempermudah pembelajaran oleh model.
d. Normalisasi dan Scaling
   Memastikan bahwa data telah dinormalisasi dan di-scaled dengan baik, sehingga algoritma pembelajaran mesin dapat lebih cepat mencapai konvergensi.

Dengan menjalankan proses preprocessing dengan benar, data menjadi lebih siap untuk digunakan dalam model

pembelajaran mesin, yang pada gilirannya dapat meningkatkan kualitas prediksi dan analisis yang dihasilkan.

D. Hasil Data Preprocessing

Hasil dari proses data preprocessing sangat penting untuk memastikan kualitas dan keandalan analisis atau model pembelajaran mesin yang akan dibangun.

1. Data yang Siap untuk Analisis atau Model Pembelajaran Mesin

a. Data Bersih dan Terstruktur
   - Setelah proses preprocessing, data menjadi bersih dari nilai yang hilang, outliers, dan kesalahan lainnya.
   - Struktur data menjadi lebih terorganisir dan mudah diakses oleh algoritma analisis atau model pembelajaran mesin.

b. Data Terstandarisasi dan Dikode
   Data telah diubah menjadi format yang konsisten dan dapat dimengerti oleh algoritma, termasuk pengkodean variabel kategorikal.

c. Data yang Diformalisasikan
   Skala data telah dinormalisasi sehingga perbedaan skala antar fitur tidak memengaruhi kinerja model.

2. Validasi Data Preprocessing

a. Pengecekan Missing Values
   Memastikan bahwa tidak ada nilai yang hilang setelah proses imputasi atau penghapusan missing values.

b. Pengecekan Kesalahan Tertulis dan Duplikasi
   Verifikasi bahwa kesalahan penulisan dan duplikasi data telah diperbaiki dengan benar dan tidak menyisakan masalah.

c. Pengecekan Format Data
   Memeriksa apakah semua data telah disesuaikan dengan format yang diharapkan.

3. Memastikan Data yang Telah Diproses Memenuhi Kriteria

a. Memeriksa Konsistensi dan Kebersihan Data
   Verifikasi bahwa data tetap konsisten dan bersih setelah preprocessing, tanpa adanya ketidaksesuaian atau kekacauan.
b. Pengecekan Validitas Data
   Menjamin bahwa data tetap valid dan memenuhi batasan atau aturan bisnis yang telah ditetapkan.
c. Evaluasi Distribusi Data
   Memeriksa distribusi data untuk memastikan bahwa distribusi tidak berubah secara signifikan dan tetap mewakili karakteristik awal.

4. Menangani Kesalahan atau Kecurangan Data yang Mungkin Muncul

a. Monitor Proses Secara Berkala
   Melakukan pemantauan secara berkala terhadap data setelah preprocessing untuk mendeteksi kemungkinan kesalahan atau kecurangan yang mungkin muncul seiring waktu.
b. Menanggapi Perubahan dalam Sumber Data
   Jika ada perubahan dalam sumber data, memastikan bahwa proses preprocessing diperbarui untuk mencerminkan perubahan tersebut.
c. Pembaruan Rutin
   Melakukan pembaruan rutin terhadap proses preprocessing untuk menanggapi perubahan kebutuhan analisis atau model.

Hasil dari data preprocessing memberikan keyakinan bahwa data yang digunakan untuk analisis atau pembelajaran mesin adalah data yang berkualitas tinggi, dapat diandalkan, dan siap digunakan untuk mendapatkan wawasan yang berharga atau melatih model yang akurat. Validasi dan pemantauan terus-menerus terhadap data setelah preprocessing sangat penting untuk memastikan keberlanjutan kualitas data sepanjang waktu.

E. Dokumentasi Data Preprocessing

1. Mencatat dan Mendokumentasikan Setiap Langkah yang Dilakukan

Mencatat setiap langkah yang dilakukan dalam proses pengolahan data atau analisis sangat penting untuk beberapa alasan utama :

a. Reproduktibilitas
   - Memungkinkan orang lain untuk mengulangi atau memvalidasi analisis Anda.
   - Membantu Anda sendiri untuk mengulangi analisis pada masa mendatang.

b. Transparansi
   Memberikan visibilitas terhadap proses kerja, memudahkan orang lain untuk memahami metodologi yang digunakan.

c. Pemecahan Masalah
   Mempermudah mengidentifikasi dan memperbaiki kesalahan atau masalah yang mungkin terjadi.

d. Kolaborasi
   Memfasilitasi kerja sama dengan tim atau rekan kerja, memungkinkan mereka untuk berkontribusi atau memahami pekerjaan Anda.

Cara Mencatat dan Mendokumentasikan :

a. Jenis Langkah yang Perlu Dicatat
   - Setiap langkah pemrosesan data.
   - Transformasi data yang dilakukan.
   - Pengaturan parameter.
   - Sumber data dan struktur asal.
   - Algoritma yang digunakan.
   - Visualisasi hasil.

b. Penjelasan Rinci
   - Sertakan penjelasan rinci untuk setiap langkah.
   - Deskripsikan tujuan dari setiap langkah.
   - Sebisa mungkin, hindari istilah atau langkah yang ambigu.

c. Gunakan Komentar di Kode (Jika Menggunakan Kode)
   - Jelaskan dengan komentar di dalam kode untuk setiap blok atau baris yang dianggap kompleks.
   - Gunakan komentar yang jelas dan deskriptif.

d. Simpan File Dokumentasi Terpisah
   Dokumentasikan langkah-langkah secara terpisah dalam file teks atau dokumen lainnya.

e. Gunakan Alat Pelacak Revisi
   Jika bekerja dalam tim atau secara berulang, gunakan alat pelacak revisi seperti Git untuk melacak perubahan.

2. Membuat Catatan Transformasi Data untuk Reproduktibilitas

Catatan transformasi data adalah dokumen atau catatan yang merekam proses perubahan data dari tahap awal hingga akhir. Ini sangat penting untuk memastikan reproduktibilitas, artinya dapat menghasilkan hasil yang sama ketika dilakukan pada data yang sama.

Langkah-langkah untuk Membuat Catatan Transformasi Data :

a. Deskripsi Data Awal
   - Jelaskan struktur dan sifat data awal.
   - Sertakan informasi seperti tipe data, kolom, dan karakteristik penting lainnya.

b. Proses Transformasi
   - Catat setiap langkah transformasi data.
   - Jelaskan mengapa setiap transformasi diperlukan.
   - Sertakan detail seperti filter, penghapusan duplikat, atau penggantian nilai yang hilang.

c. Penggunaan Fungsi atau Algoritma
   - Jelaskan fungsi atau algoritma yang digunakan untuk transformasi.
   - Sertakan parameter yang digunakan.

d. Visualisasi Transformasi (Jika Perlu)

Gunakan visualisasi, seperti grafik atau plot, untuk membantu memahami transformasi data.

e. Catatan Parameter
   Jika ada parameter atau konfigurasi khusus yang digunakan, catat informasi tersebut.
f. Referensi ke Sumber Data Eksternal
   Jika menggunakan sumber data eksternal, catat rinciannya dan pastikan dapat diakses di masa mendatang.
g. Simpan dalam Format yang Mudah Dibaca
   Dokumentasikan transformasi data dalam format yang mudah dibaca, seperti dokumen teks atau spreadsheet.
h. Update Catatan Saat Ada Perubahan
   Pastikan untuk mengupdate catatan setiap kali ada perubahan dalam proses transformasi data.

Melakukan langkah-langkah di atas akan memberikan kejelasan dan transparansi terhadap proses transformasi data, yang esensial untuk memastikan reproduktibilitas analisis data Anda.

F. RANGKUMAN

Data preprocessing adalah tahapan penting dalam analisis data yang bertujuan untuk mempersiapkan data mentah menjadi bentuk yang lebih baik dan sesuai untuk analisis.

Berikut ini adalah rangkuman dari pembahasan tentang data preprocessing :

1. **Pengertian Data Preprocessing**
   Data preprocessing adalah serangkaian langkah atau teknik yang dilakukan untuk membersihkan, mentransformasi, dan mengorganisir data mentah sehingga dapat digunakan dengan efektif dalam proses analisis.
2. **Tujuan Data Preprocessing**
   a. Mengatasi missing values dengan cara mengidentifikasi dan penanganan nilai yang hilang pada data.

b. Mendeteksi dan menangani outlier menggunakan tahapan identifikasi dan memperlakukan data yang ekstrim atau tidak sesuai.
c. Transformasi data yaitu mengubah skala atau bentuk distribusi data untuk memenuhi asumsi analisis statistik.
d. Encoding kategori dengan cara mengubah variabel kategori menjadi bentuk numerik untuk analisis lebih lanjut.
e. Normalisasi dan standarisasi dengan cara menyesuaikan skala data agar memiliki rentang atau distribusi yang konsisten.

3. Langkah-langkah Data Preprocessing

a. Pembersihan Data (*Data Cleaning*) : Mengatasi missing values, memperbaiki atau menghapus data yang tidak valid atau tidak relevan.
b. Transformasi Data (*Data Transformation*) : Melakukan normalisasi, encoding kategori, dan transformasi lainnya untuk menyesuaikan data dengan kebutuhan analisis.
c. Reduksi Dimensi (*Dimensionality Reduction*) : Mengurangi jumlah atribut atau fitur dalam data untuk mengurangi kompleksitas dan mempercepat analisis.

4. Teknik-Teknik Data Preprocessing

a. Imputasi dengan cara menggantikan nilai-nilai yang hilang dengan nilai yang diestimasi berdasarkan pola data.
b. Deteksi Outlier dengan cara mengidentifikasi dan memperlakukan nilai-nilai yang ekstrim atau tidak wajar.
c. Normalisasi dan Standarisasi dengan cara menyesuaikan skala data untuk memastikan keseragaman dan mempermudah perbandingan.
d. Encoding yaitu mengubah variabel kategori menjadi bentuk numerik untuk dapat digunakan dalam model analisis.

5. Pentingnya Data Preprocessing

a. Menjamin kualitas data : Menghilangkan ketidakakuratan dan ketidaksesuaian dalam data mentah.
b. Meningkatkan performa model : Data yang bersih dan terstruktur dapat meningkatkan hasil dari model analisis.

c. Meminimalkan bias : Pengolahan data dapat membantu mengurangi bias yang dapat memengaruhi hasil analisis.

6. Tantangan dalam Data Preprocessing

a. Kompleksitas Data : Data yang kompleks dan bervariasi dapat menimbulkan tantangan dalam preprocessing.
b. Keputusan Desain : Memilih teknik preprocessing yang tepat memerlukan pemahaman mendalam tentang data dan kebutuhan analisis.

Data preprocessing merupakan tahapan kritis dalam analisis data yang dapat membantu memastikan keberlanjutan dan akurasi hasil analisis. Melalui langkah-langkah tersebut, data dapat diolah menjadi bentuk yang lebih sesuai dan siap untuk digunakan dalam berbagai jenis analisis.

G. TES FORMATIF

Bagian 1 : Pilihan Ganda

1. Apa yang dimaksud dengan Data Preprocessing?
   a. Proses merubah data menjadi grafik
   b. Langkah awal dalam analisis data untuk membersihkan dan mempersiapkan data
   c. Langkah terakhir dalam pengolahan data
   d. Proses menyimpan data ke dalam database
2. Mengapa Data Preprocessing penting dalam analisis data?
   a. Untuk membuat data lebih kompleks
   b. Untuk mempersulit proses analisis
   c. Untuk meningkatkan kualitas dan kebersihan data
   d. Tidak ada manfaat khusus dari Data Preprocessing
3. Langkah apa yang termasuk dalam pembersihan data?
   a. Normalisasi
   b. Transformasi data
   c. Menangani nilai yang hilang dan outliers
   d. Encoding data

4. Apa fungsi utama Encoding data dalam Data Preprocessing?
   a. Mengubah data menjadi format numerik
   b. Membersihkan data dari nilai yang hilang
   c. Menormalisasi distribusi data
   d. Membagi data menjadi set pelatihan dan pengujian

Bagian 2 : Esai Singkat

1. Jelaskan mengapa penggabungan data dapat menjadi langkah kritis dalam Data Preprocessing.
2. Gambarkan perbedaan antara normalisasi dan standarisasi dalam konteks Data Preprocessing.

Bagian 3 : Studi Kasus

1. Anda diberikan dataset yang mengandung nilai yang hilang. Jelaskan tiga metode yang dapat Anda gunakan untuk menangani nilai yang hilang, dan berikan contoh situasi di mana masing-masing metode cocok digunakan.
2. Anda memiliki dataset yang mengandung atribut kategori. Pilih salah satu metode encoding data (one-hot encoding atau label encoding) dan jelaskan mengapa Anda memilih metode tersebut untuk dataset tertentu. Berikan contoh konkret.

Petunjuk

1. Waktu yang diberikan untuk tes adalah 60 menit.
2. Pastikan untuk memberikan jawaban yang rinci dan jelas pada pertanyaan esai.
3. Gunakan contoh konkret atau kasus studi untuk mendukung jawaban Anda.

H. LATIHAN

Studi Kasus :

Penggunaan Data Preprocessing dalam Analisis Kesehatan

Konteks :
Anda bekerja sebagai data scientist di sebuah rumah sakit yang memiliki dataset besar tentang rekam medis pasien. Dataset ini memiliki berbagai atribut, termasuk informasi kesehatan, riwayat penyakit, dan parameter klinis.

Tujuan Analisis :
Tujuan Anda adalah mengembangkan model prediksi untuk memperkirakan risiko penyakit jantung berdasarkan data pasien. Namun, dataset ini memerlukan beberapa tahap Data Preprocessing sebelum dapat digunakan untuk pelatihan model.

Tahapan Data Preprocessing yang Diperlukan :
1. Penanganan Nilai yang Hilang
   - Identifikasi dan analisis nilai yang hilang dalam atribut klinis dan riwayat penyakit.
   - Pilih metode yang tepat (misalnya, penggantian nilai rata-rata atau menggunakan model prediksi) untuk menangani nilai yang hilang tersebut.
2. Transformasi Data
   - Normalisasi nilai-nilai parameter klinis seperti tekanan darah dan kadar kolesterol untuk memastikan skala yang seragam.
   - Terapkan transformasi logaritma pada atribut yang memiliki distribusi yang condong.
3. Pemilihan Atribut
   - Identifikasi atribut yang mungkin tidak relevan atau memiliki korelasi tinggi.
   - Pilih subset atribut yang lebih relevan untuk model prediksi penyakit jantung.
4. Encoding Data
   Encoding atribut kategorikal seperti jenis kelamin dan status merokok menggunakan metode one-hot encoding.

5. Pembagian Data
   Bagi dataset menjadi set pelatihan (80%) dan set pengujian (20%) untuk menguji kinerja model.

Pertanyaan :

1. Penanganan Nilai yang Hilang: Bagaimana Anda memutuskan metode penanganan nilai yang hilang, dan apa konsekuensi potensial dari metode tersebut terhadap analisis kesehatan?
2. Mengapa normalisasi dan transformasi logaritma diperlukan untuk parameter klinis? Berikan alasan dan efek yang diharapkan.
3. Jelaskan mengapa pemilihan atribut penting untuk meningkatkan kinerja model. Sebutkan contoh atribut yang mungkin dihapus dan alasan di balik keputusan tersebut.
4. Mengapa one-hot encoding digunakan untuk atribut kategorikal? Apa keuntungan dan situasi di mana metode ini lebih disarankan daripada label encoding?
5. Jelaskan pentingnya pembagian data dan ukuran set pelatihan dan pengujian yang dipilih. Apa dampaknya jika pembagian tidak dilakukan dengan baik?

Catatan :

Harap jawab setiap pertanyaan dengan memberikan penjelasan dan pemahaman yang mendalam terkait dengan konteks kasus ini.

KEGIATAN BELAJAR 5

DATA MINING ROLES (Peran Data Mining)

DESKRIPSI PEMBELAJARAN

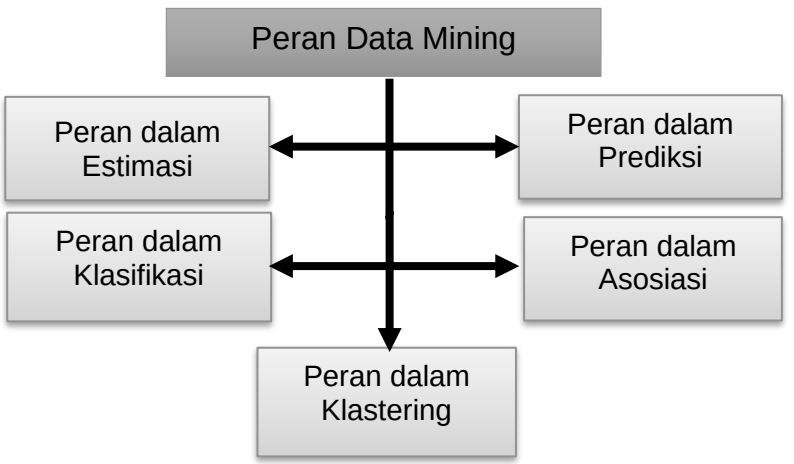
Pada bab ini mahasiswa mempelajari pengenalan dan konsep dasar teoritis Peran Data Mining dalam Estimasi, Prediksi, Klassifikasi, Klastering, Asosiasi. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk modal dasar mempelajari Data mining lebih lanjut.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

1. Mampu menjelaskan peran Data mining dalam Estimasi, Prediksi,Klassifikasi, Klastering, Asosiasi
2. Mampu menjelaskan manfaat, model evaluasi, metode, tantangan dan keuntungan peran data mining.

PETA KONSEP PEMBELAJARAN



A. PENGERTIAN DATA MINING

Data mining, atau penambangan data, adalah proses ekstraksi informasi berharga, pola, dan pengetahuan yang tersembunyi dari kumpulan data yang besar dan kompleks. Tujuan utamanya adalah untuk mengidentifikasi hubungan dan tren yang dapat digunakan untuk mendukung pengambilan keputusan strategis. Dalam esensinya, data mining merupakan teknik analisis yang menggunakan metode statistik, matematika, dan kecerdasan buatan untuk menggali pengetahuan yang belum diketahui secara otomatis dari data.

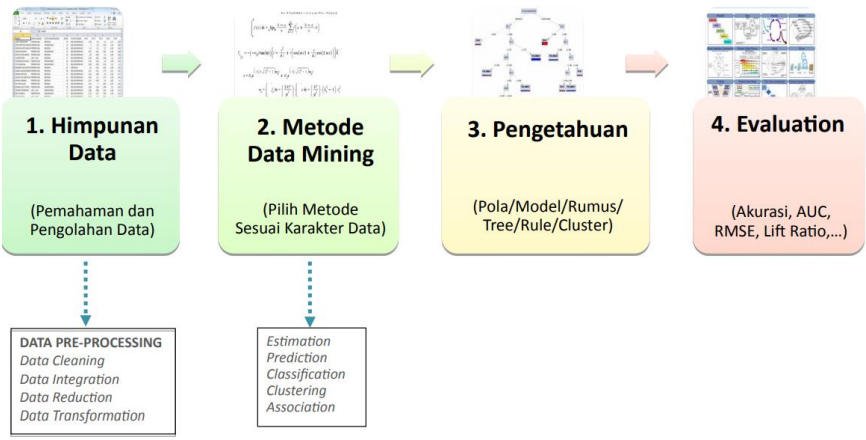
Gambar 5.1: Data Mining

Proses data mining melibatkan langkah-langkah seperti pemahaman bisnis, pemahaman data, pemilihan data relevan, transformasi data, pemodelan menggunakan algoritma khusus, evaluasi hasil, dan implementasi temuan ke dalam keputusan bisnis. Dengan mendalaminya, kita dapat mengungkapkan pola kompleks, membuat prediksi, dan mengidentifikasi informasi yang berharga dari kumpulan data yang mungkin terlalu besar atau kompleks untuk dianalisis secara manual.

Data mining berperan krusial dalam berbagai bidang seperti bisnis, keuangan, kesehatan, dan ilmu pengetahuan. Dalam era informasi ini, di mana data menjadi harta karun, data mining menjadi instrumen yang sangat penting untuk meramalkan tren

pasar, mengoptimalkan operasi bisnis, meningkatkan kepuasan pelanggan, dan mendukung pengambilan keputusan yang berbasis bukti.

Dengan terus berkembangnya teknologi dan ketersediaan data, peran data mining semakin mendalam dan berdampak besar terhadap cara kita memahami dan menggunakan informasi. Sebagai mesin pencari pengetahuan yang tak terbatas, data mining terus menjadi inti dari revolusi informasi, membawa kita menuju pemahaman yang lebih dalam tentang dunia di sekitar kita.



Gambar 1.2: Proses Data Mining

Peran penting data mining dalam konteks era informasi modern di tengah ledakan data yang terus meningkat, data mining telah menjadi tulang punggung bagi organisasi dan individu untuk menggali makna, mendeteksi pola, dan membuat keputusan yang cerdas. Dalam beberapa kata, mari kita menjelajahi esensi dan peran kritis data mining dalam membimbing kita melalui lautan data yang kompleks dan penuh potensi.adapun peran data mining berdasarkan tekniknya yaitu :

1. Classification (Peran dalam Klasifikasi)
2. Association (Peran dalam Asosiasi)
3. Clustering (Peran dalam Klastering)
4. Estimation (Peran dalam Estimasi)
5. Prediction/Forecasting (Peran dalam Prediksi/Peramalan)

B. PERAN DATA MINING DALAM KLASIFIKASI (CLASSIFICATION)

1. Definisi klasifikasi dalam Data Mining

Klasifikasi dalam konteks data mining adalah proses pengelompokan atau kategorisasi data ke dalam kelas atau kelompok berdasarkan karakteristik atau atribut tertentu. Tujuannya adalah untuk mengidentifikasi pola atau hubungan yang ada dalam data sehingga dapat diambil keputusan atau prediksi.

Klasifikasi Merupakan proses pembelajaran suatu fungsi tujuan (target) f yang memetakan tiap himpunan atribut x ke satu dari label kelas y yang didefinisikan sebelumnya.Klasifikasi ini Cocok untuk tipe data biner atau nominal

2. Jenis klasifikasi

Pemodelan deskriptif (descriptive modelling): berfungsi sebagai alat penjelasan untuk membedakan objek-objek dalam kelas-kelas yang berbeda

Pemodelan Prediktif (predictive modelling): digunakan untuk memprediksi label kelas record yang tidak diketahui

3. Pemanfaaatan Data Mining Dalam Klasifikasi

a. **Pemasaran**: Klasifikasi digunakan untuk mengelompokkan pelanggan berdasarkan perilaku pembelian atau preferensi, memungkinkan perusahaan untuk menyusun strategi pemasaran yang lebih terarah dan personal.

b. **Kesehatan**: Identifikasi penyakit berdasarkan gejala dan riwayat medis menggunakan klasifikasi dapat membantu dalam diagnosis cepat dan perencanaan perawatan yang efektif, Memprediksi sel tumor jinak atau ganas, Menggolongkan struktur protein sekunder sebagai alpha-helix, beta-sheet, atau random coil
c. **Keuangan**: Dalam industri keuangan, klasifikasi digunakan untuk mengevaluasi risiko kredit pelanggan, mendeteksi kecurangan transaksi, dan mengidentifikasi pola perubahan pasar, Menggolongkan transaksi kartu kredit sah atau curang
d. **Sumber Daya Manusia (SDM):** Klasifikasi dapat membantu dalam pengelompokan karyawan berdasarkan kinerja atau kemampuan, mendukung proses rekrutmen, dan manajemen talenta.
e. **Teknologi Informasi**: Dalam keamanan jaringan, klasifikasi dapat digunakan untuk mengenali serangan siber dan memfilter email spam.
f. **Ilmu Pengetahuan dan Penelitian**: Klasifikasi membantu dalam analisis data eksperimental, mengidentifikasi pola dalam penelitian ilmiah, dan mendukung pengembangan obat. Mengkategorikan isi berita sebagai finance, weather, entertainment, sports,dll
g. **Pendidikan:** Klasifikasi dapat digunakan untuk menilai kinerja siswa, mengidentifikasi kebutuhan belajar, dan menyusun program pendidikan yang disesuaikan.
h. **Retail dan E-commerce**: Dalam industri ini, klasifikasi membantu dalam prediksi permintaan produk, rekomendasi produk kepada pelanggan, dan analisis pola pembelian.
i. **Transportasi dan Logistik**: Klasifikasi digunakan untuk mengoptimalkan rute pengiriman, mengidentifikasi kebutuhan perawatan pada armada, dan mengelola rantai pasokan.

j. **Agriculture:** Dalam pertanian, klasifikasi dapat membantu dalam identifikasi penyakit tanaman, prediksi hasil panen, dan pengelolaan sumber daya alam.

4. Evaluasi Model (Model Pengujian)

Confusion Matrix: Accuracy, ROC Curve: Area Under Curve (AUC)

5. Metode untuk klasifikasi :

Decision Tree Induction (C4.5, ID3, dll),K-Nearest Neighbor, Naive Bayes, Neural Network, Linear Discriminant Analysis, Logistic Regression

6. Tantangan dalam Klasifikasi:

Tantangan klasifikasi melibatkan overfitting (model terlalu sesuai dengan data pelatihan), ketidakseimbangan kelas (ketidaksetaraan jumlah instans dalam setiap kelas), dan pemilihan parameter yang optimal.

7. Keuntungan Klasifikasi:

Keuntungan utama klasifikasi adalah kemampuannya untuk memberikan struktur dan urutan pada data, membuat keputusan yang cepat, dan mendeteksi pola yang tidak terlihat secara langsung oleh manusia.

C. PERAN DATA MINING DALAM ASOSIASI (ASSOCIATION)

1. Definisi Klasifikasi dalam Data Mining

Asosiasi Adalah sebuah metodologi untuk mencari relasi istimewa/menarik yang tersembunyi dalam himpunan data (data set) yang besar. Relasi yang tersembunyi ini dapat direpresentasikan dalam bentuk aturan asosiasi (association rules) atau himpunan barang yang seringkali muncul (frequent itemset) dengan contoh {roti, mentega} -> {susu} (support = 40%, confidence = 50%) Artinya “Seorang konsumen yang membeli roti dan mentega punya kemungkinan 50% untuk juga membeli susu.

Aturan ini cukup signifikan karena mewakili 40% dari catatan transaksi selama ini."

2. Pemanfaatan Data Mining Dalam Asosiasi

a. **Ritel dan E-Commerce** : Penyusunan paket produk atau rekomendasi produk berdasarkan pola pembelian pelanggan dengan cara Meletakkan barang-barang yang sering dibeli bersamaan dengan posisi berdekatan atau mudah dijangkau
b. **Manajemen Rantai Pasokan**: Mengidentifikasi hubungan antara produk dalam rantai pasokan, memprediksi permintaan, dan mengoptimalkan stok.
c. **Pemasaran**: Segmentasi pelanggan berdasarkan pola pembelian untuk merancang kampanye pemasaran yang lebih terarah dan personal.
d. **Keamanan Jaringan**: Identifikasi pola aktivitas yang mencurigakan atau serangan siber berdasarkan hubungan antara elemen jaringan.
e. **Kesehatan:** Identifikasi pola penyakit berdasarkan riwayat medis pasien atau asosiasi antara gejala tertentu.
f. **Analisis Panen**: Menemukan hubungan antara kondisi cuaca, teknik pertanian, dan hasil panen untuk mendukung keputusan pertanian.
g. **Edukasi:** Pemetaan hubungan antara metode pengajaran, tingkat partisipasi siswa, dan hasil belajar untuk meningkatkan efektivitas pembelajaran.
h. **Eksplorasi Minyak dan Gas**: Menemukan hubungan antara data seismik dan lokasi potensial untuk pengeboran minyak dan gas.
i. **Analisis Kredit dan Keuangan**: Identifikasi hubungan antara faktor keuangan dan risiko kredit dalam penilaian kredit pelanggan.
j. **Sistem Rekomendasi Online**: Menganalisis pola pembelian atau preferensi pengguna untuk memberikan rekomendasi produk atau konten yang sesuai.

k. **Teknologi dan Layanan**: Menganalisis hubungan antara penggunaan layanan dan preferensi pelanggan untuk mengoptimalkan penyediaan layanan teknologi.
l. **Penelitian Ilmiah**: Menemukan korelasi atau asosiasi antara variabel dalam data eksperimental atau penelitian ilmiah.
m. **Pemerintahan**: Menganalisis pola korupsi atau pelanggaran etika dalam data pemerintahan untuk meningkatkan transparansi dan akuntabilitas.

3. Evaluasi Model

Lift Charts: Lift Ratio, Precision and Recall (F-measure)

4. Metode untuk Asosiasi

FP-Growth, A Priori, Coefficient of Correlation, Chi Square

5. Tantangan dalam Asosiasi:

Tantangan utama melibatkan kompleksitas ekstraksi aturan dari data yang besar dan kompleks. Overfitting dan menangani data yang tidak seimbang juga dapat menjadi tantangan.

6. Keuntungan Asosiasi:

Asosiasi memungkinkan organisasi untuk memahami pola pembelian konsumen, meningkatkan strategi pemasaran, dan memberikan rekomendasi yang lebih personal.

D. PERAN DATA MINING DALAM KLASTERING (CLUSTERING)

1. Definisi Klastering dalam Data Mining

Penklusteran (clustering) digunakan untuk melakukan pengelompokan data-data kedalam sejumlah kelompok (cluster) berdasarkan kemiripan karakteristik masing-masing data pada kelompok-kelompok yang ada. Sering disebut unsupervised classification karena label diperoleh dari data. Tujuannya adalah untuk membuat kelompok yang homogen di dalamnya dan heterogen di antara kelompok-kelompok tersebut

2. Pemanfaaatan Data Mining Dalam Klastering

a. **Biologi**: taksonomi makhluk hidup: kingdom, phylum, class, order, family, genus and species
b. **Information retrieval**: pengelompokan dokumen
c. **Penggunaan lahan**: Identifikasi area penggunaan lahan yang serupa dalam database observasi bumi
d. **Pemasaran**: Membantu marketers menemukan kelompok berbeda di basis pelanggan mereka, dan kemudian menggunakan pengetahuan ini untuk mengembangkan program pemasaran
e. **Perencanaan kota**: Mengidentifikasi kelompok-kelompok rumah sesuai dengan jenis rumah, nilai, dan lokasi geografis mereka
f. **Studi gempa bumi**: Episentrum gempa bumi yang diamati harus dikelompokkan di sepanjang patahan benua
g. **Iklim:** Memahami iklim bumi, menemukan pola atmosfer dan lautan
h. **Ilmu Ekonomi**: Riset pasar

3. Evaluasi Model :

Internal Evaluation: Davies–Bouldin index, Dunn index dan untuk External Evaluation: Rand measure, Fmeasure, Jaccard index, Fowlkes–Mallows index, Confusion matrix

4. Metode untuk Klastering :

K-Means , K-Medoids, Self-Organizing Map (SOM) , Fuzzy C-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Mean Shift, Hierarchical Clustering, Agglomerative Clustering, OPTICS (Ordering Points To Identify Clustering Structure), Spectral Clustering, BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies), Affinity Propagation, CLARA (Clustering LARge Applications).

5. **Tantangan dalam Clustering:**

Tantangan utama termasuk menentukan jumlah klaster yang optimal, menangani data yang tidak terstruktur, dan memilih algoritma clustering yang sesuai dengan karakteristik data.

6. **Keuntungan Clustering:**

Clustering membantu dalam mengidentifikasi struktur tersembunyi, menyederhanakan analisis data, dan mendukung pengambilan keputusan dengan memberikan wawasan tentang hubungan antara objek dalam dataset.

E. PERAN DATA MINING DALAM ESTIMASI (ESTIMATION)

1. Definisi Estimasi dalam Data Mining

Estimasi Digunakan untuk menerka sebuah nilai yang belum diketahui, misal menerka penghasilan seseorang ketika informasi mengenai orang tersebut diketahui. Variabel target estimasi lebih ke arah numerik dari pada ke arah kategori.

2. Pemanfaaatan Data Mining Dalam Estimasi

a. **Prediksi Perilaku Konsumen**: Estimasi dapat digunakan untuk memprediksi perilaku konsumen, seperti pembelian produk atau layanan tertentu. Ini membantu perusahaan mengoptimalkan strategi pemasaran dan penawaran produk.

b. **Peramalan Penjualan**: Dengan menggunakan teknik estimasi, perusahaan dapat meramalkan penjualan di masa depan. Hal ini membantu dalam perencanaan produksi, manajemen stok, dan strategi distribusi.

c. **Optimasi Rantai Pasokan**: Estimasi membantu dalam mengoptimalkan rantai pasokan dengan memperkirakan permintaan, waktu pengiriman, dan kebutuhan persediaan. Ini dapat meningkatkan efisiensi dan mengurangi biaya operasional.

d. **Pengelolaan Risiko Keuangan**: Dalam sektor keuangan, estimasi digunakan untuk mengelola risiko, seperti memperkirakan nilai aset, menilai risiko kredit, dan memproyeksikan nilai tukar mata uang.

e. **Peramalan Cuaca**: Estimasi digunakan dalam peramalan cuaca untuk memprediksi suhu, curah hujan, dan kondisi

cuaca lainnya. Ini membantu dalam perencanaan kegiatan terkait cuaca.

f. **Pengoptimalan Energi**: Dalam sektor energi, estimasi membantu mengoptimalkan penggunaan energi dengan memprediksi konsumsi dan membangun strategi manajemen beban.
g. **Analisis Pemasaran**: Estimasi membantu dalam analisis pemasaran dengan mengidentifikasi segmen pasar, menilai efektivitas kampanye iklan, dan memprediksi respons pelanggan terhadap promosi.
h. **Manajemen Proyek:** Dalam manajemen proyek, estimasi digunakan untuk memperkirakan waktu, biaya, dan sumber daya yang dibutuhkan untuk menyelesaikan proyek.
i. **Analisis Jejak Digital**: Estimasi dapat digunakan dalam analisis jejak digital untuk memprediksi perilaku online, mengidentifikasi tren konsumen, dan meningkatkan pengalaman pengguna

3. Evaluasi Model

Error: Root Mean Square Error (RMSE), MSE, MAPE

4. Metode untuk Estimasi :

Linear Regression, Neural Network, Support Vector Machine**,** K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Gradient Boosting, Ridge Regression, Lasso Regression, Elastic Net, Metode Kriging

5. Tantangan dalam Estimasi:

Estimasi selalu melibatkan tingkat ketidakpastian dan kesalahan. Tantangan utama adalah memahami dan mengelola ketidakpastian ini untuk menghindari pengambilan keputusan yang salah.

6. **Keuntungan Estimasi** :

Estimasi memungkinkan analis untuk memahami data lebih dalam dengan mengisi celah informasi yang hilang atau tidak lengkap. Ini membantu dalam mengidentifikasi pola dan tren yang mungkin tidak terlihat secara langsung

F. PERAN DATA MINING DALAM PREDIKSI (FORECASTING)

1. Definisi Prediksi dalam Data Mining

Forecasting merupakan proses memproyeksikan nilai atau perilaku di masa depan berdasarkan pola dan tren yang ditemukan dalam data historis. Teknik Forecasting akan mengambil sederetan angka yang menunjukkan nilai yang berjalan seiring waktu dan kemudian Teknik Forecasting ini akan menghubungkan nilai masa depan dengan menggunakan bermacam-macam teknik machine-learning dan teknik statistik yang berhubungan dengan musim, trend, dan noise pada data.

2. Pemanfaaatan Data Mining Dalam Prediksi

a. **Ekonomi dan Keuangan**: Prediksi pertumbuhan ekonomi, pergerakan pasar saham, atau nilai tukar mata uang.
b. **Kesehatan**: Prediksi penyebaran penyakit dan epidemiologi, Estimasi waktu tinggal pasien di rumah sakit dan prediksi beban kerja rumah sakit.
c. **Pemasaran dan Penjualan**: Prediksi penjualan dan permintaan produk atau layanan.
d. **Transportasi dan Logistik**: Prediksi waktu kedatangan (ETA) untuk pengiriman dan transportasi.
e. **Sains Sosial dan Humaniora**: Prediksi hasil pemilihan dan perilaku pemilih, Analisis sentimen media sosial untuk memprediksi tren opini public, Estimasi dampak kebijakan sosial dan ekonomi.
f. **Pertanian:** Prediksi hasil panen dan ketersediaan sumber daya pertanian.
g. **Energi**: Prediksi konsumsi energi dan permintaan Listrik, Optimasi produksi energi dari sumber-sumber terbarukan, Prediksi harga energi dan fluktuasi pasar.
h. **Pendidikan**: Prediksi kinerja siswa dan tingkat kelulusan, Pemantauan dan prediksi kebutuhan untuk sumber daya Pendidikan, Rekomendasi karir berdasarkan pilihan dan prestasi siswa.

i. **Bisnis Ritail**: Prediksi preferensi dan perilaku pembelian pelanggan, Estimasi kebutuhan stok dan pengelolaan persediaan, Penyesuaian harga berdasarkan prediksi permintaan dan kompetisi.
j. **Manufaktur:** Prediksi ketersediaan dan permintaan produk, Perencanaan produksi dan manajemen rantai pasokan, Mengoptimalkan waktu dan penggunaan mesin dalam produksi.
k. **Hukum dan Kepolisian**: Prediksi tingkat kejahatan dan pola kejahatan, Pengidentifikasian potensi pelanggar hukum atau tindakan teroris, Prediksi kebutuhan keamanan dan penempatan personel.
l. **Pariwisata:** Prediksi tingkat kunjungan wisatawan dan permintaan akomodasi, Pengelolaan lalu lintas wisata dan pengaturan layanan, peyesuaian harga berdasarkan musim wisata dan tren permintaan.

3. **Evaluasi Model**

Error: Root Mean Square Error (RMSE) , MSE, MAPE

4. **Metode untuk Prediksi** :

Linear Regression, Neural Network, Support Vector Machine, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, ARIMA (AutoRegressive Integrated Moving Average), Gradient Boosting, Naive Bayes

5. Tantangan dalam prediksi :

prediksi selalu melibatkan ketidak lengkapan data, anomali data, ketidak seimbangan data, overfitting dan underfitting, dimensi data karena variabelyang banyak dan ketergantungan pada data historis.

6. Keuntungan Prediksi :

Prediksi dalam data mining menawarkan sejumlah keuntungan yang signifikan, membantu organisasi dan individu membuat keputusan yang lebih tepat dan merencanakan strategi dengan lebih baik

G. RANGKUMAN

Data mining adalah proses ekstraksi pola dan pengetahuan yang berharga dari dataset besar. Dalam praktiknya, beberapa tugas utama data mining mencakup estimasi, prediksi, klastering, klasifikasi, dan estimasi. Berikut adalah rangkuman singkat tentang masing-masing konsep tersebut:

Estimasi melibatkan penaksiran nilai suatu variabel berdasarkan informasi yang tersedia. Metode regresi, seperti regresi linier atau regresi logistik, sering digunakan untuk memodelkan hubungan antar variabel dan membuat perkiraan berdasarkan data historis.

Prediksi adalah tugas untuk memproyeksikan nilai atau perilaku di masa depan berdasarkan pola dan tren dalam data historis. Algoritma prediksi mencakup decision trees, neural networks, dan metode ensemble seperti random forests, yang memungkinkan pembuatan perkiraan yang akurat.

Klastering melibatkan pengelompokan objek atau data ke dalam kelompok yang homogen berdasarkan kesamaan karakteristik. Algoritma k-means dan hierarchical clustering sering digunakan untuk menemukan struktur dalam data dan membentuk kelompok yang bermakna.

Klasifikasi adalah tugas untuk mengelompokkan objek ke dalam kelas atau kategori yang telah ditentukan. Algoritma klasifikasi, seperti Naive Bayes, Support Vector Machines, dan decision trees, digunakan untuk membuat model yang dapat memprediksi kelas objek baru.

Asosiasi melibatkan identifikasi hubungan dan asosiasi antara variabel atau item dalam dataset. Algoritma seperti Apriori digunakan untuk menemukan aturan asosiasi yang menggambarkan keterkaitan antara item atau variabel.

Secara keseluruhan, data mining memungkinkan analis untuk menggali wawasan berharga dari data yang kompleks dan besar. Estimasi dan prediksi membantu dalam meramalkan dan memahami perilaku masa depan, sementara klastering dan klasifikasi memberikan struktur dan kategorisasi yang diperlukan. Asosiasi membantu mengidentifikasi pola hubungan yang dapat memberikan pemahaman lebih mendalam tentang dataset. Dengan kombinasi metode ini, data mining menjadi alat yang kuat untuk pengambilan keputusan dan inovasi di berbagai bidang.

H. TES FORMATIF

1. Apa yang dimaksud dengan "data mining"?
   a. Penyimpanan data dalam database
   b. Proses ekstraksi pola dan pengetahuan dari data
   c. Penambahan data ke dalam sistem
   d. Pembuatan data visual
   e. Proses validasi data
2. Algoritma yang digunakan untuk membuat model pohon keputusan dalam data mining disebut:
   a. K-Means
   b. Apriori
   c. Decision Tree
   d. Support Vector Machine
   e. KNN

I. LATIHAN

Definisikan konsep Data Mining dan jelaskan mengapa penting dalam konteks organisasi dan pengambilan keputusan. Berikan contoh konkretnya dalam suatu industri atau sektor.

KEGIATAN BELAJAR 6

PREDICTION AND CLASSIFICATION

DESKRIPSI PEMBELAJARAN

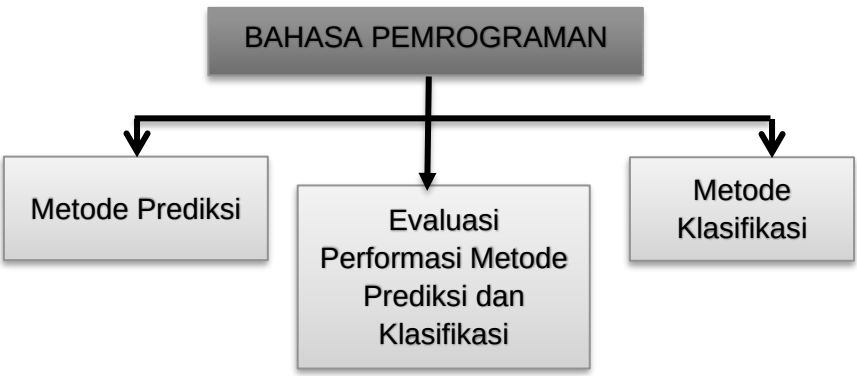
Pada bab ini mahasiswa mempelajari *predictive analytics* data mining yaitu metode prediksi dan metode klasifikasi. Diharapkan mahasiswa memiliki wawasan dan pemahaman tentang metode dan juga algoritma yang digunakan dalam melakukan prediksi dan klasifikasi dalam data mining.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan:

1. Mampu menjelaskan metode prediksi dan klasifikasi dalam data mining.
2. Mampu melakukan evaluasi algoritma metode prediksi dan algoritma klasifikasi.

PETA KONSEP PEMBELAJARAN



A. PREDICTIVE ANALYTICS

Predictive analytics atau analitik prediktif, merupakan suatu metode yang kuat untuk menganalisis data dengan tujuan memprediksi peristiwa di masa depan. Metode ini sering digunakan oleh para ahli business intelligence untuk memahami kondisi bisnis berdasarkan sejarah data.

Tujuan utama dari analitik prediktif adalah untuk memprediksi hasil di masa depan, bukan perilaku data saat ini. Ini melibatkan fungsi pembelajaran terawasi (*supervised learning*) yang digunakan untuk prediksi nilai target. Metode yang termasuk dalam kategori penambangan ini adalah metode estimasi atau prediksi dan metode klasifikasi. Pemodelan data memerlukan analisis prediktif yang bekerja dengan memanfaatkan beberapa variabel untuk mengantisipasi nilai data masa depan yang tidak diketahui untuk variabel lain.

Analisis Prediktif dijadikan sebagai alat yang menggunakan teknik statistik, pembelajaran mesin, dan penambangan data untuk menemukan fakta untuk membuat prediksi tentang peristiwa masa depan yang tidak diketahui. Model prediksi membuat perkiraan tentang nilai data yang tidak dikenal dengan menggunakan nilai yang diidentifikasi. Tentunya untuk setiap kasus prediktif harus dianalisis kebutuhannya untuk mengetahui metode yang paling cocok untuk mengekstraksi pengetahuan yang diinginkan. Analisis prediktif digunakan untuk memberikan informasi tentang "apa yang mungkin terjadi?" dan "mengapa itu bisa terjadi?". Banyak aplikasi data mining yang diperuntukkan untuk meramalkan keadaan data yang akan datang.

Dengan analitik prediktif, kita dapat mengantisipasi dan menghindari potensi masalah yang mungkin muncul serta menyusun solusi yang tepat. Ini melibatkan berbagai teknik statistik seperti pemodelan prediktif, data mining, dan

pembelajaran mesin yang memungkinkan kita untuk menganalisis aliran data historis dan meramalkan peristiwa yang belum terduga.

Perlu dicatat bahwa analitik prediktif berbeda dari peramalan, karena ia dapat memberikan prediksi yang sangat rinci dan terperinci, bahkan hingga tingkat elemen individual dalam organisasi, yang dapat menghasilkan probabilitas prediksi yang lebih akurat.

B. METODE PREDIKSI

Teknik estimasi adalah salah satu cara untuk mendapatkan nilai perkiraan dalam data. Ini melibatkan penggunaan pola-pola yang terdapat dalam kumpulan data untuk memprediksi nilai-nilai masa depan yang saat ini belum diketahui. Dalam teknik estimasi, kita menggunakan beberapa variabel untuk memprediksi nilai-nilai variabel target yang kita butuhkan. Dalam konteks ini, ada dua jenis variabel yang penting: variabel prediktor dan variabel target.

Variabel target dalam teknik estimasi adalah jenis variabel yang memiliki nilai numerik yang bersifat kontinu, bukan kategori. Estimasi nilai dari variabel target didasarkan pada nilai-nilai dari variabel prediktor atau atribut lain yang tersedia.

Selain teknik estimasi, ada juga teknik prediksi atau forecasting. Teknik ini mirip dengan teknik estimasi, tetapi berfokus pada data time series, yaitu data yang berurutan dalam rentang waktu tertentu. Meskipun istilah "prediksi" sering digunakan dalam konteks ini, itu tidak hanya berlaku untuk data time series. Teknik prediksi dapat digunakan untuk klasifikasi juga, di mana kita mencoba untuk mengklasifikasikan data ke dalam berbagai kategori berdasarkan atribut yang tersedia.

Pentingnya, semua algoritma yang digunakan dalam teknik estimasi juga dapat diterapkan dalam teknik prediksi atau forecasting. Jadi, tergantung pada jenis data dan tujuan analisis, kita dapat memilih antara teknik estimasi atau prediksi untuk memahami dan memanfaatkan data dengan lebih baik.

Sebagai gambaran, jika terdapat data mengenai lamanya waktu yang dihabiskan untuk seorang pegawai JFC untuk mengantar pesanan ke rumah pembeli. Pegawai tersebut mengendarai sepeda motor untuk mengantarkan pesanan langsung ke rumah pemesan. Data waktu sejak pegawai meninggalkan restoran hingga mencapai pintu rumah pelanggan ditampilkan pada tabel di bawah ini.

Tabel 6.1. Dataset Delivery Pemesanan JFC

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

Berdasarkan data dalam tabel, seandainya datang pesanan ke-1001 dengan jumlah pesanan 4 paket dari rumah yang jaraknya 5 km dari restoran dimana terdapat 8 traffic light yang harus dilewati, waktu pengantaran pesanan bagi pelanggan tersebut seharusnya langsung dapat diprediksi. Prediksi tersebut didasari anggapan bahwa waktu dipengaruhi oleh jarak rumah pelanggan. Apakah anggapan tersebut benar?

Contoh di atas mengilustrasikan aktivitas prediksi. Namun, bagaimana perbedaannya dengan estimasi? Estimasi adalah proses untuk mengestimasi atau menebak suatu nilai, seperti rata-rata populasi, berdasarkan data sampel yang telah dikumpulkan.

Estimasi dilakukan dengan memperhitungkan data sampel yang tersedia. Di sisi lain, prediksi melibatkan penggunaan data yang sudah ada untuk meramalkan hasil dari sebuah peristiwa yang akan datang. Dengan kata lain, estimasi digunakan untuk menghitung nilai yang tidak diketahui saat ini (seperti rata-rata populasi atau varians populasi), sementara prediksi digunakan untuk meramalkan hasil dari peristiwa yang akan terjadi di masa depan.

Selanjutnya, pertanyaannya adalah algoritma apa yang digunakan untuk teknik prediksi atau estimasi? Terdapat beberapa algoritma yang dapat digunakan dalam teknik prediksi/estimasi, seperti regresi linear, jaringan saraf tiruan (*neural network*), mesin vektor pendukung (*support vector machine*), *k-nearest neighbor*, dan banyak lainnya. Pemilihan algoritma yang paling cocok bergantung pada kebutuhan kasus yang dihadapi dan jenis data yang digunakan.

Pada teknik prediksi atau estimasi, pengetahuan yang dihasilkan adalah berupa formula atau fungsi atau rumus. Berdasarkan kasus pengantaran paket dari JFC di atas maka algoritma yang tepat adalah dengan menggunakan algoritma regresi linier berganda, dimana ada tiga variabel sebagai variabel predictor yaitu jumlah pesanan (P), jumlah traffic light (TL), dan Jarak (J). Sedangkan Waktu Tempuh (T) sebagai variabel target. Dari kasus di atas maka akan dihasilkan sebuah rumus atau formula atau fungsi regresi sebagai berikut:

$$Waktu\ Tempuh\ (T) = 0.48P + 0.23TL + 0.5J$$

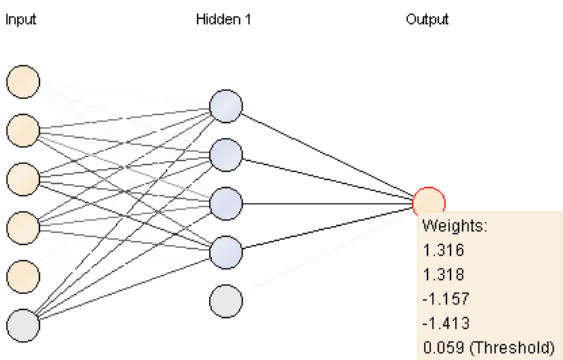
Pengetahuan yang diperoleh dari persamaan di atas adalah prediksi untuk waktu tempuh pengiriman pesanan. Waktu tempuh untuk pengiriman pesanan adalah 0.48 dikali jumlah pesanan ditambah 0.23 dikali jumlah trafik light ditambah 0.5 dikali jarak rumah dengan JFC.

Contoh kasus lain misalnya pada kasus memprediksi Harga Saham. Di tabel di bawah ini dapat dilihat dataset harga saham dalam bentuk time seris (rentet waktu), dimana variabel targetnya adalah variable harga penutupan (close).

Tabel 6.2. Dataset time series Saham

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	272010000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Algoritma yang bisa digunakan untuk melakukan pembelajaran (learning) dan training pada data di atas adalah neural network. Pengetahuan apa yang didapatkan berupa rumus neural network.



Gambar 6.1 Hasil Algoritma Neural Network

Dengan hasil training dari algoritma neural network nantinya kita bisa memprediksi berapa harga saham misalnya untuk 10 hari atau sebulan kedepan.

Gambar 6.2 Visualisasi Prediksi Harga Saham menggunakan Neural Network

C. METODE KLASIFIKASI

Klasifikasi adalah proses evaluasi data untuk mengategorikannya ke dalam salah satu dari beberapa kelas yang telah ditentukan. Dalam klasifikasi, terdapat dua tugas utama yang dilakukan: pertama, membangun model sebagai representasi prototipe yang akan disimpan sebagai referensi, dan kedua, menggunakan model tersebut untuk mengenali atau mengklasifikasikan objek data lain sehingga dapat ditempatkan dalam salah satu kelas yang ada dalam model yang telah disimpan.

Tujuan dari metode klasifikasi adalah memahami sebuah dataset sehingga kita dapat membuat aturan atau model yang mampu mengklasifikasikan data baru yang belum pernah kita lihat sebelumnya. Klasifikasi dapat dijelaskan sebagai proses untuk mengidentifikasi suatu objek data sebagai anggota dari salah satu kategori yang telah didefinisikan sebelumnya. Klasifikasi digunakan dalam berbagai aplikasi seperti profil pelanggan, deteksi kecurangan, diagnosis medis, prediksi penjualan, prediksi kelulusan mahasiswa, dan lain sebagainya.

Bagaimana kita membangun model klasifikasi? Model klasifikasi dapat dikembangkan berdasarkan pengetahuan seorang ahli, tetapi pada umumnya, model klasifikasi dibangun menggunakan teknik pembelajaran mesin (machine learning), terutama karena ukuran dataset yang besar. Proses pembelajaran otomatis pada dataset dapat menghasilkan model klasifikasi (fungsi target) yang mampu menghubungkan data masukan (input) ke salah satu kelas yang telah ditentukan sebelumnya. Dengan kata lain, proses pembelajaran memerlukan dataset pelatihan yang memiliki label (kelas) dan menghasilkan model klasifikasi sebagai output. Jenis data yang digunakan dalam klasifikasi dapat berupa data nominal, ordinal, atau kontinu, tergantung pada konteks dan tujuan analisis.

Misalkan pada dataset di bawah ini, terdapat 11000 data yang tujuannya adalah untuk menentukan apakah mahasiswa tersebut berpeluang untuk lulus tepat waktu atau tidak.

Tabel 6.3 Dataset Kelulusan Mahasiswa

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

Bagaimana cara untuk menentukan lulus tepat waktu atau tidaknya mahasiswa? Tentunya yang harus menggunakan algoritma klasifikasi diantara *Naive Bayes*, *K-Nearest Neighbor*, ID3, C4.5, CART, *Linear Discriminant Analysis*, *Logic Regression*, dan lain-lain. Penggunaan algoritma disesuaikan dengan kasus yang ditemukan. Dalam kasus di atas menggunakan algoritma klasifikasi C4.5. Pengetahuan yang dihasilkan dari metode klasifikasi adalah pohon keputusan.

Decision tree adalah salah satu metode yang sering digunakan dalam Data Mining. Ini merupakan representasi struktur berbentuk pohon di mana setiap simpul internal mewakili pengujian pada suatu atribut, cabang-cabangnya menggambarkan hasil dari pengujian tersebut, dan simpul daun mewakili kelas atau distribusi kelas. Untuk menentukan prediksi kelas suatu contoh data, kita mengikuti alur dari simpul akar ke simpul daun yang menghasilkan prediksi kelas tersebut. Decision tree juga dapat dengan mudah diubah menjadi aturan klasifikasi yang lebih mudah dimengerti.

Dengan kata lain, *Decision Tree* adalah struktur berbentuk pohon di mana setiap simpul mewakili atribut yang diuji, cabang-cabangnya menggambarkan hasil dari pengujian tersebut, dan simpul daun mewakili kelompok kelas tertentu. Biasanya, simpul paling atas dari pohon *decision tree* disebut simpul akar, dan ini adalah atribut yang memiliki pengaruh paling besar terhadap suatu kelas tertentu. *Decision tree* umumnya mengikuti strategi pencarian dari atas ke bawah (*top-down*) untuk menentukan solusi.

Ketika mengklasifikasikan data yang tidak diketahui, nilai-nilai atribut diuji dengan mengikuti jalur dari simpul akar (*root*) hingga mencapai simpul daun yang akhir, dan kemudian prediksi kelas diberikan berdasarkan simpul daun yang terpilih. Decision tree adalah salah satu metode yang populer dan dapat digunakan untuk memahami hubungan atribut dalam data dan membuat prediksi berdasarkan pengujian atribut tersebut.

Bagaimana membaca hasil dari *decision tree*? Dari hasil perhitungan C4.5 didapatkan bahwa status mahasiswa yang menjadi penentu awal apakah mahasiswa lulus tepat waktu atau tidak. Misalkan status mahasiswa adalah bekerja, maka penentu selanjutnya adalah IPS (indeks prestasi semester) 2 dimana jika IPS 2 lebih besar dari 3.590 maka mahasiswa akan lulus tepat waktu tetapi kalau misalkan IPS 2 kurang dari 3.590 maka IPS 8

yang akan menjadi pertimbangan. Jika IPS 8 lebih besar dari 2.95 maka selanjutnya yang dicek adalah IPS 3. Jika IPS3 kurang dari 2.57 maka tepat waktu sedangkan IPS 3 lebih besar dari 2.57 maka akan terlambat lulus.

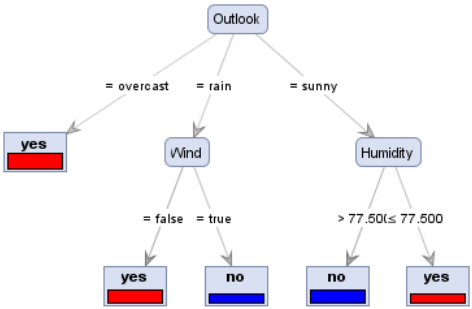
Hasil dari *decision tree* atau pohon keputusan sangat ketergantungan dengan data. Semakin detail, akurat, dan banyak datanya maka hasil decision tree akan semakin baik, begitu juga sebaliknya semakin sedikit data maka hasil *decision tree* akan semakin tidak maksimal.

Contoh kasus lain misalnya terdapat dataset yang digunakan untuk merekomendasikan apakah pemain golf bisa bermain atau tidak dengan melihat variabel-variabel berikut yaitu outlook, temperature, humidity, dan windy. Variabel targetnya adalah play untuk memutuskan yes atau no.

Tabel 6.4 Dataset Bermain Golf

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Algoritma yang digunakan pada kasus diatas bisa menggunakan ID3 atau C4.5. Hasil pengetahuan dari kasus diatas adalah sebuah pengetahuan berupa decision tree sebagai berikut:



Gambar 6.3 Hasil Klasifikasi C4.5 Dataset Bermain Golf

Dalam kasus di atas maka outlook (cuaca) adalah variabel yang dilihat pertama kali untuk menentukan apakah direkomendasikan untuk bermain atau tidak. Jika cuacanya berawan (overcast) maka keputusannya adalah bermain jika cuaca hujan maka yang dicek apakah berangin atau tidak sedangkan jika cuaca cerah maka yang kemudian dicek adalah kelembabannya.

Selain dalam bentuk pohon keputusan, bentuk output dari metode klasifikasi adalah rules (aturan) yaitu:

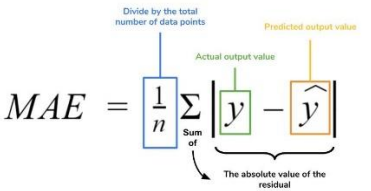
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes

D. EVALUASI METODE PREDIKSI DAN KLASIFIKASI

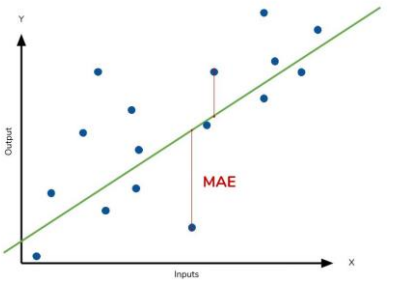
Ada dua metode yang digunakan dalam predictive data mning yaitu metode prediksi dan metode klasifikasi. Pada metode prediksi menggunakan RMSE (*Root Mean Square Error*) dan MAE (*Mean Absolute Error*) untuk evaluasi performansi dan pada metode klasifikasi menggunakan *Performance*, *Precision*, and *Recall*.

1. **Kinerja Algoritma Prediksi**

   MAE mengukur besarnya rata-rata kesalahan dalam serangkaian prediksi, tanpa mempertimbangkan arahnya. Ini adalah rata-rata dari sampel uji tentang perbedaan mutlak antara prediksi dan pengamatan aktual di mana semua perbedaan individu memiliki bobot yang sama.



   Gambar di bawah ini adalah deskripsi grafis dari MAE. Garis hijau mewakili prediksi model, dan titik biru mewakili data.



Gambar 6.4 Evaluasi menggunakan MAE

   MAE paling intuitif dari metrik karena kita hanya melihat perbedaan absolut antara data dan prediksi model. Karena kita menggunakan nilai absolut dari residu, MAE tidak menunjukkan kinerja yang kurang atau kinerja berlebih dari model (apakah model di bawah atau melebihi data aktual). Setiap residu berkontribusi secara proporsional terhadap jumlah total kesalahan, yang berarti bahwa kesalahan yang lebih besar akan berkontribusi secara linear terhadap

keseluruhan kesalahan. Seperti yang dikatakan di atas, MAE kecil menunjukkan model ini hebat dalam prediksi, sementara MAE besar menunjukkan bahwa model Anda mungkin mengalami masalah di area tertentu. MAE 0 berarti bahwa model Anda adalah prediktor yang sempurna untuk keluaran (tetapi ini hampir tidak akan pernah terjadi).

Root Mean Square Error (RMSE) adalah akar kuadrat dari MSE. Para peneliti akan sering menggunakan RMSE untuk mengubah metrik kesalahan kembali ke unit serupa, membuat interpretasi lebih mudah.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

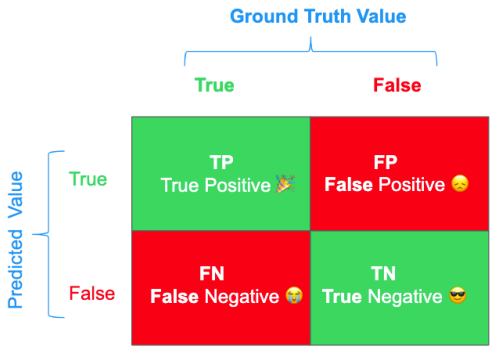
Karena RMSE kuadratkan residu, RMSE juga dipengaruhi oleh outlier. RMSE dapat dianalogikan sebagai standar deviasi dan merupakan ukuran seberapa besar residu tersebar. Baik MAE dan RMSE dapat memiliki nilai berkisar dari 0 hingga positif tidak terhingga, sehingga karena kedua langkah ini semakin tinggi, semakin sulit untuk menafsirkan seberapa baik kinerja model Anda. Cara lain kita dapat meringkas koleksi residu kami adalah dengan menggunakan persentase sehingga setiap prediksi diskalakan terhadap nilai yang seharusnya diestimasi.

2. Kinerja Algoritma Klasifikasi

Dalam mengevaluasi kinerja algoritma klasifikasi, *confusion matrix* memegang peranan penting. *Confusion Matrix* adalah matrik untuk pengukuran kinerja untuk model supervised learning khususnya klasifikasi pada machine learning.

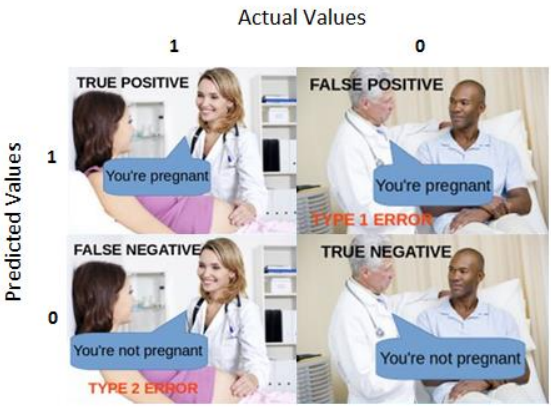
Pengukuran kinerja algoritma klasifikasi dalam machine learning dengan melihat nilai prediksi dan nilai aktual. Outcome

dari klasifikasi Terdapat 4 perbandingan antara nilai prediksi dan aktual, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).



Gambar 6.5 Confusion Matrix

TP dan TN adalah klasifikasi yang diharapkan sedangkan FP dan FN adalah error yang kemungkinan terjadi. Sebelum kita masuk ke pengukuran kinerjanya maka penting untuk memahami 2 jenis error pada klasifikasi dengan melihat gambar berikut:



Gambar 6.5 Contoh Ilustrasi Confusion Matrix

a. *True Positive*:
   Interpretasi: Anda meramalkan positif dan realitanya juga benar.

   Anda meramalkan bahwa seorang wanita hamil dan dia sebenarnya.
b. *True Negative*:
   Interpretasi: Anda memperkirakan negatif dan realitanya juga benar.

   Anda meramalkan bahwa seorang pria tidak hamil dan dia sebenarnya tidak hamil.
c. *False Positive*: (Kesalahan Tipe 1)
   Interpretasi: Anda meramalkan positif dan realitanya salah.

   Anda meramalkan bahwa seorang pria hamil tetapi sebenarnya tidak.
d. *False Negative*: (Kesalahan Tipe 2)
   Interpretasi: Anda memperkirakan negatif dan realitanya salah.

   Anda meramalkan bahwa seorang wanita tidak hamil tetapi sebenarnya dia hamil.

Akurasi, presisi, dan recall adalah tiga metrik evaluasi yang sering digunakan dalam pemodelan klasifikasi untuk mengukur seberapa baik performa suatu model klasifikasi. Mari kita jelaskan masing-masing metrik ini:

Akurasi (*Accuracy*):

- Akurasi adalah metrik yang mengukur sejauh mana model klasifikasi mampu memprediksi kelas dengan benar dari seluruh kasus yang dievaluasi.
- Rumusnya adalah (Jumlah prediksi yang benar) / (Jumlah total prediksi).

- Akurasi memberikan gambaran tentang seberapa baik model dapat mengklasifikasikan seluruh kelas dengan benar. Namun, dalam beberapa kasus di mana kelas memiliki distribusi yang tidak seimbang (imbalance class), akurasi mungkin tidak menjadi metrik yang paling informatif.

Presisi (*Precision*):

- Presisi adalah metrik yang mengukur sejauh mana prediksi positif dari model (misalnya, prediksi bahwa suatu contoh masuk ke dalam suatu kelas) adalah benar atau relevan.
- Rumusnya adalah (Jumlah prediksi positif yang benar) / (Jumlah total prediksi positif).
- Presisi memberikan gambaran tentang tingkat keakuratan prediksi positif. Ini berguna ketika kita ingin meminimalkan jumlah hasil positif palsu.

Recall (*Recall* atau *Sensitivity* atau *True Positive Rate*):

- Recall adalah metrik yang mengukur sejauh mana model mampu mengidentifikasi semua contoh sebenarnya yang masuk ke dalam suatu kelas tertentu.
- Rumusnya adalah (Jumlah prediksi positif yang benar) / (Jumlah total contoh sebenarnya yang termasuk ke dalam kelas tersebut).
- Recall berguna ketika kita ingin meminimalkan jumlah kasus yang sebenarnya positif yang terlewatkan oleh model (false negatives).

Seringkali, ada trade-off antara presisi dan recall. Peningkatan presisi biasanya mengurangi recall, dan sebaliknya. Oleh karena itu, tergantung pada tujuan aplikasi Anda, Anda mungkin harus memilih metrik yang sesuai untuk mengevaluasi model Anda. Misalnya, dalam kasus deteksi spam email, recall (agar tidak ada email penting yang terlewatkan) mungkin lebih penting daripada presisi (meskipun

ada beberapa false positives). Di sisi lain, dalam diagnosis medis, presisi (agar diagnosis positif benar-benar akurat) mungkin lebih penting daripada recall. Pemilihan metrik harus selalu disesuaikan dengan konteks aplikasi klasifikasi yang Anda hadapi.

E. RANGKUMAN

Penambangan prediktif bertujuan untuk memprediksi hasil di masa depan, bukan perilaku data saat ini. Ini melibatkan fungsi pembelajaran terawasi (*supervised learning*) yang digunakan untuk prediksi nilai target. Metode yang termasuk dalam kategori penambangan ini adalah metode estimasi atau prediksi dan metode klasifikasi. Algoritma pada metode prediksi atau estimasi antara lain *linier regression, neural network, support vector machine, k-nearest neighbor*, dan lain-lain. Algoritma pada metode klasifikasi antar lain *Naive Bayes, K-Nearest Neighbor, ID3, C4.5, CART, Linear Discriminant Analysis, Logic Regression*, dan lain-lain. Pengetahuan yang dihasilkan pada metode estimasi / prediksi adalah sebuah formula atau rumus sedangkan pada klasifikasi, pengetahuan yang dihasilkan adalah sebuah pohon keputusan atau rules.

Untuk memastikan model atau algoritma *machine learning* yang akan digunakan memiliki performansi yang baik, maka perlu dilakukan evaluasi. Dalam *predictive analytic* untuk metode prediksi atau estimasi utamanya algoritma regresi dapat menggunakan MAE dan RMSE sehingga dapat diketahui seberapa besar error yang dihasilkan dari persamaan regresi yang telah dihasilkan. Semakin kecil nilai MAE dan RMSE maka semakin baik performansi dari algoritma regresi linier. Pada metode klasifikasi, untuk mengukur kinerja dengan menggunakan *confusion matrix*. *Confusion Matrix* adalah matrik untuk pengukuran kinerja untuk model *supervised learning* khususnya

klasifikasi pada *machine learning*. Dengan menggunakan matriks ini maka dapat diketahui *accuracy, precision,* dan *recall*.

F. TES FORMATIF

1. Pengetahuan apa yang diperoleh setelah menggunakan algoritma regresi linier?
   a. Rumus atau formula
   b. Pohon keputusan
   c. Aturan asosiasi
   d. Klaster
2. Algoritma yang digunakan pada metode klasifikasi adalah sebagai berikut, kecuali …..
   a. Regresi linier, C4.5, ID3, K-NN
   b. C4.5, CART, K-Means, K-NN
   c. Bayes Classifier, C4.5, K-NN, CART
   d. Regresi Linier, K-Means, C4.5 K-NN
3. Soal no. 1 dan 2 akan menggunakan data di bawah ini:

Properti	Harga Aktual	Harga Prediksi
Rumah 2 kamar	$200.000	$230.000
Rumah 3 kamar	$300.000	$290.000
Rumah 4 kamar	$400.000	$740.000
Rumah 5 kamar	$500.000	$450.000

Berapa tingkat error dengan menggunakan MAE?
   a. $230.500
   b. $107.500
   c. $170.500
   d. $90.500

4. Jika seorang pasien tidak menderita kanker tetapi dari model yang dibuat memprediksi pasien tersebut menderita kanker, dalam confusion matrix termasuk :
   a. *True Positive*
   b. *True Negative*
   c. *False Positive*
   d. *False Negative*

G. LATIHAN

1. Tentukan masing-masing 5 contoh studi kasus untuk metode prediksi dan metode klasifikasi!
2. Lakukan evaluasi performasi klasifikasi pada kasus di bawah ini!

NIM	Nilai Aktual	Nilai Prediksi
001	Tidak DO	Tidak DO
002	Tidak DO	Tidak DO
003	Tidak DO	Tidak DO
004	Tidak DO	DO
005	Tidak DO	DO
006	DO	Tidak DO
007	DO	DO
008	DO	DO
009	DO	DO
010	DO	DO

Tentukan Data Confusion Matrixnya dan evaluasi akurasi, presisi, dan recall!

KEGIATAN BELAJAR 7

CLUSTERING ANALYSIS

DESKRIPSI PEMBELAJARAN

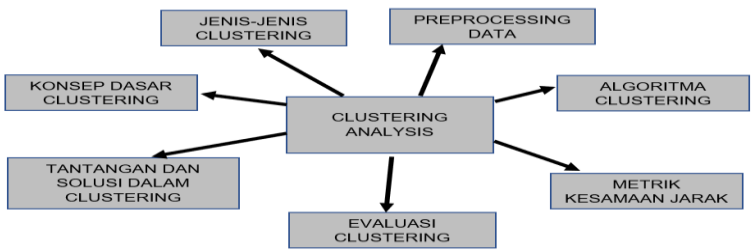
Pada bab ini mahasiswa mempelajari clustering analysis. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk modal dasar mempelajari clustering lebih lanjut. Karena pembelajaran clustering analysis melibatkan sejumlah langkah dan konsep yang harus dipahami secara mendalam.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan memiliki pengetahuan dan kemampuan :

1. Memahami apa itu clustering dan pentingnya dalam analisis data.
2. Mengenal jenis-jenis clustering: partitional, hierarchical, density-based, dan lainnya.
3. Memahami berbagai metrik seperti Euclidean, Manhattan, dan Cosine.
4. Menggunakan dataset nyata, melakukan analisis clustering, dan menyajikan temuan.
5. Mempu memanfaatkan bahasa pemrograman untuk clustering.

PETA KONSEP PEMBELAJARAN



A. KONSEP DASAR CLUSTERING

Clustering adalah metode penting dalam analisis data yang mengelompokkan sekumpulan objek ke dalam kelompok atau klaster berdasarkan kesamaan karakteristik mereka. Dalam clustering, objek dalam klaster yang sama memiliki kesamaan yang lebih tinggi dibandingkan dengan objek di klaster lain. Definisi ini mencerminkan prinsip dasar clustering, yaitu "homogenitas internal dan heterogenitas eksternal" dalam kelompok.

Tujuan utama dari clustering adalah untuk menemukan struktur tersembunyi dalam data yang tidak terstruktur atau semi-terstruktur. Dengan mengelompokkan data berdasarkan kesamaan fitur, clustering membantu mengidentifikasi pola dan hubungan yang tidak terlihat pada pengamatan kasual. Ini khususnya penting dalam data besar, di mana volume dan kompleksitas data melebihi kemampuan analisis manual.

Pentingnya clustering dalam analisis data terletak pada kemampuannya untuk memberikan wawasan yang tidak eksplisit. Clustering digunakan dalam berbagai bidang seperti pemasaran untuk segmentasi pasar, di bidang keuangan untuk pengelompokan profil risiko, dalam bioinformatika untuk analisis ekspresi gen, dan banyak lagi. Misalnya, dalam pemasaran, clustering dapat membantu mengidentifikasi kelompok pelanggan dengan preferensi serupa, memungkinkan perusahaan untuk menargetkan kampanye pemasaran mereka dengan lebih efektif.

Selain itu, clustering sering menjadi langkah awal dalam proses analisis data yang lebih besar. Dalam machine learning, misalnya, clustering bisa digunakan untuk pemilihan fitur atau reduksi dimensi sebelum menerapkan algoritma pembelajaran yang lebih kompleks. Ini membantu mengurangi kebisingan dalam data dan meningkatkan efisiensi model pembelajaran.

Salah satu kekuatan utama dari clustering adalah fleksibilitasnya. Berbagai algoritma telah dikembangkan untuk menangani jenis data dan kebutuhan analisis yang berbeda. Beberapa algoritma clustering yang paling umum meliputi K-Means, hierarchical clustering, dan DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Masing-masing memiliki keunggulan dan keterbatasannya sendiri, menjadikan pemilihan algoritma sebagai aspek penting dalam proses clustering.

K-Means, misalnya, efektif untuk data yang besar dan relatif homogen, tetapi memerlukan pengetahuan awal tentang jumlah klaster dan sensitif terhadap outlier. Di sisi lain, hierarchical clustering tidak memerlukan pengetahuan awal tentang jumlah klaster dan cocok untuk data yang memiliki struktur hierarkis alami, tetapi kurang efisien untuk dataset yang sangat besar.

Selain itu, proses evaluasi clustering juga penting. Metrik seperti Silhouette Coefficient atau Davies-Bouldin Index sering digunakan untuk menilai kualitas klaster, mengukur sejauh mana objek dalam klaster mirip satu sama lain dan berbeda dari objek di klaster lain.

Di era digital saat ini, di mana data yang dihasilkan semakin besar dan kompleks, peran clustering dalam analisis data menjadi semakin penting. Dengan kemampuan untuk mengungkap pola tersembunyi dan menyediakan wawasan yang berharga, clustering telah menjadi alat yang tak tergantikan dalam toolkit setiap analis data.

B. JENIS-JENIS CLUSTERING

Clustering adalah teknik penting dalam data mining dan analisis data yang berfokus pada pengelompokkan objek serupa ke dalam klaster. Setiap jenis clustering memiliki keunggulan dan keterbatasannya masing-masing. Pemilihan metode clustering yang tepat bergantung pada jenis data, tujuan analisis, dan

sumber daya yang tersedia. Memahami perbedaan antara berbagai jenis clustering ini penting dalam menentukan pendekatan yang paling efektif untuk analisis data tertentu. Ada berbagai jenis clustering, masing-masing dengan karakteristik dan aplikasi yang berbeda. Berikut adalah penjelasan tentang beberapa jenis clustering utama:

1. Clustering Partitional
    - Clustering partitional mengelompokkan data menjadi sejumlah klaster tanpa adanya tumpang tindih antar klaster.
    - Algoritma paling populer dalam kategori ini adalah K-Means, yang bekerja dengan menentukan centroid secara acak dan kemudian mengelompokkan objek berdasarkan kedekatan mereka dengan centroid tersebut.
    - K-Means efektif untuk dataset besar, tetapi sensitif terhadap pemilihan awal centroid dan cenderung terpengaruh oleh outlier.
    - Algoritma lainnya termasuk K-Medoids dan K-Medians, yang serupa dengan K-Means tetapi kurang sensitif terhadap outlier.

2. Clustering Hierarkis
    - Clustering hierarkis mengelompokkan data dengan cara membangun hierarki klaster.
    - Metode ini dapat dibagi menjadi dua jenis: agglomerative (bottom-up) dan divisive (top-down).
    - Dalam pendekatan agglomerative, setiap objek awalnya dianggap sebagai klaster sendiri dan klaster digabungkan secara bertahap berdasarkan kedekatan mereka. Proses ini berlanjut sampai semua objek berada dalam satu klaster atau kriteria tertentu terpenuhi.
    - Pendekatan divisive bekerja dengan cara yang berlawanan, di mana semua objek awalnya berada dalam satu klaster besar yang kemudian dibagi menjadi klaster yang lebih kecil.

- Clustering hierarkis seringkali divisualisasikan menggunakan dendrogram, yang menunjukkan bagaimana klaster dibentuk.

3. Clustering Berbasis Densitas
   - Clustering ini fokus pada identifikasi area dengan densitas tinggi yang dipisahkan oleh area dengan densitas rendah.
   - DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah algoritma populer dalam jenis clustering ini. DBSCAN tidak memerlukan penentuan jumlah klaster terlebih dahulu dan dapat mengidentifikasi outlier sebagai noise.
   - Algoritma ini cocok untuk data dengan bentuk klaster yang tidak teratur dan beragam

4. Clustering Berbasis Model
   - Clustering ini didasarkan pada pemodelan matematika dari data dan klaster.
   - Gaussian Mixture Models (GMM) adalah contoh populer, di mana setiap klaster dimodelkan sebagai distribusi Gaussian.
   - Kelebihan metode ini adalah fleksibilitas dalam bentuk klaster, tetapi memerlukan pemahaman yang baik tentang model statistik

5. Clustering Berbasis Grid
   - Data dibagi menjadi sejumlah grid dan clustering dilakukan di dalam grid tersebut.
   - Algoritma seperti STING dan CLIQUE adalah contoh dari metode ini.
   - Cocok untuk data spasial dan dapat sangat efisien dari segi komputasi.

C. PREPROCESSING DATA

Dalam analisis data, khususnya sebelum melakukan proses clustering, tahap preprocessing data menjadi sangat penting. Proses ini meliputi berbagai teknik untuk menyiapkan data agar dapat dianalisis dengan lebih efektif. Berikut adalah penjelasan tentang beberapa aspek kunci dari preprocessing data:

1. Pemilihan Fitur (Feature Selection)
    - Pemilihan fitur adalah proses memilih fitur yang paling relevan dalam dataset untuk digunakan dalam analisis.
    - Tujuannya adalah untuk mengurangi dimensi data tanpa kehilangan informasi penting. Hal ini penting karena data dengan dimensi tinggi (high-dimensional data) sering mengalami masalah seperti curse of dimensionality.
    - Teknik pemilihan fitur meliputi filter methods, wrapper methods, dan embedded methods. Filter methods, seperti chi-square test dan analysis of variance (ANOVA), menilai relevansi fitur berdasarkan karakteristik statistiknya. Wrapper methods, seperti recursive feature elimination, menggunakan model prediktif untuk menilai pentingnya fitur. Embedded methods menggabungkan pemilihan fitur sebagai bagian dari proses pembelajaran model.

2. Normalisasi dan Standarisasi
    - Normalisasi adalah proses penskalaan nilai data ke rentang tertentu, biasanya antara 0 dan 1. Teknik ini berguna ketika fitur memiliki rentang nilai yang sangat bervariasi.
    - Standarisasi, di sisi lain, adalah proses mengubah data menjadi bentuk yang memiliki rata-rata nol dan standar deviasi satu. Ini sering dilakukan dengan metode z-score normalization.
    - Kedua teknik ini penting untuk algoritma yang sensitif terhadap skala data, seperti K-Means clustering, di mana

jarak antar titik data digunakan untuk menentukan keanggotaan klaster.

3. Penanganan Missing Values
    - Data sering kali mengandung nilai yang hilang atau missing, yang dapat mempengaruhi analisis jika tidak ditangani dengan benar.
    - Beberapa teknik untuk mengatasi missing values termasuk penghapusan baris atau kolom yang mengandung nilai hilang, imputasi dengan rata-rata/median/mode, dan metode imputasi yang lebih canggih seperti k-Nearest Neighbors atau multiple imputation.
    - Pilihan metode tergantung pada jumlah dan pola missing values serta asumsi tentang alasan di balik ketiadaan data tersebut.

4. Penanganan Outliers
    - Outliers adalah data yang memiliki karakteristik sangat berbeda dari mayoritas data lainnya dan dapat mengganggu analisis.
    - Pendekatan untuk menangani outliers termasuk deteksi dan penghapusan outliers, atau transformasi data untuk mengurangi dampak outliers.
    - Teknik deteksi outliers mencakup metode statistik seperti IQR (Interquartile Range) dan z-score, serta metode yang lebih canggih seperti clustering berbasis densitas

Pentingnya preprocessing data terletak pada kemampuannya untuk meningkatkan kualitas analisis dan memastikan hasil yang lebih akurat dan dapat diandalkan. Dengan menghapus atau mengurangi dampak noise, data yang tidak relevan, atau anomali, preprocessing meningkatkan efisiensi dan efektivitas proses analisis data.

D. ALGORITMA CLUSTERING

Clustering merupakan teknik penting dalam data mining dan machine learning yang bertujuan mengelompokkan data berdasarkan kesamaan karakteristik. Algoritma clustering memainkan peran kunci dalam mengidentifikasi struktur dalam data yang tidak terlabel.

Masing-masing algoritma clustering memiliki kelebihan dan kekurangannya sendiri. Pemilihan algoritma yang tepat bergantung pada sifat data dan tujuan analisis. Penting untuk memahami karakteristik setiap algoritma untuk memastikan hasil clustering yang efektif dan akurat. Berikut adalah beberapa algoritma clustering utama dan karakteristiknya:

1. K-Means Clustering
    - K-Means adalah salah satu algoritma clustering partitional paling populer.
    - Algoritma ini bekerja dengan menentukan terlebih dahulu jumlah klaster (K) dan kemudian secara iteratif memindahkan titik tengah (centroid) setiap klaster untuk meminimalkan varians dalam klaster.
    - K-Means efektif untuk dataset besar dan cocok untuk klaster dengan bentuk globular.
    - K-Means sensitif terhadap pemilihan nilai K awal dan tidak cocok untuk klaster dengan bentuk non-globular atau memiliki ukuran dan densitas yang berbeda.

2. Hierarchical Clustering
    - Clustering hierarkis tidak memerlukan penentuan jumlah klaster pada awal algoritma dan menghasilkan dendrogram, yang memperlihatkan hubungan hierarki antar klaster. Metode ini dibagi menjadi dua tipe: agglomerative (bottom-up) dan divisive (top-down).
    - Dalam agglomerative clustering, setiap data point awalnya dianggap sebagai klaster independen dan klaster

digabungkan secara bertahap. Sedangkan dalam divisive clustering, semua data awalnya berada dalam satu klaster yang kemudian dibagi menjadi klaster yang lebih kecil.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
    - DBSCAN adalah algoritma clustering berbasis densitas yang mampu menangani noise dan outlier dengan efektif. Algoritma ini mengidentifikasi klaster sebagai area dengan kepadatan tinggi yang dipisahkan oleh area dengan kepadatan rendah.
    - DBSCAN tidak memerlukan penentuan jumlah klaster terlebih dahulu dan mampu mendeteksi klaster dengan bentuk yang beragam.
    - Kelemahannya adalah sensitivitas terhadap pemilihan parameter radius dan minimum points.

4. Mean Shift Clustering
    - Mean Shift adalah algoritma clustering berbasis kernel yang bertujuan menemukan centroid (titik tengah) dari klaster.
    - Algoritma ini tidak memerlukan penentuan jumlah klaster sebelumnya dan sangat efektif untuk klaster dengan bentuk dan ukuran yang beragam.
    - Mean Shift memiliki kompleksitas komputasional yang tinggi dan sensitif terhadap pemilihan.

E. METRIK KESAMAAN JARAK

Dalam analisis data dan machine learning, metrik kesamaan dan jarak berperan penting dalam mengukur seberapa "dekat" atau "mirip" dua objek. Metrik ini digunakan dalam berbagai algoritma, terutama dalam clustering dan sistem rekomendasi. Berikut adalah penjelasan tentang beberapa metrik utama.

1. Euclidean Distance
   Euclidean Distance adalah metrik yang paling umum digunakan. Dalam ruang dua atau tiga dimensi, ini sama dengan mengukur jarak "garis lurus" antara dua titik.

   Euclidean Distance dihitung menggunakan rumus:

$$dist = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

   dimana *xi* dan *yi* adalah koordinat dua titik dalam ruang n-dimensi.

   Metrik ini intuitif dan mudah untuk diimplementasikan, namun bisa sangat sensitif terhadap perbedaan skala dalam fitur.
2. Manhattan Distance (City Block Distance)
   Manhattan Distance, atau City Block Distance, mengukur jarak antara dua titik jika hanya pergerakan horizontal atau vertikal yang diizinkan (seperti bergerak di sekitar blok kota).

   Manhattan Distance diukur dengan rumus:

$$dist = \sum_{i=1}^{n} |x_i - y_i|$$

   Metrik ini sering digunakan dalam konteks grid seperti raster images dan urban planning, dan cenderung lebih kuat terhadap outlier dibandingkan dengan Euclidean.
3. Cosine Similarity
   Cosine Similarity mengukur kesamaan antara dua vektor dengan menghitung cosinus sudut di antara mereka.

Rumusnya adalah

$$dist = \frac{A.B}{\|A\| \, . \, \|B\|}$$

di mana A·B adalah produk dot dan ||A||, ||B|| adalah magnitudo vektor.

Metrik ini sering digunakan dalam text analysis dan information retrieval, di mana orientasi vektor lebih penting daripada magnitudonya (seperti dalam kasus TF-IDF).

4. Jaccard Similarity
   Jaccard Similarity (atau Jaccard Index) digunakan untuk mengukur kesamaan antara dua set.

   Diukur dengan rumus:

$$dist = \frac{|A \bigcap B|}{|A \bigcup B|}$$

   di mana |A∩B| adalah jumlah elemen yang sama dalam kedua set, dan |A∪B| adalah jumlah total elemen unik dalam kedua set.

Metrik ini sangat berguna dalam analisis data kategorikal dan studi keanekaragaman biologi.

Pemahaman tentang metrik-metrik ini sangat penting dalam berbagai aplikasi data science, karena pilihan metrik yang tepat dapat sangat mempengaruhi hasil dari analisis data atau model machine learning.

F. EVALUASI CLUSTERING

Evaluasi clustering adalah proses penting untuk menentukan kualitas kelompok atau klaster yang dihasilkan oleh algoritma

clustering. Evaluasi ini dapat dilakukan melalui metode internal dan eksternal, serta melalui proses validasi. Berikut adalah penjelasan tentang masing-masing metode tersebut:

1. **Evaluasi Internal (Contoh: Silhouette Coefficient)**
   Evaluasi internal mengukur kualitas clustering berdasarkan data yang digunakan untuk clustering itu sendiri, tanpa memerlukan data eksternal atau label sebelumnya.

   Silhouette Coefficient adalah salah satu metrik evaluasi internal yang populer. Nilai ini mengukur seberapa serupa objek dengan klaster yang mereka miliki dibandingkan dengan klaster lain. Skor Silhouette berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan klaster yang lebih baik.

   Rumusnya melibatkan perhitungan jarak rata-rata objek ke semua objek lain dalam klaster yang sama (a) dan jarak rata-rata ke objek di klaster terdekat lainnya (b), kemudian menggunakan formula

$$\frac{b-a}{max(a,b)}$$

2. **Evaluasi Eksternal (Contoh: Adjusted Rand Index)**
   Evaluasi eksternal mengukur kualitas clustering dengan membandingkannya dengan partisi yang telah diketahui sebelumnya (biasanya ground truth atau benchmark).

   Adjusted Rand Index (ARI) adalah metrik yang menyesuaikan kesepakatan murni dengan kesempatan. ARI memiliki nilai antara -1 dan 1, di mana nilai yang lebih tinggi menunjukkan kesesuaian yang lebih baik antara hasil clustering dan partisi referensi.

ARI dihitung berdasarkan jumlah pasangan elemen yang dikelompokkan sama dalam kedua partisi (hasil clustering dan partisi referensi) dibandingkan dengan jumlah total pasangan.

3. **Validasi**

   Validasi clustering adalah proses memeriksa keandalan dan kegeneralan hasil clustering. Hal ini sering melibatkan pengujian stabilitas klaster terhadap variasi dalam metode clustering atau parameter yang digunakan.

   Salah satu pendekatan adalah menggunakan dataset yang berbeda atau subset dari dataset asli untuk melihat apakah hasil klaster konsisten.

   Teknik lain termasuk cross-validation dan penggunaan dataset sintetis dengan properti yang diketahui untuk memeriksa apakah algoritma clustering dapat menemukan struktur yang diharapkan.

Evaluasi clustering, baik internal maupun eksternal, bersama dengan validasi, memberikan wawasan penting tentang kualitas dan keandalan kelompok yang dihasilkan oleh algoritma clustering. Ini penting tidak hanya untuk penelitian tetapi juga untuk aplikasi praktis di mana keputusan bisnis atau ilmiah mungkin bergantung pada hasil analisis clustering.

G. TANTANGAN DAN SOLUSI DALAM CLUSTERING

Clustering adalah teknik analisis data yang penting, tetapi penerapannya sering menghadapi berbagai tantangan. Tantangan ini termasuk menentukan jumlah klaster yang optimal, mengatasi sensitivitas terhadap outliers, dan memastikan skalabilitas dan efisiensi algoritma. Berikut adalah pembahasan tentang tantangan-tantangan ini beserta solusi potensialnya:

1. Menentukan Jumlah Klaster

Salah satu tantangan utama dalam clustering adalah menentukan jumlah klaster yang tepat. Salah memilih jumlah klaster dapat mengakibatkan interpretasi yang salah tentang data.

- Metode Elbow: Salah satu cara untuk menentukan jumlah klaster adalah menggunakan metode Elbow, yang melibatkan plot varians data terhadap jumlah klaster. Titik di mana peningkatan jumlah klaster tidak lagi memberikan penurunan varians yang signifikan dianggap sebagai 'siku' dan menunjukkan jumlah klaster yang tepat.
- Silhouette Score: Metode ini mengukur seberapa serupa objek dengan klaster mereka sendiri dibandingkan dengan klaster lain, dan dapat digunakan untuk menilai kualitas klaster pada jumlah klaster yang berbeda.

2. Sensitivitas terhadap Outliers
   Banyak algoritma clustering, seperti K-Means, sangat sensitif terhadap outliers. Outliers dapat memengaruhi penentuan centroid dan hasil akhir clustering.
   - Pembersihan Data: Langkah awal yang dapat dilakukan adalah pembersihan data, yaitu dengan mengidentifikasi dan menghapus outliers sebelum melakukan clustering.
   - Algoritma Robust: Menggunakan algoritma yang lebih tahan terhadap outliers, seperti DBSCAN atau algoritma clustering hierarkis, juga dapat menjadi solusi.

3. Skalabilitas dan Efisiensi
   Dalam menghadapi dataset yang sangat besar, efisiensi komputasi dan skalabilitas algoritma clustering menjadi penting.
   - Penggunaan Algoritma Efisien: Algoritma seperti K-Means atau hierarchical clustering dengan pendekatan divisive dapat lebih efisien untuk dataset besar.

- Penggunaan Teknik Sampling: Menggunakan sampel dari dataset yang lebih besar dapat mengurangi waktu komputasi tanpa mengorbankan terlalu banyak akurasi.
- Penggunaan Platform Big Data: Platform seperti Apache Hadoop atau Spark dapat memanfaatkan komputasi terdistribusi untuk meningkatkan skalabilitas dan efisiensi clustering.

Setiap tantangan dalam clustering memerlukan pendekatan yang berbeda dan sering kali bergantung pada konteks data dan tujuan analisis. Memahami berbagai solusi untuk tantangan ini memungkinkan peneliti dan praktisi untuk menerapkan teknik clustering secara lebih efektif dan akurat.

H. KASUS

Sebuah Toko akan mengelompokkan pembelian pelanggan kedalam tiga kelompok, yaitu: Banyak, Sedang, dan Sedikit. Pengelompokkan berdasarkan “Total Belanja” dan “Frekuensi Pembelian” seperti tabel berikut.

No	Total Belanja	Frekuensi Pembelian
1	8270	12
2	1860	7
3	6390	4
4	6191	9
5	6734	3
6	7265	5
7	1466	3
8	5426	7

9	6578	5
10	9322	9
11	2685	7
12	1769	2
13	7949	4
14	3433	9
15	6311	12
16	6051	14
17	7420	2
18	2184	10
19	5555	9
20	4385	10
21	7396	5
22	9666	2
23	3558	4
24	8849	12
25	3047	12
26	3747	7
27	1189	12
28	3734	13
29	4005	8
30	5658	3

Total belanja dalam ribuan

Untuk menjalankan perhitungan manual algoritma K-Means dengan jumlah klaster 3 pada dataset 30 pelanggan, kita akan mengikuti langkah-langkah berikut:

1. **Inisialisasi Centroid:** Pilih 3 titik sebagai centroid awal secara acak atau berdasarkan intuisi dari dataset.
   C1 (1189; 2)
   C2 (5606.5; 7)
   C3 (9666; 14)
2. **Penugasan Klaster:** Untuk setiap titik data, hitung jaraknya ke masing-masing centroid dan tugaskan titik data ke klaster dengan centroid terdekat.

   Jarak data pelanggan pertama dengan pusat Cluster pertama

   $$d_{11} = \sqrt{(8270 - 1189)^2 + (12 - 2)^2} = 7081.01$$

   Jarak data pelanggan pertama dengan pusat Cluster kedua

   $$d_{12} = \sqrt{(8270 - 5606.5)^2 + (12 - 7)^2} = 2663.50$$

   Jarak data pelanggan pertama dengan pusat Cluster ketiga

   $$d_{13} = \sqrt{(8270 - 9666)^2 + (12 - 14)^2} = 1396.00$$

   Hasil perhitungan secara lengkap dan posisi cluster pada iterasi pertama seperti pada tabel berikut.

No	Total Belanja	Frekuensi Pembelian	Jarak			Cluster		
			C1	C2	C3	C1	C2	C3
1	8270	12	7081,01	2663,50	1396,00			*
2	1860	7	671,02	3746,50	7806,00	*		
3	6390	4	5201,00	783,51	3276,02		*	
4	6191	9	5002,00	584,50	3475,00		*	
5	6734	3	5545,00	1127,51	2932,02		*	
6	7265	5	6076,00	1658,50	2401,02		*	
7	1466	3	277,00	4140,50	8200,01	*		
8	5426	7	4237,00	180,50	4240,01		*	
9	6578	5	5389,00	971,50	3088,01		*	
10	9322	9	8133,00	3715,50	344,04			*
11	2685	7	1496,01	2921,50	6981,00	*		
12	1769	2	580,00	3837,50	7897,01	*		
13	7949	4	6760,00	2342,50	1717,03			*
14	3433	9	2244,01	2173,50	6233,00		*	
15	6311	12	5122,01	704,52	3355,00		*	
16	6051	14	4862,01	444,56	3615,00		*	
17	7420	2	6231,00	1813,51	2246,03		*	
18	2184	10	995,03	3422,50	7482,00	*		
19	5555	9	4366,01	51,54	4111,00		*	
20	4385	10	3196,01	1221,50	5281,00		*	
21	7396	5	6207,00	1789,50	2270,02		*	
22	9666	2	8477,00	4059,50	12,00			*
23	3558	4	2369,00	2048,50	6108,01		*	
24	8849	12	7660,01	3242,50	817,00			*
25	3047	12	1858,03	2559,50	6619,00	*		
26	3747	7	2558,00	1859,50	5919,00		*	
27	1189	12	10,00	4417,50	8477,00	*		
28	3734	13	2545,02	1872,51	5932,00		*	
29	4005	8	2816,01	1601,50	5661,00		*	
30	5658	3	4469,00	51,66	4008,02		*	

Perbarui Centroid: Hitung ulang centroid untuk setiap klaster berdasarkan titik data yang telah ditugaskan.

Pada cluster pertama terdapat 7 data, yaitu data ke-2, 7, 11, 12, 18, 25, dan 27. Sehingga

C_{11} = (1860 + 1466 + 2685 + 1769 + 2184 + 3047 + 1189)/7
= 2028.57

C_{12} = (7 + 3 + 7 + 2 +10 + 12 + 12)/7
= 7.57

Pada cluster kedua terdapat 18 data, yaitu data ke-3, 4, 5, 6, 8, 9, 14, 15, 16, 17, 19, 20, 21, 23, 26, 28, 29 dan 30. Sehingga

C_{21} = (6390 + 6191 + 6734 + 7265 + 5426 + 6578 + 3433 + 6311 + 6051 + 7420 + 5555 + 4385 + 7396 + 3558 + 3747 + 3734 + 4005 + 5628)/18
= 5546.50

C_{22} = (4 + 9 + 3 + 5 + 7 + 5 + 9 + 12 + 14 + 2 + 9 + 10 + 5 + 4 + 7 + 13 + 8 + 3)/18
= 7.17

Pada cluster ketiga terdapat 5 data, yaitu data ke-1, 10, 13, 22, dan 24. Sehingga

C_{31} = (8270 + 9322 + 7949 + 9666 + 8849)/5
= 8811.20

C_{32} = (12 + 9 + 4 + 2 + 12)/5
= 7.80

3. **Iterasi:** Ulangi langkah 2 dan 3 hingga penugasan klaster tidak lagi berubah.

   Pada kasus ini cluster tidak berubah pada iterasi keempat. Hasil perhitungan secara lengkap dan posisi cluster pada iterasi keempat seperti pada tabel berikut.

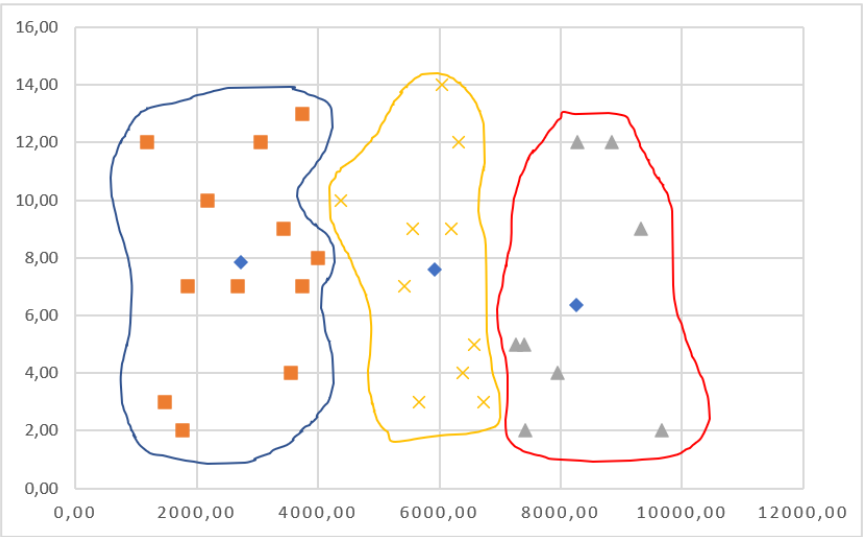
No	Total Belanja	Frekuensi Pembelian	Jarak			Cluster		
			C1	C2	C3	C1	C2	C3
1	8270	12	5546,92	2342,10	6,32			*
2	1860	7	863,08	4067,90	6407,13	*		
3	6390	4	3666,92	462,11	1877,13		*	
4	6191	9	3467,92	263,10	2076,13		*	
5	6734	3	4010,92	806,11	1533,13		*	
6	7265	5	4541,92	1337,10	1002,13			*
7	1466	3	1257,09	4461,90	6801,13	*		
8	5426	7	2702,92	501,90	2841,13		*	
9	6578	5	3854,92	650,11	1689,13		*	
10	9322	9	6598,92	3394,10	1054,88			*
11	2685	7	38,09	3242,90	5582,13	*		
12	1769	2	954,10	4158,90	6498,13	*		
13	7949	4	5225,92	2021,10	318,13			*
14	3433	9	709,92	2494,90	4834,13	*		
15	6311	12	3587,92	383,13	1956,13		*	
16	6051	14	3327,92	123,27	2216,14		*	
17	7420	2	4696,92	1492,11	847,14			*
18	2184	10	539,09	3743,90	6083,13	*		
19	5555	9	2831,92	372,90	2712,13		*	
20	4385	10	1661,92	1542,90	3882,13		*	
21	7396	5	4672,92	1468,10	871,13			*
22	9666	2	6942,92	3738,10	1398,88			*
23	3558	4	834,93	2369,90	4709,13	*		
24	8849	12	6125,92	2921,10	581,90			*
25	3047	12	323,94	2880,90	5220,13	*		
26	3747	7	1023,92	2180,90	4520,13	*		
27	1189	12	1534,09	4738,90	7078,13	*		
28	3734	13	1010,93	2193,91	4533,13	*		
29	4005	8	1281,92	1922,90	4262,13	*		
30	5658	3	2934,92	269,94	2609,13		*	

Dari hasil cluster diperoleh kelompok sebagai berikut.

a. Cluster pertama memiliki pusat cluster (2723.08; 7.83) yang dapat diartikan sebagai kelompok pelanggan belanja sedikit, dengan jumlah 12 data.

b. Cluster kedua memiliki pusat cluster (5927.90; 7.60) yang dapat diartikan sebagai kelompok pelanggan belanja sedang, dengan jumlah 10 data.
c. Cluster ketiga memiliki pusat cluster (8267.13; 6.38) yang dapat diartikan sebagai kelompok pelanggan belanja banyak, dengan jumlah 8 data.

Untuk lebih jelas melihat kelompok setiap cluster dapat dilihat pada gambar berikut.



Gambar 7.1 Hasil Cluster Tiap Kelompok

I. RANGKUMAN

Clustering adalah proses mengelompokkan set data ke dalam kelompok atau klaster berdasarkan kesamaan atau pola yang ada di dalam data. Dalam clustering, data yang mirip ditempatkan dalam satu klaster, sementara data yang berbeda dikelompokkan ke dalam klaster lain. Tujuan utamanya adalah untuk menemukan struktur tersembunyi, mengekstraksi pola yang berguna, dan mengidentifikasi kelompok atau segmen dalam data.

1. Jenis-Jenis Clustering
    - Partitional Clustering: Mengelompokkan data menjadi sejumlah klaster yang tidak tumpang tindih. Contoh: K-Means, K-Medoids.
    - Hierarchical Clustering: Membangun hirarki klaster, bisa dilakukan secara agglomerative (bottom-up) atau divisive (top-down). Hasilnya sering disajikan dalam bentuk dendrogram.
    - Density-Based Clustering: Klaster dibentuk berdasarkan area kepadatan data yang tinggi. Contoh: DBSCAN, OPTICS.
    - Model-Based Clustering: Membentuk klaster berdasarkan model statistik. Contoh: Gaussian Mixture Models.

2. Tahapan Clustering
    - Preprocessing Data: Termasuk normalisasi, penanganan missing values, dan outliers.
    - Pemilihan Algoritma: Memilih algoritma clustering yang sesuai dengan jenis dan karakteristik data.
    - Penentuan Jumlah Klaster: Menggunakan metode seperti Elbow Method, Silhouette Coefficient untuk menentukan jumlah klaster yang optimal.
    - Penerapan dan Evaluasi: Menerapkan algoritma clustering pada data dan mengevaluasi hasilnya menggunakan metrik seperti Silhouette Coefficient atau Davies-Bouldin Index.

3. Kegunaan Clustering
   Clustering digunakan dalam berbagai bidang seperti pemasaran untuk segmentasi pelanggan, biologi untuk analisis genetik, keuangan untuk deteksi penipuan, dan banyak lagi. Dalam bisnis, clustering membantu dalam pengambilan keputusan strategis dan targeting pemasaran.

4. Tantangan dalam Clustering
    - Menentukan Jumlah Klaster yang Tepat: Kesulitan dalam menentukan jumlah klaster ideal.
    - Sensitivitas terhadap Outliers: Sebagian besar algoritma clustering sensitif terhadap outliers.
    - Skalabilitas dan Efisiensi: Tantangan dalam mengelola dataset besar.

5. Aplikasi Praktis
   Clustering sering digunakan untuk analisis data eksploratif, sebagai langkah awal dalam proses analisis data kompleks. Ini membantu dalam menemukan pola tersembunyi dan mengambil insight dari kumpulan data yang besar dan kompleks.

J. TES FORMATIF

1. Apa itu Clustering dalam konteks Data Mining?
    a. Proses pengurutan data
    b. Proses pengelompokkan objek serupa
    c. Proses penghapusan data
    d. Proses penggabungan data dari sumber yang berbeda
2. Algoritma K-Means termasuk dalam kategori clustering apa?
    a. Hierarchical Clustering
    b. Density-Based Clustering
    c. Partitional Clustering
    d. Grid-Based Clustering
3. DBSCAN merupakan singkatan dari...
    a. Density-Based Spatial Clustering of Applications with Noise
    b. Data-Based Spatial Computing of Applications with Numbers
    c. Density-Bound Spatial Clustering of Applied Networks
    d. Data-Bound Spatial Clustering of Applications with Noise

4. Apa perbedaan utama antara Hierarchical Clustering dan K-Means Clustering?
    a. Hierarchical Clustering tidak memerlukan penentuan jumlah klaster di awal
    b. K-Means Clustering menggunakan jarak Euclidean
    c. Hierarchical Clustering lebih cepat daripada K-Means
    d. K-Means Clustering tidak dapat digunakan untuk data berdimensi tinggi
5. Apa tujuan utama dari metode Elbow dalam konteks clustering?
    a. Menemukan centroid terbaik untuk setiap klaster
    b. Menentukan jumlah klaster yang optimal
    c. Mengukur jarak antar titik dalam satu klaster
    d. Mengidentifikasi outlier dalam dataset
6. Apa fungsi Silhouette Coefficient dalam analisis clustering?
    a. Mengukur tingkat kepadatan sebuah klaster
    b. Mengukur seberapa baik objek telah dikelompokkan
    c. Menghitung jumlah klaster dalam sebuah dataset
    d. Mengidentifikasi variabel penting dalam sebuah klaster
7. Manakah dari berikut ini yang merupakan contoh penggunaan clustering?
    a. Mengklasifikasikan email menjadi 'spam' atau 'bukan spam'
    b. Mengelompokkan pelanggan berdasarkan kebiasaan belanja
    c. Menghitung total penjualan di sebuah toko
    d. Memprediksi harga saham di masa depan
8. Pada algoritma clustering, apa itu 'Centroid'?
    a. Titik data dengan nilai tertinggi dalam sebuah klaster
    b. Titik tengah geometris dari sebuah klaster
    c. Objek terjauh dalam sebuah klaster
    d. Algoritma untuk menghitung jarak antar objek

9. Apa perbedaan antara Hard Clustering dan Soft Clustering?
    a. Hard Clustering mengalokasikan setiap objek ke satu klaster, sementara Soft Clustering memungkinkan objek menjadi bagian dari lebih dari satu klaster
    b. Hard Clustering menggunakan jarak Euclidean, sementara Soft Clustering menggunakan jarak Manhattan
    c. Hard Clustering digunakan untuk data berdimensi rendah, sedangkan Soft Clustering untuk data berdimensi tinggi
    d. Hard Clustering lebih cepat dari Soft Clustering

K. LATIHAN

Deskripsikan prinsip dasar dari clustering, serta apa itu centroid dan bagaimana peranannya dalam proses clustering pada metode K-Means Clustering?

KEGIATAN BELAJAR 8

TEORI DASAR *ASSOCIATION RULES*

DESKRIPSI PEMBELAJARAN

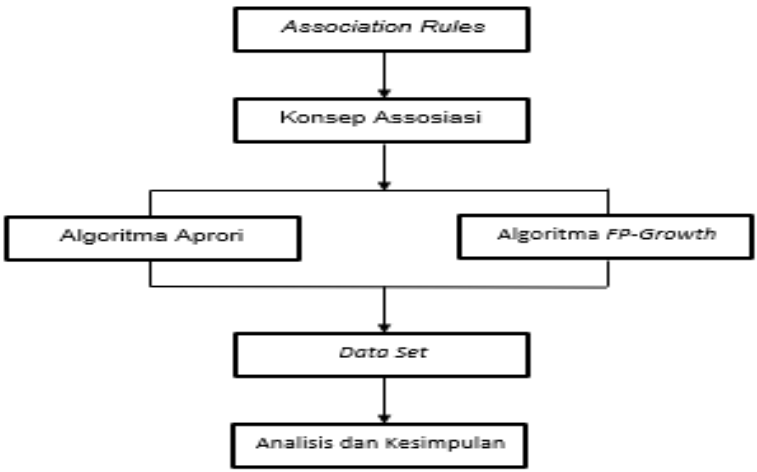
Pada bab ini mahasiswa mempelajari pengenalan dan teori dasar *Association Rules*. Diharapkan mahasiswa memiliki kemampuan untuk mengidentifikasi hubungan dan pola asosiasi antara item atau variabel dalam suatu dataset.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

1. Mampu menjelaskan konsep asosiasi
2. Mempu menjelaskan algoritma Apriori dan menerapkannya dalam data set
3. Mampu menjelaskan algoritma *FP-Growth* dan menerapkannya dalam data set

PETA KONSEP PEMBELAJARAN



A. PENDAHULUAN

Dalam era digital ini, data menjadi aset berharga yang melimpah di berbagai sektor, mulai dari bisnis hingga penelitian. Sebagai pemahaman atas pola dan hubungan dalam data semakin penting, metode analisis data seperti *Association Rules* muncul sebagai alat yang kuat dan efektif.

Association Rules adalah metode analisis data yang bertujuan mengidentifikasi hubungan dan pola asosiasi antara item atau variabel dalam dataset. Dengan kata lain, metode ini membantu kita menemukan aturan atau korelasi yang mungkin tersembunyi di tengah-tengah data yang kompleks, baik untuk memahami perilaku konsumen, meningkatkan efisiensi operasional, atau mendukung pengambilan keputusan.

Dalam pengembangan metode ini, konsep seperti *support, confidence*, dan *lift* menjadi landasan utama. *Support* mengukur seberapa sering suatu kombinasi item muncul dalam dataset, *confidence* menilai sejauh mana hubungan antara item-item tersebut, sementara *lift* mengukur kekuatan asosiasi relatif terhadap frekuensi itemset yang diharapkan secara acak.

Buku ini menyajikan dua algoritma utama dalam *Association Rules* yaitu Algoritma Apriori dan *FP-Growth*. Selain itu, disajikan juga parameter-parameter kunci dalam analisis dan bagaimana melakukan evaluasi serta seleksi rules dengan baik. Melalui pemahaman yang mendalam terhadap teori dasar *Association Rules*, kita akan siap menghadapi tantangan analisis data yang kompleks.

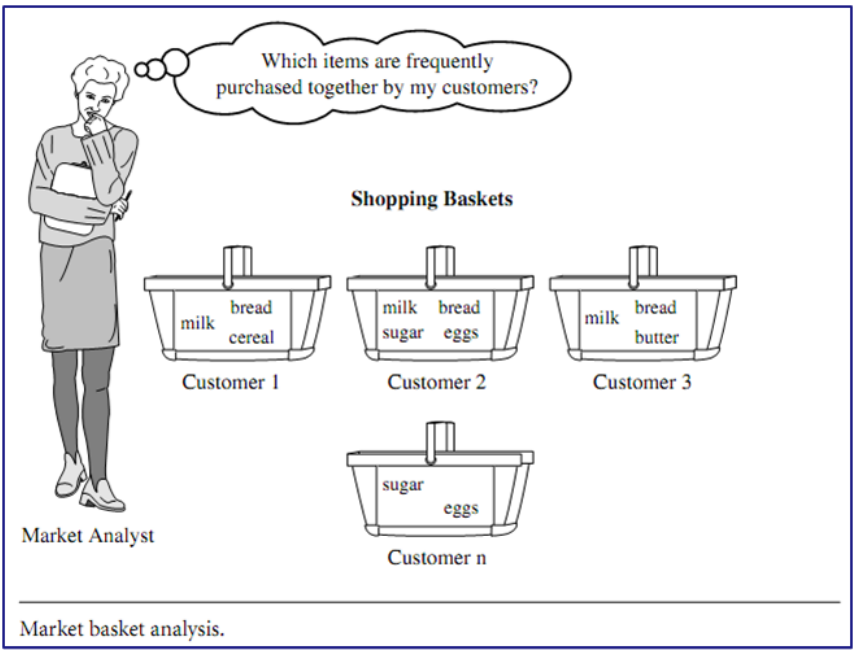
Association Rules dapat dimanfaatkan secara luas dalam proses bisnis diantaranya dalam proses penjualan. Ketersediaan detail informasi transaksi pelanggan mendorong pengembangan teknik yang secara otomatis mencari hubungan antara *item* dalam data

di database. Sebagai contoh data didapat dari scanner *barcode* di supermarket. Database penjualan menyimpan jumlah *record* transaksi penjualan yang sangat besar. Setiap *record* memberikan daftar item barang yang dibeli oleh pelanggan dalam satu transaksi.

Dari database tersebut dapat diketahui pola belanja konsumen yang nantinya akan dimanfaatkan oleh perusahaan. Manager dapat menggunakan data tersebut dalam pengaturan *layout* toko untuk meletakkan item barang secara optimal dengan keterkaitan satu dengan lainnya, dapat pula digunakan dalam promosi, atau dalam desain katalog dan untuk mengidentifikasi segmen pelanggan berdasar pola pembelian.

Sebagai contoh, jika pembeli membeli susu, bagaimana kemungkinan mereka juga akan membeli roti dalam waktu yang sama di *supermarket*? Informasi seperti itu akan membantu meningkatkan penjualan dengan membantu retailer melakukan pemasaran yang selektif dan merencanakan layout mereka. Sebagai contoh, meletakan susu dan roti dengan posisi dekat kemungkinan akan meningkatkan penjualan item tersebut bersama dalam sebuah penjualan.

Salah satu penerapan *Association Rules* seperti uraian di atas adalah menggunakan Algoritma Apriori. Algoritma Apriori merupakan algoritma yang sangat terkenal untuk menemukan pola frekuensi tinggi. Pola frekuensi tinggi adalah pola-pola *item* di dalam suatu database yang memiliki frekuensi atau support di atas ambang batas tertentu yang disebut dengan istilah *minimum support*. Pola frekuensi tinggi ini digunakan untuk menyusun aturan assosi dan juga beberapa teknik *data mining* lainnya. Aturan yang menyatakan asosiasi antara beberapa atribut sering disebut sebagai *market basket analysis.* Contoh *market basket analysis* diberikan pada Gambar 8.1.

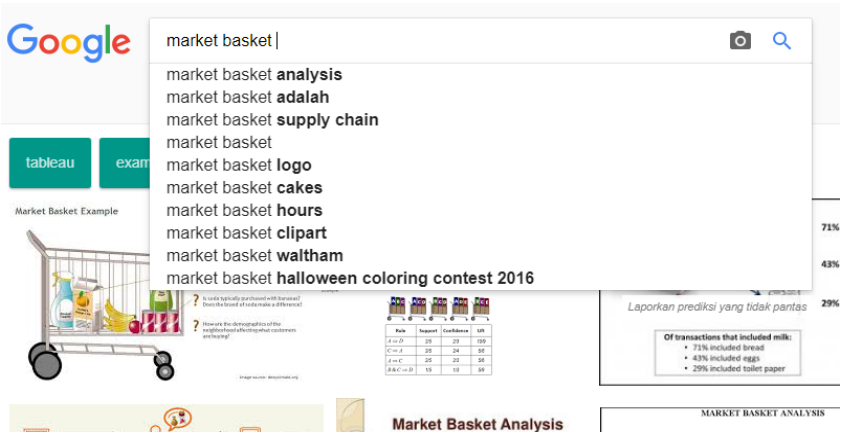


Gambar 8.1. *Market Basket Analysis*
(Sumber: yahoo.com)

Selain diterapkan dalam pencarian pola belanja konsumen, contoh pemakaian lain dilakukan Amazon.com yang mengembangkan perekomendasi (recommender). Perekomendasi merupakan sebuah program untuk merekomendasikan barang-barang lain kepada pembeli pada saat pembeli melakukan *browsing* atau membeli suatu barang berdasarkan tingkat keyakinan (*confidence*). Gambar 8.2 menunjukkan hasil ketika seorang konsumer menuliskan judul sebuah buku.

Gambar 8.2. Contoh Hasil Perekomendasi Amazon.com

Google mengembangkan fitur *auto-complete*, yaitu saat pemakai mengetikkan suatu kata, program akan menampilkan daftar kata-kata berikutnya, yang paling banyak memiliki asosiasi pada kata yang diketik sebagaimana Gambar 8.3.



Gambar 8.3. Contoh Hasil Fitur *Auto-complete* oleh Google

B. KONSEP ASOSIASI

Konsep asosiasi dalam data mining merujuk pada hubungan dan pola keterkaitan antara satu atau lebih item atau atribut dalam suatu dataset. Asosiasi ditemukan melalui identifikasi aturan

asosiasi, yang dapat memberikan wawasan tentang bagaimana item-item atau atribut-atribut tertentu cenderung muncul bersamaan. Berikut adalah beberapa konsep kunci terkait asosiasi: konsep seperti *support, confidence,* dan *lift* menjadi landasan utama. *Support* mengukur seberapa sering suatu kombinasi item muncul dalam dataset, *confidence* menilai sejauh mana hubungan antara item-item tersebut, sementara *lift* mengukur kekuatan asosiasi relatif terhadap frekuensi itemset yang diharapkan secara acak.

a. Sebuah item adalah sebuah atribut dalam suatu keranjang.

1. Contoh: Pena atributnya *pen.*
2. Sebuah transaksi berisi semua item dalam suatu keranjang yang dibeli bersama.
3. Sebuah data set transaksi adalah himpunan seluruh transaksi.
4. E adalah himpunan yang tengah dibicarakan.
5. Contoh: *{Asparagus, Beans, …, Tomatoes}.*
6. D adalah himpunan seluruh transaksi yang dibicarakan.
7. Contoh: *{Transaksi 1, transaksi 2, …, transaksi 14}.*
8. Item set adalah himpunan dari beberapa item atau item-item di E.
9. Contoh: Ada suatu himpunan *E={a,b,c,}*, Item setnya adalah *{a};{b}:{c};{a,b};{a,c};{b,c}*
10. *k-item set* adalah Item set yang terdiri dari k-buah item yang ada pada E. Contoh: 2-item set adalah item yang berisi 2 unsur, misalnya *{a,b},{a,c},{b,c}.*
11. *Candidate k-itemset* (Ck) : calon k-itemset dari data transaksi.
12. *Frequent k-itemset* (Lk) : itemset yang memiliki frekuensi kemunculan lebih dari nilai minimum yang telah ditentukan.

Lift ratio adalah nilai yang menunjukkan kevalidan proses transaksi dan memberikan informasi apakah benar produk A dibeli bersamaan dengan produk B.

Pengertian *support, confidence*, dan *lift ratio* diberikan di bawah ini.

Support *Support* merupakan persentase kombinasi item muncul di dalam keseluruhan basis data yang ada. Nilai *support* dapat diketahui dari seberapa sering suatu kombinasi muncul dalam keseluruhan transaksi yang terjadi. *Support* digunakan untuk menentukan seberapa sering suatu itemset muncul dalam dataset, dan itemset-itemset yang memiliki *support* di atas ambang batas (*MinSupport*) akan dianggap sebagai pola frequent. Berikut rumus *support*.

$$Support(X) = \frac{Frekuensi(X)}{Jumlah\ Keseluruhan\ Data}$$

Confidence *Confidence* merupakan persentase kuatnya hubungan produk. *Confidence* menunjukkan kemungkinan antar produk untuk diambil bersamaan. Sebagai contoh, kemungkinan produk pensil diambil saat produk bolpoin diambil. Dari nilai *confidence* inilah dapat diketahui seberapa kuat hubungan antar produk. Pada umumnya, *FP-Growth* tidak menggunakan rumus *confidence* secara langsung seperti pada Algoritma Apriori. Sebagai gantinya, setelah pola frequent ditemukan, aturan asosiasi dapat dihasilkan dengan menggabungkan itemset pada tingkat yang berbeda dalam pohon *FP-Tree*. Nilai *confidence* dapat dicari dengan persamaan berikut.

$$\boldsymbol{Confidence}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{Support}(\boldsymbol{X} \cup \boldsymbol{Y})}{\boldsymbol{Support}(\boldsymbol{X})}$$

Lift Ratio *Lift ratio* menunjukkan adanya tingkat kekuatan rule atas kejadian acak dari antecedent dan consequence berdasarkan pada *support*nya masing-masing. Nilai *lift* sendiri memiliki rentang antara 0 hingga tak terhingga. Nilai tersebut menyatakan keterkaitan suatu produk. Apabila hasil perhitungan berada di bawah 1 maka item-item tersebut tidak menunjukkan adanya saling keterkaitan antara antecedent dan consequent.

Lift ratio adalah suatu ukuran untuk mengetahui kekuatan aturan asosiasi (*association* rule) yang telah terbentuk. Nilai *lift* ratio biasanya digunakan sebagai penentu apakah aturan asosiasi valid atau tidak valid. Untuk menghitung *lift* ratio dengan menggunakan rumus sebagai berikut.

$$Lift\ ratio = \frac{Confidence\ (A,B)}{Benchmark\ Confidence(A,B)}$$

dimana untuk mendapatkan nilai *benchmark confidence* dapat dihitung menggunakan rumus berikut.

$$Benchmark\ Confidence = \frac{NC}{N}$$

dengan

NC = jumlah transaksi dengan item yang menjadi consequent
N = jumlah transaksi basis data

C. ALGORITMA APRIORI

Algoritma apriori terbagi menjadi dua tahap, yaitu:

1. Analisis pola frekuensi tinggi

Tahap ini mencari kombinasi item yang memenuhi syarat minimum dari nilai support dalam database. Nilai support sebuah item diperoleh dengan rumus berikut:

$$Support(A) = \frac{\sum transaksi\ mengandung\ A}{\sum transaksi}$$

Sedangkan nilai support dari 2 item diperoleh dari rumus berikut:

$$Support(A,B) = P(A \cap B) = \frac{\sum transaksi\ mengandung\ A\ dan\ B}{\sum transaksi}$$

2. **Pembentukan Aturan Asosiasi**

   Setelah semua pola frekuensi tinggi ditemukan, barulah dicari aturan assosiatif yang memenuhi syarat minimum untuk confidence dengan menghitung confidence aturan assosiatif A → B

   Nilai confidence dari aturan A →B diperoleh dari rumus berikut

$$Confidence = P(B|A) = \frac{\sum transaksi\ mengandung\ A\ dan\ B}{\sum transaksi\ mengandung\ A}$$

Langkah-Langkah Langkah-langkah Algoritma Apriori

1. Tentukan minimum support.
2. Iterasi 1: Sistem men-scan database untuk mendapat kandidat 1-itemset (himpunan item yang terdiri dari 1 item) dan menghitung nilai *support*nya. Kemudian nilai support tersebut dibandingkan dengan *minimum support* (*MinSupport*) yang telah ditentukan, jika nilainya lebih besar atau sama dengan *minimum support* maka itemset tersebut termasuk dalam *large itemset*. *Itemset* yang tidak termasuk dalam *large itemset* tidak diikutkan dalam iterasi selanjutnya.
3. Iterasi 2: Sistem akan menggunakan hasil large *itemset* pada iterasi pertama (L1) untuk membentuk kandidat *itemset* kedua (L2). Pada iterasi selanjutnya sistem akan menggunakan hasil *large itemset* pada iterasi selanjutnya akan menggunakan hasil *large itemset* pada iterasi sebelumnya (Lk-1) untuk membentuk kandidat *itemset* berikut (Lk). Sistem akan menggabungkan Lk-1 dengan Lk-1 untuk mendapatkan Lk, seperti pada iterasi sebelumnya sistem akan menghapus kombinasi itemset yang tidak termasuk dalam *large itemset*
4. Setelah dilakukan operasi gabung, maka pasangan itemset baru hasil proses join tersebut dihitung *support*nya
5. Proses pembentuk kandidat yang terdiri dari proses join dan pangkas akan terus dilakukan hingga himpunan kandidat

itemsetnya null, atau sudah tidak ada lagi kandidat yang akan dibentuk.

6. Setelah itu, dari hasil *frequent itemset* tersebut dibentuk *Association Rules* yang memenuhi nilai *support* dan *confidence* yang telah ditentukan.

Contoh Terapan Misalnya dipunyai data penjualan Smartphone sebagaimana diberikan di Tabel 8.1 berikut.

Tabel 8.1. Data Penjualan Smartphone

No.	Itemset
1	Samsung, Advance, Lenovo
2	Samsung, Apple, Lenovo
3	Huawei, Advance, Lenovo
4	Samsung, Apple, Advance
5	Samsung, Apple, Lenovo
6	Samsung, Huawei, Lenovo
7	Samsung, Apple, Lenovo
8	Samsung, Apple, Huawei
9	Apple, Huwei, Advance
10	Samsung, Huawei, Advance
11	Samsung, Advance, Lenovo
12	Apple, Huawei, Advance

Format tabular data transaksi tampak seperti Tabel 8.2 berikut ini.

Tabel 8.2. Data Transaksi Mingguan dalam Format Tabular

Minggu	Samsung	Apple	Huawei	Advance	Lenovo
1	1	0	0	1	1
2	1	1	0	0	1
3	0	0	1	1	1
4	1	1	1	1	0
5	1	0	1	1	1
6	1	0	0	0	1
7	1	1	0	0	1

8	1	1	0	0	0
9	0	1	1	1	0
10	1	0	1	1	0
11	1	0	1	1	1
12	0	1	1	1	0

Misalkan kita tentukan *minimum support* = 30% dan *minimum confidence* = 60%

1. Support

Analisis Pola Frekuensi Tinggi

a. 1-Itemset

Proses pembentukan C1 atau disebut dengan 1 itemset dengan jumlah minimum support = 30%. Berikut merupakan perhitungan pembentukan 1 itemset:

$$Support(samsung) = \frac{\sum transaksi\ yang\ mengandung\ Samsung}{\sum transaksi}$$
$$= \frac{9}{12}$$
$$= 0{,}75$$
$$Support(apple) = \frac{\sum transaksi\ yang\ mengandung\ apple}{\sum transaksi}$$
$$= \frac{6}{12}$$
$$= 0{,}5$$
$$Support(huawei) = \frac{\sum transaksi\ yang\ mengandung\ huawei}{\sum transaksi}$$
$$= \frac{6}{12}$$
$$= 0{,}5$$
$$Support(advance) = \frac{\sum transaksi\ yang\ mengandung\ advance}{\sum transaksi}$$
$$= \frac{8}{12}$$
$$= 0{,}667$$
$$Support(lenovo) = \frac{\sum transaksi\ yang\ mengandung\ lenovo}{\sum transaksi}$$
$$= \frac{7}{12}$$
$$= 0{,}583$$

Tabel 8.3. 1-Itemset dengan Item Support lebih dari 0,3

Item ID	*Item*	*Support Count*	*Support*
13	Samsung	9	75%
14	Apple	6	50%
15	Huawei	6	50%
16	Advance	8	66,7%
17	Lenovo	7	58%

b. 2-Itemset

Proses selanjutnya adalah pembentukan C2 atau disebut dengan *2-itemset* dengan jumlah *minimum support* adalah 30%. Berikut merupakan perhitungan pembentukan C2 atau *2-itemset*.

$$Support(Samsung, Apple) = \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Apple}{\sum transaksi} = \frac{4}{12} = 0{,}333$$

$$Support(Samsung, Huawei) = \frac{\sum transaksi\ mengandung\ Samsung\ \&\ Huawei}{\sum transaksi} = \frac{2}{12} = 0{,}167$$

$$Support(Samsung, Advance) = \frac{\sum transaksi\ mengandung\ Samsung\ \&\ Advance}{\sum transaksi} = \frac{5}{12} = 0{,}417$$

$$Support(Samsung, Lenovo) = \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Lenovo}{\sum transaksi} = \frac{6}{12}$$

$= 0{,}5$

$$Support(Apple, Huawei) = \frac{\sum transaksi\ mengandung\ Apple\ dan\ Huawei}{\sum transaksi}$$

$$= \frac{3}{12}$$

$= 0{,}25$

$$Support(Apple, Advance) = \frac{\sum transaksi\ \ mengandung\ Apple\ dan\ Advance}{\sum transaksi}$$

$$= \frac{3}{12}$$

$= 0{,}25$

$$Support(Apple, Lenovo) = \frac{\sum transaksi\ mengandung\ Apple\ dan\ Lenovo}{\sum transaksi}$$

$$= \frac{2}{12}$$

$= 0{,}167$

$$Support(Huawei, Advance) = \frac{\sum transaksi\ \ mengandung\ Huawei\ dan\ Advance}{\sum transaksi}$$

$$= \frac{4}{12}$$

$= 0{,}333$

$$Support(Huawei, Lenovo) = \frac{\sum transaksi\ mengandung\ Huawei\ dan\ Lenovo}{\sum transaksi}$$

$$= \frac{2}{12}$$

$= 0{,}167$

$$Support(Advance, Lenovo) = \frac{\sum transaksi\ mengandung\ Advance\ dan\ Lenovo}{\sum transaksi}$$

$$= \frac{4}{12}$$

$= 0{,}25$

Tabel 3 berikut menunjukkan hasil perhitungan *Support Count* dan *Support.*

Tabel 8.4. 2-Itemset

Item ID	*Itemset*	*Support Count*	*Support*
{13,14}	Samsung, Apple	4	33,3%
{13,15}	Samsung, Huawei	2	~~16,7%~~
{13,16}	Samsung, Advance	5	41,7%
{13,17}	Samsung, Lenovo	6	50%
{14,15}	Apple, Huawei	3	~~25%~~
{14,16}	Apple, Advance	3	~~25%~~
{14,17}	Apple, Lenovo	2	~~16,7%~~
{15,16}	Huawei, Advance	4	33,3%
{15,17}	Huawei, Lenovo	2	~~16,7%~~
{16,17}	Advance, Lenovo	4	33,3%

Minimum support yang ditentukan adalah 30%, jadi kombinasi 2 itemset yang tidak memenuhi *minimum support* akan dihilangkan, sehingga diperoleh hasil sebagaimana Tabel 8.5 di bawah.

Tabel 8.5. *2-Itemset* dengan *Item Support* lebih dari 0,3

Item ID	*Itemset*	*Support Count*	*Support*
{13,14}	Samsung, Apple	4	33,3%

{13,16}	Samsung, Advance	5	41,7%
{13,17}	Samsung, Lenovo	6	50%
{15,16}	Huawei, Advance	4	33,3%
{16,17}	Advance, Lenovo	4	33,3%

c. *3-Itemset*

Pembentukan C3 atau disebut dengan *3-itemset* dengan jumlah *minimum support* = 30%. Berikut merupakan perhitungan pembentukan C3 atau *3-itemset.*

Tabel 8.5. *3-itemset*

Item ID	*Itemset*	*Support Count*	*Support*
{13,14,15}	Samsung, Apple, Huawei	1	8,3%
{13,14,16}	Samsung, Apple, Advance	1	8,3%
{13,14,17}	Samsung, Apple, Lenovo	3	25%
{13,15,16}	Samsung, Huawei,Advance	1	8,3%
{13,15,17}	Samsung, Huawei, Lenovo	1	8,3%
{13,16,17}	Samsung,Advance, Lenovo	3	25%

Karena kandidat *3-itemset* tidak ada yang memenuhi *minimum support,* maka tidak digunakan untuk pertimbangan asosiasi. Dalam kasus ini, kita hanya mempertimbangan asosiasi *2-itemset.*

2. *Confidence*

Nilai *confidence* dari aturan A→B diperoleh dengan rumus:

$$Confidence = P(B|A) = \frac{\sum transaksi\ mengandung\ A\ dan\ B}{\sum transaksi\ mengandung\ A}$$

$$Confidence(Apple \rightarrow Samsung)$$
$$= \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Apple}{\sum transaksi\ mengandung\ Samsung}$$
$$= \frac{4}{9} = 0{,}444$$
$$Confidence(Samsung \rightarrow Apple) =$$
$$\frac{\sum transaksi\ mengandung\ Samsung\ dan\ Apple}{\sum transaksi\ mengandung\ Apple}$$
$$= \frac{4}{6} = 0{,}667$$
$$Confidence(Advance \rightarrow Samsung)$$
$$= \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Advance}{\sum transaksi\ mengandung\ Samsung}$$
$$= \frac{5}{9} = 0{,}556$$
$$Confidence(Samsung \rightarrow Advance)$$
$$= \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Advance}{\sum transaksi\ mengandung\ Advance}$$
$$= \frac{5}{8} = 0{,}625$$
$$Confidence(Lenovo \rightarrow Samsung)$$
$$= \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Lenovo}{\sum transaksi\ mengandung\ Samsung}$$
$$= \frac{6}{9} = 0{,}667$$
$$Confidence(Samsung \rightarrow Lenovo)$$
$$= \frac{\sum transaksi\ mengandung\ Samsung\ dan\ Lenovo}{\sum transaksi\ mengandung\ Lenovo}$$
$$= \frac{6}{7} = 0{,}857$$

$$Confidence(Huawei \rightarrow Advance)$$
$$= \frac{\sum transaksi\ mengandung\ Huawei\ dan\ Advance}{\sum transaksi\ mengandung\ Advance}$$
$$= \frac{4}{8} = 0{,}5$$
$$Confidence(Advance \rightarrow Huawei)$$
$$= \frac{\sum transaksi\ mengandung\ Huawei\ dan\ Advance}{\sum transaksi\ mengandung\ Huawei}$$
$$= \frac{4}{6} = 0{,}667$$
$$Confidence(Advance \rightarrow Lenovo)$$
$$= \frac{\sum transaksi\ mengandung\ Advance\ dan\ Lenovo}{\sum transaksi\ mengandung\ Lenovo}$$
$$= \frac{4}{7} = 0{,}571$$

$$Confidence(Lenovo \rightarrow Advance)$$
$$= \frac{\sum transaksi\ mengandung\ Advance\ dan\ Lenovo}{\sum transaksi\ mengandung\ Advance}$$
$$= \frac{4}{8} = 0{,}5$$

Tabel 8.6. Hasil Perhitungan *Confidence* untuk *Asosiasi 2-itemset*

Aturan Asosiasi	*Confidence*
Apple → Samsung	~~44,4%~~
Samsung→ Apple	66,7%
Advance → Samsung	~~55,6%~~
Samsung→ Advance	62,5%
Lenovo→Samsung	66,7%
Samsung→Lenovo	85,7%
Huawei → Advance	~~50%~~
Advance → Huawei	66,7%
Advance → Lenovo	~~57,1%~~

Lenovo → Advance	~~50%~~

Aturan assosiasi yang tidak memenuhi *minimum confidence* 60% dihilangkan, sehingga diperoleh hasil yang disajikan di Tabel 8.7 berikut.

Tabel 8.7. Aturan Asosiasi *2-itemset* yang Memenuhi *Minimum Confidence*

Aturan Asosiasi	*Confidence*
Samsung→ Apple	66,7%
Samsung→ Advance	62,5%
Lenovo→Samsung	66,7%
Samsung→Lenovo	85,7%
Advance → Huawei	66,7%

Berdasarkan Tabel 8.7 kita bisa menyimpulkan seorang konsumen membeli Samsung dan membeli Apple dengan kemungkinan sebesar 66,7%. Hal yang sama ditunjukkan untuk pembelian barang yang lain.

D. ALGORITMA *FP-Growth*

Frequent Pattern-Growth disingkat *FP-Growth* adalah salah satu alternatif algoritma yang digunakan untuk menentukan himpunan data yang paling sering muncul (*frequent itemset*) dalam sebuah kumpulan data. Algoritma ini menggunakan konsep pembangunan tree, yang disebut *FP-Tree*, untuk mencari frequent itemsets tanpa menggunakan generate candidate seperti yang dilakukan pada algoritma Apriori. Dengan menggunakan *FP-Tree, FP-Growth* dapat langsung memperoleh *frequent itemset*, membuatnya lebih cepat daripada algoritma Apriori. Proses penentuan *frequent itemset* melibatkan dua tahap: pembuatan *FP-Tree* dan penerapan algoritma *FP-Growth* untuk menemukan *frequent*

itemset. *FP-Growth* menggunakan pendekatan yang berbeda dari paradigma yang digunakan pada algoritma Apriori.

Struktur data yang digunakan untuk mencari *frequent itemset* adalah perluasan dari pohon *prefix*, atau *FP-Tree*, yang memungkinkan algoritma *FP-Growth* mengekstrak *frequent itemset* dengan prinsip divide and conquer. Penerapan *FP-Growth* telah digunakan dalam berbagai konteks, seperti menentukan tata letak barang dalam bisnis retail, aplikasi prediksi persediaan sepeda motor, dan penentuan *cross-selling* produk. *FP-Growth* secara efisien memanfaatkan struktur data tree untuk mencari himpunan data yang sering muncul dari sekumpulan data.

Algoritma *FP-Growth*

1. **Membentuk *FP-Tree*** Penggalian frequent itemset dengan menggunakan algoritma *FP-Growth* dilakukan dengan cara membangkitkan struktur *data tree* atau disebut dengan *FP-Tree*. *FP-Tree* merupakan struktur penyimpanan data yang dimampatkan. *FP-Tree* dibangun dengan memetakan setiap data transaksi ke dalam setiap lintasan tertentu dalam *FP-Tree*.
2. **Membentuk *Conditional Pattern Base*** Hasil pembentukan *FP-Tree* akan digunakan untuk membangkitkan *subdata* yang berisi *prefix path* (lintasan awal) dan *suffix path* (pola akhiran). *Subdata* inilah yang dinamakan dengan Conditional Pattern Base.
3. **Membentuk *Conditional Tree*** *Conditional Pattern Base* mempunyai *support count* di setiap *item*nya yang akan dijumlahkan, lalu ketika item memiliki jumlah *support* count lebih besar atau sama dengan minimum *support* count maka dibangkitkan dengan conditional *FP-Tree*.
4. ***Association Rule Frequent Itemset*** *Association* rule merupakan suatu proses pada data mining untuk menentukan semua aturan asosiatif yang memenuhi syarat minimum untuk *support* (*minSup*) dan *confidence* (*minConf*) pada sebuah

database. Kedua syarat tersebut akan digunakan untuk *Association Rules* dengan dibandingkan dengan batasan yang telah ditentukan, yaitu *minSup* dan *minConf*. *Association rule* adalah suatu prosedur untuk mencari hubungan antar item dalam suatu *dataset.* Dimulai dengan mencari *frequent itemset,* yaitu kombinasi yang paling sering terjadi dalam suatu *itemset* dan harus memenuhi *minSup*.

5. ***Frequent Itemset*** Diperoleh dari kombinasi item untuk setiap *conditional FP-Tree* jika conditional *FP-Tree* merupakan *single path* (lintasan tunggal). Apabila *conditional FP-Tree* merupakan *non-single path* (bukan lintasan tunggal) akan dilakukan pembangkitan rekursif oleh *FP-Growth* yaitu *FP-Growth* akan memanggil dirinya sendiri.

Contoh Terapan Misalkan sebuah *platform* belanja *online* yang menyimpan data transaksi pelanggan. *Dataset* ini berisi informasi tentang barang-barang yang dibeli oleh pelanggan dalam beberapa transaksi. Dengan menggunakan algoritma *FP-Growth*, *platform* tersebut dapat mengidentifikasi pola pembelian yang kuat, membantu dalam menyusun strategi pemasaran yang lebih efektif dan menyajikan rekomendasi produk yang lebih baik. Data diberikan pada Tabel 8.8.

Tabel 8.8. Data Penjualan *Online*

No.	*Itemset*
1	Laptop, Printer, Mouse
2	Laptop, Headset, Mouse
3	Smartphone, Charger
4	Smartphone, Laptop, Mouse
5	Printer, Headset, Mouse, Charger

Langkah-langkah penyelesaian menggunakan *FP-Growth*.

1) Menghitung *support*

a) Menghitung dukungan untuk setiap item tunggal dan buang item yang tidak mencapai tingkat dukungan minimum.
b) Misalnya dengan tingkat dukungan minimum 2 transaksi: Laptop (3), Printer (2), Mouse (4), Headset (2), Smartphone (2), Charger (2).

2) Mengurutkan item berdasarkan *support*
   a) Mengurutkan *item* berdasarkan tingkat dukungan secara menurun: {Mouse, Laptop, Printer, Smartphone, Headset, Charger}.
3) Membangun pohon *FP-Tree*
   a) Melakukan pemrosesan transaksi dan konstruksi pohon FP-Tree.
   b) Sebagai contoh, transaksi 1: {Laptop, Printer, Mouse} → FP-Tree: (Root) - [Mouse, Laptop, Printer].
   c) Dan seterusnya
4) Memproses ulang transaksi dan pohon
   a) Mengulang transaksi, misalnya transaksi 2: {Laptop, Headset, Mouse} → *FP-Tree*: (Root) - [Mouse, Laptop, Printer] - [Headset].
5) Mengekstraksi pola asosiasi
   a) Dengan *FP-Tree* yang telah dibangun, kita dapat mengekstrak pola asosiasi dengan menggabungkan cabang pohon yang memiliki tingkat dukungan mencukupi.
   b) Memperoleh aturan asosiasi. Sebagai contoh diperoleh aturan asosiasi: {Laptop} -> {Mouse} dengan tingkat kepercayaan 66.67%.

Kesimpulan Dengan menggunakan algoritma FP-Growth, platform belanja *online* dapat menemukan pola-pola pembelian yang signifikan. Misalnya, dari contoh di atas, platform dapat menyimpulkan bahwa ketika pelanggan membeli Laptop, mereka cenderung juga membeli Mouse dengan tingkat kepercayaan tertentu. Hal ini memungkinkan platform untuk menyusun promosi

bundel atau menyesuaikan rekomendasi produk, meningkatkan pengalaman belanja pelanggan dan meningkatkan penjualan.

E. PENERAPAN PADA R

Di bawah ini diberikan contoh penggunaan R untuk menyelesaikan masalah asosiasi. Beberapa library harus diinstall terlebih dahulu.

```
library(arules)
library(Matrix)
library(grid)
library(arulesViz)
data=list(c("A"),c("A","B"))
data
data1=as(data,"transactions")
data1
dim(data1)
class(data1)
crossTable(data1)
data2=apriori(data1,parameter= list(supp=0.2, conf=0.2))
inspect(sort(data2))
as(data2,"data.frame")
plot(data2,method="graph")
```

```
> library(arules)
Loading required package: Matrix

Attaching package: 'arules'

The following objects are masked from 'package:base':

    abbreviate, write

> library(Matrix)
> library(grid)
> library(arulesViz)
> data=list(c("A"),c("A","B"))
> data
[[1]]
[1] "A"

[[2]]
[1] "A" "B"

> data1=as(data,"transactions")
> data1
transactions in sparse format with
 2 transactions (rows) and
 2 items (columns)
> dim(data1)
[1] 2 2
> class(data1)
[1] "transactions"
attr(,"package")
[1] "arules"
```

```
> crossTable(data1)
  A B
A 2 1
B 1 1
> data2=apriori(data1,parameter= list(supp=0.2, conf=0.2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.2    0.1    1 none FALSE            TRUE       5     0.2      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 0

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[2 item(s), 2 transaction(s)] done [0.00s].
sorting and recoding items ... [2 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [4 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> inspect(sort(data2))
    lhs    rhs support confidence lift count
[1] {}  => {A} 1.0     1.0        1    2
[2] {}  => {B} 0.5     0.5        1    1
[3] {B} => {A} 0.5     1.0        1    1
[4] {A} => {B} 0.5     0.5        1    1
> as(data2,"data.frame")
       rules support confidence lift count
1  {} => {B}     0.5        0.5    1     1
2  {} => {A}     1.0        1.0    1     2
3 {B} => {A}     0.5        1.0    1     1
4 {A} => {B}     0.5        0.5    1     1
```

F. RANGKUMAN

Associations Rules merupakan konsep dalam *data mining* yang digunakan untuk menemukan pola atau hubungan antara *item-item* dalam *dataset*. Tujuannya adalah untuk mengidentifikasi asosiasi atau korelasi antara *item-item* tersebut, sehingga dapat memberikan wawasan tentang hubungan yang mungkin terjadi.

Metode yang umum digunakan untuk menemukan *Association Rules* adalah metode Apriori. Metode ini berfokus pada identifikasi *itemset* yang sering muncul bersama dalam *dataset*. Langkah-langkahnya melibatkan pencarian *itemset* yang memenuhi tingkat

dukungan (*support*) minimum yang ditentukan, dan kemudian menghasilkan *Association Rules* berdasarkan *itemset* tersebut.

Algoritma *FP-Growth* (*Frequent Pattern Growth*) adalah metode yang efisien untuk menemukan pola sering muncul (*frequent patterns*) dalam *dataset*. *FP-Growth* memberikan pendekatan inovatif dalam penemuan pola sering muncul dan memanfaatkan struktur FP-Tree untuk mengoptimalkan proses tersebut.

G. TES FORMATIF

1. Jelaskan apa yang dimaksud dengan *Association Rules*!
2. Jelaskan apa yang dimaksud dengan *support, confidence, minimum support, minimum confidence,* dan *lift!* Berikan contoh!
3. Berikan contoh perhitungan *support, confidence, dan lift ratio!*
4. Dalam sebuah tabel transaksi terdiri dari 20 transaksi dan 10 item. Jika dilakukan analisa asosiasi, berapa banyaknya kemungkinan *rule* yang terbentuk dari tabel transaksi tersebut!
5. Berikan Langkah-langkah algoritma Apriori!
6. Berikan Langkah-langkah algoritma *FP-Growth*!
7. Di superstore tersedia bermacam-macam barang. Misalkan X dan Y adalah dua macam barang yang berbeda. Pegawai superstore menyusun barang-barang menggunakan algoritma Apriori. Apa yang harus dia ketahui tentang Confidence($X \rightarrow Y$)?
8. Jelaskan syarat dan arti Confidence($X \rightarrow Y$)!
9. Kesimpulan apa yang diperoleh jika digunakan algoritma Apriori?
10. Kesimpulan apa yang diperoleh jika digunakan algoritma *FP-Growth*?

H. LATIHAN

1. Jika dipunyai himpunan transaksi apa tujuan utama kita melakukan penambangan himpunan tersebut dengan aturan asosiasi! Jelaskan!
2. Apakah yang dimaksud dengan kompleksitas waktu komputasi? Sebutkan faktor-faktor yang mempengaruhi kompleksitas waktu komputasi!
3. Dalam himpunan transaksi jika penambangan dilakukan dengan menggunakan aturan asosiasi *Frequent Itemset Generation*, secara perhitungan masih mahal. Jelaskan alasannya! Berapakah kompleksitasnya!
4. Jelaskan apa yang dimaksud dengan *frequent itemset*! Berikan contohnya!
5. Jelaskan apa yang dimaksud dengan *minSupp* dan *minConf*!
6. Perhatikan Tabel di bawah ini. Jelaskan informasi yang diberikan dalam Tabel tersebut!

Tabel. Daftar transaksi item

TID	Daftar Item
1	Baby Shoap, Nuts, Diaper
2	Baby Shoap, Coffee, Diaper
3	Baby Shoap, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

7. Carilah *support count* semua *frequent itemset* dengan panjang k=2!
8. Berapakah jumlah aturan asosiasi yang dapat dibentuk dari data tabel transaksi di atas!
9. Apabila minSupp=40% dan minConf=60%, dengan menggunakan algoritma apriori, carilah pola aturan yang

terbaik dengan bantuan software R! Berikan penjelasan apa yang telah Anda peroleh!

10. Apabila minSupp=30% dan minConf=40%, dengan menggunakan algoritma *FP-Growth*, carilah pola aturan yang terbaik dengan bantuan software R! Berikan penjelasan apa yang telah Anda peroleh!

KEGIATAN BELAJAR 9

TEXT DATA MINING

DESKRIPSI PEMBELAJARAN

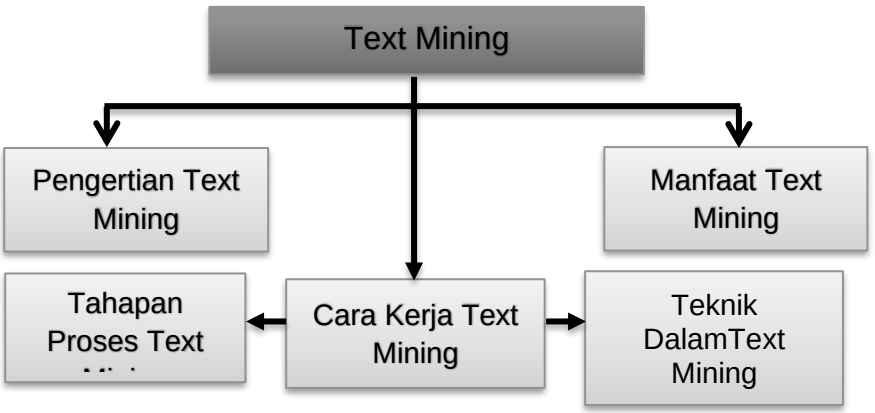
Pada bab ini mahasiswa mempelajari pengenalan dan konsep dasar teoritis text mining. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk modal dasar mempelajari Data mining dan text mining lebih lanjut.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan :

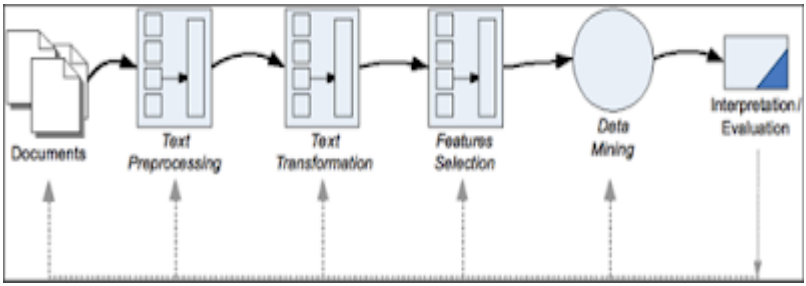
1. Mampu menguraikan definisi text mining dalam data mining.
2. Mempu menjelaskan fungsi dan manfaat text mining.
3. Mampu menjelaskan tahapan proses, cara kerja dan teknik dalam text mining.

PETA KONSEP PEMBELAJARAN



A. PENGERTIAN TEXT MINING

Data mining merupakan proses penggalian untuk menyelesaikan masalah kebutuhan informasi dengan menerapkan teknik data mining, machine learning, natural language processing, pencarian informasi dan manajemen pengetahuan. Text mining melibatkan praproses dokumen seperti kategorisasi teks, ekstraksi informasi dan ekstrasi kata. Metode ini digunakan untuk mengekstrasi informasi dari sumber data melalui identifikasi dan eksplorasi pola yang menarik.



Gambar 9.1. Tahapan Proses Text Mining (sumber: Erni Sitas Blog)

Teks mining dapat dilihat sebagai proses dua tahap yang dimulai dengan penerapan struktur ke sumber data teks dan dilanjutkan dengan ekstraksi pengetahuan dan informasi yang terkait dengan data teks terstruktur menggunakan metode dan alat yang sama.

Secara khusus, tujuan dari text mining dapat dibagi menjadi dua:

1. Pengkategorisasian data teks (text categorization) Dalam pengkategorisasian, text mining dipergunakan sebagai alat untuk menemukan kategori yang sesuai dengan kelas yang telah ditentukan (supervised learning)
2. Pengelompokan data teks (text clustering) Pada pengelompokan, text mining berfungsi sebagai alat untuk mengelompokkan data teks berdasarkan kesamaan karakteristik, dan clustering dapat digunakan untuk

memberikan label pada kelas yang belum diketahui (unsupervised learning)

Tujuan text mining adalah sama dengan tujuan data mining yaitu menemukan pola pada data agar dapat dimanfaatkan manusia untuk membantu pekerjaannya. Karena data teks belum terstruktur maka pada text mining terdapat proses-proses tambahan yang harus dilakukan sebelum dilakukan operasi penambangan. Proses tambahan itu adalah preprocessing yang bertujuan untuk membersihkan teks. Kemudian proses ekstraksi fitur yang mengubah data teks menjadi data terstruktur untuk diproses oleh operasi penambangan dengan algoritma data mining. Selain itu pada text mining dapat dilakukan analisis sentimen dengan berbasis lexicon.

Untuk mempermudah melakukan proses-proses tersebut bisa digunakan bahasa pemrograman R, karena bahasa ini adalah salah satu yang umum digunakan dalam riset data mining atau text mining. Tantangan umum yang dihadapi pada text mining adalah bahasa. Pada implementasinya bahasa mempengaruhi metode atau teknik yang digunakan.

Text mining dapat didefenisikan secara luas sebagai proses intensif pengetahuan di mana pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis. Penambangan terks berusaha untuk mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan identifikasi dan eksplorasi pola yang menarik. Dalam kasus penambangan teks, bagaimanapun, sumber data adalah kumpulan dokumen, dan pola yang menarik ditemukan bukan di antara catatan database yang di formalkan tetapi dalam data tekstual yang tidak terstruktur dalam koleksi dokumen. Oleh karena itu penambangan teks dan sistem penambangan data menunjukkan banyak kesamaan arsitektur tingkat tinggi. Contohnya, kedua jenis sistem ini bergantung pada rutinitas preprocessing, algoritma penemuan pola, dan elemen lapisan

presentasi seperti alat visualisasi untuk meningkatkan penelusuran set jawaban.

B. MANFAAT TEXT MINING

Text mining adalah salah satu bidang khusus dalam data mining yang memiliki defenisi menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata – kata yang dapat mewakili isi dari dokumen sehingga dapat menganalisa dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui bagaimana mereka berhubungan dengan variabel

Penerapan penggunaan yang paling umum dilakukan text mining dengan analitik untuk mendapatkan wawasan tentang sentimen pelanggan dapat membantu perusahaan mendeteksi masalah produk dan bisnis dan kemudian mengatasinya sebelum menjadi masalah besar yang mempengaruhi penjualan.

Teks mining dalam ulasan dan komunikasi pelanggan juga dapat mengidentifikasi fitur baru yang diperlukan untuk teks mining yang diperlukan untuk membantu memperkuat penawaran produk.Dalam setiap kasus, teknologi menghadirkan peluang untuk meningkatkan pengalaman pelanggan secara keseluruhan, yang diharapkan dapat menigkatkan pendapatan dan keuntungan.

Sebagai bagian dari program manajemen pelanggan dan pemasaran, text mining juga dapat membantu memprediksi pergantian pelanggan, memungkinkan perusahaan mengambil tindakan untuk mencegah potensi pembelotan ke pesaing komersial. Deteksi penipuan manajemen risiko, iklan online,dan manajemen konten web adalah fungsi lain yang

dapat memanfaatkan alat text mining. Dalam perawatan kesehatan, teknologi mungkin dapat membantu mendiagnosa penyakit dan kondisi medis berdasarkan gejala yang dilaporkan pasien. Penggunaan umum lainnya dari penambangan teks termasuk menyaring pelamar kerja berdasarkan kata-kata di resume, memblokir spam, mengklasifikasikan konten situs web, menandai klaim asuransi yang berpontensi penipuan, menganalisis deskripsi gejala medis untuk membantu diagnosis, dan memeriksa dokumen perusahaan sebagai bagian dari proses penemuan elektronik.

Perangkat lunak penambangan teks juga menyediakan kemampuan pencarian informasi yang serupa dengan yang disediakan oleh mesin pencari dan platform pencarian perusahaan, tetapi ini sering kali hanya merupakan elemen dari aplikasi penambangan teks tingkat tinggi dan tidak berguna dengan sendirinya. Penerapan yang paling umum dilakukan text mining saat ini misalnya penyaringan spam, analisa mengukur preferensi pelanggan, meringkas pengelompokan topik penelitian, dan banyak lainnya.

C. TAHAPAN PROSES TEXT MINING

Tahapan proses text mining berdasarkan ketidakaturan struktur data teks, maka proses text mining memerlukan beberapa tahap awal yang berfungsi mempersiapkan agar teks dapat diubah menjadi lebih terstruktur.

Berikut merupakan proses dari Text Mining.

1. Dokumen - Plain text, format elemen (XML, email, HTML, RTF, OTD, dsb) dan format biner (PDF, DOC, dsb)
2. Text Preprocessing – Process pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur

3. Text Transformation – Pembentukan atribut mengacu pada proses untuk mendapatkan representasi dokumen yang diharapkan
4. Features Selection – Pemilihan fitur merupakan tahapan lanjut dari pengurangan dimensi pada proses transformasi teks.
5. Data Mining – Penemuan pola atau pengetahuan dari keseluruhan data teks.
6. Interpretation/Evaluation – Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.

Algoritma Text Mining, antara lain:
1. Information Extraction from Text Data
2. Text Summarization
3. Unsupervised Learning Methods from Text Data
4. LSI and Dimensionality Reduction for Text Mining
5. Supervised Learning Methods for Text Data
6. Transfer Learning with Text Data
7. Probabilistic Technique for Text Mining
8. Mining Text Streams
9. Opinion Mining from Text Data
10. Text Mining in Social Media
11. Text Mining from Biomedical Data.

D. CARA KERJA TEXT MINING

Text mining mirip dengan data mining, tetapi berfokus pada teks daripada bentuk data yang lebih terstruktur. Namun, langkah pertama dalam proses text mining adalah mengatur dan menyusun data dengan cara tertentu sehingga dapat digunakan untuk analisis kualitatif.

Melakukannya sering kali melibatkan penggunaan teknik pemrosesan bahasa alami yang menerapkan prinsip-prinsip

linguistik komputasional untuk mengurai dan menginterpretasikan kumpulan data.

Pekerjaan awal termasuk mengklasifikasikan, mengelompokkan, dan memberi label teks; menggabungkan kumpulan data; membuat taksonomi; dan mengekstrak informasi tentang frekuensi istilah dan hubungan antara entitas data. Model analitis kemudian dijalankan untuk menghasilkan temuan yang membantu mendorong strategi bisnis dan tindakan operasional.

Gambar 9.2: Teknik Text Mining dan contoh penerapannya (sumber: dasboard kompas)

E. TEKNIK DALAM TEXT MINING

1. Klasifikasi Text: Teknik Klasifikasi Teks melibatkan penggunaan algoritma seperti Naïve bayes, Super Vecor Machine (SVM), atau Deep Learning untuk mengklasifikasikan teks kedalam kategori atau label yang relevan, misalnya klasifikasi teks dapat digunakan untuk memprediksi apakah sbuah email adalah spam atau bukan.
2. Pengelompokan Text (Clustering): Teknik pengelompokan teks melibatkan pengelompokan dokumen yang memiliki kesamaan berdasarkan pola atau karakteristik yang ada dalam

teks.Metode seperti K-Means atau Hierarchical Clustering dapat digunakan untuk mengelompokkan dokumen yang serupa. Misalnya pengelompokan teks dapat digunakan untuk mengelompokkan berita-berita berdasarkan topik yang sama.

3. Analisis Sentimen: Teknik analisis sentimen digunakan untuk menganalisis sentiment atau sikap opini dalam teks, apakah bersifat positif, negatif atau atau netral. Metode seperti analisis sentiment lexicon, atau Machine Learning dapat digunakan untuk menganalisis sentiment dalam taks, mialnya untuk melihat respon penggunan terhadap produk atau layanan.
4. Ekstraksi Informasi: Teknik ekstraksi informasi bertujuan untuk menemukan dan mengestrak informasi tertentu dari teks, seperti entitas nama, tanggal, lokasi, atau angka. Metode seperti Named Entity Recognition (NER) atau Reguler Expression (RegEx) dapat di gunakan untuk mengestrak informasi tersebut. Contohnya, ekstaksi informasi dapat digunakan untuk mengestrak nama-nama produk dari ulasan pelanggan.
5. Pemrosesan Bahasa Alami (NLP): Teknik pemrosesan bahasa alami melibatkan penerapan teknik dan model yang dirancang untuk memahami, menginterpretasi, dan menghasilkan teks yang lebih terstruktur dan bermakna. Pemrosesan bahasa alami dapat mencakup pemahaman teks, analisis sintaks, analisis semantik, atau generasi teks / penerapan NLP dapat di gunakan dalam berbagi tugas seperti chatbat, pemahaman pertanyaan atau terjemahan otomatis.

Rumus Dalam Text Mining :

1. Term Frequency (TF): Rumus TF digunakan untuk mengukur seberapa sebuah kata muncul dalam sebuah dokumen atau korpus teks. Rumus TF adalah jumlah kemunculan kata dalam dokumen di bagi dengan jumlah kata dalam dokumen tersebut. TF = (jumlah kemunculan kata dalam dokumen) / (jumlah kata dalam dokumen).

2. Inverse Document Frequency (IDF) : Rumus IDF digunakan untuk mengukur pentingnya sebuah kata dalam korpus teks secara keseluruhan. Rumus IDF adalah logaritma dari jumlah dokumen dalam korpus di bagi dengan jumlah dokumen yang mengandungkata tersebut. IDF = Log((jumlah dokumen dalam korpus) / (jumlah dokumen yang mengandung kata))
3. TF-IDF: Rumus TF-IDF digunakan untuk menggabungkan informasi dari TFdan IDF sehingga dapat menemukan bobot kata yang tinggi dalam sebuah dokumen, tetapi rendah dalam korpus keseluruhan.Rumus TF-IDF adalah hasil perkalian antara nilai TF dengan nilai IDF. TF-IDF = TF * IDF.
4. Cosine Similarity : Rumus Cosine Similarity digunakan untuk mengukur kemiripan antara dua vector representasi teks.Rumus Cosine Similarity adalah hasil perkalian dot product antara dua vector representasi teks, dibagi dengan norma vector dari kedua vector tersebut. Cosine Similarity = (A-B) / (II A II * II B II).

F. RANGKUMAN

Dalam era digital yang penuh dengan data teks, kemampuan untuk memproses dan menganalisis informasi yang tersembunyi dalam teks semakin penting. Dalam hal ini text mining menjadi kunci untuk menggali informasi yang berharga dari teks dalam skala besar. Text mining melibatkan analisis dan ekstraksi informasi dari teks menggunakan teks dan rumus khusus.

Berdasarkan uraian di atas di mulai dari pengertian test mining, manfaat text mining, tahapan proses text mining, cara kerja text mining sampai ke teknik dalam text mining yang mempunyai tujuan text mining adalah sama dengan tujuan data mining yaitu menemukan pola pada data agar dapat dimanfaatkan manusia untuk membantu pekerjaannya. Karena data teks belum terstruktur maka pada text mining terdapat proses-proses tambahan yang harus dilakukan sebelum dilakukan operasi

penambangan. Penambangan teks dan sistem penambangan data menunjukkan banyak kesamaan arsitektur tingkat tinggi bergantung pada rutinitas preprocessing, algoritma penemuan pola, dan elemen lapisan presentasi seperti alat visualisasi untuk meningkatkan penelusuran set jawaban.

G. TES FORMATIF

1. Perangkat Lunak Penambangan Teks menyediakan :
   a. Kemampuan pencarian informasi yang serupa dengan yang disediakab oleh mesin pencari
   b. Kemampuan menciptakan kata di platform
   c. Kemampuan melihat platform
   d. Kemampuan menghilangkan informasi di platform
   e. Semua benar
2. Mengklasifikasikan, mengelompokkan, dan memberi label teks; menggabungkan kumpulan data; membuat taksonomi; dan mengekstrak informasi tentang frekuensi istilah dan hubungan antara entitas data. Merupakan cara kerja dari :
   a. Klasifikasi
   b. Data
   c. Data Mining
   d. Text Mining
   e. Entitas informasi

H. LATIHAN

1. Jelaskan apa itu Text mining dan mengapa diperlukan.
2. Uraikan dan jelaskan perbedaan Text Mining, Data Mining dan Web Mining.
3. Uraikan dan jelaskan tahapan Text Mining.

KEGIATAN BELAJAR 10

METODE EKSTRAKSI DAN SELEKSI FITUR

DESKRIPSI PEMBELAJARAN

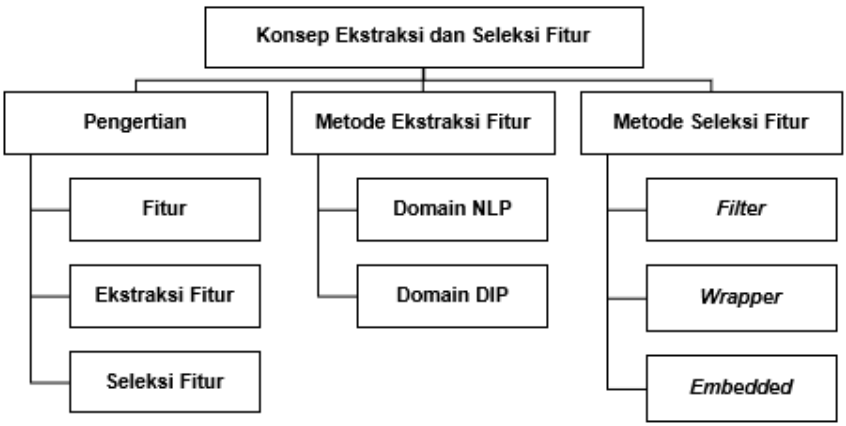
Pada bab ini mahasiswa mempelajari pengenalan dan konsep dasar mengenai ekstraksi dan seleksi fitur pada data mining. Diharapkan mahasiswa memiliki wawasan dan pemahaman untuk modal dasar mempelajari data mining lebih lanjut.

KOMPETENSI PEMBELAJARAN

Setelah mengikuti perkuliahan ini diharapkan mahasiswa dan mahasiswi memiliki pengetahuan dan kemampuan:

1. Mampu menguraikan definisi ekstraksi dan seleksi fitur.
2. Mampu menjelaskan fungsi dan tujuan ekstraksi dan seleksi fitur.
3. Mampu menjelaskan metode ekstraksi dan seleksi fitur.

PETA KONSEP PEMBELAJARAN



A. PENGERTIAN EKSTRAKSI DAN SELEKSI FITUR

Ekstraksi dan seleksi fitur merupakan salah satu tahapan yang dilakukan pada proses penambangan data (*data mining*). Tahapan ini memiliki peran penting, karena menjadi awal berhasil tidaknya proses penambangan data secara keseluruhan. Sebelum dikaji lebih dalam, ada baiknya dipahami dulu terkait fitur, ekstraksi fitur, serta seleksi fitur. Berikut pemaparan dari masing-masing istilah tersebut.

1. Fitur

Fitur dalam konteks data mining mengacu pada atribut atau variabel individu. Fitur yang dimiliki sebuah data bisa merepresentasikan jenis warna, frekuensi suara, tekstur, bentuk, hingga teks. Melalui fitur ini, proses pemahaman dan analisis data menjadi lebih mudah dilakukan, terlebih dalam menemukan pola yang terkandung di dalamnya.

Sebagai contoh dalam sebuah tabel pelanggan, maka fitur yang mungkin bisa digali seperti usia, jenis kelamin, pola barang yang dibeli, dan lainnya. Begitu juga misalnya dalam membedakan jenis buah dalam sebuah citra, maka fitur yang mungkin bisa dianalisis dari sisi warna buah, bentuk buah, serta tekstur dari kulit buah.

Fitur dapat dibedakan menjadi dua, yakni fitur natural dan fitur buatan. Fitur natural merupakan fitur yang menjadi bagian dari data, seperti tepi objek pada citra, ukuran frekuensi pada suara, dan sejenisnya. Fitur buatan merupakan fitur yang diperoleh setelah adanya intervensi dari operasi-operasi tertentu. Contohnya nilai histogram data, pengaturan nilai frekuensi, dan lainnya.

2. Ekstraksi Fitur

Sesuai definisi fitur yang telah dijelaskan, maka untuk bisa mendapatkan fitur tersebut diperlukan adanya teknik khusus untuk menggalinya. Teknik ini dikenal dengan istilah ekstraksi

fitur. Ekstraksi fitur dapat dikatakan sebagai proses identifikasi serta mengekstrak fitur penting yang ada di dalam data.

Ekstraksi fitur ini merupakan tahap penting dalam pengolahan data, karena berpengaruh pada tingkat akurasi model *machine learning* yang dikembangkan. Setelah fitur-fitur data diekstrak, maka model *machine learning* dapat dilatih menggunakan fitur tersebut untuk menghasilkan nilai akurasi yang optimal.

Mengingat pentingnya peran fitur, maka dalam proses ektraksi fitur perlu dilakukan dengan cermat. Kesalahan dalam proses ekstraksi, akan menghasilkan model *machine learning* yang buruk. Begitu juga jika fitur diekstrak dengan tepat, maka model *machine learning* yang dihasilkan juga akan baik.

3. Seleksi Fitur

Proses ekstraksi fitur seperti yang sudah dijelaskan, bahwa memiliki peran penting dalam menggali fitur-fitur yang dimiliki sebuah objek. Dalam kondisi ideal tentunya diharapkan fitur yang diperoleh bisa memberikan tingkat akurasi model *machine learning* yang tinggi. Namun, tidak jarang juga fitur yang dihasilkan kurang memberikan hasil akurasi yang optimal. Oleh karena itu diperlukan adanya pemilihan fitur, atau yang dikenal dengan istilah seleksi fitur.

Seleksi fitur dalam penambangan data mengacu pada proses pemilihan fitur atau variabel yang paling relevan dan signifikan dalam dataset. Pemilihan fitur ini diharapkan mampu meningkatkan performa model *machine learning* yang dirancang. Oleh karena karena itu, seleksi fitur dapat dikatakan sebagai teknik untuk mengurangi jumlah fitur dalam dataset tanpa mengorbankan kinerja model *machine learning*.

B. METODE EKSTRAKSI FITUR

Metode dalam proses ekstraksi fitur sangat beragam sesuai domain penelitian yang dikembangkan. Pada kesempatan ini metode ekstraksi fitur yang dibahas berdasarkan domain NLP dan domain pengolahan citra digital.

1. Domain *Natural Language Processing* (NLP)

Proses ekstraksi fitur pada domain NLP bertujuan untuk mengubah teks mentah menjadi format angka, sehingga algoritma dari *machine learning* dapat memahaminya. Beberapa metode yang sering digunakan yakni sebagai berikut.

a. ***Bag of Word* (BoW)**

Metode BoW mengasumsikan bahwa sebuah dokumen teks atau kalimat sebagai kumpulan kata yang saling independen. Dalam metode ini, setiap kata dalam teks dianggap sebagai fitur dan dihitung frekuensinya.

Contohnya terdapat kalimat berikut: "***Saya suka belajar data mining dan suka belajar data warehouse***". Dari contoh kalimat tersebut, maka BoW yang terbentuk adalah {***"Saya": 1, "suka": 2, "belajar": 2, "data": 2, "mining": 1, "dan": 1, "warehouse": 1***}

Keunggulan metode ini relatif sederhana dan mudah diimplementasikan. Selain itu bekerja dengan baik pada dokumen pendek, serta mampu menggambarkan frekuensi kata dalam dokumen.

b. ***Term Frequency-Inverse Document Frequency* (TF-IDF)**

Metode TF-IDF mengukur seberapa penting suatu kata dalam sebuah dokumen teks. Metode ini menghitung bobot untuk tiap kata berdasarkan frekuensi kata dalam dokumen teks tersebut, serta kebalikannya dengan frekuensi kata dalam seluruh korpus.

Perhitungan nilai TF dilakukan dengan membagi jumlah kemunculan suatu kata dalam sebuah dokumen dengan total kata dalam dokumen tersebut.

Misalkan kata "***data***" muncul sebanyak **50 kali** dalam satu dokumen, dan di dokumen tersebut terdapat **250 kata**, maka nilai **TF** kata "***data***" adalah **50/250 = 0,2**.

Selanjutnya perhitungan nilai IDF dilakukan berdasarkan nilai logaritmik dari pembagian jumlah seluruh dokumen dengan jumlah dokumen yang mengandung suatu kata tertentu.

Misalkan dari kata "***data***" sebelumnya, diketahui muncul dalam **5 dokumen**, dengan **total dokumen** yang ada adalah **15 dokumen**, maka nilai **IDF** adalah **log(15/5) = 0,48**.

Perhitungan terakhir TF-IDF dilakukan dengan mengalikan nilai TF dan IDF, sehingga nilai **TF-IDF** untuk contoh kata "***data***" adalah **0,2 x 0,48 = 0,096**. Perlu dicatat bahwa semakin jarang sebuah kata muncul, maka semakin besar nilai IDF-nya dan begitu juga sebaliknya.

Keuntungan penggunaan metode TF-IDF adalah bobot kata-kata umum yang tidak memberikan banyak informasi dapat dikurangi. Selain itu metode ini mampu menggambarkan kepentingan kata dalam dokumen teks atau korpus kata.

c. ***Word Embedding***

Metode *word embedding* memetakan setiap kata dalam teks ke dalam ruang vektor dengan dimensi tertentu. Beberapa metode *word embedding* yang cukup populer yakni *Word2Vec*, *GloVe*, dan *FastText*.

Keuntungan penggunaan metode *word embedding* yakni dapat digunakan melihat hubungan semantic antar kata, dapat mengenali sinonim dan analogi antar kata, serta mampu bekerja dengan baik dalam konteks pemrosesan bahasa alami.

d. **N-*grams***

N-*grams* adalah kumpulan kata yang berdampingan antar dalam suatu dokumen teks. Jumlah kata berdampingan ini tergantung pada nilai N yang digunakan. Misalnya *unigram* (N=1), *bigram* (N=2), *trigram* (N=3), dan seterusnya.

Secara umum, metode ini memprediksi probabilitas N-*gram* tertentu dalam setiap urutan kata dalam bahasa. Model yang dihasilkan dengan N-*gram* dikatakan baik apabila mampu memprediksi kata berikutnya dalam kalimat.

Memilih metode ekstraksi fitur untuk tugas NLP tentunya memerlukan pemahaman yang baik tentang metode-metode yang ada, serta disesuaikan juga dengan studi kasus. Berikut beberapa *key-points* yang bisa dijadikan pertimbangan.

- ✓ **Jenis tugas**; jenis tugas NLP yang dikerjakan sangat berpengaruh pada pemilihan metode. Seperti misalnya klasifikasi teks, maka metode BoW dan TF-IDF sekiranya relatif sesuai untuk digunakan.
- ✓ **Jenis data**; jenis data yang dimiliki juga penting dijadikan pertimbangan dalam penentuan metode ekstraksi fitur. Contohnya data yang dimiliki berupa teks yang sangat panjang, maka yang bisa dipertimbangkan adalah metode *word embedding*. Hal ini karena dapat menangkap konteks dan makna semantik.
- ✓ **Kinerja model**; uji coba beberapa metode ekstraksi fitur dan membandingkannya dapat dilakukan untuk menentukan kinerja model terbaik.
- ✓ **Performa komputasi**; beberapa metode ekstraksi fitur kemungkinan memerlukan *resource* yang besar, sehingga

perlu dipertimbangkan keterbatasan komputasi yang dimiliki.

- ✓ **Interpretasi**; beberapa metode mungkin lebih mudah diinterpretasikan daripada metode lainnya. Misalnya BoW dan TI-IDF lebih mudah diinterpretasikan dibandingkan metode *word embedding*.

2. Domain *Digital Image Processing* (DIP)

Pada domain pengolahan citra atau DIP, teknik ekstraksi fitur didasarkan pada kondisi objek citra, serta bergantung pada tujuan dari pengolahan yang dilakukan. Secara umum teknik ekstraksi fitur yang dapat dilakukan dibagi menjadi tiga jenis fitur, yakni fitur bentuk, tekstur, dan warna.

Berikut beberapa metode yang bisa digunakan untuk mengekstraksi fitur bentuk, tekstur, atapun warna sebuah citra.

a. Invariant Moment

Invarian dalam metode ini mengacu pada kondisi bahwa nilai citra tidak akan berubah atau hanya berubah sedikit jika dikenakan proses transformasi (*rotation, flipping, scaling*). Metode ini diciptakan oleh Hu (Theoridis dan Koutoumbas, 2006) sejumlah tujuh parameter *moment* seperti yang tertera pada Persamaan (1)

$$
\begin{aligned}
I_1 &= \eta_{20} + \eta_{02} \\
I_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
I_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
I_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2
\end{aligned}
\qquad (1)
$$

$$I_5 = (\eta_{30} - 3\eta_{12})\,(\eta_{30} + \eta_{12})\,[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})\,(\eta_{21} + \eta_{03})\,[3(\eta_{30}+\eta_{12})^2 - (\eta_{21}+\eta_{03})^2]$$

$$I_6 = (\eta_{20} - \eta_{02})\,[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}\,(\eta_{30} + \eta_{12})\,(\eta_{21} + \eta_{03})$$

$$I_7 = (3\eta_{21} - \eta_{03})\,(\eta_{30} + \eta_{12})\,[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})\,(\eta_{21} + \eta_{03})\,[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

b. Zernike Moment

Zernike moment diperkenalkan oleh F. Zernike pada tahun 1934, dan digunakan dalam pengolahan citra digital pertama kali oleh M.R. Teague pada tahun 1980 (Chen, dkk., 2005), dengan hasil berupa *Zernike Moment Descriptors* (ZMD). Momen ini memiliki kelebihan sebagai berikut (Mingqiang, dkk. 2008).

1. Bersifat independent terhadap rotasi
2. Andal terhadap derau dan variasi minor dalam bentuk objek
3. Memiliki redundansi informasi minimum

Meskipun demikian, momen ini memliki kelemahan sebagai berikut.

1. Perlu normalisasi ruang koordinat
2. Perlu penggunaan hampiran penjumlahan mengingat aslinya menggunakan integral. Hal ini yang memberikan pengaruh terhadap terhadap sifat ketidakbergantungan pada rotasi
3. Perlu dilakukan normalisasi untuk translasi dan penyekalaan

c. Convex Hull dan Solidity

Convex Hull mengacu pada himpunan konveks atau cembung, mencakup semua titik yang menghubungkan

dua titik yang berada dalam himpunan. Himpunan konveks merupakan bentuk polygon terkecil yang melingkupi objek.

Sebuah objek dikatakan konveks apabila seluruh pasangan dua titik yang terkandung didalamnya, dibentuk oleh garis yang seluruhnya berada dalam himpunan. Bersama soliditas, konveksitas merupakan fitur morfologi himpunan konveks yang menyatakan sebuah objek terkategori konveks atau tidak. Konveksitas dihitung berdasarkan nilai keliling objek, sedangkan soliditas dihitung berdasarkan nilai luas objek.

d. 1^{st} order Statistical

Metode ekstraksi fitur statistik orde satu didasarkan atas karakteristik histogram citra. Fitur ini digunakan untuk membedakan tesktur makrostruktur, yakni pengulangan pola lokal secara periodik. Melalui fitur ini, diperoleh informasi dasar terkait distribusi intensitas citra, seperti **rerata intensitas**, **simpangan baku** (ukuran sebaran data. Jika nilai ini dikuadratkan disebut varian, yang dapat digunakan untuk menghitung nilai *smoothness*), ***skewness*** (tingkat kecondongan relatif kurva distribusi terhadap rerata intensitas suatu data), ***energy*** (distribusi intensitas piksel terhadap jangkauan aras keabuan), ***entropy*** (fitur untuk mengukur keacakan distribusi intensitas, serta mengindikasikan tingkat kompleksitas data), dan ***smoothness*** (mengacu pada tingkat kehalusan intensitas citra).

e. Gray Level Co-occurence Matrix (GLCM)

GLCM merupakan metode ekstraksi fitur yang digunakan untuk menganalisis tekstur citra. Metode ini menggambarkan hubungan antara dua piksel yang saling bertetangga, memiliki intensitas keabuan, jarak, dan sudut. Terdapat empat sudut yang dapat digunakan pada GLCM, yakni 0^o, 45^o, 90^o, 135^o. Masing-masing nilai dari tiap sudut

ini selanjutnya dirata-ratakan sehingga diperoleh nilai fitur GLCM secara global.

Beberapa nilai statistik yang bisa diekstrak oleh GLCM yakni: ***angular second moment***/ASM (untuk mengukur sifat homogenitas data), ***inverse difference moment***/IDM (mengukur homogenitas citra yang berderajat keabuan sejenis), ***correlation*** (mengukur linieritas data), ***contrast*** (mengukur frekuensi spasial citra dan ukuran penyebarannya), dan ***entropy*** (mengukur kompleksitas atau varians data).

f. Color Moment

Color moment merupakan ukuran yang merepresentasikan distribusi warna dalam citra dengan cara yang sama seperti momen pusat merepresentasikan distribusi probabilitas secara unik. *Color moment* ini digunakan untuk tujuan indeks warna sebagai fitur dalam aplikasi pembandingan citra dari segi warna.

Biasanya terdapat empat parameter yang dijadiakn acuan sebagai *color moment*, yakni: ***mean*** (menyatakan warna rata-rata dalam citra), **simpangan baku** (diperoleh dari akar kuadrat varians distribusi warna), ***skewness*** (mengukur tingkat asimetris distribusi warna, serta informasi tentang bentuk distribusi warna), dan **kurtosis** (mirip dengan *skewness*, berfungsi memberikan bentuk distribusi warna).

C. METODE SELEKSI FITUR

Metode seleksi fitur dapat diterapkan ketika fitur telah diperoleh sesuai hasil tahap ekstraksi fitur. Beberapa metode yang bisa digunakan untuk melakukan seleksi fitur yakni sebagai berikut.

1. *Filter Methods*

Metode *filter* merupakan metode seleksi fitur yang tidak mempertimbangkan akibat dari seleksi fitur terhadap kinerja algoritma *machine learning*. Metode ini menggunakan pengukuran statistik untuk menilai bobot setiap fitur. Seluruh fitur selanjutnya akan diurutkan berdasarkan nilai atau ranking yang diperoleh, untuk selanjutnya dipertimbangkan akan dipertahankan atau dihilangkan dari dataset.

Metode seleksi fitur ini memiliki kelebihan lebih cepat dari sisi komputasi jika dibandingkan dengan metode *wrapper*, serta dapat meminimalisir terjadinya kasus *over fitting*. Hanya saja memiliki kelemahan relatif kurang akurat dalam memetakan atau mempertimbangkan hubungan antar fitur. Contoh metode yang bisa digunakan yakni *correlation coefficient*, *chi-square test*, ANOVA, *information gain*.

2. *Wrapper Methods*

Wrapper merupakan metode seleksi fitur yang menggunakan algoritma *machine learning* sebagai bagian dari fungsi evaluasi. Algoritma ini digunakan sebagai semacam "*black box*" saat proses pemilihan fitur. Metode ini melakukan pendekatan *greedy search* dengan mengevaluasi semua kombinasi fitur terhadap kriteria evaluasi. Contoh metode *wrapper* yakni *recursive feature elimination* (RFE), *sequential feature selection* (SFS), *genetic algorithm* (GA).

Metode *wrapper* secara umum dapat dibagi menjadi tiga kategori, yakni *forward selection*, *backward elimination*, dan *recursive feature elimination*.

- ***Forward Selection***: metode seleksi dengan proses iteratif, dengan fitur kosong sebagai inisasi awal pada model. Setiap iterasi, fitur yang memiliki pengaruh paling signifikan ditambahkan ke dalam koleksi fitur. Kemudian dilanjutkan dengan penambahan variabel baru yang tidak meningkatkan kinerja model.

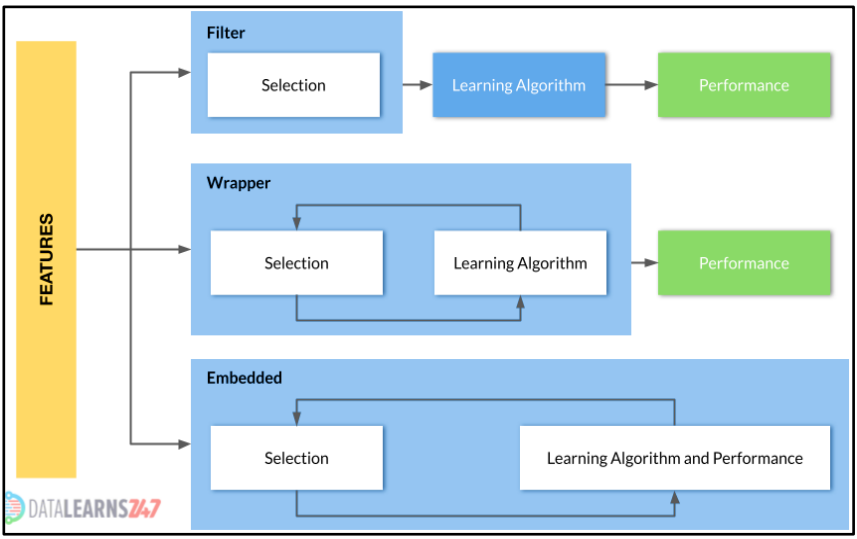
- ***Backward Elimination***: kebalikan dari metode *forward selection*, metode ini diinisiasi awal dengan semua fitur yang ada. Selanjutnya, setiap tahap iterasi dilakukan pengurangan fitur yang dianggap tidak relevan dengan model, hingga seluruh fitur yang tersisa itulah yang dianggap fitur paling optimal.
- ***Recursive Feature Elimination***: metode ini merupakan optimasi dari algoritma *greedy* yang bertujuan untuk menentukan subset fitur yang memiliki performa baik. Setiap iterasinya, metode ini membentuk model yang dimulai dari fitur paling kiri hingga semua fitur diperiksa. Metode ini tidak akan memperhatikan fitur yang memiliki pengaruh baik ataupun buruk dalam setiap iterasinya. Namun metode ini akan memberikan peringkat fitur berdasarkan urutan eliminasinya.

3. Embedded Methods

Metode *embedded* merupakan metode yang menggabungkan keunggulan yang dimiliki metode *filter* dan *wrapper*. Secara kinerja, pemilihan fitur dilakukan dalam algoritma *machine learning* itu sendiri. Namun tidak semua algoritma memiliki fungsi ini secara intrinsik.

Metode *embedded* ini melakukan seleksi fitur selama proses *training* terjadi, sekaligus mencari fitur yang paling relevan untuk model tersebut. Hal inilah yang membuat metode *embedded* lebih efisien dibandingkan dengan metode *wrapper* yang mencoba semua kombinasi fitur, selain itu tetap lebih efektif daripada metode *filter* yang mengevaluasi fitur secara independen.

Gambar 10.1 merupakan ilustrasi perbandingan antara metode *filter*, *wrapper*, dan *embedded*.



Gambar 10.1: Perbandingan Metode Seleksi Fitur Berbasis *Filter*, *Wrapper*, dan *Embedded*

Sumber: Website https://www.datalearns247.com/mengenal-feature-selection-dalam-machine-learning-69

D. RANGKUMAN

Secara umum, ekstraksi dan seleksi fitur merupakan dua aspek penting dalam *machine learning*. Ekstraksi fitur merupakan proses penggalian informasi penting yang terkandung di dalam suatu data. Informasi penting ini dikenal dengan istilah fitur. Fitur-fitur ini digunakan untuk membentuk model yang dapat digunakan untuk tahap berikutnya, seperti memprediksi ataupun mengklasifikasikan data baru. Penerapan ekstraksi fitur ini kebanyakan dimanfaatkan untuk pengolahan teks (bidang NLP) serta pengolahan citra (bidang DIP).

Seleksi fitur merujuk pada tahap pra-pengolahan untuk mereduksi fitur, khususnya fitur-fitur yang dinilai kurang relevan terhadap model yang dibentuk. Secara umum terdapat tiga jenis seleksi

fitur, yakni *filter*, *wrapper*, dan *embedded selector*. Metode *filter* mengevaluasi setiap fitur secara bebas dari *classifier*, kemudian memberikan peringkat pada fitur tersebut. Metode *wrapper* memiliki ciri khas yakni memerlukan bantuan sebuah algoritma *machine learning*, dan menggunakannya sebagai evaluator dalam menentukan fitur-fitur yang berpengaruh signifikan. Kemudian untuk metode *embedded selector* mengacu pada hibridisasi metode *filter* dan *wrapper*.

E. TES FORMATIF

1. Manakah metode yang bukan termasuk dalam ekstraksi fitur untuk domain NLP?
   a. BoW
   b. *Invariant Moment*
   c. N-*grams*
   d. TF-IDF
   e. *FastText*
2. Manakah yang termasuk metode *filter* untuk proses seleksi fitur?
   a. Genetic Algorithms
   b. *Correlation-based Feature Selection*
   c. *Recursive Feature Elimination*
   d. *Sequential Feature Selection*
   e. Forward and Backward Selection

F. LATIHAN

Berikan penjelasan kapan seseorang sebaiknya menggunakan metode *filter*, *wrapper*, dan *embedding* untuk tahap seleksi fitur, sertakan contoh kasusnya untuk masing-masing metode.

DAFTAR PUSTAKA

Prahendratno, A., Mahendra, G. S., Zebua, R. S. Y., Zaebabe, H., Sepriano, Handika, I. P. S., Rahayu, P. W., & Sudipa, I. G. I. (2023). Business Intelegent (Pengantar Business Intelligence dalam Bisnis). PT. Sonpedia Publishing Indonesia.

Putra, R. F., Zebua, R. S. Y., Budiman, Rahayu, P. W., Bangsa, Mhd. T. A., Zulfadhilah, M., Choirina, P., Wahyudi, F., & Andiyan, A. (2023). Data Mining : Algoritma dan Penerapan. Dalam Efitra & Sepriano (Ed.), PT. Sonpedia Publishing Indonesia.

Saluky, Pamungkas, S. P. A., Saputra, P. S., Nurhayati, S., Putra, A. I., Ardiada, I. M. D., Mandias, G. F., Cokrowibowo, S., Ismail, Pratama, P. A., Rahayu, P. W., & Mahardika, F. (2023). Ilmu Komputer. PT. Literasi Nusantara Abadi Grup.

Afifuddin, A., & Hakim, L. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4. 5. Jurnal Krisnadana, 3(1), 25–33.

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013. https://doi.org/10.1109/ICMIRA.2013.45

Anjelita, M., Windarto, A. P., Wanto, A., & Sudahri, I. (2020). Pengembangan Datamining Klastering Pada Kasus Pencemaran Lingkungan Hidup. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) , 309–313.

Arifin, A., Djumat, I., Nicolas, D. G., Syam, A. S. M., & Saputra, N. (2023). Metaverse in Education; Innovation Strategy,

Learning Acceleration, and Optimization. Journal of Namibian Studies: History Politics Culture, 34, 1470–1485.

Arifin, A., Haryanto, H., Basri, M., & Ansari, A. (2018). Multicultural Approach in Developing Instructional Learning Material at Indonesian Senior High School. PROCEEDINGS OF THE 65th TEFLIN INTERNATIONAL CONFERENCE, 65(02).

Arifin, A., Prajayanti, E., Hasby, M., Taufik, M., & Anggarini, D. T. (2023). The Unex Application as An English Interactive Learning Media: A Feasibility Study. Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran, 9(2).

Asana, I. M. D. P., Sudipa, I. G. I., Mayun, A. A. T. W., Meinarni, N. P. S., & Waas, D. V. (2022). Aplikasi Data Mining Asosiasi Barang Menggunakan Algoritma Apriori-TID. INFORMAL: Informatics Journal, 7(1), 38–45.

Damanik, S. F., Wanto, A., & Gunawan, I. (2022). Penerapan Algoritma Decision Tree C4. 5 untuk Klasifikasi Tingkat Kesejahteraan Keluarga pada Desa Tiga Dolok. Jurnal Krisnadana, 1(2), 21–32.

Dewantara, R., & Giovanni, J. (2023). Analisis Peramalan Item Penjualan dalam Optimalisasi Stok Menggunakan Metode Least Square. Jurnal Krisnadana, 3(1), 59–66.

Dewi, I. G. A. M. P., Parwita, W. G. S., & Setiawan, I. M. D. (2021). Algoritma Decision Tree untuk Klasifikasi Calon Debitur LPD Desa Adat Anggungan. Jurnal Krisnadana, 1(1), 23–36.

Hariyono, R. C. S., Kuntarto, G. P., Sudipa, I. G. I., Juliandy, C., Kharisma, L. P. I., Hartati, S., Aryuni, M., Lestari, W. S., Saragih, Y. M., & Ulina, M. (2023). BUKU AJAR

PENGANTAR BASIS DATA. PT. Sonpedia Publishing Indonesia.

Hayadi, B. H., Sudipa, I. G. I., & Windarto, A. P. (2021). Model Peramalan Artificial Neural Network pada Peserta KB Aktif Jalur Pemerintahan menggunakan Artificial Neural Network Back-Propagation. MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer, 21(1), 11–20.

Kwintiana, B., Nengsih, T. A., Baradja, A., Harto, B., Sudipa, I. G. I., Handika, I. P. S., Adhicandra, I., & Gugat, R. M. D. (2023). DATA SCIENCE FOR BUSINESS: Pengantar & Penerapan Berbagai Sektor. PT. Sonpedia Publishing Indonesia.

Mahendra, G. S., Hariyono, R. C. S., Purnawati, N. W., Hatta, H. R., Sudipa, I. G. I., Hamali, S., Sarjono, H., & Meilani, B. D. (2023). BUKU AJAR SISTEM PENDUKUNG KEPUTUSAN. PT. Sonpedia Publishing Indonesia.

Mahendra, G. S., Wardoyo, R., Pasrun, Y. P., Sudipa, I. G. I., Putra, I. N. T. A., Wiguna, I. K. A. G., Aristamy, I. G. A. A. M., Kharisma, L. P. I., Sutoyo, M. N., & Sarasvananda, I. B. G. (2023). IMPLEMENTASI SISTEM PENDUKUNG KEPUTUSAN: Teori & Studi Kasus. PT. Sonpedia Publishing Indonesia.

Muhammad Wali, S. T., Efitra, S., Kom, M., Sudipa, I. G. I., Kom, S., Heryani, A., Sos, S., Hendriyani, C., Rakhmadi Rahman, S. T., & Kom, M. (2023). Penerapan & Implementasi Big Data di Berbagai Sektor (Pembangunan Berkelanjutan Era Industri 4.0 dan Society 5.0). PT. Sonpedia Publishing Indonesia.

Prahendratno, A., Mahendra, G. S., Zebua, R. S. Y., Tahir, R., Sepriano, S., Handika, I. P. S., Rahayu, P. W., Sudipa, I. G. I., & Efitra, E. (2023). BUSINESS INTELEGENT: Pengantar

Business Intelligence dalam Bisnis. PT. Sonpedia Publishing Indonesia.

Putri, N. P. M. E., Sudipa, I. G. I., Wiguna, I. K. A. G., Sarasvananda, I. B. G., & Sunarya, I. W. (2024). Decision Making Model for Temple Revitalization in Bali Using Fuzzy-SMARTER Combination Method. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 9(1), 61–74.

Putri, R. M. A., Parwita, W. G. S., Handika, I. P. S., Sudipa, I. G. I., & Santika, P. P. (2024). Evaluation of Accounting Information System Using Usability Testing Method and System Usability Scale. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 9(1), 32–43.

Radhitya, M. L., & Sudipa, G. I. (2020). PENDEKATAN Z-SCORE DAN FUZZY DALAM PENGUJIAN AKURASI PERAMALAN CURAH HUJAN. SINTECH (Science and Information Technology) Journal, 3(2), 149–156.

Risqi Ananda, M., Sandra, N., Fadhila, E., Rahma, A., & Nurbaiti, N. (2023). Data Mining dalam Perusahaan PT Indofood Lubuk Pakam. Comit: Communication, Information and Technology Journal, 2(1), 108–119. https://doi.org/10.47467/comit.v2i1.124

Rony, Z. T., Lestari, T. S., Ismaniah, Yasin, M., & Lubis, F. M. (2023). The complexity of leadership competence in universities in the 21st century. Cogent Social Sciences, 9(2), 2276986.

Safitri, N., & Bella, C. (2022). PENERAPAN DATA MINING UNTUK ANALISIS POLA PEMBELIAN PELANGGAN (STUDI KASUS : TOKO DIENGVA BANDAR JAYA). 2(1), 1–8.

Saputra, I. K. D. A., Satwika, I. P., & Utami, N. W. (2022). Analisis Transaksi Penjualan Barang Menggunakan Metode Apriori pada UD. Ayu Tirta Manis. Jurnal Krisnadana, 1(2), 11–20.

Simanjuntak, D. S. M., Gunawan, I., Sumarno, S., Poningsih, P., & Sari, I. P. (2023). Penerapan Algoritma K-Medoids Untuk Pengelompokkan Pengangguran Umur 25 tahun Keatas Di Sumatera Utara. Jurnal Krisnadana, 2(2), 289–309.

Sudipa, I. G. I., Aditama, P. W., & Yanti, C. P. (2022). Evaluation of Lontar Prasi Bali Application based on Augmented Reality Using User Experience Questionnaire. East Asian Journal of Multidisciplinary Research, 1(9), 1845–1854.

Sudipa, I. G. I., Asana, I. M. D. P., Atmaja, K. J., Santika, P. P., & Setiawan, D. (2023). Analisis Data Kepuasan Pengguna Layanan E-Wallet Gopay Menggunakan Metode Naïve Bayes Classifier Algorithm. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer Dan Manajemen), 4(3), 726–735.

Sudipa, I. G. I., Riana, R., Putra, I. N. T. A., Yanti, C. P., & Aristana, M. D. W. (2023). Trend Forecasting of the Top 3 Indonesian Bank Stocks Using the ARIMA Method. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 8(3), 1883–1893.

Sudipa, I. G. I., Sarasvananda, I. B. G., Prayitno, H., Putra, I. N. T. A., Darmawan, R., & WP, D. A. (2023). Teknik Visualisasi Data. PT. Sonpedia Publishing Indonesia.

Sudipa, I. G. I., Wardoyo, R., Hatta, H. R., Sagena, U., Gunawan, I. M. A. O., Zahro, H. Z., & Adhicandra, I. (2023). MULTI CRITERIA DECISION MAKING: Teori & Penerapan Metode Pengambilan Keputusan dengan MCDM. PT. Sonpedia Publishing Indonesia.

Suryawan, I. G. T., Arimbawa, I. K. S., & Sudipa, I. G. I. (2023). Implementation of Naive Bayes Method for Granting Fisherman Business Credit. Jurnal Info Sains: Informatika Dan Sains, 13(01), 24–32.

Wahyuddin, S., Sudipa, I. G. I., Putra, T. A. E., Wahidin, A. J., Syukrilla, W. A., Wardhani, A. K., Heryana, N., Indriyani, T., & Santoso, L. W. (2023). Data Mining. Global Eksekutif Teknologi.

Wiguna, I. K. A. G., Utami, N. L. P. A. C., Parwita, W. G. S., Udayana, I. P. A. E. D., & Sudipa, I. G. I. (2023). Rainfall Forecasting Using the Holt-Winters Exponential Smoothing Method. Jurnal Info Sains: Informatika Dan Sains, 13(01), 15–23.

Card, S. K., & Mackinlay, J. 1997. The structure of the information visualization design space. Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium, 92–99.

Few, S. 2006. Multivariate analysis using parallel coordinates. Perceptual Edge, 1–9.

GRINSTEIN, G. G. 2002. PATRICK E. HOFFMAN. Information Visualization in Data Mining and Knowledge Discovery, 47.

Han, J., Kamber, M., & Mining, D. 2006. Concepts and techniques. Morgan Kaufmann, 340, 93205–94104.

Heinrich, J., & Weiskopf, D. 2013. State of the Art of Parallel Coordinates. Eurographics (State of the Art Reports), 95–116.

Hoffman, P., & Grinstein, G. 1997. Visualizations for high dimensional data mining-table visualizations. Citeseer.

Im, J.-F., McGuffin, M. J., & Leung, R. 2013. GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2606–2614.

Jeovano, J. 2020. 2D Data Visualization Tools Menggunakan Flask dan AngularJS. INSYST: Journal of Intelligent System and Computation, 2(2), 91–97.

Kopanakis, I. 2003. Visual Data Mining Models for Enhancing the Knowledge Extraction from Data Mining Outcomes. The University of Manchester (United Kingdom).

Mulyana, S., Winarko, E., Studi, P., Komputer, I., Matematika, F., Ilmu, D., & Alam, P. 2009. TEKNIK VISUALISASI DALAM DATA MINING. Seminar Nasional Informatika, 23–2009.

Nocke, T., Schlechtweg, S., & Schumann, H. 2005. Icon-based visualization using mosaic metaphors. Ninth International Conference on Information Visualisation (IV'05), 103–109.

Nurirwan Saputra, P. T. I. U. P. Y. (n.d.). Pengenalan Data Mining.

Pleuss, A., Rabiser, R., & Botterweck, G. 2011. Visualization techniques for application in interactive product configuration. Proceedings of the 15th International Software Product Line Conference, Volume 2, 1–8.

Purwati, N., Pedliyansah, Y., Kurniawan, H., Karnila, S., & Herwanto, R. 2023. Komparasi Metode Apriori dan FP-Growth Data Mining Untuk Mengetahui Pola Penjualan. Jurnal Informatika: Jurnal Pengembangan IT, 8(2), 155–161.

Sastry, S. H., Babu, P., & Prasada, M. S. 2013. Implementation of CRISP methodology for ERP systems. ArXiv Preprint ArXiv:1312.2065.

Schnorr, L. M., Navaux, P. O. A., & Huard, G. n.d.-a. Visualization Techniques for Grid Environments: a Survey and Discussion.

Schnorr, L. M., Navaux, P. O. A., & Huard, G. n.d.-b. Visualization Techniques for Grid Environments: a Survey and Discussion.

Scrucca, L., & Raftery, A. 2015. Improved initialisation of model-based clustering using Gaussian hierarchical partitions. Advances in Data Analysis and Classification, 9, 447–460. https://doi.org/10.1007/s11634-015-0220-z

Siirtola, H., & Räihä, K.-J. 2006. Interacting with parallel coordinates. Interacting with Computers, 18(6), 1278–1309. https://doi.org/https://doi.org/10.1016/j.intcom.2006.03.006

Wahyuddin, S., Sudipa, I. G. I., Putra, T. A. E., Wahidin, A. J., Syukrilla, W. A., Wardhani, A. K., Heryana, N., Indriyani, T., & Santoso, L. W. 2023. Data Mining. Global Eksekutif Teknologi.

Wang, H., Ni, Y., Sun, L., Chen, Y., Xu, T., Chen, X., Su, W., & Zhou, Z. 2021. Hierarchical visualization of geographical areal data with spatial attribute association. Visual Informatics, 5(3), 82–91.

Wang, W. B., Huang, M. L., Nguyen, Q. V., Huang, W., Zhang, K., & Huang, T.-H. 2016. Enabling decision trend analysis with interactive scatter plot matrices visualization. Journal of Visual Languages & Computing, 33, 13–23.

Ardhiansyah, F., & Ratih, S. W. W. (2020). Data Mining Berdasarkan Analisis Runtun Waktu untuk Pembuatan Model Prediksi Pasien Terjangkit COVID-19 dan Pasien

Meninggal karena COVID-19 di Indonesia. Jakarta: Universitas Gunadarma.

Aziz, A. R., et al. (2021). Pengaruh Transformasi Data Pada Metode Learning Vector Quantization Terhadap Akurasi Klasifikasi Diagnosis Penyakit Jantung. Jurnal Gaussian, 10(1), 21-30.

Fadilah, Z. R., & Wijayanto, A. W. (2023). Perbandingan Metode Klasterisasi Data Bertipe Campuran: One-Hot-Encoding, Gower Distance, dan K-Prototype Berdasarkan Akurasi (Studi Kasus: Chronic Kidney Disease Dataset). Journal of Applied Informatics and Computing, 7(1), 63-73.

Irnawan, F. D., et al. (2021). Metode Imputasi pada Data Debit Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta. Jurnal Nasional Teknik Elektro dan Teknologi Informasi| Vol, 10(4).

Lutfi, M., et al. (2023). Penanganan Data Tidak Seimbang Menggunakan Hybrid Method Resampling Pada Algoritma Naive Bayes Untuk Software Defect Prediction. INFORMAL: Informatics Journal, 8(2), 119-127.

Mintoro, S., & Afandi, A. (2021). implementasi algoritma k-means dan algoritma apriori optimasi kinerja ecu (study kasus mobil avanza dan xenia). Jurnal Informasi dan Komputer, 9(2), 81-88.

Nazatushima, M. A. (2003). Pembersihan data untuk gudang dan perlombongan data (Cleaning data for warehousing and mining)/Nazatushima Mohd Arshad (Doctoral dissertation, University of Malaya).

Nisa, A., et al. (2019). Analisis Sentimen Menggunakan Naive Bayes Classifier dengan Chi-Square Feature Selection

Terhadap Penyedia Layanan Telekomunikasi. eProceedings of Engineering, 6(2).

Rizal, S. (2019). Development of Big Data Analytics Model. ITEJ (Information Technology Engineering Journals), 4(1), 14-25.

Setiawan, Z., et al. (2023). BUKU AJAR DATA MINING. PT. Sonpedia Publishing Indonesia.

Syahyadi, A. I., et al. (2023). integrasi data akademik pergururuan tinggi dengan pangkalan data dikti menggunakan sistem integrasi feeder terbaru (SIFEEKA). Jurnal INSTEK (Informatika Sains dan Teknologi), 8(1), 112-121.

Wahyuddin, S., et al. (2023). Data Mining. Global Eksekutif Teknologi.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-54.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6), 866-883.

Wulandari, Novi (2021). Data Mining : Teknik Data Mining. [Scribdlink]. Diakses pada 06 Januari 2024 dari https://www.scribd.com/presentation/505383662/Data-Mining-5-Teknik-Data-Mining.

Ridwan, A., Andono, P., & Supriyanto, C. (2018). Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma Naive Bayes Classification Adaboost. Jurnal Teknologi Informasi. http://research.pps.dinus.ac.id/index.php/Cyberku/article/download/76/72

Connoly, Thomas M., Begg, Carolyn E., and Strachan, Anne D. 1999. Database System. A practical pproach to Design, Implementation, and Management, Addison Wesley Company.

Ian H. Witten, Frank Eibe, Mark A. Hall. 2011. Data mining: Practical Machine Learning Tools and Techniques 3rd Edition, Elsevier.

Jiawei Han and Micheline Kamber. 2012. Data Mining: Concepts and Techniques Third Edition, Elsevier

Kristanto, Andri. 2004. Kecerdasan Buatan. Yogyakarta : Graha Ilmu.

Kusumadewi, Sri. 2003. Artificial Intelligence (Teknik dan Aplikasinya). Yogyakarta : Graha Ilmu.

Kuswadi, Son, 2007. Kendali Cerdas (Teori dan Aplikasi Praktisnya). Yogyakarta : Penerbit ANDI.

Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin. 2015. Introduction to Data Mining 2nd edition, Pearson Education, Inc

Wahono, Romi Satrio. 2020. Data Mining. https://romisatriawahono.net/dm/ diakses pada tanggal 15 Desember 2023

Berkhin, P. 2006. A Survey of Clustering Data Mining Techniques. In Grouping Multidimensional Data. Springer.

Buulolo, E. 2020, Data Mining Untuk Perguruan Tinggi. Deepublish Yogyakarta.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-1996 Proceedings.

García, S., Luengo, J., & Herrera, F. 2015. Data Preprocessing in Data Mining. Springer.

Han, J., Pei, J., & Kamber, M. 2011. Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hubert, L., & Arabie, P. 1985. "Comparing Partitions". Journal of Classification.

Jain, A. K., Murty, M. N., & Flynn, P. J. 1999. Data Clustering: A Review. ACM Computing Surveys.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. 2006. Data Preprocessing for Supervised Learning. International Journal of Computer Science.

Rismawan, T dan Kusumadewi, S. 2008. Aplikasi K-Means Untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass Index (BMI) dan Ukuran Kerangka. Seminar Nasional Aplikasi Teknologi Informasi.

Rousseeuw, P.J. 1987. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Journal of Computational and Applied Mathematics.

Swastika, R., dkk. 2023, Implementasi Data Mining (Clastering, Assocation, Prediction, Estimation, Classification). Adanu Abimata Jawa Barat.

Xu, R., & Wunsch, D. 2005. Clustering. IEEE Press Series on Computational Intelligence.

A. Ardianto dan D. Fitrianah. (2019). Penerapan Algoritma FP-Growth Rekomendasi Trend Penjualan ATK Pada CV. Fajar Sukses Abadi, J. Tek. Inform., Vol. 9, No. 1, April 2019.

Aryasa, K. (2015). Big Data: Challenges and Opportunities. In Workshop Big Data Puslitbang Aptika dan IKP, tanggal 19 Mei 2015. Puslitbang Aptika dan IKP.

Bramer, M. (2013). Principles of Data Mining, second edition, London: Springer.

Buulolo, E. (2020). Implementasi Algoritma Apriori pada Sistem Persediaan Obat. Medan: STIMIK Budi Darma

D. Listriani, A. H. Setyaningrum, dan F. E. M. A .(2016). Penerapan Metode Asosiasi Menggunakan Algoritma Apriori Pada Aplikasi Analisa Pola Belanja Konsumen (Studi Kasus Toko Buku Gramedia Bintaro), J. Tek. Inform., Vol. 9, No. 2, pp. 120–127.

F. Fatihatul, A. Setiawan, and R. Rosadi. (2011). Asosiasi Data Mining Menggunakan Algoritma FP Growth Untuk Market Basket Analysis, pp. 1–8.

F. Rumaisa. (2012). Penentuan Association rule Pada Pemilihan Program Studi Kasus Pada Universitas Widyatama Bandung, Semin. Nas. Apl. Teknol. Inf., Vol. 1, No. Snati, pp. 15–16

Hastie, T., Tibshirani, R., Friedman, J. (2013). The Elements of Statistical Learning, New York : Springer,

F.A. Hermawati. (2013). Data Mining.Yogyakarta: ANDI.

I. Ali. (2015). Penerapan Data Mining dengan Algoritma FP-Growth untuk Mendukung Strategi Promosi Pendidikan . Jurnal Saintikom, 14(3).

L. Y. Dewi. (2015). Penerapan Data Mining Menggunakan Algoritma FP-Tree dan FP-Growth

E. T. Luthfi. (2009). Algoritma Data Mining. Yogyakarta: ANDI.

A. B. Mark and D. Laney. (2012). The importance of 'Big Data': A Definition. Gartner,

A. P. Narendra. (2015). Data Besar, Data Analisis, dan Pengembangan Kompetensi Pustakawan. Fakultas Teknologi Informasi UKSW Salatiga.

Rainer, R., Kelly, and C. G. Cegielski. (2009). Introduction to Information Systems. John Wiley & Sons (Asia) Pte Ltd.

Rathi, R. dan Lohiya, S. (2014). Big Data and Hadoop. International Journal of Advanced Research in Computer Science and Technology (IJARCST 2014), 2(2), pp. 214 – 217.

S. Sidhu. (2014). FP Growth Algorithm Implementation, Vol. 93, No. 8, pp. 6–10, 2014.

E. R. E. Sirait. (2016). Implementasi Teknomogi Big Data di Lembaga Pemerintahan Indonesia. Jurnal Penelitian Pos dan Informatika

N.P., Tan, M., Steinbach, V. Kumar. (2006). Introduction to Data Mining, Essex : Pearson International.

Syafira Salsabila. 2022. Modul Data Mining Text Mining. Univerditas Esa Unggul.

Sancha Diwandari, Adityo Permana Wibowo. 2022. Modul Praktikum Pemrosesan Teks. Yogyakarta : Universitas Teknologi Yogyakarta.

Sitti Harlina, dkk. 2022. Klasifikasi Sentimen Tweet Mengenai Covid-19 Pada Twitter di Indonesia Dengan Metode Vector Space Model, CogiTo: Universitas Klabat Manado.

Mohammad Reza Faisal, Dwi KartiniHariyanto,2022. Belajar Data Science Text Mining untuk Pemula I .Penerbit : Scripta Cendekia.

Muhammad Zidane Zukhrifa. Text Mining. https://bisa.ai/portofolio/detail/MzE4Nw, diakses tanggal 30 Desember 2023.

kolonginfo.com/pengertian-dan-manfaat-text-mining/ diakses tanggal 29 desember 2023.

NoviaIdy, 2020. Expert System Of Text Mining To Analyze Student Interaction In FTKI . OnlineLectures: Mantik

Novia Lestari, 2023. Implementation of Tex Mining and PATTERN Discovery dengan Algoritma Naïve bayes Untuk Klasifikasi Dokumen Teks. Jurnal: Teknologi Informasi dan Komunikasi

Edio da Costa.2018. DText Mining For Pest Disease Identification on Rice Farming With Interactive Text Message : Sepuluh November Institute Of Technology.

Jaka Harjanta, 2015. Preprocessing Text untuk Meminalisir Kata yang Tidak Berarti Dalam proses Text mining. Jurnal Informatika Upgris .

Bird, S., Klein, E., and Loper, E. 2009. Natural Language Processing with Python. California: O'Reilly Media.

Chen, Q., Yang, X., and Zhao, J. 2005. Robust Image Watermaking with Zernike Moments. on Proc. of the IEEE CCECE/CCGEI, pp. 1340-1343.

Kadir, A., dan Susanto, A. 2012. Teori dan Aplikasi Pengolahan Citra. Yogyakarta: CV. Andi Offset.

Mengenal Feature Selection dalam Machine Learning. 2021. https://www.datalearns247.com/mengenal-feature-selection-dalam-machine-learning-69

Mingqiang, Y., Kidiyo, K., and Joseph, R. 2008. A Survey of Shape Feature Extraction Techniques. on Pattern Recognition Technique, Technology and Application, pp. 43-90.

Putra, Darma. 2010. Pengolahan Citra Digital. Yogyakarta: CV. Andi Offset.

Theoridis, S. and Koutroumbas, K. 2006. Pattern Recognition. 3rd Edition. San Diego: Academic Press.

TENTANG PENULIS

Prastyadi Wibawa Rahayu, S.Kom., M.Kom.

Seorang Penulis dan Dosen Prodi Teknik Informatika Fakultas Teknologi dan Informatika Universitas Dhyana Pura Bali. Lahir di tahun 1994 Bali. Penulis merupakan anak kedua dari dua bersaudara dari pasangan bapak I Made Ambara Wijaya dan Wening Sejati. ia menamatkan Pendidikan program Sarjana (S1) di STMIK STIKOM Bali prodi Sistem Informasi dan menyelesaikan program Pasca Sarjana (S2) di Universitas Pendidikan Ganesha prodi Ilmu Komputer konsentrasi di bidang Sistem Informasi.

I Gede Iwan Sudipa, S.Kom., M.Cs.

Penulis lahir di Singaraja, Bali. Penulis menyelesaikan pendidikan Strata I pada STMIK AKAKOM Yogyakarta dan Pendidikan Magister (S2) bidang Ilmu Komputer di Universitas Gadjah Mada Yogyakarta. Penulis menjadi Dosen tetap program studi Teknik Informatika pada Institut Bisnis dan Teknologi Indonesia (INSTIKI). Penulis juga aktif dalam menulis Karya Tulis Ilmiah dan Buku yang telah ditulis yaitu : Basis data : teori dan perancangan, Logika informatika, Green Technology : Penerapan Teknologi Ramah Lingkungan Berbagai Bidang, Inovasi & tren layanan digital berbagai sektor : optimalisasi dan otomatisasi digital untuk dunia kerja & bisnis, Metode penelitian bidang ilmu informatika : teori & referensi berbasis studi kasus, Teknik

Visualisasi Data, MULTI CRITERIA DECISION MAKING : Teori & Penerapan Metode Pengambilan Keputusan dengan MCDM, Sistem pendukung keputusan, Penerapan Sistem Informasi di Berbagai Bidang, TEKNOLOGI INFORMASI & SDGs (Peranan Teknologi Informasi di Berbagai Bidang Dalam Mendukung Sustainable Development Goals), DATA SCIENCE FOR BUSINESS : Pengantar & Penerapan Berbagai Sektor, Pemanfaatan teknologi informasi di berbagai sektor pada masa society 5.0, TEKNIK PENULISAN KARYA ILMIAH : Cara membuat Karya Ilmiah yang baik dan benar, Sistem Informasi: Pengantar Komprehensif, Rekayasa perangkat lunak, Metode penelitian berbagai bidang keilmuan : panduan & referensi, IMPLEMENTASI SISTEM PENDUKUNG KEPUTUSAN : Teori & Studi Kasus, BUSINESS INTELEGENT : Pengantar Business Intelligence dalam Bisnis, Penerapan decision support system (DSS) dalam berbagai bidang, FENOMENA ARTIFICIAL INTELLIGENCE (AI), Buku Ajar Sistem Pendukung Keputusan

Email : *iwansudipa@instiki.ac.id*

Suryani, S.Kom., M.T.

Lahir di Maros, pada 4 januari 1987. Tercatat sebagai lulusan Universitas Dipa Makassar dan Universitas Hasanuddin Makassar. Wanita yang kerap disapa Surya ini adalah anak pertama dari pasangan Zainuddin Dg. Ajang (ayah) dan Nurjannah Dg. Ngai (ibu). Suryani berprofesi sebagai salah satu dosen Universitas Dipa Makassar mulai 1 Maret 2016 hingga saat ini. Isteri dari Sunarya Suardy, Amd.Kom., S.E. ini bermula mengecap pendidikan di perguruan tinggi program Diploma Tiga (D3) Jurusan Manajemen Informatika di Universitas Dipa Makassar ex STMIK Dipanegara pada tahun 2008, kemudian mendapatkan

beasiswa wisudawan terbaik utama melanjutkan study program Strata Satu (S1) Jurusan Teknik Informatika di kampus yang sama, dan selesai dengan predikat Terbaik Utama atau Cum Laude dengan IPK Tertinggi yaitu 4.00. Dengan mendapatkan Beasiswa Unggulan (BU) Calon Dosen dari LLDIKTI, Ia menyelesaikan study program Pasca Sarjana atau S2 Program Study Teknik Elektro Jurusan Teknik Informatika di Universitas Hasanuddin makassar. Ibu dari anak perempuan yang bernama Shaqueena Nur Panrita ini, selain hobby memasak, mengunjungi tempat wisata alam, menulis buku, ia juga senang melakukan riset di bidang artificial intelligence, expert system, decision support systems dan lain sebagainya.

Email Penulis: suryani187@undipa.ac.id

Scopus.ID: 57212150032; WoS.ID: AFS-0346-2022;

ORIC.ID: https://orcid.org/ 0000-0001-6540-2607 ;

googlescholar.ID: 1TJzgREAAAAJ; SintaID: 6100069;

Arie Surachman, M.Kom.

Lahir di Jakarta, pada 01 Januari 1984. Menyelesaikan S1 Sistem Informasi di STMIK Muhammad Husni Thamrin Jakarta dan S2 Magister Komputer di STMIK Eresha Jakarta. Riwayat Pengalaman menjadi Manager Marketing, Mutu, HRM & General Affairs di PT. Mutumed Prima Services, Pengalaman Mengajar di Prodi Kebidanan Universitas MH Thamrin, Fakultas Teknologi Informasi Universitas Respati Indonesia, dan STMIK Islam International Jakarta, dan saat ini sebagai Penulis & Editor Buku Bersertifikat BNSP, dan merupakan Dosen Tetap di Program Studi Teknik Informatika, Universitas Indraprasta PGRI Jakarta, NIDN : 0301018409, Email: ariesurachmanmkom@gmail.com. Buku

yang telah ditulis dan terbit berjudul di antaranya : *Kesehatan Reproduksi Remaja, Sistem Informasi Surveilans, Manajemen Pendidikan.* Buku yang telah disunting sebagai Editor dan terbit berjudul di antaranya : Pengantar Jaringan Komputer.

Achmad Ridwan, S.Kom., M.Kom.
Seorang penulis dan dosen tetap Prodi Sistem Informasi Fakultas Sains, Teknologi dan Matematika Universitas Muhammadiyah Kudus. Lahir di desa Jekulo, 28 Juni 1984 Jawa tengah.. Pendidikan program Serjana (S1) Universitas Muria Kudus Prodi Sistem Informasi dan menyelesaikan program Pasca Sarjana (S2) di Universitas Dian Nuswantoro "Udinus" Semarang Program Studi Teknik Informatika konsentrasi di bidang Sistem Informasi. Menjabat sebagai Ketua Program Studi Sistem Informasi Universitas Muhammadiyah Kudus. Buku yang pernah ditulis dan terbit diantaranya : Kompeten di dunia kerja dengan Ms.Word 2016. Daftar Paten yang pernah dibuat diantaranya : Karya cipta EC00202036349 10 Juni 2020 Learning Management System UMKU, Karya Cipta EC00202108184 25 Januari 2021 Aplikasi Antrian Wisuda Secara Drive Thru . Daftar Publikasi Artikel Jurnal Diantaranya : Developing Mathematical Exercise Software For Visually Impaired Students, The Comparison Of Accuracy Between Naïve Bayes Clasifier And C4. 5 Algorithm In Classifying Toddler Nutrition Status Based On Anthropometry Index, Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus, Penerapan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Pada Algoritma C4. 5 Dalam Menentukan Blogger Profesional, Text Mining Sentimen Analisis Pengguna Aplikasi Marketplace Tokopedia Berdasar Rating dan Komentar Pada Google Play Store, Application of the C4. 5

Algorithm for Early Cervical Cancer Classification, Classification of Heart Failure using the Naïve Bayes Algorithm dll.

I Gede Mahendra Darmawiguna, S.Kom., M.Sc..

Seorang penulis dan dosen tetap Prodi Sistem Informasi, Jurusan Teknik Informatika, Fakultas Teknik dan Kejuruan, Universitas Pendidikan Ganesha, Bali, Indonesia. Lahir di Kota Singaraja, 4 Januari 1985. Penulis menempuh pendidikan S1 *Information Technology* di President University, dan menyelesaikan S2 *Computer Science*, di University of Mysore, India. Buku yang telah ditulis dan terbit berjudul di antaranya: Data Mining Menemukan Pengetahuan dalam Data, Mari Belajar Pemrograman C#; 56 Kode Program Siap untuk Dicoba, dan Book Chapter The Augmented Reality Story Book Project: A Collection of Balinese Miths and Legends.

Ir. Muh. Nurtanzis Sutoyo, S.Kom., M.Cs.,IPP

Seorang penulis dan dosen tetap Prodi Sistem Informasi Fakultas Teknologi Informasi Universitas Sembilanbelas November Kolaka Sulawesi Tenggara. Lahir di Wuluhan, 21 Juni 1984 Jember Jawa Timur. Pendidikan program Serjana (S1) di STIMIK Bina Bangsa Kendari Program Studi Sistem Komputer tahun 2008 dan menyelesaikan program Pasca Sarjana (S2) di Universitas Gadjah Mada Yogyakarta Program Studi Ilmu Komputer tahun 2015, serta Profesi Insinyur (Ir) di Universitas Hasanuddin Makassar tahun 2022.

ISNANDAR SLAMET

Memperoleh gelar Sarjana Matematika dengan bidang konsentrasi Statistika dari Jurusan Matematika FMIPA UGM pada tahun 1989. Pada tahun 2000, memperoleh gelar Master dari *Department of Mathematics and Statistics, Curtin University, Western Australia.* Menerima gelar Doktor dari universitas yang sama yaitu *Curtin University* pada tahun 2013.

Menjadi Pengajar di Program Studi Matematika, Jurusan Teknik Sipil, Fakultas Teknik, Universitas Sebelas Maret sejak tahun 1992. Mengabdikan diri sebagai Dosen Jurusan Matematika sejak Fakultas MIPA UNS dibuka. Menjabat sebagai Sekretaris Jurusan Matematika pada rentang tahun 2007-2009. Menjabat sebagai Kepala Prodi Statistika sejak dibukanya Program Studi Statistika pada tahun 2014 sampai saat tahun 2023.

Aktif di dalam kegiatan pengajaran, penelitian, dan pengabdian kepada masyasrakat. Mengampu mata kuliah di tingkat sarjana yaitu Teori Antrean, Pengantar Data Mining, Komputasi Statistika, Metode Survai Sampel, Probabilitas, Statistika Matematika, dan Kuliah Magang Mahasiswa. Mengampu mata kuliah di tingkat magister yaitu adalah Teknik Analisis Dasar dan Statistik Terapan. Menjadi Staf Akademik *(Seasonal Staff)* untuk mata kuliah *Statistics Data Analysis* 201 dan *Statistics Data Analysis* 501 di *Department of Mathematics and Statistics, Curtin University, Western Australia* pada tahun 1998-1999 dan 2011-2012*.* Menjadi Pegawai Tidak Tetap di *School of Curriculum and Standards Authority, Western Australia* pada tahun 2011-2012. Memperoleh *scholarship* untuk program pengembangan kurikulum di *la Rochelle University*, la Rochelle, Perancis. Menulis buku berjudul **Dominasi Stokastik: Teori dan Aplikasi** dan **Pengantar Komputasi Statistik dengan R**.

Membimbing skripsi lebih dari 80 mahasiswa S1 dan membimbing tesis untuk lebih dari 40 mahasiswa S2. Memberikan motivasi

kepada para mahasiswa di beberapa kampus di dalam negeri maupun luar negeri.
Aktif dalam kegiatan penelitian dan pertemuan ilmiah di bidang Statistika. Menerima berbagai dana penelitian dengan beberapa skema. Menjadi *visiting research fellow* di *Curtin University,* Perth, Western Australia dan *visiting research scholar* di *Flinders University,* Newcastle, Australia. Menulis lebih dari 90 tulisan yang dipublikasikan di berbagai jurnal dan prosiding terindeks Scopus, dan banyak di jurnal nasional dan internasional bereputasi. Di bidang penelitian, menjadi ketua grup riset di Grup Riset Pemodelan Stokastik (2016 sampai dengan 2018), Statistika Industri (2018 sampai dengan 2019), dan Statistika dan Sains Data Bidang Industri dan Ekonomi (2019 sampai dengan sekarang), dan sebagai anggota di Grup Riset Statistika Terapan dan Inferensi (2016 sampai dengan 2018) dan Grup Riset Statistika Ekonomi, Bisnis, dan Finansial (2018 sampai dengan 2019).
Aktif dalam kegiatan pengabdian kepada masyarakat. Menjadi dosen pembimbing lapangan bagi mahasiswa yang mengikuti program Kuliah Kerja Nyata di berbagai kabupaten di Jawa Tengah. Menjadi dosen pembimbing bagi para mahasiswa yang mengikuti kegiatan MBKM.
Dapat dihubungi dengan E-mail: isnandarslamet@staff.uns.ac.id

Sitti Harlina, SE.,M.Kom.
Lahir di Limbung, Makassar Sulawesi Selatan pada 27 Maret 1975. Beliau menyelesaikan Pendidikan Ahli Madya pada Jurusan Administrasi Niaga Politeknik Universitas Hasanuddin pada tahun 1997, melanjutkan ke jenjang pendidikan Strata Satu di STIEM Bongaya Makassar jurusan Manajemen Keuangan dan Perbankan. Beliau melanjutkan ke jenjang Magister dan tercatat sebagai lulusan Universitas Dian

Nuswantoro Semarang tahun 2017. Wanita yang kerap disapa Lina ini adalah anak dari pasangan Hanapi (ayah, alm) dan St.Hadrah (ibu. alm). Saat ini menjadi salah satu tenaga pengajar di Universitas Dipa Makassar.

I Made Dendi Maysanjaya, S.Pd., M.Eng.
Lahir di Singaraja, Bali, pada tanggal 15 Mei 1990. Penulis menyelesaikan pendidikan sarjana strata-1 (S1) di Program Studi Pendidikan Teknik Informatika, Universitas Pendidikan Ganesha (Undiksha), Singaraja, Bali, pada tahun 2012. Kemudian melanjutkan pendidikan sarjana strata-2 (S2) di Program Studi Teknik Elektro Konsentrasi Teknologi Informasi, Universitas Gadjah Mada (UGM), Yogyakarta, pada tahun 2015. Saat ini berstatus sebagai dosen tetap Prodi Sistem Informasi, Fakultas Teknik dan Kejuruan, Undiksha. Bidang penelitian yang ditekuni dari sejak S1 seputar pengolahan citra digital, kecerdasan buatan, sistem pakar dan sistem pendukung keputusan. Dari tahun 2020, bergabung dalam kelompok riset Virtual Vision, Image, and Pattern/VVIP-RG (https://research.undiksha.ac.id/vvip-rg/). Mata kuliah yang diampu meliputi Algoritma dan Pemrograman, Pemrograman dan Struktur Data, Desain dan Analisis Algoritma, Logika Informatika, Matematika Diskrit, Perancangan Basis Data, Manajemen Basis Data, Testing dan Implementasi SI, Pengolahan Citra Digital, Mikroprosesor dan Dasar Robotika, Pengantar Kecerdasan Buatan, Sistem Pakar dan Sistem Pendukung Keputusan.

Penerbit :

PT. Sonpedia Publishing Indonesia

Buku Gudang Ilmu, Membaca Solusi Kebodohan, Menulis Cara Terbaik Mengikat Ilmu. Everyday New Books

Redaksi :

Jl. Kenali Jaya No 166
Kota Jambi 36129
Tel +6282177858344
Email: sonpediapublishing@gmail.com
Website: www.buku.sonpedia.com