ISBN 978-623-125-183-1

DATA MINING DAN MANAJEMEN PENGETAHUAN

Aisyah Mutia Dawis, Irmawati, Hermila A, Nahya Nur, Sulfayanti,
Siti Aulia Rachmini, Farid Wajidi, Mutmainnah Muchtar,
Wawan Firgiawan, Nurhikma Arifin, Chairi Nur Insani,
Frans Mikael Sinaga, Nurul Afifah Arifuddin, Yulita Salim,
A. Amirul Asnan Cirua

DATA MINING DAN MANAJEMEN PENGETAHUAN

Aisyah Mutia Dawis
Irmawati
Hermila A
Nahya Nur
Sulfayanti
Siti Aulia Rachmini
Farid Wajidi
Mutmainnah Muchtar
Wawan Firgiawan
Nurhikma Arifin
Chairi Nur Insani
Frans Mikael Sinaga
Nurul Afifah Arifuddin
Yulita Salim
A. Amirul Asnan Cirua

GET PRESS INDONESIA

DATA MINING DAN MANAJEMEN PENGETAHUAN

Penulis :
Aisyah Mutia Dawis
Irmawati
Hermila A
Nahya Nur
Sulfayanti
Siti Aulia Rachmini
Farid Wajidi
Mutmainnah Muchtar
Wawan Firgiawan
Nurhikma Arifin
Chairi Nur Insani
Frans Mikael Sinaga
Nurul Afifah Arifuddin
Yulita Salim
A. Amirul Asnan Cirua

ISBN : 978-623-125-183-1

Editor : Ari Yanto., M.Pd
Penyunting : Tri Putri Wahyuni., S.Pd
Desain Sampul dan Tata Letak : Atyka Trianisa, S.Pd

Penerbit : GET PRESS INDONESIA
Anggota IKAPI No. 033/SBA/2022

Redaksi :
Jln. Palarik Air Pacah No 26 Kel. Air Pacah
Kec. Koto Tangah Kota Padang Sumatera Barat
Website : www. getpress.co.id
Email : adm.getpress@gmail.com

Cetakan pertama, Mei 2024

Hak cipta dilindungi undang-undang
Dilarang memperbanyak karya tulis ini dalam bentuk dan
dengan cara apapun tanpa izin tertulis dari penerbit.

KATA PENGANTAR

Segala Puji dan syukur atas kehadirat Allah SWT dalam segala kesempatan. Sholawat beriring salam dan doa kita sampaikan kepada Nabi Muhammad SAW. Alhamdulillah atas Rahmat dan Karunia-Nya penulis telah menyelesaikan Buku Data Mining Dan Manajemen Pengetahuan ini.

Buku Ini Membahas Konsep Data Mining, Konsep Knowledge Management, Definisi Data, Informasi, Dan Pengetahuan, Jenis Data Untuk Data Mining, Jenis Pola Dalam Data Mining, Tipe-tipe Atribut Data Pada Data Mining, Statistik Deskriptif Yang Dibutuhkan Dalam Data Mining, Kegunaan Data Preprocessing, Tugas-tugas Dalam Data Preprocessing, Konsep Dasar Dari Pola Untuk Data Mining, Konsep Dasar Dari Klasifikasi Data Mining, Pendekatan Klasifikasi, Konsep Cluster Analysis, Jenis-jenis Knowledge, Fase Pada Knowledge Management Cycle, Model Knowledge Management, Metode Pengukuran Kinerja Knowledge Management Dalam Organisasi.

Proses penulisan buku ini berhasil diselesaikan atas kerjasama tim penulis. Demi kualitas yang lebih baik dan kepuasan para pembaca, saran dan masukan yang membangun dari pembaca sangat kami harapkan.

Penulis ucapkan terima kasih kepada semua pihak yang telah mendukung dalam penyelesaian buku ini. Terutama pihak yang telah membantu terbitnya buku ini dan telah mempercayakan mendorong, dan menginisiasi terbitnya buku ini. Semoga buku ini dapat bermanfaat bagi masyarakat Indonesia.

Padang, Mei 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
KONSEP DATA MINING, KONSEP KNOWLEDGE MANAGEMENT, DEFINISI DATA, INFORMASI, DAN PENGETAHUAN

Oleh Aisyah Mutia Dawis

1.1 Konsep Data Mining

Data mining merupakan proses penemuan pola tersembunyi atau informasi yang berguna dari sekumpulan besar data. Ini melibatkan analisis yang mendalam terhadap data untuk mengidentifikasi tren, korelasi, atau pola yang mungkin tidak terlihat secara langsung. Dengan kemajuan teknologi dan peningkatan besar dalam pengumpulan data, data mining menjadi kunci dalam mengungkap wawasan berharga dari informasi yang terkandung dalam dataset yang kompleks.

Gambar 1.1. Illustrasi Data Mining
(Sumber : https://www.freepik.com/)

Pentingnya data mining terletak pada kemampuannya untuk mengubah data mentah menjadi wawasan yang dapat dimanfaatkan secara strategis. Proses ini melibatkan beberapa langkah, termasuk pemilihan data, preprocessing, penerapan teknik analisis, dan interpretasi hasil (Nikmatun and Waspada, 2019).

Pertama-tama, pemilihan data yang tepat sangat penting. Ini melibatkan identifikasi sumber data yang relevan dan penting untuk tujuan analisis. Data yang diperoleh kemudian dipersiapkan melalui proses *preprocessing* (Wicaksono, Purbolaksono and Faraby, 2023) . Langkah ini melibatkan pembersihan data dari kecacatan, penghilangan data yang tidak relevan atau duplikat, dan transformasi data ke format yang sesuai untuk analisis.

Setelah data dipersiapkan, berbagai teknik data mining dapat diterapkan. Ini termasuk klastering untuk mengelompokkan data serupa, klasifikasi untuk memprediksi kategori atau label dari data baru, asosiasi untuk menemukan hubungan di antara variabel, dan analisis regresi untuk memahami hubungan antara variabel. Teknik-teknik ini membantu dalam mengeksplorasi dan menganalisis data dengan lebih dalam.

Data mining memiliki aplikasi yang luas di berbagai bidang. Di industri, hal ini digunakan untuk analisis pasar, deteksi pola fraud, pengoptimalan rantai pasokan, dan pengelolaan inventaris. Di bidang kesehatan, data mining membantu dalam identifikasi tren penyakit, diagnosis dini, dan pengembangan obat. Bahkan, di sektor keuangan, data mining digunakan untuk memprediksi perilaku pasar, mengidentifikasi risiko, dan memahami perilaku konsumen.

Namun, dengan penggunaan yang semakin luas, ada tantangan yang perlu diatasi dalam praktik data mining. Salah satunya adalah privasi dan etika dalam penggunaan data pribadi. Perlindungan terhadap informasi pribadi menjadi sangat penting dalam penggunaan data mining yang bertanggung jawab.

Secara keseluruhan, data mining memegang peranan kunci dalam mengubah data menjadi wawasan berharga. Dengan pendekatan yang tepat, penggunaan data mining tidak hanya membantu dalam pengambilan keputusan yang lebih baik, tetapi

juga membuka pintu menuju inovasi dan perkembangan di berbagai bidang kehidupan.

Berikut ini merupakan beberapa contoh yang merupakan data mining :

1.1.1 *Market Basket Analysis*

Data mining adalah proses ekstraksi informasi yang relevan dan menarik dari sejumlah besar data yang belum terstruktur atau terstruktur (Alfarizi, Rizqy and Ghufroni, 2023). Ini melibatkan penggunaan berbagai teknik, algoritma, dan perangkat lunak untuk menemukan pola, hubungan, atau tren yang tersembunyi dalam data. Misalnya, sebuah perusahaan *e-commerce* menggunakan data mining untuk menganalisis pola pembelian pelanggan dan memprediksi preferensi mereka di masa mendatang. Dengan memanfaatkan teknik seperti *clustering* atau *association* rules, mereka dapat mengidentifikasi hubungan antara produk yang dibeli dan menyesuaikan strategi pemasaran mereka.

Para profesional data mining harus memiliki pemahaman yang mendalam tentang statistik, *machine learning*, dan analisis data. Mereka menggunakan algoritma seperti *decision trees*, *neural networks*, atau *clustering algorithms* untuk memproses data yang besar dan kompleks.

Market Basket Analysis (MBA) merupakan teknik data mining yang digunakan untuk menemukan pola atau asosiasi di antara item-item yang sering dibeli bersamaan oleh pelanggan (Dio, Dermawan and Putera, 2023). Ini digunakan secara luas dalam ritel dan perdagangan elektronik untuk memahami perilaku pembelian pelanggan.

Misalkan Anda memiliki data transaksi dari sebuah toko. Setiap entri dalam data tersebut mencatat barang-barang yang dibeli oleh seorang pelanggan pada satu transaksi. MBA akan mengeksplorasi data ini untuk menemukan hubungan antara item-item yang sering dibeli bersama dalam transaksi yang sama.

Gambar 1.2. *Market Basket Analysis*
(Sumber : https://www.freepik.com/)

Contohnya, dari analisis ini, toko dapat menemukan asosiasi seperti: "**Jika seseorang membeli roti, mereka juga cenderung membeli mentega**." Informasi ini dapat digunakan untuk meningkatkan strategi penjualan, seperti menempatkan roti dan mentega dekat satu sama lain di rak toko, membuat promosi bundel, atau memberikan diskon bersamaan untuk mendorong penjualan yang lebih baik.

Teknik yang sering digunakan dalam *Market Basket Analysis* adalah Algoritma Apriori, yang secara sistematis mencari asosiasi antara item-item dalam data transaksi dengan menetapkan batas minimum untuk *support* dan *confidence*. *Support* mengukur seberapa sering kombinasi item muncul dalam transaksi, sementara *confidence* mengukur seberapa sering kombinasi item tersebut dibeli bersama.

1.1.2 Prediksi Risiko Kredit

Prediksi risiko kredit merupakan salah satu aplikasi penting dari data mining dalam industri keuangan. Proses ini menggunakan teknik analisis data untuk mengevaluasi risiko yang terkait dengan memberikan pinjaman kepada pelanggan atau peminjam potensial.

Biasanya, institusi keuangan seperti bank mengumpulkan data historis tentang pelanggan mereka, termasuk riwayat pinjaman, pembayaran, penghasilan, dan faktor-faktor lain yang relevan. Data ini kemudian dianalisis secara teliti untuk membangun model prediksi yang dapat mengidentifikasi potensi risiko kredit dari calon peminjam.

Model prediksi risiko kredit ini didasarkan pada algoritma machine learning dan teknik statistik yang kompleks. Mereka menggunakan data historis untuk mengidentifikasi pola dan tren yang dapat mengindikasikan kemungkinan pembayaran kembali pinjaman secara tepat waktu atau potensi gagal bayar.

1.1.3 Pengelompokan (*Clustering*)

Pengelompokan atau *clustering* adalah teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang serupa berdasarkan karakteristik atau atribut yang ada. Tujuan utamanya adalah menemukan pola-pola alami dalam data yang mungkin tidak terlihat secara langsung.

Proses pengelompokan dimulai dengan dataset yang besar dan kompleks. Algoritma *clustering* akan mencari pola atau kesamaan antara data-data tersebut (Fitri, Suryono and Wantoro, 2023). Ini dilakukan dengan mempertimbangkan jarak atau kesamaan antara titik-titik data di ruang fitur yang telah ditentukan. Contohnya, jika kita memiliki kumpulan data pelanggan dari sebuah toko, algoritma pengelompokan dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku belanja mereka. Ini bisa menjadi pelanggan yang cenderung membeli produk-produk tertentu, pelanggan yang memiliki jumlah belanja yang serupa, atau pola pembelian lainnya.

1.1.4 Penambangan Web (Web Mining)

Penambangan web (Web Mining) merupakan proses ekstraksi informasi yang bernilai dari *World Wide Web* dengan menggunakan teknik analisis data dan teknologi data mining. Ini melibatkan pengumpulan data dari berbagai sumber online, analisis struktur dan konten halaman web, serta penemuan pola atau informasi yang bermanfaat dari data yang diambil.

Pertama, tahap pengumpulan data melibatkan *crawling*, yaitu proses dimana bot atau crawler secara otomatis menjelajahi internet untuk mengumpulkan informasi dari halaman web (Aji Susanto, 2023). Data ini bisa berupa teks, gambar, audio, atau bahkan metadata seperti tag HTML, struktur situs, dan link antar halaman.

Setelah pengumpulan data, langkah berikutnya adalah pre-processing, di mana data yang dikumpulkan dianalisis untuk membersihkan, memfilter, dan mempersiapkannya untuk analisis lebih lanjut. Ini melibatkan menghilangkan duplikasi, mengelompokkan data ke dalam kategori tertentu, dan mungkin menerapkan teknik pemrosesan bahasa alami untuk memahami dan mengekstraksi makna dari teks.

Kemudian, tahap analisis dimulai, yang bisa mencakup beberapa teknik data mining. Contohnya:

1. Pengelompokan (*Clustering*): Mengelompokkan halaman web ke dalam kelompok berdasarkan kesamaan topik atau konten.
2. Klasifikasi: Mengklasifikasikan halaman web ke dalam kategori atau label berdasarkan karakteristik tertentu. Misalnya, mengklasifikasikan berita menjadi berbagai topik seperti politik, olahraga, atau bisnis.
3. Asosiasi: Mencari hubungan dan pola antara halaman web, seperti halaman mana yang sering kali diakses bersamaan.
4. Analisis Sentimen: Menganalisis teks dari ulasan atau konten web untuk mengetahui sentimen atau perasaan yang terkandung di dalamnya.

Hasil dari penambangan web dapat sangat bermanfaat. Misalnya, perusahaan bisa menggunakan informasi yang diperoleh untuk memahami tren pasar, meningkatkan strategi pemasaran online, mengoptimalkan pengalaman pengguna, atau bahkan untuk tujuan riset akademis dalam berbagai bidang seperti ilmu komputer, ilmu informasi, atau sains sosial.

1.2 Konsep Knowledge Management

Knowledge Management (KM) merupakan konsep yang menjadi pilar bagi organisasi modern yang berorientasi pada pengetahuan. Di era di mana informasi menjadi aset berharga, keberhasilan sebuah organisasi sangat tergantung pada kemampuannya dalam mengelola, menyimpan, membagikan, dan menggunakan pengetahuan dengan efektif.

Secara sederhana, *Knowledge Management* mengacu pada praktik, kebijakan, dan inisiatif yang bertujuan untuk menciptakan, menyimpan, membagikan, dan mengelola pengetahuan dalam sebuah organisasi (Wijaya, 2023). Ini melibatkan segala aspek yang berkaitan dengan pengetahuan, termasuk informasi, pengalaman, keahlian, dan wawasan yang dimiliki oleh individu atau entitas.

Mengapa Knowledge Management Penting?

1. **Inovasi Berkelanjutan**
   Knowledge Management (KM) memungkinkan organisasi untuk memanfaatkan ide-ide baru dan menerapkan inovasi secara konsisten, mengarah pada pengembangan produk dan layanan yang lebih baik.
2. **Peningkatan Efisiensi**
   Dengan akses yang lebih baik terhadap pengetahuan, proses bisnis dapat dioptimalkan, menghemat waktu dan sumber daya.
3. **Meningkatkan Kualitas Pengambilan Keputusan**
   Informasi yang terorganisir dengan baik membantu dalam pengambilan keputusan yang lebih baik, karena berdasarkan pada analisis data yang akurat dan relevan.
4. **Pemberdayaan Karyawan**
   Dengan KM, karyawan merasa didukung dalam pembelajaran kontinu dan pengembangan pribadi, yang memotivasi mereka untuk berkontribusi lebih baik (Haryanto and Saputra, 2023).

Komponen Knowledge Management

1. **Pengidentifikasian Pengetahuan**
   Pertama-tama, organisasi harus mengidentifikasi pengetahuan yang dimilikinya, termasuk dalam bentuk

dokumen, basis data, atau pengetahuan yang tersimpan dalam pikiran individu.

2. **Pengorganisasian dan Penyimpanan**
   Pengetahuan perlu diatur dengan baik agar mudah diakses. Sistem penyimpanan dan pengindeksan, termasuk basis data dan sistem manajemen dokumen, sangat penting.
3. **Pembagian Pengetahuan**
   Membagikan pengetahuan kepada mereka yang membutuhkannya merupakan elemen penting. Ini dapat dilakukan melalui platform kolaboratif, database yang dapat diakses, atau sesi pelatihan dan mentoring.
4. **Pemanfaatan Pengetahuan**
   Pengetahuan tidak berguna jika tidak digunakan. Organisasi harus mendorong penggunaan pengetahuan ini dalam pengambilan keputusan, inovasi, dan dalam meningkatkan proses internal.

Strategi Implementasi Knowledge Management

1. **Kultur Organisasi yang Mendorong Pembelajaran**
   Organisasi harus menciptakan lingkungan yang mempromosikan berbagi pengetahuan, menghargai pembelajaran, dan mendorong kolaborasi.
2. **Teknologi yang Mendukung**
   Penggunaan teknologi yang tepat adalah kunci. Sistem manajemen pengetahuan, platform kolaboratif, dan tools analitik dapat membantu mengelola dan memanfaatkan pengetahuan dengan lebih efektif.
3. **Pelatihan dan Pengembangan Karyawan**
   Investasi dalam pelatihan yang terkait dengan KM membantu karyawan untuk memahami pentingnya pengetahuan dalam konteks pekerjaan mereka dan bagaimana cara terbaik untuk mengelolanya.
4. **Pengukuran dan Evaluasi**
   Menetapkan metrik untuk mengukur keberhasilan program KM adalah langkah krusial (Rofianto *et al.*, 2023). Evaluasi teratur membantu organisasi mengetahui sejauh mana mereka berhasil dalam mengelola pengetahuan.

Tantangan dalam Implementasi Knowledge Management

1. **Resistensi terhadap Perubahan**
   Adopsi KM seringkali dihadapi dengan resistensi dari individu yang cenderung mempertahankan cara kerja lama.
2. **Kesulitan dalam Menemukan Pengetahuan yang Tepat**
   Terlalu banyak informasi bisa menjadi bumerang. Organisasi harus memiliki strategi untuk menemukan pengetahuan yang tepat di antara banjir informasi.
3. **Kesulitan dalam Pengukuran Efektivitas**
   Mengukur dampak langsung dari KM bisa menjadi tugas yang rumit. Pengukuran efektivitas adalah langkah kritis dalam mengevaluasi sejauh mana suatu proses, kegiatan, atau strategi telah berhasil mencapai tujuan yang ditetapkan.
4. **Kesulitan Meningkatkan Keterlibatan Karyawan**
   Tidak semua orang mungkin merasa nyaman dalam berbagi pengetahuan atau berpartisipasi dalam inisiatif KM (Yusnaldi *et al.*, 2023). Meningkatkan keterlibatan karyawan adalah tujuan krusial bagi setiap organisasi yang ingin menciptakan lingkungan kerja yang produktif, berinovasi, dan mempertahankan bakat terbaik.

Peran Knowledge Management di Masa Depan

Dalam era yang semakin terhubung dan berbasis teknologi, *Knowledge Management* (KM) akan menjadi kunci utama bagi kesuksesan organisasi (Putu Kawiana *et al.*, 2023). Integrasi kecerdasan buatan, analisis big data, dan teknologi lainnya akan memperkuat praktik KM, memungkinkan organisasi untuk mengambil keputusan yang lebih baik dan berevolusi dengan cepat sesuai dengan perubahan lingkungan.

Pentingnya *Knowledge Management* tidak hanya terletak pada ketersediaan informasi, tetapi juga pada cara organisasi mengelolanya, membagikannya, dan menggunakannya untuk pertumbuhan dan keberhasilan jangka panjang (Puji Ratwiyanti and Harisno, 2023). Menerapkan strategi yang tepat, menciptakan budaya yang mendukung, dan memanfaatkan teknologi adalah

langkah penting dalam membangun fondasi yang kuat untuk KM dalam sebuah organisasi.

1.3 Definisi Data, Informasi, dan Pengetahuan

Data, informasi, dan pengetahuan adalah konsep yang saling terkait, namun memiliki makna dan tingkat kompleksitas yang berbeda dalam konteks penggunaannya.

Data

1. Data merujuk pada fakta mentah atau elemen-elemen yang tidak memiliki konteks tertentu atau makna langsung (Ramadhan and Astutik, 2020). Data dapat berupa angka, teks, gambar, atau segala bentuk yang dapat diukur atau dihitung. Contohnya, angka 5, nama seseorang, atau gambar berwarna-warni semua merupakan contoh data.
2. Namun, data sendiri belum memiliki makna atau interpretasi yang jelas tanpa proses pengolahan lebih lanjut. Misalnya, angka "5" sendiri tidak memberikan informasi yang berarti tanpa konteksnya. Data ini perlu diolah atau diinterpretasikan agar menjadi informasi yang bermanfaat.

Informasi

1. Informasi adalah data yang telah diolah, diorganisir, atau diinterpretasikan sehingga memiliki makna dan relevansi tertentu (Febri Putra Raharjo and Restu Putra, 2023). Proses transformasi data menjadi informasi melibatkan konteks, struktur, dan pengaturan yang membuatnya dapat dipahami dan berguna. Misalnya, ketika angka "5" dikaitkan dengan informasi bahwa itu adalah suhu dalam derajat Celsius, maka itu menjadi informasi yang bermanfaat dalam konteks suhu.
2. Informasi dapat memberikan pemahaman atau penjelasan tentang suatu situasi, peristiwa, atau kondisi (Erwan Effendy *et al.*, 2023). Ini memungkinkan pengambilan keputusan yang lebih baik karena informasi memberikan konteks yang diperlukan. Dalam dunia digital dan bisnis, informasi adalah elemen penting dalam mengelola proses, menganalisis tren, dan membuat keputusan strategis.

Pengetahuan

1. Pengetahuan adalah pemahaman yang lebih dalam atau tingkat pemrosesan mental yang melibatkan informasi. Ini adalah hasil dari interpretasi dan pengalaman yang dimiliki seseorang terhadap informasi. Pengetahuan tidak hanya tentang memahami fakta-fakta atau informasi, tetapi juga tentang bagaimana fakta-fakta ini dihubungkan, diterapkan, dan digunakan dalam konteks tertentu (Supriatna *et al.*, 2023).
2. Pengetahuan mencakup pengalaman, wawasan, intuisi, dan keterampilan yang diperoleh dari pemahaman yang mendalam terhadap informasi. Ini memungkinkan seseorang untuk membuat analisis yang kompleks, berpikir kritis, dan mengambil keputusan yang lebih tepat dalam situasi yang kompleks atau tidak terduga.

Hubungan Antar Konsep

1. Data, informasi, dan pengetahuan saling terkait dan membentuk tahapan yang berkaitan satu sama lain. Data merupakan bahan mentah yang diperlukan untuk membuat informasi. Informasi, di sisi lain, adalah hasil dari pengolahan data yang memberikan makna dan konteks. Sedangkan pengetahuan adalah hasil dari memahami, menggunakan, dan menghubungkan informasi.
2. Misalnya, dalam dunia medis, data dapat berupa nilai-nilai laboratorium pasien. Ketika data ini diolah dan diinterpretasikan dengan mengaitkannya dengan riwayat medis pasien, itu menjadi informasi yang membantu dokter untuk membuat diagnosis. Namun, pengetahuan seorang dokter adalah kumpulan dari pengalaman, pelatihan, dan pemahaman mendalam yang memungkinkan mereka untuk mengaitkan informasi ini dengan kemungkinan diagnosis dan rencana pengobatan yang efektif.

Pentingnya Memahami Perbedaan Ini

1. Pemahaman tentang perbedaan antara data, informasi, dan pengetahuan penting dalam banyak aspek kehidupan, termasuk bisnis, pendidikan, ilmu pengetahuan, dan pengambilan

keputusan. Dalam bisnis, manajemen data yang efektif dapat membantu dalam menghasilkan informasi yang relevan, yang kemudian dapat digunakan untuk membangun pengetahuan yang berguna bagi pengambilan keputusan strategis.

2. Di bidang pendidikan, pengolahan informasi yang tepat dari data dapat membantu siswa membangun pengetahuan yang mendalam. Di dunia ilmu pengetahuan, penggunaan data dan informasi yang akurat adalah kunci untuk membuat pengetahuan baru yang memajukan peradaban manusia.

Data, informasi, dan pengetahuan membentuk suatu hierarki yang saling melengkapi. Data adalah bahan mentah, informasi adalah hasil pengolahan data, dan pengetahuan adalah pemahaman mendalam yang dibangun dari informasi. Memahami perbedaan dan hubungan antara ketiganya sangat penting dalam konteks mengelola informasi, membuat keputusan, dan membangun pemahaman yang mendalam tentang dunia di sekitar kita.

DAFTAR PUSTAKA

Aji Susanto, I.A.D. (2023) 'Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest', 9(20), pp. 59–65.

Alfarizi, M., Rizqy, M. and Ghufroni, R.I. (2023) 'Analisis Sentimen Persepsi Publik Terhadap Kasus Bullying Siswa Cilacap Menggunakan Pendekatan Machine Learning', 4(3), pp. 265–276.

Dio, R., Dermawan, A.A. and Putera, D.A. (2023) 'Application of Market Basket Analysis on Beauty Clinic to Increasing Customer's Buying Decision', *Sinkron*, 8(3), pp. 1348–1356. Available at: https://doi.org/10.33395/sinkron.v8i3.12421.

Erwan Effendy *et al.* (2023) 'Konsep Informasi Konsep Fakta Dan Informasi', *Jurnal Pendidikan dan Konseling*, 5 Nomor 2(Vol. 5 No. 2 (2023): Jurnal Pendidikan dan Konseling), pp. 1–7.

Febri Putra Raharjo, A. and Restu Putra, D. (2023) 'Sistem Informasi Profil Perusahaan Berbasis Web Pada Toko Dua Arah Dengan Metode Extreme Programming', *Jurnal Informatika MULTI*, 1(4), pp. 389–398.

Fitri, E.M., Suryono, R.R. and Wantoro, A. (2023) 'Klasterisasi Data Penjualan Berdasarkan Wilayah Menggunakan Metode K-Means Pada Pt Xyz', *Jurnal Komputasi*, 11(2), pp. 157–168. Available at: https://jurnal.fmipa.unila.ac.id/komputasi/article/view/12582.

Haryanto, F. and Saputra, D.P. (2023) 'BUDAYA ORGANISASI DAN DAMPAKNYA PADA PERILAKU DAN KINERJA KARYAWAN'.

Nikmatun, I.A. and Waspada, I. (2019) 'Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor', *Jurnal SIMETRIS*, 10(2), pp. 421–432.

Puji Ratwiyanti and Harisno (2023) 'Pengembangan Arsitektur Knowledge Management System Untuk Model Bisnis Pendidikan Tinggi', *JTIM : Jurnal Teknologi Informasi dan Multimedia*, 5(1), pp. 34–47. Available at: https://doi.org/10.35746/jtim.v5i1.332.

Putu Kawiana, I.G. *et al.* (2023) 'Peran Komitmen Organisasi Pada Pengaruh Knowledge Management dan Motivasi Kerja Terhadap Kinerja Karyawan di PT Limajari Interbhuana Bali', *Jesya*, 6(2), pp. 2024–2040. Available at: https://doi.org/10.36778/jesya.v6i2.1258.

Ramadhan, M.G. and Astutik, A.P. (2020) 'Implementasi Budaya Religius Dalam Penanaman Adab Siswa', 5(July), pp. 1–23. Available at: https://doi.org/10.19109/pairf.v5i3.

Rofianto, D. *et al.* (2023) 'DESAIN SISTEM PENDUKUNG KEPUTUSAN STATUS KESTABILAN KAPAL ROLL ON-ROLL OVER (RORO) BERBASIS FUZZY SUGENO', 7(2), pp. 270–276.

Supriatna, M.N. *et al.* (2023) 'Analisis Perbandingan Kurikulum KTSP, K13 dan Kurikulum Merdeka di Sekolah Dasar', *Journal on Education*, 06(01), pp. 9163–9172.

Wicaksono, M.H., Purbolaksono, M.D. and Faraby, S. Al (2023) 'Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily', *eProceedings of Engineering*, 10(3), pp. 3591–3600. Available at: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20631/19944%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20631.

Wijaya, A.G. dan M.I. (2023) 'Penerapan Knowledge Management Pada Pt. Xyz Dengan Model Seci Dalam Upaya Meningkatkan Kinerja Karyawan', *Journal of Digital Business and Innovation Management*, 2(1), pp. 1–16. Available at: https://doi.org/10.1234/jdbim.v2i1.51795.

Yusnaldi, E. *et al.* (2023) 'Efektivitas Model Pembelajaran Discovery Learning untuk Meningkatkan Pemahaman Siswa dalam Mata Pelajaran IPS di SD', 7, pp. 28528–28532.

BAB 2
JENIS DATA UNTUK DATA MINING

Oleh Irmawati

2.1 Jenis Data

Data yang digunakan dalam data mining merujuk pada kumpulan informasi atau fakta yang dapat dianalisis untuk menghasilkan pola, tren, atau pengetahuan yang bermanfaat. Data mining adalah proses ekstraksi pola atau pengetahuan yang tersembunyi dalam data, dan data yang digunakan dalam konteks ini dapat berasal dari berbagai sumber (Han, Pei and Tong, 2022).

Beberapa karakteristik data yang penting untuk data mining meliputi:

1. Volume:
   Data mining sering melibatkan jumlah data yang besar. Semakin besar volume data, semakin kompleks analisis yang dapat dilakukan, dan semakin besar potensi penemuan pola yang berharga.
2. Varietas:
   Data mining dapat melibatkan berbagai jenis data, termasuk teks, gambar, suara, dan data terstruktur lainnya. Varietas data ini dapat memberikan wawasan yang lebih komprehensif jika dianalisis bersama.
3. Kecepatan:
   Beberapa aplikasi data mining memerlukan analisis real-time atau setidaknya analisis yang cepat. Oleh karena itu, kecepatan pengaksesan dan pemrosesan data dapat menjadi faktor penting.
4. Keterpaduan:
   Data mining sering melibatkan penggabungan data dari berbagai sumber untuk mendapatkan pemahaman yang lebih lengkap. Oleh karena itu, integrasi dan keterpaduan data menjadi penting.

5. Keakuratan:
   Data yang digunakan dalam data mining harus akurat. Kesalahan atau ketidakakuratan dalam data dapat menghasilkan analisis yang tidak dapat diandalkan dan menyesatkan.
6. Ketidakpastian:
   Beberapa data mungkin memiliki tingkat ketidakpastian, dan data mining harus dapat menangani ketidakpastian ini untuk menghasilkan hasil yang berguna dan dapat diandalkan.
7. Dimensi:
   Dimensi mengacu pada jumlah atribut atau fitur dalam setiap entitas data. Data dengan dimensi tinggi (banyak atribut) dapat menantang dalam analisis, dan teknik khusus mungkin diperlukan untuk mengatasi kompleksitas tersebut.

Data mining melibatkan proses ekstraksi informasi yang berharga dari data dengan menggunakan berbagai teknik seperti clustering, klasifikasi, regresi, asosiasi, dan lainnya. Data yang baik untuk data mining harus relevan dengan tujuan analisis dan dipersiapkan dengan baik untuk memastikan hasil yang akurat dan bermanfaat (Aggarwal, 2015).

Dalam konteks data mining, data dapat diklasifikasikan menjadi beberapa jenis berdasarkan karakteristik dan sifatnya. Berikut adalah beberapa jenis data yang sering dihadapi dalam data mining:

1. Data Terstruktur:
   Data yang diatur dalam format yang terstruktur, seperti tabel relasional dengan kolom-kolom tertentu.
   Contoh: Basis data SQL, spreadsheet Excel.
2. Data Semi-Terstruktur:
   Data yang memiliki struktur yang lebih fleksibel daripada data terstruktur, tetapi masih memiliki beberapa format terstruktur.
   Contoh: Dokumen XML, JSON.

3. Data Tidak Terstruktur:
   Data tanpa struktur yang jelas, sulit untuk diorganisir atau diinterpretasikan secara langsung.
   Contoh: Teks bebas, audio, video.
4. Data Numerik:
   Data yang terdiri dari angka dan dapat diukur secara kuantitatif.
   Contoh: Nilai suhu, pendapatan tahunan.
5. Data Kategorikal:
   Data yang terdiri dari kategori atau label, bukan nilai numerik.
   Contoh: Jenis produk, warna mobil.
6. Data Temporal:
   Data yang berkaitan dengan waktu atau urutan kejadian.
   Contoh: Tanggal transaksi, waktu klik pada suatu situs web.
7. Data spasial atau Geografis:
   Data yang terkait dengan lokasi geografis atau spasial.
   Contoh: Koordinat GPS, peta.
8. Data Textual:
   Data dalam bentuk teks atau urutan karakter.
   Contoh: Artikel berita, tweet.
9. Data Grafik:
   Data yang mewakili hubungan antar entitas dalam bentuk grafik.
   Contoh: Jaringan sosial, diagram alur.

Memahami jenis data yang dihadapi sangat penting, karena metode data mining yang efektif dapat bervariasi tergantung pada sifat data tersebut.

2.2 Klasifikasi Jenis Data

Klasifikasi jenis data membantu dalam mengembangkan strategi analisis yang sesuai dan memudahkan pengelolaan data sesuai dengan kebutuhan. Pemahaman yang baik tentang jenis data juga membantu dalam menentukan metode analisis yang tepat dan mengoptimalkan nilai informasi yang dapat diambil dari data tersebut. Data dapat dibedakan berdasarkan karakteristiknya,

memungkinkan kita untuk memahami sifat dan kegunaannya dengan lebih baik.

2.2.1 Data Numerik

Data numerik adalah jenis data yang berisi angka atau nilai numerik. Data ini dapat diukur dan dihitung menggunakan operasi matematika. Jenis data numerik umumnya digunakan dalam statistik, analisis data, dan ilmu matematika lainnya. Data numerik dapat dibagi menjadi dua kategori utama: data diskrit dan data kontinu. Data diskrit terdiri dari nilai terpisah, sedangkan data kontinu dapat mengambil nilai di antara dua nilai tertentu (Provost and Fawcett, 2013).

Contoh data numerik melibatkan nilai atau angka konkret. Berikut adalah beberapa contoh:

1. Usia seseorang: 25 tahun
2. Jumlah barang yang terjual dalam sehari: 150 unit
3. Temperatur ruangan: 28,5 derajat Celsius
4. Tinggi badan seseorang: 175 centimeter
5. Pendapatan bulanan: Rp. 3,000

2.2.2 Data Kategorikal

Data kategorikal merupakan jenis data yang menggambarkan kategori atau kelompok tertentu dan tidak memiliki tingkatan atau urutan yang dapat diukur secara numerik. Karakteristik utama dari data ini adalah sifat diskrit dan ketidakberurutan, serta ketidakmampuannya mengandung skala interval atau rasio. Data kategorikal cenderung bersifat kualitatif, memperlihatkan keanggotaan suatu entitas dalam suatu kategori tanpa adanya implikasi hubungan sekuensial antar-nilai (Agresti, 2012).

Karakteristik Data Kategorikal:

1. Diskrit dan Tidak Berurutan:
   Data kategorikal bersifat diskrit, yang berarti hanya mengambil nilai-nilai terbatas dan tidak ada nilai di antara. Selain itu, tidak ada urutan yang bermakna di antara nilai-nilai tersebut.

2. Tidak Memiliki Skala Interval atau Rasio:
   Data kategorikal tidak memiliki skala interval atau rasio. Artinya, perbedaan antara kategori tidak dapat diukur secara kuantitatif dengan cara yang bermakna.

Contoh Data Kategorikal:
1. Jenis Kelamin: Pria atau Wanita.
2. Jenis Darah: A, B, AB, O.
3. Warna Mata: Biru, Coklat, Hijau.
4. Status Pernikahan: Menikah, Belum Menikah, Cerai.
5. Kategori Produk: Elektronik, Pakaian, Makanan.

Data kategorikal memainkan peran penting dalam berbagai bidang analisis, termasuk ilmu sosial, ekonomi, dan kedokteran. Teknik analisis yang umum digunakan melibatkan uji chi-square untuk mengidentifikasi hubungan antar-kategori.

2.2.3 Data Temporal

Data temporal adalah data yang berkaitan dengan waktu atau memiliki dimensi waktu. Karakteristik utama dari data temporal adalah adanya informasi waktu yang terkandung dalam setiap entitas data. Beberapa karakteristik kunci dari data temporal melibatkan perubahan nilai-nilai data sepanjang waktu, urutan waktu yang signifikan, dan kemungkinan adanya tren atau pola berdasarkan waktu. Berikut adalah beberapa karakteristik data temporal:

1. Dimensi Waktu:
   Data temporal memiliki dimensi waktu yang mencerminkan perubahan nilai-nilai data seiring berjalannya waktu. Contohnya termasuk data cuaca harian, penjualan bulanan, atau catatan waktu peristiwa.
2. Tren dan Pola:
   Data temporal sering menunjukkan tren atau pola tertentu sepanjang waktu. Analisis tren ini dapat membantu dalam meramalkan atau memahami perubahan yang mungkin terjadi di masa depan.
3. Sejarah dan Urutan:

Informasi waktu memberikan urutan dan sejarah pada data. Ini membantu dalam memahami bagaimana nilai-nilai data berkembang sepanjang waktu dan bagaimana peristiwa di masa lalu dapat mempengaruhi situasi saat ini.

4. Ketergantungan Waktu:
   Data temporal memiliki ketergantungan waktu, yang berarti bahwa nilai-nilai pada suatu waktu dapat dipengaruhi oleh nilai-nilai pada waktu sebelumnya. Contoh ini dapat ditemukan dalam analisis keuangan atau produksi.
5. Fenomena Musiman:
   Beberapa jenis data temporal dapat menunjukkan pola musiman, seperti peningkatan penjualan pada musim liburan atau fluktuasi suhu yang teratur dalam setahun.
6. Kecepatan Perubahan:
   Kecepatan perubahan data temporal dapat bervariasi. Beberapa data dapat berubah dengan cepat, seperti data saham, sementara yang lain mungkin berubah secara lebih lambat, seperti data demografi.

Penerapan data temporal melibatkan pengumpulan, penyimpanan, analisis, dan interpretasi data berdasarkan dimensi waktu. Beberapa contoh penerapan data temporal termasuk:

1. Analisis Keuangan: Memantau perubahan harga saham sepanjang waktu untuk membuat keputusan investasi.
2. Prediksi Cuaca: Menganalisis data cuaca harian dan musiman untuk meramalkan kondisi cuaca di masa depan.
3. Manajemen Persediaan: Melacak perubahan dalam permintaan produk sepanjang waktu untuk mengelola persediaan dengan efisien.
4. Analisis Lalu Lintas: Memantau pola lalu lintas jalan raya pada jam sibuk dan hari tertentu untuk meningkatkan manajemen lalu lintas.

Penerapan data temporal memerlukan pemahaman yang baik tentang konsep waktu dan metodologi analisis data temporal seperti time series analysis.

2.3 Pengumpulan dan Pengolahan Data

Pengumpulan dan pengolahan data merupakan dua tahapan kunci dalam proses data mining. Data mining adalah suatu proses untuk menemukan pola atau informasi yang bermanfaat dari sekumpulan data yang besar dan kompleks (Ratner, 2017). Berikut adalah penjelasan tentang pengumpulan dan pengolahan data pada data mining:

1. Pengumpulan Data (Data Collection):
    a. Sumber Data: Data dapat berasal dari berbagai sumber, termasuk basis data perusahaan, data transaksi, data sensor, data media sosial, dan lainnya.
    b. Ekstraksi Data: Data yang relevan diekstrak dari sumber-sumber tersebut. Proses ini melibatkan penentuan data apa yang diperlukan untuk analisis dan pemahaman tujuan data mining.
    c. Pembersihan Data (Data Cleaning): Langkah ini melibatkan identifikasi dan penanganan data yang hilang, tidak lengkap, atau tidak valid. Pembersihan data diperlukan agar hasil analisis lebih akurat dan dapat diandalkan.
    d. Integrasi Data: Data dari berbagai sumber dapat diintegrasikan untuk membuat satu set data yang utuh dan konsisten. Melibatkan penyesuaian format dan representasi data agar sesuai dengan kebutuhan analisis.
2. Pengolahan Data (*Data Processing*):
    a. Transformasi Data: Data sering kali perlu diubah atau diolah agar sesuai dengan format atau struktur yang diperlukan untuk analisis. Ini bisa melibatkan normalisasi, penggabungan, atau transformasi variabel.
    b. Reduksi Dimensi (*Dimensionality Reduction*): Proses ini melibatkan pengurangan jumlah variabel atau atribut dalam dataset. Hal ini berguna untuk mengatasi masalah "kutukan dimensi" dan meningkatkan efisiensi analisis.
    c. Pemilihan Fitur (*Feature Selection*): Menentukan atribut atau fitur yang paling relevan dan signifikan untuk

analisis. Pemilihan fitur membantu mengurangi kompleksitas model dan meningkatkan interpretabilitas hasil.

d. *Discretization* dan *Binning*: Beberapa algoritma data mining memerlukan data yang bersifat diskrit atau tersegmentasi. Oleh karena itu, proses ini melibatkan pembagian nilai-nilai kontinu menjadi kelompok-kelompok diskrit.

3. Penyiapan Data (*Data Preparation*):
   a. Pembagian Data (Data Splitting): Dataset biasanya dibagi menjadi subset pelatihan dan subset pengujian untuk melatih dan menguji model.
   b. Normalisasi dan Standarisasi: Memastikan bahwa variabel-variabel memiliki skala yang seragam untuk mencegah atribut dengan magnitudo besar mendominasi proses analisis.
   c. Penanganan Data Tidak Seimbang: Jika dataset tidak seimbang, di mana kelas-kelas tertentu memiliki frekuensi yang sangat berbeda, diperlukan strategi penanganan khusus.

Pengumpulan dan pengolahan data yang cermat adalah kunci untuk memastikan bahwa model data mining dapat memberikan hasil yang akurat dan bermakna. Setelah data dipersiapkan dengan baik, model-model data mining seperti decision trees, clustering algorithms, dan neural networks dapat diterapkan untuk mengidentifikasi pola, tren, atau informasi yang berguna dari data yang ada.

2.4 Teknik Pengolahan Data

Teknik Pengolahan Data dalam data mining merupakan langkah-langkah yang dilakukan untuk membersihkan, mengorganisir, dan mengubah data mentah menjadi bentuk yang dapat digunakan untuk analisis lebih lanjut. Proses ini memainkan peran kritis dalam mendukung keberhasilan proses data mining, karena data yang baik dan berkualitas akan menghasilkan hasil analisis yang lebih akurat dan bermakna (Verma *et al.*, 2019).

Beberapa teknik pengolahan data yang umum digunakan dalam data mining antara lain:

1. *Cleaning* (Pembersihan Data):
   Proses ini melibatkan identifikasi dan penanganan data yang hilang, inkonsisten, atau tidak valid. Contohnya, mengatasi nilai-nilai yang hilang atau outliers yang mungkin mempengaruhi hasil analisis.
2. *Integration* (Integrasi Data):
   Menggabungkan data dari berbagai sumber atau sumber yang berbeda format ke dalam satu format atau struktur data yang bersifat konsisten. Misalnya, menggabungkan data dari database, spreadsheet, dan file teks ke dalam satu dataset.
3. *Transformation* (Transformasi Data):
   Mengubah data ke dalam format atau struktur yang lebih sesuai untuk analisis. Contohnya, mengubah format tanggal, normalisasi data, atau mengganti nilai-nilai kategorikal menjadi representasi numerik.
4. *Reduction* (Reduksi Data):
   Mengurangi kompleksitas data dengan memilih subset relevan dari atribut atau variabel yang paling berpengaruh untuk analisis. Teknik ini dapat membantu mengatasi masalah dimensi tinggi (high-dimensional data).
5. *Discretization* (Discretisasi): Mengubah data kontinu menjadi data diskrit. Contohnya, mengelompokkan umur menjadi kelompok-kelompok tertentu seperti "anak-anak," "remaja," dan "dewasa."

Contoh aplikasi dari teknik pengolahan data ini bisa ditemukan dalam berbagai studi kasus data mining, seperti dalam analisis penjualan ritel, prediksi churn pelanggan, atau identifikasi pola anomali dalam data keamanan. Namun, penting untuk diingat bahwa setiap proyek data mining memiliki kebutuhan khusus, dan teknik pengolahan data yang digunakan dapat bervariasi.

DAFTAR PUSTAKA

Aggarwal, C.C. (2015) *Data mining: the textbook*. Springer.

Agresti, A. (2012) *Categorical data analysis*. John Wiley & Sons.

Han, J., Pei, J. and Tong, H. (2022) *Data mining: concepts and techniques*. Morgan kaufmann.

Provost, F. and Fawcett, T. (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking*. ' O'Reilly Media, Inc.'

Ratner, B. (2017) *Statistical and machine-learning data mining:: Techniques for better predictive modeling and analysis of big data*. books.google.com.

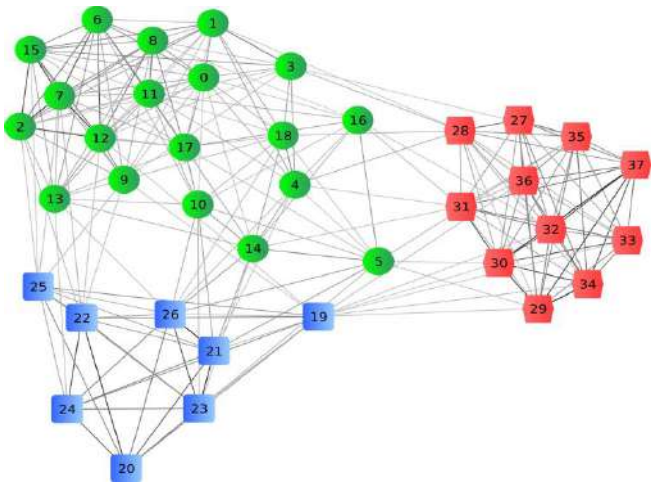
Verma, K. *et al.* (2019) 'Latest tools for data mining and machine learning', *International Journal of Innovative Technology and Exploring Engineering*, 8(9S), pp. 18–23.

BAB 3
JENIS POLA DALAM DATA MINING

Oleh Hermila A.

3.1 Pendahuluan

Pola adalah istilah yang digunakan dalam data mining untuk menggambarkan struktur atau informasi yang dapat ditemukan dalam kumpulan data yang sangat besar dan kompleks. Pola-pola ini dapat membantu menemukan tren, hubungan, atau karakteristik yang tersembunyi dalam data, yang pada gilirannya dapat digunakan untuk memahami fenomena yang lebih baik atau membuat keputusan yang lebih baik. *Algoritme Typical Patterns Mining (TPM)* secara efisien menambang pola-pola umum dari basis data, memberikan representasi yang ringkas dan sesuai dengan set data asli dalam situasi praktis (1)(2).



Gambar 3.1. Pola Data Mining
(Sumber: https://mustakimtelematika.wordpress.com(3))

Data mining membuat model 1 cluster dengan menggunakan data yang ada dan relevan untuk menemukan pola di antara atribut

dataset. Pola-pola di antara berbagai atribut objek (seperti pelanggan) dalam dataset diidentifikasi oleh model, yang merupakan penyajian matematis yang terdiri dari persamaan linear sederhana dan/atau persamaan kompleks yang sangat tidak linear. Beberapa pola ini bersifat deskriptif, menunjukkan hubungan atau persamaan dan kesamaan di antara berbagai atribut tersebut, sementara yang lain adalah prediktif, memprediksi apa yang akan terjadi pada atri

3.2 Jenis Pola Data Mining

Alat dan teknik yang digunakan untuk menemukan pola dalam data bervariasi tergantung pada jenis pola yang ingin ditemukan dan sifat dari data itu sendiri. Beberapa teknik yang umum digunakan termasuk algoritma asosiasi seperti Apriori, algoritma klastering seperti K-Means, algoritma pemodelan sekuensial seperti algoritma Markov, dan berbagai teknik visualisasi data untuk membantu pemahaman pola yang ditemukan. Terdapat beberapa jenis pola yang umum ditemui dalam data mining, di antaranya:

3.2.1 Pola Asosiasi

Pola asosiasi data mining adalah teknik penambangan data yang mengidentifikasi asosiasi antara kelompok item dalam data supermarket (4). Pola asosiasi mengacu pada hubungan statistik yang kuat antara satu set item dalam data. Algoritma seperti Apriori digunakan dalam data mining untuk menemukan pola asosiasi ini dengan menganalisis frekuensi kemunculan kombinasi item dalam dataset. Pola asosiasi mencerminkan hubungan antara item dalam kumpulan data. Contohnya adalah aturan asosiasi, seperti "Jika pelanggan membeli roti, kemungkinan besar mereka juga akan membeli mentega." Pola asosiasi sering digunakan dalam analisis pasar, rekomendasi produk, dan analisis pola pembelian. Berikut lebih lengkap cara kerja pola asosiasi dalam data mining:

1. Prinsip dasar pola asosiai dalam yaitu bertujuan untuk menemukan hubungan dan ketergantungan antara item dalam sebuah kumpulan data. Ini menunjukkan item yang sering muncul bersama atau memiliki hubungan yang kuat

satu sama lain. Sebagai contoh, ditemukan bahwa ketika konsumen membeli barang A, mereka juga cenderung membeli barang B.

2. Jumlah kemunculan kombinasi item dalam kumpulan data disebut support. Jika kombinasi item A dan B muncul bersama secara teratur dalam kumpulan data, maka support untuk kombinasi tersebut tinggi. Support dapat dihitung dengan membagi jumlah kemunculan kombinasi item tersebut dengan total jumlah transaksi.
3. Confidence mengukur seberapa sering aturan asosiasi tersebut benar. Dalam hal ini, jika pelanggan membeli produk A, seberapa sering mereka juga membeli produk B. Confidence dihitung sebagai jumlah transaksi yang memuat kedua item tersebut dibagi dengan jumlah transaksi yang memuat item pertama.
4. Lift mengukur seberapa besar peningkatan dalam kemungkinan pembelian item B ketika item A dibeli, dibandingkan dengan kemungkinan pembelian item B secara keseluruhan. Lift yang lebih besar dari 1 menunjukkan bahwa pembelian item A meningkatkan kemungkinan pembelian item B, sementara lift kurang dari 1 menunjukkan kebalikannya.
5. Salah satu algoritma yang paling banyak digunakan untuk menemukan pola asosiasi adalah Apriori. Ini dilakukan dengan membagi dataset menjadi subset yang lebih kecil. Kemudian, algoritma ini menghitung dukungan untuk setiap item dan kombinasi item dan kemudian menggunakan aturan pruning untuk menyingkirkan item dan kombinasi item yang tidak memenuhi batasan dukungan dan kepercayaan yang telah ditentukan sebelumnya.
6. Langkah-langkah Algoritma Apriori (1) Hitung support untuk setiap item tunggal dalam dataset. (2) Buat aturan asosiasi dengan satu item pada bagian kiri dan satu item pada bagian kanan, lalu hitung confidence untuk setiap aturan. (3) Gunakan aturan pruning untuk menghilangkan aturan yang tidak memenuhi batasan support dan confidence yang ditentukan. (4) Ulangi langkah-langkah di

atas dengan menambahkan item-item baru ke aturan yang ada untuk mencari pola asosiasi yang lebih kompleks.

7. Setelah pola asosiasi telah diidentifikasi, langkah terakhir adalah menerjemahkan dan menerapkan pola tersebut dalam konteks aplikatif yang relevan. Ini dapat mencakup strategi pemasaran yang disesuaikan, penyusunan produk yang lebih efektif, atau perbaikan layanan pelanggan.

3.2.2 Pola Sekuensial

Pola Sekuensial data mining mengungkapkan urutan yang sering terjadi dalam database, dengan aplikasi dalam pembelian pelanggan dan pola akses web, tetapi menantang karena banyaknya urutan antara (5). Pola yang ditemukan dalam urutan waktu atau peristiwa menunjukkan bagaimana satu peristiwa mengarah ke yang lain dalam rentang waktu tertentu. Pola sekuensial, misalnya, dapat ditemukan dalam data transaksi kartu kredit, seperti pola pembelian berulang atau perilaku pelanggan yang berubah dari waktu ke waktu. Berikut lebih lengkap cara kerja pola sekuensial dalam data mining:

1. Prinsip Dasar pola sekuensial dalam data mining bertujuan untuk menemukan pola atau urutan peristiwa yang terjadi secara berurutan dalam data. Ini membantu dalam mengidentifikasi tren, kebiasaan, atau pola berulang yang muncul dari data yang diurutkan berdasarkan waktu atau urutan lainnya.
2. Pola sekuensial sering ditemui dalam analisis transaksi, analisis aktivitas pengguna, analisis log web, dan berbagai bidang lainnya di mana data disusun dalam urutan waktu atau urutan peristiwa.
3. Untuk menemukan pola sekuensial, beberapa metode yang umum digunakan adalah **Metode Brute-Force**, Metode ini membutuhkan pemindaian seluruh kumpulan data untuk menemukan pola sekuensial yang memenuhi spesifikasi tertentu. Namun, karena kompleksitasnya, metode ini mungkin tidak efektif untuk kumpulan data yang besar. **Algoritma Apriori**, Algoritma Apriori awalnya dirancang untuk menemukan pola asosiasi, tetapi dapat diubah untuk

menemukan pola sekuensial dengan menambahkan batasan urutan atau waktu pada aturan yang dibuat. **Algoritma GSP (Generalized Sequential Pattern)**, Algoritma ini dibuat khusus untuk menemukan pola sekuensial dalam dataset besar. Pola sekuensial yang signifikan ditemukan melalui pendekatan penambangan bertahap.

4. Parameter utama yaitu (1) Support: Seberapa sering suatu urutan peristiwa muncul dalam dataset dapat dihitung dengan pola sekuensial, di mana jumlah kemunculan urutan dibagi dengan jumlah total urutan dalam dataset. (2)Length Constraint: Batasan waktu atau jarak antara peristiwa dalam pola sekuensial membantu mengontrol kompleksitas pola yang ditemukan. (3)Window Constraint: Batasan waktu maksimum di mana pola sekuensial diidentifikasi. Confidence: Seberapa sering urutan peristiwa yang terjadi bersama-sama berulang.
5. Langkah Algoritma meliputi (1) Membagi dataset menjadi urutan peristiwa yang lebih kecil. (2) Menghitung support untuk setiap urutan peristiwa. (3) Menerapkan aturan pruning untuk menghilangkan urutan peristiwa yang tidak memenuhi batasan support dan constraint lainnya. (4) Menerapkan langkah-langkah penambangan bertahap untuk mengidentifikasi pola sekuensial yang signifikan.
6. Setelah pola sekuensial telah diidentifikasi, langkah terakhir adalah menerjemahkan dan menerapkan pola tersebut dalam konteks aplikatif yang relevan. Ini dapat mencakup pengembangan strategi pemasaran yang lebih terarah, personalisasi pengalaman pengguna, atau pengoptimalan proses bisnis.

3.2.3 Pola Klaster

Pola Klaster adalah teknik penggalian data yang lebih efektif untuk mengidentifikasi kelompok kecil akun yang menunjukkan perilaku khas dalam data kredit, dibandingkan dengan metode pengelompokan tradisional. penggalian pola berurutan mengungkapkan sekuens yang sering terjadi dalam basis data, dengan aplikasi dalam pembelian pelanggan dan pola akses web,

tetapi menantang karena banyaknya sekuens perantara (6). Pola klaster berhubungan dengan pengelompokan data menjadi kelompok-kelompok yang serupa berdasarkan fitur tertentu. Kelompok data yang memiliki karakteristik yang serupa disebut pola klaster. Algoritma klastering digunakan untuk mengelompokkan data dengan karakteristik yang serupa, seringkali tanpa label atau klasifikasi sebelumnya. Penggunaan pola klaster membantu dalam menemukan pola yang muncul dalam kelompok data yang homogen. Pola klaster ini ditemukan dengan algoritma klastering seperti K-Means dan DBSCAN. Berikut lebih lengkap cara kerja pola klaster:

1. Prinsip dasar pola klaster Data mining menggunakan pola klaster untuk mengelompokkan data ke dalam kelompok yang memiliki karakteristik yang sama. Tujuan utama pembentukan kelompok ini adalah untuk menemukan pola yang tersembunyi dalam data dan memahami struktur dalamannya.
2. Metode pengelompokan digunakan untuk menemukan pola klaster. K-Means, Hierarchical Clustering, DBSCAN, dan Model Mixture Gaussian adalah beberapa algoritma klastering yang paling umum digunakan untuk menemukan pola klaster dalam data.
3. Parameter utama yaitu (1) Jumlah klaster yang dihasilkan oleh algoritma. Jumlah ini dapat ditentukan secara empiris melalui teknik validasi klaster atau berdasarkan pengetahuan domain sebelumnya. (2) Jarak Euclidean, jarak Manhattan, dan jarak Minkowski adalah fungsi jarak yang digunakan untuk mengukur jarak antara titik data. (3) Metode untuk menginisialisasi pusat klaster pada awal iterasi algoritma. Inisialisasi yang buruk dapat berdampak pada konvergensi algoritma dan kualitas klaster yang dihasilkan.
4. Setelah pola klaster telah diidentifikasi, langkah terakhir adalah menerjemahkan dan menerapkan pola tersebut dalam konteks aplikatif yang relevan. Ini dapat mencakup segmentasi pasar, pemahaman perilaku konsumen,

identifikasi anomali, atau pengelompokan data untuk analisis lebih lanjut.

3.2.4 Pola Anomali

Pola anomali adalah pola data yang sangat berbeda dari pola umum dataset. Menemukan pola anomali sangat penting untuk menemukan kejadian langka atau perilaku tidak biasa yang mungkin menunjukkan masalah, seperti kecurangan atau kejadian yang tidak diharapkan. Metode kami secara akurat mendeteksi pola anomali dalam kumpulan data kategorikal, mendeteksi peningkatan signifikan dalam aktivitas anomali di rumah sakit dunia nyata, pengiriman kontainer, dan data intrusi jaringan (7). Pola anomali, juga dikenal sebagai outlier, adalah pola data yang sangat berbeda dari pola umum dalam kumpulan data. Penemuan pola anomali sangat penting untuk deteksi fraud, pemeliharaan prediktif, dan analisis keamanan jaringan.

1. Prinsip dasar pola anomali data mining memanfaatkan pola anomali untuk menemukan data atau observasi yang signifikan yang berbeda dari pola umum kumpulan data. Ini membantu menemukan peristiwa yang tidak biasa, kejadian langka, atau perilaku yang mencurigakan.
2. Terdapat beberapa metode yang umum digunakan untuk mendeteksi pola anomali yaitu (1)Statistik, Metode ini menggunakan pendekatan statistik untuk menentukan apakah suatu observasi tergolong sebagai anomali berdasarkan deviasi dari distribusi normal atau pola umum lainnya. (2)Pemodelan Probabilistik, Metode ini menggunakan model probabilistik untuk mengevaluasi kemungkinan suatu observasi sebagai anomali berdasarkan probabilitasnya dalam distribusi data. (3) Jarak, Metode ini memanfaatkan pengukuran jarak antara observasi dan titik data lainnya. Observasi yang berjarak jauh dari titik data lain mungkin dianggap sebagai anomali. (4) Pemodelan Pembelajaran Mesin: Metode ini melibatkan pembelajaran dari pola umum dalam data dan kemudian mengidentifikasi observasi yang signifikan berbeda dari pola yang telah dipelajari. Pendekatan Kombinasi: Beberapa pendekatan

menggabungkan metode di atas untuk meningkatkan akurasi deteksi anomali.

3. Parameter utama yaitu (1)**Threshold**, aAmbang batas yang digunakan untuk menentukan apakah suatu observasi dianggap sebagai anomali. Observasi yang melewati ambang batas ini dianggap sebagai anomali. (2)**Metrik anomali**, metrik yang digunakan untuk mengukur seberapa signifikan perbedaan observasi dari pola umum dalam data. Metrik ini bisa berupa deviasi standar, nilai p-nilai, atau skor anomali khusus yang dihasilkan oleh algoritma.
4. Langkah terakhir setelah menemukan pola anomali adalah menerjemahkannya dan menerapkannya dalam konteks aplikatif yang sesuai. Deteksi penipuan, pemeliharaan prediktif, penemuan masalah operasional, atau perlindungan keamanan data adalah beberapa contohnya.

3.2.5 Pola Prediktif

Pola prediktif dalam data mining mengacu pada proses mengidentifikasi pola atau hubungan dalam data yang memungkinkan untuk membuat prediksi tentang perilaku atau kejadian di masa depan. Ini merupakan salah satu aspek penting dari data mining yang banyak digunakan dalam berbagai aplikasi, termasuk prediksi bisnis, analisis keuangan, ilmu pengetahuan, kesehatan, dan lainnya. Pola ini digunakan untuk memprediksi nilai-nilai di masa depan dengan menggunakan data sebelumnya. Misalnya, dalam analisis prediksi penjualan, pola prediktif dapat digunakan untuk menemukan tren dan pola musiman dalam penjualan produk. Berikut lebih lengkap cara kerja pola prediktif

1. Pola prediktif bergantung pada gagasan bahwa data historis memiliki informasi yang dapat digunakan untuk memprediksi dan menemukan pola dan tren yang mungkin terjadi di masa depan. Tujuan utama pola prediktif adalah untuk membuat model yang dapat digunakan untuk membuat prediksi yang akurat pada data baru dengan menggunakan data sebelumnya.
2. Untuk mengembangkan model pola prediktif, berbagai teknik pembelajaran mesin digunakan. Regresi, digunakan

ketika variabel target (target) adalah variabel kontinu. Tujuannya adalah untuk memodelkan hubungan antara variabel input dan variabel target. Klasifikasi, digunakan ketika variabel target adalah variabel kategorikal. Tujuannya adalah untuk memprediksi kategori atau kelas dari suatu observasi berdasarkan atribut-atributnya. Jaringan Saraf Tiruan (*Neural Networks*), model yang terinspirasi dari struktur jaringan saraf manusia yang digunakan untuk memodelkan hubungan kompleks antara variabel input dan output. Pohon Keputusan (*Decision Trees*), model yang berbentuk struktur pohon yang digunakan untuk membuat keputusan dengan membagi data menjadi subset yang lebih kecil berdasarkan fitur-fitur tertentu. Metode Pembelajaran Ensemble, teknik yang menggabungkan prediksi dari beberapa model untuk meningkatkan kinerja prediksi.

3. Setelah model prediktif dikonstruksi dan divalidasi, langkah terakhir adalah menerjemahkan hasilnya dan menerapkannya dalam konteks aplikatif yang relevan. Ini bisa berupa pengambilan keputusan berbasis prediksi, pengoptimalan proses bisnis, atau identifikasi peluang bisnis baru.

3.2.6 Pola spasial dan temporal

Pola spasial dan temporal dalam data mining adalah tentang menemukan pola, hubungan, atau tren dalam data yang berkaitan dengan lokasi geografis (spasial) dan waktu (temporal). Pola spasial dan temporal sangat penting dalam berbagai bidang, termasuk geografi, ilmu lingkungan, ilmu sosial, transportasi, dan lainnya. Pola-pola ini terkait dengan analisis data spasial dan temporal, di mana kita mencari pola di ruang dan waktu. Contohnya adalah pola distribusi populasi di suatu wilayah geografis atau pola pergerakan pelanggan di toko ritel. Cara kerja pola spasial dan temporal dalam data mining meliputi:

1. Pola spasial dan temporal melibatkan analisis data yang melibatkan dimensi waktu dan lokasi. Ini membantu untuk mengidentifikasi pola-pola yang berkaitan dengan evolusi fenomena dalam waktu dan ruang, seperti pergerakan

populasi, penyebaran penyakit, pola lalu lintas, atau distribusi geografis dari suatu peristiwa.

2. Data spasial biasanya mencakup koordinat geografis (misalnya, lintang dan bujur) yang terkait dengan lokasi tertentu di permukaan bumi. Data temporal mencakup informasi waktu, seperti tanggal, waktu, atau interval waktu di mana suatu peristiwa terjadi.
3. Berbagai metode yang digunakan dalam analisis pola spasial dan temporal seperti Pemetaan Spasial, memetakan data spasial ke dalam peta atau visualisasi geografis lainnya untuk mengidentifikasi pola spasial dan distribusi geografis dari fenomena tertentu. Analisis Klaster Spasial, mengidentifikasi klaster atau kelompok dari titik data yang berdekatan secara spasial untuk menemukan pola-pola yang signifikan. Analisis Jaringan, menggunakan representasi jaringan dari hubungan spasial antara entitas (misalnya, jalan, pipa, atau koneksi internet) untuk mengidentifikasi pola-pola dalam koneksi atau hubungan antara entitas tersebut. Model Spasial-Temporal, membangun model yang memperhitungkan kedua dimensi spasial dan temporal untuk memahami perubahan dan pola dalam data seiring waktu dan ruang. Prediksi Spasial-Temporal, menerapkan teknik prediktif untuk memprediksi nilai-nilai spasial dan temporal di masa depan berdasarkan data historis.
4. Setelah pola spasial dan temporal telah diidentifikasi, langkah terakhir adalah menerjemahkan dan menerapkan pola tersebut dalam konteks aplikatif yang relevan. Ini bisa berupa perencanaan kota, pemantauan lingkungan, pengendalian epidemi, atau perencanaan rute transportasi.

3.2.7 Pola Multivariat

Pola-pola ini melibatkan berbagai variabel atau karakteristik dalam data, yang membantu memahami hubungan kompleks antara berbagai faktor dalam dataset. Misalnya, dalam analisis kesehatan, kita dapat menemukan pola multivariat tentang hubungan antara variabel seperti diet, olahraga, dan genetika dengan kesehatan seseorang. Dalam pengolahan data multivariat, identifikasi dan

analisis pola atau hubungan antara dua atau lebih variabel yang terdapat dalam kumpulan data diperlukan. Analisis hubungan simultan antara variabel yang berbeda juga termasuk, yang dapat memberikan pemahaman yang lebih dalam tentang data daripada analisis variabel tunggal (variabel tunggal). Cara kerja pola multivariat meliputi:

1. Pola multivariat melibatkan analisis variabel yang lebih dari satu pada waktu yang sama. Tujuannya adalah untuk mengidentifikasi pola, hubungan, atau struktur yang ada di antara variabel-variabel ini. Pola-pola ini dapat berupa korelasi, asosiasi, atau perbedaan di antara grup variabel.
2. Data multivariat terdiri dari beberapa variabel yang diamati bersama-sama untuk setiap observasi dalam kumpulan data. Variabel dapat berupa numerik (misalnya, suhu dan kelembaban udara), kategorikal (misalnya, jenis tanah dan jenis tanaman), atau kombinasi keduanya.
3. Metode analisis yang umum diigunakan yaitu korelasi, regresi, PCA (*Principal Component Analysis*) dan Jalur (*Path Analysis*)
4. Setelah pola multivariat telah diidentifikasi, langkah terakhir adalah menerjemahkan dan menerapkan pola tersebut dalam konteks aplikatif yang relevan. Ini bisa berupa pemahaman lebih dalam tentang hubungan antarvariabel dalam data, prediksi nilai-nilai masa depan, atau pengambilan keputusan yang lebih baik.

DAFTAR PUSTAKA

Hu HL, Chen YL. Mining typical patterns from databases. Inf Sci. 2008 Oct 1;178(19):3683–96.

Delibašić B, Kirchner K, Ruhland J. A Pattern Based Data Mining Approach. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data Analysis, Machine Learning and Applications. Berlin, Heidelberg: Springer; 2008. p. 327–34.

Mustakimtelematika ~. Cara Kerja Data Mining [Internet]. Mustakim Telematika. 2015 [cited 2024 Apr 21]. Available from: https://mustakimtelematika.wordpress.com/2015/03/21/cara-kerja-data-mining/

Aggarwal CC. Association Pattern Mining. In: Aggarwal CC, editor. Data Mining: The Textbook [Internet]. Cham: Springer International Publishing; 2015 [cited 2024 Apr 21]. p. 93–133. Available from: https://doi.org/10.1007/978-3-319-14142-8_4

Shen W, Wang J, Han J. Sequential Pattern Mining. In: Aggarwal CC, Han J, editors. Frequent Pattern Mining [Internet]. Cham: Springer International Publishing; 2014 [cited 2024 Apr 21]. p. 261–82. Available from: https://doi.org/10.1007/978-3-319-07821-2_11

Adams NM, Hand DJ, Till RJ. Mining for classes and patterns in behavioural data. J Oper Res Soc. 2001 Sep 1;52(9):1017–24.

Das K, Schneider J, Neill DB. Anomaly pattern detection in categorical datasets. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2008 [cited 2024 Apr 22]. p. 169–76. (KDD '08). Available from: https://doi.org/10.1145/1401890.1401915

BAB 4
TIPE-TIPE ATRIBUT DATA PADA DATA MINING

Oleh Nahya Nur

4.1 Dataset

Dataset atau himpunan data merupakan kumpulan data yang terorganisir biasanya disajikan dalam bentuk tabel yang memiliki struktur yang terdiri atas baris dan kolom. Setiap baris mewakili satu entitas sedangkan setiap kolom mewakili satu atribut. Setiap entitas dalam dataset dideskripsikan berdasarkan nilai atribut yang berkaitan dengannya (Ha et al., 2011). Contohnya, setiap entitas mahasiswa akan memiliki atribut seperti NIM, Nama, IPK, IPS, jumlah SKS, dan sebagainya.

Atribut

Data

ID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes

Gambar 4.1. Data

4.2 Atribut

Dalam konteks data mining, atribut merujuk pada karakteristik yang digunakan untuk menggambarkan entitas atau

objek dalam sebuah dataset. Terdapat beberapa istilah lain yang sering digunakan untuk merepresentasikan atribut, diantaranya fitur, variabel, maupun dimensi. Istilah dimensi biasanya digunakan dalam *data warehouse*. Istilah fitur cenderung digunakan pada literatur dalam bidang machine. Sedangkan istilah variabel biasanya digunakan dalam bidang statistika. Meskipun isitilah yang digunakan berbeda-beda tetapi istilah-istilah tersebut sama-sama menunjukkan karakteristik dari objek. Atribut dapat berupa berbagai jenis informasi, seperti angka, teks, kategori, waktu, lokasi, dan sebagainya. Setiap baris atau entitas dalam dataset biasanya memiliki nilai untuk setiap atribut yang ada.

Atribut sangat penting dalam proses data mining karena merupakan aspek yang akan dianalisis dan dieksplorasi untuk menemukan pola atau informasi yang berguna dalam data. Tujuan utama dari data mining adalah untuk mengidentifikasi hubungan, tren, atau pola dalam dataset, dan atribut adalah elemen kunci yang digunakan dalam proses ini (Ha et al., 2011).

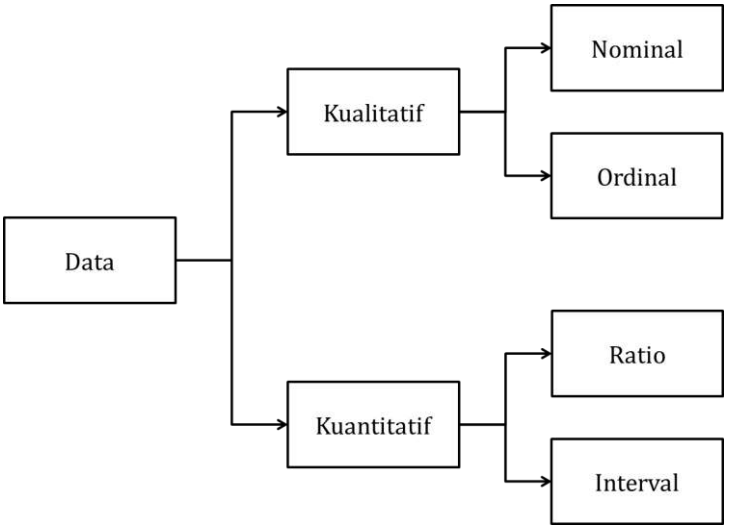
Pemahaman yang baik tentang atribut dalam dataset merupakan langkah penting dalam persiapan data sebelum menerapkan algoritma data mining. Beberapa faktor mengapa pemahaman terkait atribut sangat diperlukan diantaranya (Sudipa et al., 2024)

1. Penyesuaian metode analisis
   Dengan mengetahui tipe dari atribut pada data, pemilihan algoritma dapat dilakukan dengan lebih efisien. Berbagai teknik dan algoritma data mining dirancang untuk bekerja dengan jenis data tertentu. Misalnya, beberapa algoritma seperti *regresi linier* atau *k-means clustering* cocok untuk data numerik, sementara algoritma seperti *decision tree* atau *naive Bayes* cocok untuk data kategorikal (Kellstedt et al., 2022).
2. Pemrosesan data yang tepat
   Jenis atribut juga mempengaruhi langkah-langkah *preprocessing* data yang diperlukan sebelum menerapkan algoritma data mining. Misalnya, data numerik mungkin memerlukan normalisasi untuk mengatasi perbedaan skala, sementara data kategorikal mungkin perlu dikonversi

menjadi representasi biner atau *one-hot encoding*. Dengan mengetahui tipe atribut, kita dapat melakukan *preprocessing* data yang sesuai (Kellstedt et al., 2022).

3. Interpretasi hasil analisis
   Hasil dari analisis data mining seringkali harus diinterpretasikan dan dipahami oleh seseorang yang tidak memiliki latar belakang teknis yang mendalam. Memahami tipe atribut dapat membantu dalam menjelaskan hasil analisis dengan lebih baik. Misalnya, jika hasil klasifikasi menunjukkan bahwa jenis kelamin adalah atribut penting, kita dapat menjelaskan bahwa perbedaan antara laki-laki dan perempuan memiliki pengaruh yang signifikan terhadap hasil tersebut (Kellstedt et al., 2022).
4. Pencegahan kesalahan dalam analisis
   Salah mengidentifikasi tipe atribut dapat menghasilkan analisis yang tidak akurat atau bahkan tidak berguna. Misalnya, menerapkan *regresi linier* pada data kategorikal atau melakukan analisis klastering pada data yang tidak dinormalisasi dapat menghasilkan kesimpulan yang salah. Dengan mengidentifikasi tipe atribut dengan benar, kita dapat mencegah kesalahan semacam ini.
5. Efisiensi komputasi
   Pemilihan metode analisis yang sesuai dengan tipe atribut juga dapat meningkatkan efisiensi komputasi. Algoritma yang dirancang khusus untuk tipe atribut tertentu biasanya lebih efisien dalam memproses data dan menghasilkan hasil dengan waktu yang lebih cepat.

Secara umum, jenis atribut dibedakan menjadi dua bagian besar yaitu atribut kualitatif dan atribut kuantitatif. Pembagian secara lebih luas terkait dua jenis atribut tersebut dapat ditunjukkan pada gambar 4.2 berikut.



Gambar 4.2. Pembagian jenis atribut

Salah satu cara sederhana yang dapat digunakan untuk menentukan tipe suatu atribut adalah dengan mengidentifikasi sifat-sifat bilangan yang sesuai dengan sifat-sifat yang mendasari atribut tersebut. Misalnya, atribut seperti panjang memiliki banyak properti angka. Hal tersebut memungkinkan untuk membandingkan dan mengurutkan objek berdasarkan panjangnya jika membahas terkait perbedaan dan perbandingan panjang. Properti (operasi) angka berikut biasanya digunakan untuk mendeskripsikan atribut.

1. *Distinctness* = dan ≠
2. *Order* <, ≤, >, dan ≥
3. *Addition* + dan −
4. *Multiplication* ∗ dan /

Dengan adanya sifat-sifat ini, kita dapat mendefinisikan empat jenis atribut: nominal, ordinal, interval, dan rasio. Tabel 4.1 menunjukkan definisi dari masing-masing jenis atribut tersebut, beserta informasi tentang operasi statistik yang valid untuk setiap jenis. Setiap tipe atribut memiliki semua properti dan operasi dari tipe atribut di atasnya. Konsekuensinya, setiap properti atau operasi yang berlaku untuk atribut nominal, ordinal, dan interval juga berlaku untuk atribut rasio. Dengan kata lain, definisi tipe atribut

bersifat kumulatif. Namun hal ini tidak berarti bahwa operasi yang sesuai untuk satu atribut juga cocok untuk jenis atribut di atasnya (Kellstedt et al., 2022).

1. Atribut Nominal : *distinctness*
2. Atribut Ordinal : *distinctness* dan *order*
3. Atribut Interval : *distinctness, order,* dan *addition*
4. Atribut Rasio : keseluruhan 4 properti

Beberapa perbedaan mendasar dari tipe-tipe atribut dapat dilihat pada tabel 4.1 berikut :

Tabel 4.1. Perbedaan mendasar tipe atribut

Tipe Atribut		**Deskripsi**	**Contoh**
Kategorikal (Kualitatif)	Nominal	Nilai atribut nominal hanya terletak pada penamaan yang berbeda; dengan kata lain, nilai nominal hanya memberikan informasi yang cukup untuk membedakan satu objek dengan objek lainnya. (= , ≠)	- kode pos - ID mahasiswa - warna mata - jenis kelamin
	Ordinal	Nilai atribut ordinal sudah cukup memberikan informasi untuk mengurutkan objek. (< , >)	- Tingkat kekerasan mineral; {bagus, lebih baik, terbaik}, - Indeks Nilai; {A, B, C, D, E}
Numerik (Kuantitatif)	Interval	Untuk atribut interval, perbedaan antar nilai memiliki makna. (+, -)	- Tanggal kalender - Suhu dalam Celsius atau Fahrenheit
	Ratio	Untuk variabel rasio, selisih dan rasio memiliki makna yang berarti. (*, /)	- Suhu dalam Kelvin, - umur, - massa, - panjang, - arus listrik

Sumber: (Ville, 2001)

Atribut nominal dan ordinal secara kolektif disebut sebagai atribut kategorikal atau kualitatif. Seperti namanya, atribut kualitatif, contohnya ID karyawan. Meskipun direpresentasikan dalam angka, seperti bilangan bulat, ID karyawan harus diperlakukan lebih seperti simbol. Dua jenis atribut lainnya, interval dan rasio, secara kolektif disebut sebagai atribut kuantitatif atau numerik. Atribut kuantitatif diwakili oleh angka dan memiliki sebagian besar properti angka. Perhatikan bahwa atribut kuantitatif dapat bernilai integer atau kontinu (Ville, 2001).

4.3 Atribut Kualitatif

Atribut kualitatif merupakan atribut yang tidak dapat diukur menggunakan angka atau numerik, tetapi lebih diformulasikan dalam kategori atau deskripsi. Atribut pada tipe kualitatif dapat dibedakan menjadi atribut nominal dan ordinal.

4.3.1 Nominal

Nilai atribut nominal berupa simbol atau sesuatu yang berkaitan dengan nama. Setiap nilai mewakili beberapa jenis kategori, kode, kota, dan sebagainya dimana atribut nominal juga disebut sebagai atribut kategorikal. Nilai-nilainya dalam atribut nominal sifatnya setara atau tidak memiliki urutan atau hierarki tertentu diantara datanya. Pada bidang ilmu komputer, nilai tersebut juga dikenal sebagai enumerasi.

Pada penjelasan sebelumnya, disebutkan bahwa atribut nominal diperuntukkan untuk mendefinisikan sesuatu yang berkaitan dengan simbol maupun penamaan, akan tetapi nilai pada atribut ini juga dapat representasikan dalam bentuk angka. Selain itu, beberapa algoritma hanya dapat memproses data dalam bentuk numerik (Nur et al., 2023). Sebagai contoh atribut warna rambut seperti hitam, cokelat, dan sebagainya dapat direpresentasikan sebagai 0 untuk warna hitam, 1 untuk warna cokelat, dan seterusnya. Meski demikian, angka-angka tersebut tidak dimaksudkan untuk digunakan secara kuantitatif. Dengan kata lain, operasi matematika untuk nilai pada atribut nominal tidak dapat dilakukan meskipun dituliskan dalam bentuk angka. Karena nilainya tidak dapat diurutkan dan bukan merupakan nilai kuantitatif

sehingga tidak dapat digunakan untuk menghitung nilai rata-rata maupun nilai tengah dari atribut tipe ini (Ha et al., 2011).

Pada atribut nominal, terdapat satu bentuk khusus yaitu atribut biner. Sesuai dengan namanya, atribut nominal dalam bentuk atribut biner hanya memiliki dua nilai seperti ya dan tidak, *true* dan *false,* dan sebagainya (Aggarwal, 2015). Contoh penggunaan atribut biner seperti dalam merepresentasikan status pasien yang hanya memiliki dia kemungkinan nilai. Misalnya nilai 1 menyatakan hasil tes pasien positif, sedangkan nilai 0 berarti hasil yang diperoleh negatif.

Sebuah atribut biner dikatakan simetris jika kedua nilai pada atribut tersebut dianggap sama pentingnya dan memiliki bobot (*jumlah tuple)* yang seimbang. Dengan kata lain tidak ada preferensi nilai yang mana yang harus dikodekan 1 atau 0, seperti atribut *gender.* Untuk atribut asimetris, nilai atribut yang bukan nol merupakan nilai yang dianggap lebih penting. Misalnya terdapat suatu dataset dimana setiap objek adalah mahasiswa dan memiliki atribut yang merepresentasikan mahasiswa mengambil mata kuliah tertentu. Pada mahasiswa tertentu, atribut akan bernilai 1 jika mengambil mata kuliah dam akan bernilai 0 jika sebaliknya. Karena mahasiswa hanya mengambil sebagian kecil dari semua mata kuliah yang tersedia, maka sebagian besar nilai dalam dataset tersebut adalah 0. Oleh karena itu, akan lebih bermakna dan efisien jika berfokus pada nilai yang bukan nol. Dalam mengkodekan atribut asimetris, kode 1 diberikan kepada nilai yang dianggap lebih penting atau yang menjadi perhatian dalam pengamatan (Ville, 2001). Contoh atribut nominal dapat ditunjukkan pada tabel 4.2.

Tabel 4.2. Contoh dataset dengan atribut nominal

No	Nama	Jenis Kelamin	Pekerjaan	Kode Pos
1.	Ani	Perempuan	Dokter	91311
2.	Budi	Laki-laki	Programmer	91214
3.	Cindi	Perempuan	Dosen	91315
4.	Deni	Laki-laki	Petani	91421
5.	Enni	Perempuan	Guru	92312

Atribut nama, jenis kelamin, pekerjaan, dan kode pos merupakan beberapa contoh atribut nominal. Meskipun atribut nominal merupakan salah satu jenis atribut kualitatif, atribut jenis ini juga dapat direpresentasikan dalam bentuk angka.

4.3.2 Ordinal

Atribut ordinal merupakan jenis atribut di mana nilai-nilainya memiliki urutan atau peringkat yang bermakna. Akan tetapi, besaran perbedaan antara nilai-nilai yang berurutan tersebut tidak diketahui. Atribut ordinal juga dapat diperoleh dari proses diskritisasi dari nilai numerik dengan membagi rentang nilai menjadi sejumlah kategori terurut yang terbatas (Ha et al., 2011).

Tabel 4.3. Contoh dataset dengan atribut ordinal

No	Nama	Nilai Skripsi	Predikat Kelulusan
1	Ani	A	Cumlaude
2	Budi	C	Cukup memuaskan
3	Cindi	B	Memuaskan
4	Deni	A	Cumlaude
5	Enni	B	Sangat memuaskan

Pada Tabel 4.3, nilai skripsi dan predikat kelulusan merupakan contoh dari atribut ordinal, dimana nilainya dapat ditentukan yang mana yang lebih baik. Sebagai contoh, Nilai A lebih baik dari nilai B dan C. Selain itu, predikat *cumlaude* lebih bagus dibandingkan dengan predikat lainnya.

4.4 Atribut Kuantitatif

Atribut kuantitatif adalah atribut yang dapat diukur secara numerik. Atribut pada tipe ini dapat diwakili dalam bilangan bulat atau real. Atribut kuantitatif memungkinkan kita untuk melakukan analisis lebih lanjut, melakukan perhitungan matematis, dan membandingkan secara objektif antara berbagai kondisi atau objek.

Atribut kuantitatif dapat dibedakan menjadi atribut berskala interval dan atribut berskala ratio.

4.4.1 Interval

Nilai-nilai dari atribut berskala interval memiliki urutan dan dapat bernilai positif, 0, atau negatif. Jadi, sebagai tambahan untuk memberikan peringkat nilai, atribut tersebut dimungkinkan untuk dibandingkan dan diukur perbedaan antar nilainya.

Salah satu contoh atribut berskala interval adalah temperatur dalam satuan Celcius dan Fahrenfeit dimana pengukuran temperatur pada satuan tersebut tidak memiliki titik referensi nol, artinya nilai 0^o C dan 0^o F tidak menunjukkan bahwa tidak ada temperatur atau dengan kata lain bukan merupakan nilai terendah dari pengukuran yang dimaksud.

4.4.2 Ratio

Atribut dengan jenis ratio merupakan salah satu atribut numerik yang memiliki titik referensi nol. Jika suatu pengukuran berskala rasio, kita dapat mengatakan suatu nilai sebagai kelipatan (atau rasio) dari nilai lain. Selain itu, nilainya dapat diurutkan, dan perbedaan antara nilai, serta mean, median, dan modus dapat dihitung.

Tidak seperti Celcius dan Fahreinheit, temperatur dalam Kelvin merupakan salah satu contoh atribut berskala ratio sebab temparatur pada skala tersebut memiliki titik referensi nol, 0^oK yang menunjukkan ketiadaan energi termal. Contoh lain dari atribut ini, diantaranua pengukuran berat, tinggi, latitude, longitude, dan sebagainya.

4.5 Atribut Diskrit dan Kontinu

Dalam melakukan analisis data statistika, atribut dapat dikategorikan menjadi dua (Rabbi Radliya, 2016), yaitu

1. Atribut Diskrit
   Atribut diskrit merupakan jenis atribut yang memiliku nilai yang terbatas dalam suatu himpunan. Biasanya, atribut diskrit ditemukan pada atribut kategorikal yang hanya mempunyai beberapa variasi nilai, seperti jenis kelamin

(perempuan, laki-laki), predikat kelulusan mahasiswa (*cumlaude*, sangat memuaskan, memuaskan, cukup memuaskan, kurang), 0/1, dan sebagainya

2. Atribut Kontinu
   Atribut kontinu memiliki jangkauan nilai real dan memiliki rentang yang tidak terbatas. Misalnya, variabel seperti panjang, tinggi, dan berat sering kali direpresentasikan dalam bentuk angka desimal (*floating point*). Meskipun nilai-nilai ini menggunakan representasi real, tingkat presisi dengan jumlah angka di belakang koma biasanya tetap diperhitungkan.

Pemahaman terkait perbedaan antara atribut diskrit dan atribut kontinu penting dalam analisis statistik dan pemodelan data, karena metode analisis yang digunakan sering kali berbeda tergantung pada jenis atribut yang dihadapi. Misalnya, dalam regresi linear, atribut kontinu sering kali lebih cocok daripada atribut diskrit, sementara untuk pemanfaatan algoritma klasifikasi, atribut diskrit cenderung lebih sesuai.

DAFTAR PUSTAKA

Aggarwal, C. C. (2015). Data Mining: The Textbook. In *Springer International Publishing*. https://doi.org/10.1007/978-3-319-14142-8_10

Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Kellstedt, P. M., Whitten, G. D., & Tuch, S. A. (2022). Getting to Know Your Data. *The Fundamentals of Social Research*, 114–130. https://doi.org/10.1017/9781316415399.008

Nur, N., Wajidi, F., Sulfayanti, S., & Wildayani, W. (2023). Implementasi Algoritma Random Forest Regression untuk Memprediksi Hasil Panen Padi di Desa Minanga. *Jurnal Komputer Terapan*, *9*(1), 58–64. https://doi.org/10.35143/jkt.v9i1.5917

Rabbi Radliya, N. (2016). *DATA MINING* (Issue 321, pp. 1–15).

Sudipa, I. G. I., Darmawiguna, I. G. M., Dendi, I. M., & Sanjaya, M. (2024). *Buku ajar data mining* (Issue January).

Ville, B. de. (2001). Introduction to Data Mining. In *Microsoft Data Mining*. https://doi.org/10.1016/b978-155558242-5/50003-6

BAB 5
STATISTIK DESKRIPTIF YANG DIBUTUHKAN DALAM DATA MINING

Oleh Sulfayanti

5.1 Statistik Deskriptif

Pemrosesan data dapat dilakukan dengan baik dengan memanfaatkan gambaran umum data yang dimiliki. Statistik deskriptif dasar digunakan untuk menyelesaikan hal ini dengan mengidentifikasi properti-properti data lalu menunjukkan nilai data yang di anggap sebagai noise (gangguan) atau outlier (poin data yang berbeda secara signifikan dari mayoritas). Pengetahuan terkait statistik deskriptif meliputi pengukuran tendensi sentral, penilaian penyebaran data, serta grafik untuk menampilkan visualisasi hasil inspeksi statistik deskriptif dasar terhadap data. Pengukuran tendensi sentral yaitu pengukuran digunakan untuk mengetahui lokasi tengah atau pusat dari distribusi data, sedangkan pengukuran penyebaran data digunakan untuk mengetahui seberapa dekat elemen data satu sama lain dan untuk mengidentifikasi noise atau outlier (Han, Kamber and Pei, 2012; Anderson and Semmelroth, 2017).

5.2 Pengukuran Tendensi Sentral

Metode yang dapat digunakan untuk mengetahui pusat dari data, diantaranya yaitu: mean, median, modus dan midrange. Misalkan terdapat beberapa atribut X, seperti pendapatan yang mengandung sekumpulan unsur objek. Maka $x_1, x_2, \cdots, x_N$ dapat digunakan untuk menyatakan sejumlah N nilai untuk X. Nilai-nilai ini menunjukkan himpunan data dari X. Pengamatan terhadap data yang paling banyak muncul dari data pendapatan dapat dilakukan dengan pengukuran pusat data seperti mean, median, modus dan midrange (Han, Kamber and Pei, 2012).

5.2.1 Mean

Di dalam statistik, mean bersinonim dengan rata-rata. Metode pengukuran ini adalah yang paling umum dan efektif untuk menentukan pusat data. Perhitungan mean dilakukan dengan menjumlahkan semua elemen dalam himpunan data lalu kemudian membaginya dengan jumlah banyaknya elemen atau data (Anderson and Semmelroth, 2017). Jika $x_1, x_2, \cdots, x_N$ adalah himpunan N nilai atau observasi untuk beberapa atribut numerik X seperti pendapatan. Maka, secara umum mean dari himpunan nilai-nilai tersebut adalah

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{5.1}$$

dimana $\bar{x}$ (dibaca x bar) adalah mean dari data, i adalah indeks yang menunjukkan label bilangan untuk setiap elemen dalam data dalam rentang 1 sampai N dan x_i adalah nilai elemen dalam data. Mean dari sampel data biasa dituliskan menggunakan $\bar{x}$ sedangkan mean dari populasi data dituliskan dengan μ (alpabet Yunani mu).

Contoh 5.1: Misalkan nilai-nilai yang terdapat pada *salary*(dalam nilai ribuan dollar) dituliskan dalam urutan dari yang terkecil ke terbesar seperti berikut: $25, 26, 31, 43, 47, 47, 51, 55, 65, 65, 76, 105$. Berdasarkan persamaan (5.1) diperoleh nilai mean sebagai berikut

$$\bar{x} = \frac{25+26+31+43+47+47+51+55+65+65+76+105}{12}$$
$$= \frac{636}{12} = 53$$

Nilai di atas menunjukkan mean *salary* sebesar \$53.000.

Nilai x_i dalam sebuah himpunan kadangkala berasosiasi dengan sebuah bobot w_i untuk $i = 1, \cdots, N$. Bobot merefleksikan tingkat signifikansi, kepentingan, atau frekuensi kemunculan untuk setiap nilai. Dalam kasus ini, nilai mean dapat dihitung menggunakan persamaan berikut

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \tag{5.2}$$

Persamaan di atas disebut juga sebagai *weighted arithmetic mean* atau *weighted average*.

Masalah yang seringkali ditemukan pada *mean* adalah sensitivitasnya terhadap nilai ekstrim. Misalkan terdapat sebuah

bilangan kecil yang merupakan nilai outlier, hal ini dapat memperburuk nilai *mean* yang diperoleh. Untuk mengimbangi efek yang disebabkan oleh sejumlah nilai outlier, metode *trimmed mean* (mean terpangkas) dapat digunakan yaitu dengan menghitung nilai mean setelah memotong/membuang nilai-nilai ekstrim yang tinggi ataupun rendah (Han, Kamber and Pei, 2012).

5.2.2 Median

Median dalam data merujuk kepada nilai tengah data. Untuk data yang asimetris, lebih baik menggunakan median sebagai pengukuran pusat data karena nilai tengah data dicari dalam himpunan nilai-nilai yang telah diurutkan. Nilai tengah yang diperoleh adalah nilai yang memisahkan kumpulan data yang memiliki nilai tinggi-nilai di atas median dengan kumpulan data yang memiliki nilai rendah-nilai dibawah median (Wilcox, 2010; Anderson and Semmelroth, 2017).

Dalam probabilitas dan statistik, kumpulan data sejumlah N nilai untuk atribut X diurutkan dari nilai terkecil ke yang terbesar. Jika N ganjil, maka median adalah nilai tengah dari himpunan terurut. Jika N genap, mediannya tidak unik tapi merupakan nilai yang berada di antara dua nilai paling tengah. Jika X adalah atribut numerik, maka median diperoleh dari rata-rata dari dua nilai paling tengah (Han, Kamber and Pei, 2012).

Contoh 5.2: Misalkan dilakukan pengukuran median dari data pada Contoh 5.1. Data telah terurut dari yang terkecil ke terbesar dan total berjumlah genap yaitu 12 sehingga median tidak unik. Nilai median berada di antara dua nilai paling tengah dari data yaitu data ke-6 dan ke-7 yaitu 47 dan 51. Nilai rata-rata dari kedua nilai paling tengah inilah yang merupakan nilai median yaitu $\frac{47+51}{2} = \frac{98}{2} = 49$ sehingga mediannya adalah \$49.000.

Misalkan jika data yang diberikan hanya terdiri dari 11 nilai awal. Data akan berjumlah ganjil dan mediannya adalah nilai yang berada paling tengah. Dalam hal ini, nilai paling tengah berada pada data ke-6 yaitu \$47.000.

Perhitungan median akan sulit dilakukan pada data observasi yang berjumlah banyak. Namun untuk atribut numerik, perhitungan nilai median tetap memungkinkan untuk dilakukan

dengan mudah dengan asumsi bahwa data di grupkan dalam interval-interval berdasarkan nilai-nilai data x_i dan frekuensi (jumlah nilai data) untuk masing-masing interval yang diketahui. Interval yang mengandung frekuensi median menjadi *interval median*. Perhitungan median dari seluruh data dapat diketahui dengan melakukan interpolasi menggunakan persamaan berikut:

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width, \qquad (5.3)$$

dimana L_1 adalah batas bawah dari interval median, N adalah jumlah nilai-nilai dalam keseluruhan data, $(\sum freq)_l$ adalah jumlah frekuensi dari semua interval yang lebih rendah dari *interval median*, $freq_{median}$ adalah frekuensi dari interval median, dan $width$ adalah panjang dari interval median (Han, Kamber and Pei, 2012).

5.2.3 Modus

Modus (atau *mode*) adalah pengukuran pusat sentral dengan mengambil nilai yang paling sering muncul dalam himpunan data. Pengukuran modus dapat dilakukan dengan mengurutkan data secara menaik lalu menghitung masing-masing frekuensi nilai dalam data (Bernstein and Bernstein, 1999; Anderson and Semmelroth, 2017). Besar kemungkinan bahwa beberapa nilai berbeda memiliki frekuensi yang sama, sehingga dapat diperoleh lebih dari satu modus. Himpunan-himpunan data dengan satu, dua, atau tiga modus masing-masing disebut sebagai unimodal, bimodal, dan trimodal. Sebuah himpunan data dengan dua atau lebih modus disebut sebagai multimodal. Dalam kasus yang lain, jika nilai data hanya muncul dengan frekuensi satu atau jumlah frekuensi sama maka tidak terdapat modus dalam data tersebut (Han, Kamber and Pei, 2012).

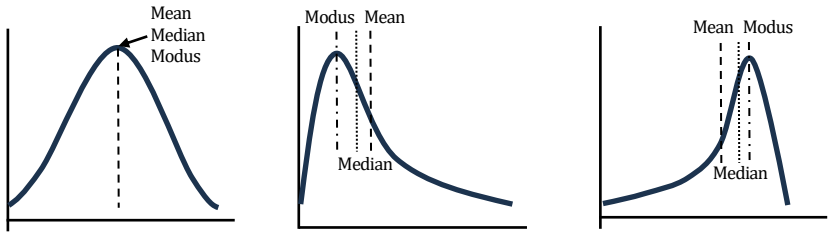
Contoh 5.3: Berdasarkan data contoh 5.1, himpunan data tersebut adalah bimodal. Dua modus tersebut adalah $47.000 dan $65.000.

5.2.4 Midrange

Midrange merupakan nilai rata-rata yang diperoleh menggunakan nilai terbesar dan terkecil dalam himpunan data (Bernstein and Bernstein, 1999).

Contoh 5.4: Midrange dari data Contoh 5.1 adalah $\frac{25.000+105.000}{2} =$ \$65.000.

Gambar 5.1 menunjukkan posisi mean, median, modus pada masing-masing kurva frekuensi unimodal dengan distribusi data simetrik dan asimetrik (Han, Kamber and Pei, 2012).



(a) Data simetrik (b) Data asimetrik cenderung positif (c) Data asimetrik cenderung negatif

Gambar 5.1. Mean, median, dan mode pada data simetrik dan data asimetrik
Sumber: (Han, Kamber and Pei, 2012)

Berikut ini adalah beberapa hal yang dpat disimpulkan terkait metode pengukuran pusat data (Setiawan, 2024):

1. Median atau modus lebih baik digunakan pada kondisi dimana distribusi frekuensi data tidak normal atau tidak simetris atau apabila terdapat nilai-nilai outlier, baik yang lebih kecil ataupun besar, sedangkan
2. Secara umum, mean, median dan modus dapat digunakan jika distribusi data normal atau simetris. Namun, mean lebih sering memenuhi persyaratan untuk pengukuran pusat yang baik.
3. Midrange dapat digunakan jika data yang diamati berkaitan dengan laju, kecepatan atau harga. Sedangkan midrange (rata-rata geometrik) lebih baik digunakan pada data yang mengalami perubahan relatif, seperti dalam kasus pertumbuhan bakteri, pembelahan sel dan sebagainya.

5.3 Perhitungan Penyebaran Data

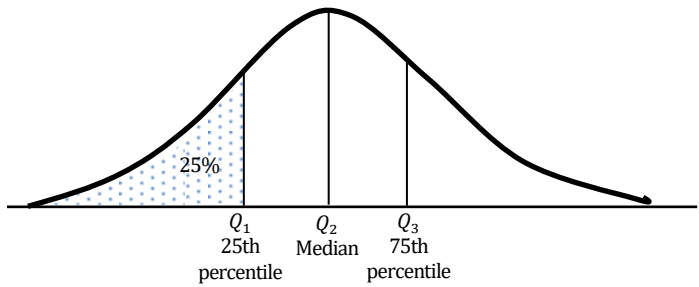
Pengukuran penyebaran data numerik dapat dilakukan melalui perhitungan range, quantile, quartile, percentil dan range interquartile. Kelima nilai ini dapat di tampilkan dalam sebuah boxplot yang berguna untuk mengidentifikasi noise. Variansi dan standar deviasi juga untuk mengidentifikasikan penyebaran dalam distribusi data (Han, Kamber and Pei, 2012).

5.3.1 Range, Quartile, dan Interquartile Range

Misalkan $x_1, x_2, \cdots, x_N$ adalah himpunan data untuk beberapa atribut, X. Range adalah perbedaan antara nilai terbesar(maksimum) dengan nilai terkecil (minimum).

Misalkan data pada atribut X telah diurutkan dari yang terkecil ke yang terbesar dan akan dipilih suatu titik data untuk membagi distribusi data dalam ukuran yang sama sebagaimana Gambar 5.2. Titik-titik data ini yang disebut dengan quantile, yaitu nilai yang diambil pada interval reguler dari distribusi data sehingga membagi data ke dalam himpunan data yang berturut-turut berukuran sama. (Han, Kamber and Pei, 2012).

Quantile-2 adalah titik data yang membagi dua distribusi data menjadi data bagian bawah dan bagian atas. Quantile-2 bersesuaian dengan median. Quantile-4 adalah tiga titik data yang membagi distribusi data menjadi empat bagian yang sama; setiap bagian merepresentasikan seperempat distribusi data. Quantile-4 biasa juga disebut sebagai quartile. Quantile-100 lebih sering disebut sebagai percentile, dimana distribusi data dibagi menjadi 100 set berturut-turut dan berukuran sama. Median quartile dan percentile adalah bentuk quantile yang paling banyak digunakan (Han, Kamber and Pei, 2012).



Gambar 5.2. Distribusi data pada beberapa atribut *X*

Quartile mampu menggambarkan pusat, persebaran, dan bentuk distribusi. Quartile pertama, dinotasikan dengan Q_1, adalah persentil ke-25. Quartile ini mengambil 25% data. Quartile ketiga, dinotasikan dengan Q_3, merupakan persentil ke-75. Quartile ini mengambil sekitar 75%(atau 25% teratas) dari data. Quartile kedua adalah persentil ke-50. Quartile ini sekaligus sebagai median yang dapat memberikan pusat distribusi data (Han, Kamber and Pei, 2012).

Jarak antara quartile pertama dan ketiga adalah ukuran penyebaran sederhana yang memberikan rentang yang mencakup separuh tengah data. Jarak ini disebut sebagai interquartile range(IQR) dan didefinisikan sebagai

$$IQR = Q_3 - Q_1 \tag{5.5}$$

Interquartile range adalah tiga nilai yang membagi himpunan data terurut menjadi empat bagian yang sama (Han, Kamber and Pei, 2012).

Contoh 5.5: Berdasarkan Contoh 5.1 terdapat 12 data observasi yang telah diurutkan dan dibagi ke dalam empat bagian. Quartile untuk data ini adalah nilai data ke-3, ke-6, dan ke-9, sehingga $Q_1 = \$31.000$ dan $Q_3 = \$65.000$. Interquartile range yaitu $IQR = 65.000 - 31.000 = \$34.000$.

5.3.2 Five-Number Summary, Boxplots, and Ouliers

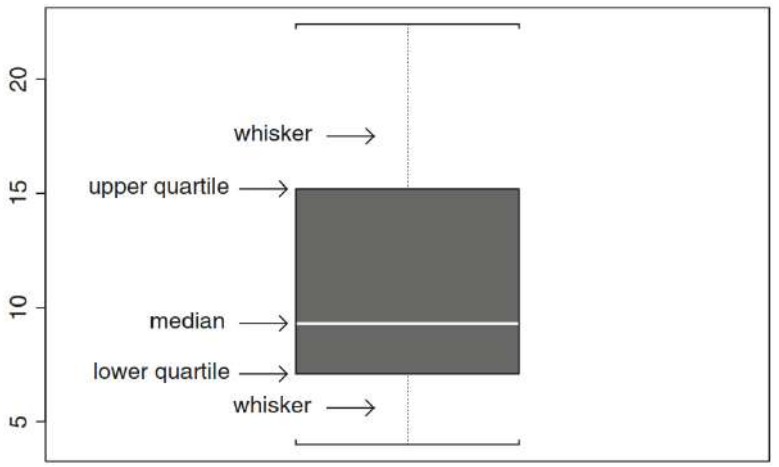
Gambar 5.1 akan lebih informatif jika disediakan dua quartile Q_1 dan Q_3 bersama dengan median. Aturan umum untuk mengidentifikasi outlier adalah dengan memilih nilai-nilai yang

beradasetidaknya $1{,}5 \times IQR$ di atas quartile ketiga atau di bawah quartile pertama.

Karena Q_1, median, dan Q_3 tidak memiliki informasi tentang titik akhir data, ringkasan bentuk distribusi data yang lebih lengkap dapat diperoleh dengan memberikan nilai data terendah dan juga data tertinggi. Ini dikenal sebagai ringkasan lima angka (Five-number summary). Five-number summary suatu distribusi terdiri dari median (Q_2), quartile bawah Q_1 dan quartile atas Q_3, serta observasi nilai terkecil dan terbesar, jika ditulis berdasarkan urutan yaitu: Nilai minimum, Q_1, Median, Q_3, Maksimum pertama (Han, Kamber and Pei, 2012).

Boxplot merupakan cara yang banyak digunakan untuk memvisualisasikan distribusi menggunakan sebuah grafik. *Boxplot* menggabungkan Five-number summary sebagai berikut:

1. Biasanya ujung-ujung kotak berada pada quartile sehingga panjang kotak merupakan rentang antar quartile.
2. Median ditandai dengan garis di dalam kotak.
3. Dua garis yang disebut whiskers, di luar kotak meluas ke pengamatan terkecil (Minimum) dan terbesar (Maksimum).



Gambar 5.3. Contoh Boxplot
Sumber: (Wilcox, 2010)

5.3.3 Variansi dan Standar Deviasi

Variansi dan standar deviasi adalah pengukuran penyebaran data yang dapat menunjukkan seberapa tersebarnya distribusi data. Variansi adalah rata-rata aritmatika dari kuadrat deviasi mean data. Variansi dari sejumlah N data observasi $x_1, x_2, \cdots, x_N$ untuk atribut numerik X adalah

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2, \qquad (5.6)$$

dimana $\bar{x}$ adalah nilai mean data observasi yang didefinisikan dalam persamaan 5.1. Standar deviasi, σ, adalah akar dari variansi σ^2.

Standar deviasi yang rendah berarti bahwa data observasi cenderung mendekati mean, sedangkan standar deviasi yang tinggi menunjukkan bahwa data tersebar pada rentang nilai yang luas (Bernstein and Bernstein, 1999; Han, Kamber and Pei, 2012).

Contoh 5.6: Pada contoh 5.1, diketahui $\bar{x} = \$53.000$ dengan menggunakan persamaan 5.1 untuk mean. Untuk menentukan variansi dan standar deviasi data dari contoh tersebut, ditetapkan $N = 12$ dan dengan menggunakan persamaan 5.6 diperoleh

$$\begin{aligned}\sigma^2 &= \frac{1}{12}(25^2 + 26^2 + \cdots + 105^2) - 53^2 \\ &\approx 882{,}83 \\ \sigma &\approx \sqrt{882{,}83} \approx 29{,}71\end{aligned}$$

Karateristik dasar dari standar deviasi, σ, sebagai ukuran penyebaran adalah sebagai berikut:

1. σ mengukur sebaran di sekitar mean dan harus dipertimbangkan hanya ketika mean dipilih sebagai ukuran pusat.
2. $\sigma = 0$ hanya bila tidak ada penyebaran, yaitu bila semua data observasi mempunyai nilai yang sama. Jika tidak, $\sigma > 0$.

5.4 Grafik Statistik Deskriptif Dasar

Visualisasi grafis dari deskripsi statistik dasar yang akan dibahas mencakup quantile plot, quantile-quantile plot, histogram, dan scatter plot. Grafik-grafik ini berguna untuk pemeriksaan visual data, yang berguna dalam pemrosesan awal data. Tiga plot pertama untuk visualisasi distribusi univariat (yaitu, data untuk satu atribut),

sedangkan scatter plot untuk visualisasi distribusi bivariat (yaitu, melibatkan dua atribut).

5.4.1 Quantile plot

Untuk mendapatkan Quantile plot, pertama, plot akan menampilkan semua data untuk atribut tertentu. Kedua, informasi quantile akan diplot. Misal x_i, untuk $i = 1$ sampai N, adalah data yang diurutkan secara menaik sehingga x_1 adalah observasi terkecil dan x_N adalah observasi terbesar untuk beberapa atribut ordinal atau numerik X. Setiap observasi, x_i, dipasangkan dengan persentase, f_i, yang menunjukkan bahwa sekitar $f_i \times 100\%$ data berada di bawah nilai, x_i. Terdapat 0,25 persentil mewakili quartile Q_1, 0,50 persentil adalah median, dan 0,75 persentil adalah Q_3. Sehingga

$$f_i = \frac{i-0,5}{N} \tag{5.7}$$

Angka-angka ini meningkat secara bertahap sebesar $1/N$. Pada plot quantile, x_i dibuat grafiknya terhadap f_i. Hal ini memungkinkan untuk membandingkan distribusi yang berbeda berdasarkan quantilenya (Han, Kamber and Pei, 2012).

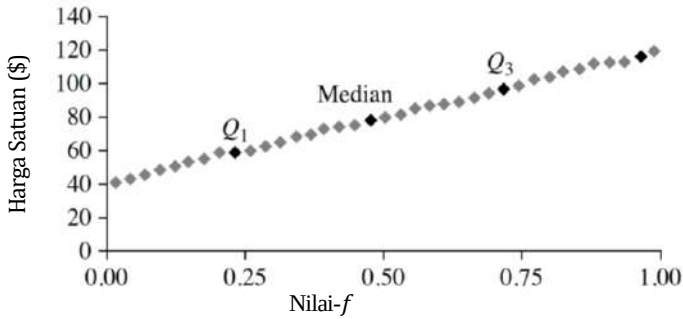
Contoh 5.7: Diberikan data sebagai berikut:

Tabel 5.1. Himpunan data harga untuk item yang terjual pada salah satu cabang Toko Elektronik

Harga Satuan($)	Jumlah Item yang Terjual
40	275
43	300
47	250
-	-
74	360
75	515
78	540
-	-
115	320
117	270
120	350

Sumber: (Han, Kamber and Pei, 2012)

Quantile plot dari Tabel 5.1 terdapat pada Gambar 5.4:

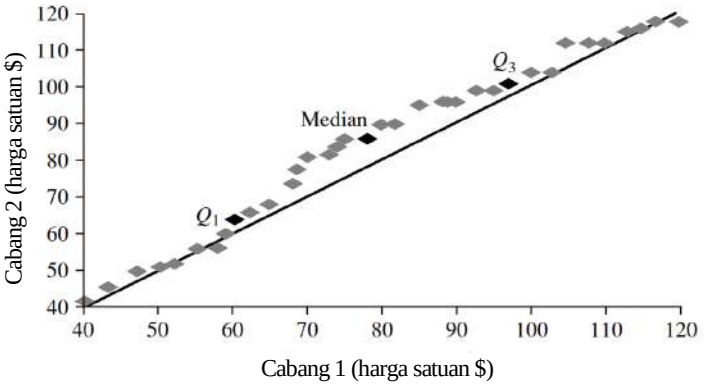


Gambar 5.4. Quantile plot Data Tabel 5.1
Sumber: (Han, Kamber and Pei, 2012)

5.4.2 Quantile-quantile plot

Sebuah quantile-quantile plot (q-q plot), menggambarkan grafik quantile dari satu distribusi univariat terhadap quantile yang bersesuaian di distribusi univariat lainnya. Misalkan kita mempunyai dua himpunan data observasi untuk atribut atau variabel harga satuan, yang diambil dari dua lokasi cabang yang berbeda. Misal $x_1, \cdots, x_N$ adalah data dari cabang pertama, dan $y_1, \cdots, y_m$ adalah data dari cabang kedua, yang setiap kumpulan datanya diurutkan secara menaik. Jika $M = N$ (yaitu, jumlah titik-titik pada masing-masing himpunan adalah sama), maka cukup memplot y_i terhadap x_i, dimana y_i dan x_i keduanya $(i - 0{,}5)/N$ quantile dari kumpulan datanya masing-masing. Jika $M < N$ (yaitu, cabang kedua mempunyai observasi yang lebih sedikit dibandingkan cabang pertama), maka yang ada hanya titik M pada q-q plot. Di sini, y_i adalah $(i - 0{,}5)/M$ quantile y data, yang di plot terhadap $(i - 0{,}5)/M$ quantile x data (Han, Kamber and Pei, 2012).
Contoh 5.8: Gambar 5.5 menunjukkan plot quantile-quantile untuk data harga satuan barang yang dijual di dua cabang Toko elektronik selama periode waktu tertentu. Setiap titik sesuai dengan quantile yang sama untuk masing-masing himpunan data dan menunjukkan harga satuan barang yang dijual di cabang 1 dan cabang 2. Garis lurus mewakili kasus di mana, untuk setiap quantile tertentu, harga

satuan di setiap cabang adalah sama. Titik yang lebih gelap masing-masing sesuai dengan data untuk Q_1, median, dan Q_3.



Gambar 5.5. q-q plot untuk data harga item 2 cabang Toko elektronik
Sumber: (Han, Kamber and Pei, 2012)

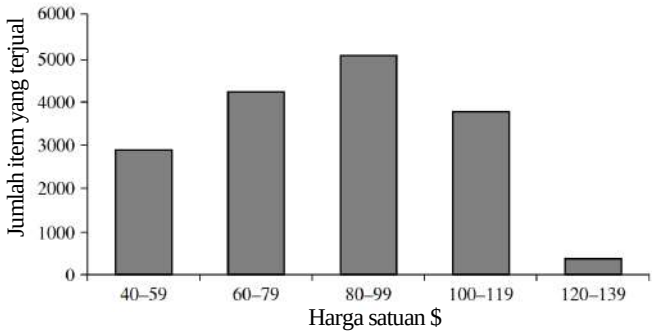
Dapat dilihat, Misalnya, bahwa pada Q_1, harga satuan barang yang dijual di cabang 1 sedikit lebih rendah dibandingkan harga satuan di cabang 2. Dengan kata lain, 25% barang yang dijual di cabang 1 kurang dari atau sama dengan \$60, sedangkan 25% item terjual pada Cabang 2 lebih sedikit dari atau sama dengan \$64. Secara umum, terdapat pergeseran distribusi Cabang 1 terhadap Cabang 2 dimana harga satuan barang yang dijual di Cabang 1 cenderung lebih rendah dibandingkan dengan harga satuan di Cabang 2.

5.4.3 Histogram

Histogram (atau histogram frekuensi) banyak digunakan secara luas. "Histos" berarti tiang, dan "gram" berarti bagan, jadi histogram adalah bagan dari tiang. Histogram adalah metode grafis untuk meringkas distribusi suatu atribut tertentu. Setiap bar merepresentasikan kategori tunggal yang mengandung sebuah nilai atau range nilai. Tinggi bar merepresentasikan frekuensi dari masing-masing kategori (Anderson and Semmelroth, 2017).

Contoh 5.9: Gambar 5.6 berikut menunjukkan sebuah histogram dari data yang terdapat pada Tabel 5.1, dimana setaip bar

ditentukan oleh rentang lebar yang sama yang mewakili kenaikan $20 dan frekuensinya adalah jumlah barang yang terjual.



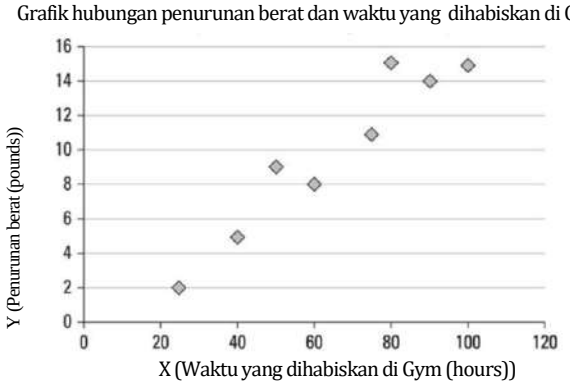
Gambar 5.6. Histogram untuk data harga satuan item dari Tabel 5.1
Sumber: (Han, Kamber and Pei, 2012)

Meskipun histogram banyak digunakan, namun histogram tidak seefektif metode plot quantile, q-q plot, dan boxplot untuk membandingkan grup-grup dalam observasi univariat.

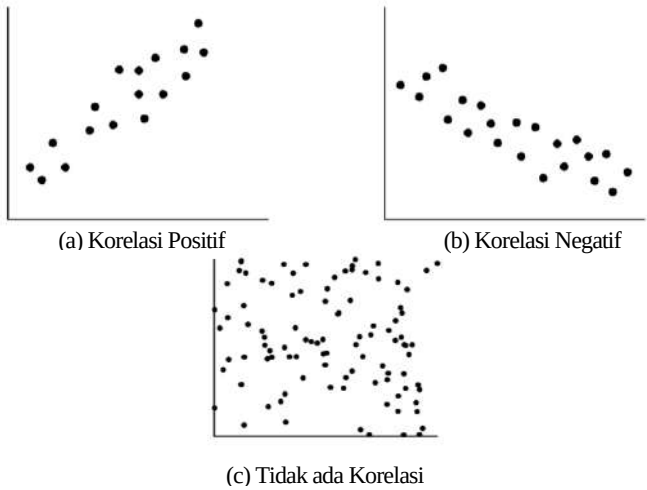
5.4.4 Scatter Plot dan korelasi data

Scatter plot adalah salah satu metode visualisasi grafis paling efektif untuk menentukan adanya hubungan, pola, atau tren antara dua atribut numerik. Untuk membuat scatter plot, setiap pasangan nilai diperlakukan sebagai pasangan koordinat dalam aljabar dan diplot sebagai titik-titik pada bidang. Gambar 5.7 menunjukkan contoh scatter plot.

Scatter plot adalah metode yang berguna untuk memberikan pandangan pertama pada data bivariat untuk memperoleh klaster titik dan outlier, atau untuk mengeksplorasi kemungkinan hubungan korelasi. Dua atribut, X dan Y, berkorelasi jika satu atribut mengimplikasikan atribut lainnya. Korelasi bisa positif, negatif, atau nol (tidak berkorelasi).



Gambar 5.7. Contoh scatter plot untuk data
Sumber: (Anderson and Semmelroth, 2017)



Gambar 5.8. Scatter plot yang dapat digunakan untuk menemukan ada tidaknya korelasi diantara atribut
Sumber: (Han, Kamber and Pei, 2012)

Gambar 5.8 menunjukkan contoh korelasi positif dan negatif antara dua atribut. Jika pola titik yang diplot miring dari kiri bawah ke kanan atas, ini berarti nilai *X* meningkat seiring dengan meningkatnya nilai Y, menunjukkan korelasi positif (Gambar 5.8a). Jika pola titik-titik yang diplot miring dari kiri atas ke kanan bawah, nilai *X* meningkat seiring dengan penurunan nilai *Y*, menunjukkan

korelasi negatif (Gambar 5.8b). Garis yang paling sesuai dapat ditarik untuk mempelajari korelasi antar variabel. Gambar 5.9 menunjukkan kasus ketiga dimana tidak ada hubungan korelasi antara dua atribut di masing-masing himpunan data yang diberikan.

Sebagai kesimpulan, deskripsi data dasar (misalnya ukuran tendensi sentral dan ukuran penyebaran) dan tampilan statistik grafis (misalnya plot quantile, histogram, dan plot sebar) memberikan wawasan berharga tentang perilaku data secara keseluruhan. Dengan membantu mengidentifikasi noise dan outlier, keduanya sangat berguna untuk pembersihan data.

DAFTAR PUSTAKA

Anderson, A. and Semmelroth, D. (2017) *Statistics for Big Data for Dummies*, *John Wiley & Sons, Inc.* New Jersey: John Wiley & Sons, Inc.

Bernstein, S. and Bernstein, R. (1999) *Schaum's Outline of Theory and Problems of Elements of Statistics*.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*, *Morgan Kaufmann*. Massachusetts. Available at: https://doi.org/10.1016/C2009-0-61819-5.

Setiawan, A. (2024) *Ukuran Pemusatan Data: Memahami Mean, Medi*an, Mode, Smartstat.

Wilcox, R.R. (2010) *Fundamentals of Modern Statistical Methods*, *Springer*. California: Springer.

BAB 6
KEGUNAAN DATA PRAPROCESSING

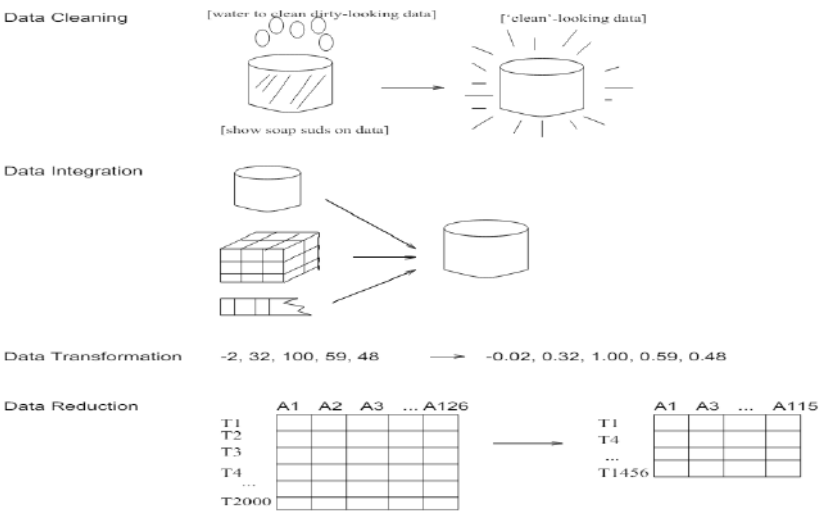
Oleh Siti Aulia Rachmini

Praprocessing merupakan sebuah teknik yang digunakan untuk mengolah data mentah menjadi format yang dapat digunakan dalam proses data mining. Pra processing penting dilakukan sebelum data tersebut digunakan dalam proses data mining untuk mendapatkan data dengan kualitas yang baik sehingga tingkat akurasi dalam pemanfaatan data tersebut dapat bernilai tinggi. Pra processing memiliki beberapa kegunaan antara lain :

1. Memastikan data memiliki kualitas baik dan seragam
   Fungsi ini digunakan untuk membersihkan data dari kontaminasi atau *noise* sehingga data yang dihasilkan memiliki kualitas yang baik untuk digunakan dalam proses data mining
2. Memastikan data bersih dari missing value, noise dan inkosistensi
   Fungsi ini ditujukan untuk menghilangkan kesalahan data seperti duplikasi data serta data yang hilang atau tidak valid
3. Menangani masalah kekurangan data
   Fungsi ini digunakan untuk menangani dan memperbaiki data yang hilang atau tidak lengkap
4. Memperbaiki inkonsistensi data
   Fungsi ini digunakan untuk memastikan data yang ada ada memiliki format dan tipe yang sama
5. Memperkecil dimensi data
   Fungsi ini digunakan untuk memperkecil jumlah aribut atau fitur data yang akan digunakan dalam proses data mining.
6. Mengubah data sesuai dengan algoritma data mining
   Fungsi ini digunakan untuk mengubah data mentah menjadi bentuk yang lebih sesuai dengan algoritma pada data mining seperti proses normalisasi atau standarisasi

7. Mengintegrasikan data dari sumber yang berbeda
   Mengumpulkan serta Menggabungkan data dari sumber yang berbeda menjadi satu data yang dapat digunakan dalam proses data mining

Berikut ini merupakan gambar dari proses pra processing data :



Gambar 6.1. *Pra processing* data
(sumber : The Morgan Kaufman Series 2006)

Dalam proses *pra processing*, ada beberapa hal yang dapat dilakukan sehingga data yang dihasilkan menjadi baik dan dapat diproses lebih lanjut dalam proses data mining :

1. Pembuangan Outlier

Penghapusan atau pembuangan outlier adalah proses menghilangkan atau mengabaikan data yang berbeda secara signifikan dari nilai lain dalam suatu kumpulan data. Outlier adalah nilai yang berbeda secara signifikan dari nilai lain dalam data dan muncul sebagai pengecualian terhadap fungsi normal. Outlier dapat terjadi karena berbagai alasan, termasuk kesalahan pengukuran, kejadian yang tidak biasa, atau penyebab lain yang tidak terduga. Terlebih lagi, outlier adalah sumber data yang memiliki nilai yang sama sekali

berbeda dengan orang. Meskipun definisi ini tampak sederhana, namun mengidentifikasi sumber data outlier sebenarnya sangat sulit, tergantung pada penelitian yang dilakukan dan jumlah data yang dikumpulkan. Sebelum membahas lebih jauh, mari kita pahami perbedaan antara variabel internal dan eksternal. Permasalahan lingkungan seringkali muncul karena kesalahan pengukuran, pengumpulan data yang tidak akurat, atau entri data yang salah. dll.

Ada beberapa cara untuk mengidentifikasi outlier:

a. Gunakan logika Dengan data yang sudah kita ketahui (seperti tinggi badan seseorang), kita dapat menggunakan logika untuk menemukan outlier yang disebabkan oleh kesalahan pencatatan. Misalnya kita mengetahui bahwa 618 cm bukanlah standar tinggi badan manusia, namun 168 cm adalah standar tinggi badan.
b. Penyajian informasi Ada banyak cara untuk melihat gambar atau visualisasi, antara lain:
    1) Box Plot kotak adalah diagram yang menunjukkan sebaran data menggunakan median dan kuartil bawah dan atas. Kita dapat dengan mudah mendeteksi jalan keluar dari kotak, dimana setiap titik berada di atas atau di bawah kumis yang mewakili bagian luar. Disebut juga metode tunggal karena di sini kita menggunakan analisis eksternal terhadap satu variabel saja.
    2) Histogram Outlier juga dapat diidentifikasi menggunakan histogram; Mayoritas penelitian berpihak pada satu pihak, dan beberapa penelitian tampak jauh dari kelompok utama, yang kita sebut outlier.
    3) Distribusi Untuk data yang berbeda, penyebarannya bisa sangat baik. Diagram sebar menunjukkan kumpulan data yang sumbu x (horizontal) mewakili variabel bebas dan sumbu y (vertikal) mewakili variabel terikat. Distribusi dapat dengan mudah

menampilkan "tinggi 618 cm" sebagai orang selain data yang ditentukan.

Ada banyak situasi di mana kita perlu membuang data yang tidak perlu, terutama ketika outlier muncul karena kesalahan dalam proses input atau pengukuran, ketika kita memiliki banyak data dengan jumlah outlier yang sedikit, dan ketika ada hal-hal tertentu yang memungkinkan kita untuk mengulang kembali. mengumpulkan data. data. dimana data dan outlier sama sekali tidak mewakili data yang ada

Ada banyak metode untuk mengidentifikasi outlier seperti metode IQR (*Interquartile Range*), metode persentil, dan metode varians. Metode IQR menggunakan kuartil pertama dan ketiga (Q1 dan Q3) untuk menghitung nilai IQR kemudian menghitung batas bawah (Q1 – 1.5IQR) dan batas atas (Q3 + 1.5IQR). Jika data berada di luar kisaran ini, maka data tersebut dapat dianggap internal. Penghapusan outlier tidak boleh dilakukan secara langsung karena outlier juga dapat mencerminkan kesalahan pengumpulan data atau kesalahan pengukuran. Sebelum mengabaikan outlier, penting untuk mengevaluasinya terlebih dahulu untuk menentukan alasan di balik outlier tersebut. Jika dapat ditentukan bahwa data eksternal bukan merupakan kesalahan pengumpulan atau pengukuran data, data tersebut dapat dibuang.

Kami akan memberikan contoh kasus penyelesaian dari pembuangan outlier. Data ini merupakan data pendapatan bulanan seorang pekerja di sebuah perusahaan.

Tabel 6.1. Data Pendapatan Bulanan :

Bulan	Pendapatan
Januari	2000
Februari	3000
Maret	2500

Bulan	Pendapatan
April	5000
Mei	4000
Juni	6000
Juli	7000
Agustus	8000
September	9000
Oktober	10000
November	12000
Desember	15000

Sumber : https://pacmann.io/blog/cara-mendeteksi-dan-menangani-outlier-saat-melakukan-data-analysis

Disini kami menggunakan metode IQR untuk mendeteksi outlier dalam data tersebut.
Menghitung kuartil pertama (Q1) dan kuartil ketiga (Q3):
Q1 = 2500
Q3 = 8000
Menghitung nilai IQR:
IQR = Q3 - Q1
= 8000 - 2500
= 5500
Menghitung batas bawah dan batas atas:
Batas bawah = Q1 - 1.5 * IQR = 2500 - 1.5 * 5500 = -2775
Batas atas = Q3 + 1.5 * IQR = 8000 + 1.5 * 5500 = 13500
Menghapus data yang di luar batas bawah dan batas atas:
Januari: 2000 (diluar batas bawah)

Juli: 7000 (diluar batas atas)

Tabel 6.2. Data pendapatan bulanan yang tidak diluar batas bawah dan batas atas:

Bulan	Pendapatan
Februari	3000
Maret	2500
April	5000
Mei	4000
Juni	6000
Aguustus	8000
September	9000
Oktober	10000
November	12000
Desember	15000

Sumber : https://jagostat.com/machine-learning/cara-mengatasi-outlier#google_vignette

Dalam contoh ini, kita sudah menghapus data pendapatan bulanan Januari dan Juli, yang dianggap sebagai outlier. Setelah pembuangan outlier, data pendapatan bulanan yang tersisa dapat digunakan untuk analisis statistik dan model prediksi yang lebih akurat. Menghapus data akan membuat jumlah amatan yang diamati menjadi berkurang.

Mengatasi outlier dengan menghapus data outlier sebaiknya menjadi pilihan terakhir karena sekarang sudah tersedia alternatif-alternatif metode yang bisa digunakan dengan menyertakan data outlier.

contoh :

110, 165, 174, 171, 174, 167, 175, 166, 170, 165

dari data tinggi badan 10 siswa, ada 1 siswa yang memilki tinggi badan outlier. Rata-rata tinggi badan ke 10 siswa adalah 163,70 cm, padahal sebagian besar populasi (selain outlier) memiliki tinggi badan 165-175 cm.

Dengan menghapus 1 siswa tersebut dari kelompok data, maka data akan kembali normal, dan nilai rata-ratanya menjadi 169,67 cm. Nilai rata-rata ini sesuai dengan gambaran populasi yang ada. Namun, konsekuensinya adalah data yang awalnya berjumlah 10 siswa, sekarang hanya menjadi 9 siswa.

2. Normalisasi Data

Tujuan utama normalisasi data adalah menghilangkan redudansi data (pengulangan) dan menstandarisasi informasi dalam data mining. Hal ini agar mengubah nilai kolom numerik dalam dataset menjadi skala sehingga berada pada rentang yang lebih kecil, seperti -1 hingga 1 atau 0 hingga 1, ini memungkinkan analisis data yang lebih akurat dan efisien. Tujuan lain dari Normalisasi data adalah meningkatkan kualitas data, sehingga proses data mining dapat menghasilkan pengetahuan baru yang lebih baik.

Data yang telah dinormalisasi dapat meningkatkan performa model dan meningkatkan aakurasi model. Hal ini dapat membantu algoritma yang mengandalkan metrik jarak, seperti K-NN atau support vector machine, dengan mencegah fitur dengan skala lebih besar mendominasi proses pembelajaran.

Normalisasi meningkatkan kestabilan dalam proses optimisasi dan mempercepat konvergensi saat melatih model berbasis gradien. Ini mengurangi risiko masalah seperti gradien yang hilang atau meledak, memungkinkan model mencapai solusi optimal dengan lebih efisien. Data yang dinormalisasi juga lebih mudah diinterpretasikan, memudahkan pemahaman. Ketika semua fitur dalam kumpulan data memiliki skala yang serupa, hubungan antara fitur dapat lebih mudah diidentifikasi dan divisualisasikan, memungkinkan perbandingan yang lebih bermakna.

Ambil contoh sederhana: kita ingin memprediksi harga rumah berdasarkan luas bangunan, jumlah kamar tidur, dan jarak ke supermarket. Fitur-fitur ini memiliki skala yang berbeda, misalnya luas bangunan berkisar antara 500 hingga 5.000 kaki persegi, jumlah kamar tidur berkisar dari 1 hingga 5, dan jarak ke supermarket berkisar antara 0,1 hingga 10 mil. Tanpa normalisasi, algoritma mungkin memberi bobot lebih pada fitur dengan skala yang lebih besar, seperti luas bangunan, sementara mengabaikan fitur-fitur lainnya. Ini bisa mengakibatkan performa model menjadi kurang optimal dan prediksi menjadi bias.

Dengan normalisasi, setiap fitur berkontribusi secara proporsional terhadap pembelajaran model. Ini memungkinkan model untuk mempelajari pola di seluruh fitur dengan lebih efektif, menghasilkan representasi yang lebih akurat dari hubungan mendasar dalam data.

Ada berbagai metode normalisasi data, bergantung pada karakteristik data Anda dan masalah yang Anda pecahkan. Tiga metode normalisasi data yang paling umum digunakan:

a. Decimal Scaling

Decimal scaling atau decimal place normalization digunakan untuk tipe data numerik. Biasanya pada Excel, terdapat dua digit angka setelah tanda koma. Decimal scaling berfungsi mengatur berapa banyak angka yang ingin diletakkan setelah koma.

Cara melakukan decimal scaling adalah dengan membagi nilai data dengan 10 pangkat bilangan tertentu sesuai dengan nilai maksimum data.

Rumus decimal scaling sebagai berikut:

Decimal scaling (Vi') = Vi10j……..(Pers 1)

Contoh dan penyelesainnya :

diketahui nilai data adalah -10, 201, 301, -401, 501, 601, 701. Berdasarkan data, nilai maksimum data adalah 701, maka bentuk desimalnya harus memiliki tiga angka di belakang koma. Artinya, setiap data harus dibagi dengan 10 pangkat 3.

Maka, cara menghitungnya:

Decimal scaling (Vi') = -10103 , dan seterusnya pada seluruh data.
Hasil yang didapatkan yaitu -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701.

b. Z-score normalization

Normalisasi Z-score merupakan teknik normalisasi yang hasilnya diperoleh dari mean dan deviasi standar data. Metode ini memiliki nilai outlier yang stabil atau nilai yang lebih besar dari MaxA atau kurang dari MinA. Normalisasi Zscore dapat dihitung dengan menggunakan rumus berikut :

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$ .....(pers 2)

$$s = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n-1}}$$ .....(pers 3)

Keterangan:

σ	= standar deviasi untuk populasi
s	= standar deviasi untuk populasi
x	= nilai di dalam atribut yang akan dihitung
$\bar{X}$	= nilai rata-rata
N	= ukuran banyaknya data populasi
n	= ukuran banyaknya data sampel

c. Min-Max Normalization

Min-max normalization atau disebut juga dengan linear normalization digunakan untuk mempertahankan bentuk distribusi dan nilai pasti dari data minimum dan maksimum.
Dalam teknik ini, normalisasi dilakukan pada data asli dengan rumus:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$ .....(pers 4)

Keterangan:

x = atribut yang akan dihitung.
x_{new} = ukuran atribut baru yang akan dihasilkan
x_{old} = ukuran atribut lama yang dimiliki
x_{min} = ukuran minimal yang dimiliki oleh atribut
x_{max} = ukuran maksimal yang dimiliki oleh atribut

3. *Missing Value* / Data Yang Salah

Ketika kita berbicara tentang missing value, kita merujuk pada keadaan di mana data dalam sebuah dataset tidak lengkap atau tidak terisi. Kondisi ini bisa memiliki dampak serius terhadap analisis data dan pembangunan model, karena ketiadaan informasi tersebut bisa menyebabkan bias atau mengurangi keakuratan hasil yang diperoleh.

Kekurangan data, atau nilai kosong, adalah situasi ketika data pada variabel tertentu dalam kumpulan data tidak tersedia atau tidak tercatat. Hal ini dapat terjadi karena berbagai faktor, seperti kesalahan pengukuran, ketidaklengkapan input, atau bahkan sengaja dibiarkan kosong. Dalam proses analisis data, keberadaan missing value dapat berdampak signifikan pada hasil analisis karena dapat menimbulkan bias dan menurunkan akurasi. Pada beberapa kasus, missing value mungkin juga mengandung informasi penting. Oleh karena itu, penting untuk menangani missing value dengan cermat, baik dengan mengisi nilainya dengan estimasi yang logis, menghapus baris atau kolom yang mengandung missing value, atau menggunakan metode model untuk memprediksi nilai atau data yang hilang berdasarkan data yang ada.

Dalam menangani missing value, penting untuk mempertimbangkan berbagai strategi yang tersedia agar dapat menjaga integritas dan akurasi analisis data. Salah satu pendekatan yang umum digunakan adalah imputasi, di mana nilai yang hilang digantikan dengan estimasi yang masuk akal berdasarkan data yang tersedia. Metode imputasi dapat bervariasi, mulai dari menggunakan nilai rata-rata atau median variabel yang relevan hingga teknik yang lebih canggih seperti

imputasi berbasis model, di mana model statistik atau machine learning digunakan untuk memprediksi nilai yang hilang.

Perlu diingat bahwa imputasi harus dilakukan dengan hati-hati, karena estimasi yang tidak akurat dapat menghasilkan hasil analisis yang bias. Selain itu, terkadang mempertimbangkan untuk menghapus baris atau kolom yang mengandung missing value juga merupakan opsi, terutama jika jumlah missing value relatif kecil dan tidak signifikan secara statistik. Namun, penghapusan data dapat mengurangi jumlah sampel yang tersedia untuk analisis, sehingga mempengaruhi keakuratan model yang dibangun. Dalam beberapa kasus, missing value mungkin mengandung informasi penting yang tidak boleh diabaikan. Oleh karena itu, penting untuk mempertimbangkan konteks dan karakteristik data secara keseluruhan sebelum mengambil keputusan tentang bagaimana menangani missing value dalam analisis data. Dengan menggunakan pendekatan yang tepat, kita dapat meminimalkan dampak missing value dan meningkatkan kualitas hasil analisis secara keseluruhan.

Jenis jenis missing value

Dalam missing value atau data yang hilang pada sebuah dataset atau apapun itu terdapat beberapa jenis missing value yang bis akita tahu jenisnya yaitu sebagai berikut:

a. ***Missing completely at random*** **(MCAR)**

Adalah jenis missing value yang dimana distribusi data yang hilang tidak memiliki pola atau alasan yang dapat di jelaskan yang berarti Jenis missing value ini bisa terjadi secara acak, dan data variable yang tidak terkait dengan data variable itu sendiri seperti hilangnya data tanpa disengaja kejadian ini bisa terjadi apabila responden lupa memasukkan data dalam pertanyaan wawancara atau survey dan jika dalam sebuah kegiatan survey apabila melakukan pengumpulan data dan terjadi kesalahan pada saat proses penginputan data atau kerusakan suatu perangkat lunak untuk pengumpulan sebuah data, maka dapat mengakibatkan data yang tidak terinput atau hilang itu dapat disimpulkan data missing value dengan jenis MCAR.

b. ***Random Missing Values*** **(MAR)**

Dalam jenis missing value ini, kehadiran sebuah nilai atau data yang hilang dapat terkait dengan jenis variabel yang diamati , tetapi data yang hilang hilang tidak terjadi secara langsung dengan sendirinya. Yang berarti meskipun ada sebuah korelasi antara data atau nilai yang hilang dan variabel lain dalam sebuah dataset. Sebagai contoh jiak dalam sebuah kegiatan survey yang Dimana kita dapat pendapatan responden dengan pendapatan yang tinggi kebanyakan dari responden yang seperti itu tidak mau memberikan informasi tentang pendapatan yang telah diperoleh dan Keputusan tersebut dapat menyembunyikan infromasi yang telah didapatkan tidak terkait dengan data atau nilai yang hilang dan ini bisa terjadi karena sebuah alasan tertentu yang berkaitan dengan data lain dalam sebuah dataset, Misal ketika kita mewawancarai seseorang yang mungkin tidak nyaman dengan pertanyaan yang harus ada pada dataset yang orang tersebut dapatkan sehingga seeorang tersebut melewatkan pertanyaan untuk pengumpulan data, hal tersebut dapat menimbulkan missing values pada sebuah dataset.

c. ***Non-Random Missing Values*** **(MNAR)**

Dalam kasus Jenis missing value ini bisa terjadi data atau nilai yang hilang dengan data yang tidak diamati atau tidak terdeteksi dalam sebuah dataset dan ini berarti bahwa data yang hilang akan bergantung pada data atau nilai yang sebenarnya, oleh karena itu kejadian ini tidak dapat disimpulkan sebagai kejadian yang kebetulan atau hasil dari sebuah pola yang ada pada data sebagai contoh , dalam sebuah kegiatan survey mengenai Kesehatan mental Dimana sang responden menderita atau mengalami gangguan mental yang cukup serius yang mungkin tidak mau memberikan sebuah informasi yang jelas, sehingga dapat menyebabkan data mengenai kesehatan mental ini hilang secara tidak menyebar sehingga kita tidak dapat mendapatkan datanya.

4. Seleksi Atribut

Seleksi atribut yang juga dikenal sebagai pemilih fitur adalah proses memilih subset atribut yang paling relevan dan informatif dari kumpulan data yang lebih besar. Tujuannya adalah untuk meningkatkan performa model pembelajaran mesin dengan mengurangi dimensi ruang atribut dan menghilangkan atribut yang tidak relevan atau redundan.

Metode seleksi atribut:

Ada banyak metode seleksi atribut yang berbeda, yang dapat dikategorikan menjadi tiga kelompok utama:

a. **Filter**

Metode filter mengevaluasi setiap atribut secara independen dan memilih atribut berdasarkan skor yang dihitung. Skor ini dapat didasarkan pada berbagai kriteria, seperti gain informasi, korelasi, atau varians.

b. Wrapper

Metode wrapper mengevaluasi subset atribut secara keseluruhan dengan membungkusnya di dalam model pembelajaran mesin dan mengukur performanya. Subset atribut yang menghasilkan kinerja terbaik kemudian dipilih.

c. **Embedded**

Metode embedded mengintegrasikan seleksi atribut ke dalam proses pembelajaran model. Contohnya adalah algoritma pohon keputusan, yang secara alami memilih atribut yang paling informatif selama proses pembangunan pohon.

5. Pengukuran *Separable Class*

Pengukuran kelas separable (atau "separable class measurement" dalam bahasa Inggris) adalah sebuah konsep dalam statistik yang mengacu pada kemampuan untuk mengukur variabel-variabel yang terpisah secara jelas. Dalam konteks analisis statistik, variabel-variabel tersebut dapat dipisahkan dengan jelas ke dalam kelompok-kelompok atau kelas-kelas yang berbeda.

Pengukuran kelas separable seringkali diinginkan dalam analisis statistik karena memungkinkan untuk memperoleh pemahaman yang lebih baik tentang hubungan antara variabel-variabel tersebut. Misalnya, jika melakukan eksperimen untuk mengukur efek dua jenis obat terhadap penyakit tertentu, mungkin ingin mengukur setiap efek secara terpisah untuk memahami perbedaan efektivitas di antara keduanya.

Dengan memiliki variabel yang dapat dipisahkan dengan jelas ke dalam kelas-kelas yang berbeda, analisis statistik dapat dilakukan dengan lebih tepat dan menghasilkan kesimpulan yang lebih dapat diandalkan. Ini memungkinkan untuk mengidentifikasi pola atau tren yang mungkin tidak terlihat jika variabel-variabel tersebut bercampur atau tidak dapat dipisahkan dengan jelas.

Berikut ini adalah penjelasan tambahan tentang pengukuran kelas separable:

1. **Pemisahan Variabel:**
   Konsep ini bermula dari gagasan bahwa ada hubungan yang jelas antara variabel-variabel yang diamati dan bahwa variabel-variabel tersebut dapat dipisahkan dengan jelas ke dalam kelompok-kelompok atau kelas-kelas yang berbeda. Misalnya, jika Anda memiliki variabel X dan Y, pengukuran kelas separable akan mencoba untuk menentukan apakah ada pemisahan yang jelas antara dua variabel tersebut ke dalam kelompok-kelompok atau kelas-kelas tertentu.
2. **Analisis Klasifikasi:**
   Pengukuran kelas separable seringkali melibatkan analisis klasifikasi, di mana observasi-observasi atau data dikelompokkan ke dalam kategori-kategori yang berbeda berdasarkan nilai-nilai dari variabel yang diamati. Proses ini bertujuan untuk menemukan pola atau perbedaan yang signifikan antara kelompok-kelompok ini.
3. **Pemodelan Statistik :**
   Dalam konteks pemodelan statistik, pengukuran kelas separable dapat digunakan untuk membangun model yang

memprediksi kelas atau kategori dari suatu observasi berdasarkan nilai-nilai dari variabel yang diamati. Misalnya, dalam klasifikasi biner, model dapat digunakan untuk memprediksi apakah suatu observasi termasuk dalam kelas tertentu atau tidak.

4. **Penerapan dalam Berbagai Bidang:**
   Konsep pengukuran kelas separable memiliki aplikasi luas di berbagai bidang, termasuk dalam ilmu sosial, ilmu alam, kedokteran, dan teknik. Contohnya termasuk penggunaannya dalam klasifikasi pasien berdasarkan gejala penyakit, identifikasi pola perilaku konsumen berdasarkan preferensi, atau analisis genetik untuk mengklasifikasikan organisme ke dalam kelompok-kelompok tertentu.
5. **Evaluasi dan Validasi:**
   Penting untuk melakukan evaluasi dan validasi terhadap hasil pengukuran kelas separable. Hal ini dilakukan dengan memeriksa sejauh mana pengelompokan atau klasifikasi hasil dari analisis tersebut konsisten dan relevan dengan data yang diamati. Metode-metode seperti validasi silang atau validasi menggunakan data yang independen sering digunakan untuk tujuan ini.

Pengukuran kelas separable menjadi alat yang berguna dalam statistik untuk memahami dan menggambarkan hubungan antara variabel-variabel yang diamati dengan lebih baik. Dengan menganalisis variabel-variabel ini secara terpisah ke dalam kelompok-kelompok yang jelas, kita dapat mengungkap pola-pola yang mungkin tidak terlihat jika kita hanya melihat data secara keseluruhan.

Contoh Kasus Pengukuran Separable Class

1. Pengenalan Kelas yang Dapat Dipisahkan:
   Kelas yang dapat dipisahkan adalah konsep dalam pembelajaran mesin di mana dua kelas data dapat dipisahkan dengan garis lurus atau hiperplan dalam ruang fitur. Artinya, data dari dua kelas tersebut memiliki sebaran yang berbeda dan tidak saling tumpang tindih.

2. Klasifikasi Spam Email:
   Misalkan dua kelasnya adalah email spam dan email non-spam. Fitur-fiturnya dapat berupa kata-kata dalam email, alamat email pengirim, dan informasi lainnya. Algoritma pembelajaran mesin dapat dilatih untuk memisahkan email spam dari email non-spam dengan menemukan garis lurus atau hiperplan yang memisahkan kedua kelas data tersebut.
   Klasifikasi Gambar Kucing dan Anjing: Bayangkan dua kelasnya adalah gambar kucing dan gambar anjing. Fitur-fiturnya dapat berupa piksel gambar, tekstur, dan bentuk. Algoritma pembelajaran mesin dapat dilatih untuk memisahkan gambar kucing dari gambar anjing dengan menemukan garis lurus atau hiperplan yang memisahkan kedua kelas data tersebut.
3. Pentingnya Pengukuran Separable Class:
   Pengukuran kelas yang dapat dipisahkan penting dalam pembelajaran mesin karena dapat membantu menentukan apakah algoritma pembelajaran mesin akan dapat belajar secara efektif dari data. Jika data tidak terpisah (not separable), algoritma pembelajaran mesin mungkin kesulitan menemukan garis lurus atau hiperplan yang memisahkan dua kelas data, sehingga akurasinya mungkin rendah.
4. Metode Pengukuran Separability Class:
   Ada beberapa metode untuk mengukur separability class, antara lain:
   a. **Margin**: Margin adalah jarak antara garis lurus atau hiperplan yang memisahkan dua kelas data dan titik data terdekat dari setiap kelas. Margin yang besar menunjukkan bahwa data lebih terpisah.
   b. **Entropy**: Entropy adalah ukuran ketidakpastian dalam data. Entropy yang rendah menunjukkan bahwa data lebih terpisah.
   c. **Visualisasi data**: Memvisualisasikan data dapat membantu untuk melihat apakah data tersebut dapat dipisahkan secara visual.

Pengukuran kelas yang dapat dipisahkan adalah alat penting dalam pembelajaran mesin untuk menentukan

apakah algoritma pembelajaran mesin akan dapat belajar secara efektif dari data. Ada beberapa metode untuk mengukur separability class, dan metode terbaik untuk digunakan tergantung pada situasi spesifiknya.

DAFTAR PUSTAKA

Jain, S., Shukla, S., & Wadhvani, R. (2018). *Dynamic selection of normalization techniques using data complexity measures*. Expert Systems with Applications, 106, 252-262.

Patro, S., & Sahu, K. K. (2015). *Normalization: A preprocessing stage*. arXiv preprint arXiv:1503.06462.

Kusnaidi, M. R., Gulo, T., & Aripin, S. (2022). *Penerapan Normalisasi Data Dalam Mengelompokkan Data Mahasiswa Dengan Menggunakan Metode K-Means Untuk Menentukan Prioritas Bantuan Uang Kuliah Tunggal*. Journal of Computer System and Informatics (JoSYC), 3(4), 330–338.

Patro, S. G. K., & Sahu, K. K. (2015). *Normalization: A Preprocessing Stage*.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2019). *Multivariate Data Analysis (8th ed).* Cengage Learning.

Moch. Lutfi, Mochamad Hasyim. "*Penanganan Data Missing Value pada Kualitas Produksi Jagung dengan Menggunakan Metode K-NN Imputation pada Algoritma C4.5.*" Universitas Yudharta Pasuruan, Jl. Yudharta No.7, Kembangkuning, Sengonagun, Purwosari, Pasuruan, Jawa Timur 67162

Wan, G., Li, W. K., & Geng, Z. (2012). *Handling Missing Values in Data: Algorithms and Methodologies*. Knowledge and Information Systems, 31(3), 561-607. DOI: 10.1007/s10115-011-0460-2.

BAB 7
TUGAS-TUGAS DALAM DATA PREPROCESSING

Oleh Farid Wajidi

7.1 Mengapa *Prerocesssing* Diperlukan?

Kualitas data dalam implementasi basis data modern seringkali terhambat oleh berbagai masalah, seperti *noise* (gangguan), *missing values* (data hilang), dan ketidakkonsistenan. Hal ini, terutama pada data berukuran besar (gigabyte atau lebih), dapat menghambat proses analisis dan pemodelan yang efektif. Oleh karena itu, *preprocessing* data menjadi langkah penting untuk memastikan kualitas data sebelum diolah.

Analogikan dengan proses memasak. Seorang koki perlu membersihkan, mengupas, dan memotong bahan makanan sebelum diolah menjadi hidangan lezat. *Preprocessing* data berperan serupa, yaitu membersihkan data dari noise, missing values, outlier, dan variabel tidak relevan, sehingga data siap dianalisis atau dimodelkan.

Preprocessing data terbukti meningkatkan kualitas data untuk analisis dan pemodelan yang efektif(Jassim and Abdulwahid, 2021). Tahapan ini meliputi pembersihan data, integrasi data, dan tranformasi data untuk mengoptimalkan kualitas data(Jamshed *et al.*, 2019).

Preprocessing data menawarkan berbagai manfaat penting untuk analisis dan pemodelan yang efektif, di antaranya, membantu mengidentifikasi dan menangani masalah seperti missing values, outlier, dan variabel tidak relevan. Hal ini penting untuk memastikan bahwa data yang digunakan dalam analisis dan pemodelan akurat, lengkap, dan konsisten. Data yang berkualitas tinggi akan menghasilkan wawasan yang lebih andal dan dapat ditindaklanjuti. Tahapan *preprocessing* juga dapat Meningkatkan akurasi Analisis. Data yang kotor dan tidak diolah dengan baik dapat

menyebabkan bias dan kesalahan dalam proses analisis, sehingga menghasilkan kesimpulan yang keliru. *Preprocessing* data membantu meminimalkan bias dan kesalahan, sehingga analisis menjadi lebih akurat dan dapat dipercaya. Selain itu *preprocessing* membantu untuk Meningkatkan efisiensi dan efektivitas. Data yang telah melalui *preprocessing* lebih mudah dipahami, dianalisis, dan dimodelkan. Hal ini membantu meningkatkan efisiensi dan efektivitas proses analisis dan pemodelan, sehingga menghemat waktu dan sumber daya. Dalam pengolahan data, model analisis dan pemodelan seringkali memiliki persyaratan data tertentu, seperti skala, dimensi, dan format data. *Preprocessing* data memungkinkan penskalaan data, pengurangan dimensi, dan transformasi agar sesuai dengan kebutuhan model, sehingga model dapat bekerja dengan optimal(Han, Kamber and Pei, 2012).

Secara keseluruhan, *preprocessing* data merupakan langkah penting dan tak terpisahkan dalam proses analisis dan pemodelan data. Dengan memastikan kualitas data sebelum diolah, kita dapat menghasilkan analisis yang lebih akurat, andal, dan berwawasan, yang pada akhirnya dapat membantu pengambilan keputusan yang lebih baik.

7.2 Data *Preparation* (Persiapan Data)

Persiapan data merupakan langkah awal yang sangat penting dalam proses penambangan data (data mining). Langkah ini melibatkan serangkaian teknik yang bertujuan untuk mengubah data mentah menjadi data yang siap digunakan dalam algoritma penambangan data. Tanpa persiapan data yang memadai, algoritma penambangan data mungkin tidak dapat beroperasi dengan baik atau bahkan dapat menghasilkan kesalahan selama runtime. Dalam kasus terbaik, algoritma mungkin berfungsi, tetapi hasil yang diberikan mungkin tidak masuk akal atau tidak dianggap sebagai pengetahuan yang akurat.

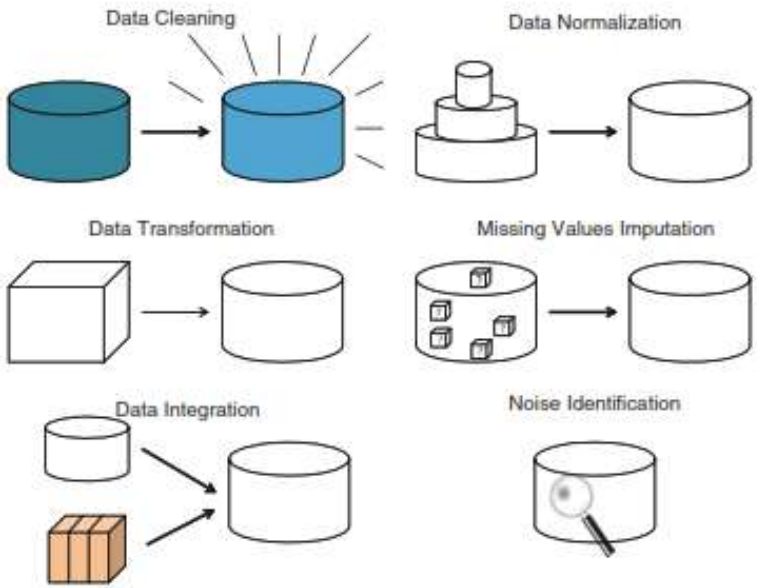
Persiapan data tidak hanya penting untuk memastikan bahwa data siap digunakan dalam algoritma penambangan data tetapi juga untuk meningkatkan kualitas hasil yang diperoleh. Data di dunia nyata sering kali tidak murni (impure), yang berarti bahwa data tersebut mungkin mengandung nilai-nilai yang keluar dari

jangkauan, kombinasi data yang tidak mungkin, nilai yang hilang, dan lain-lain. Menganalisis data yang belum disaring dengan hati-hati untuk masalah-masalah tersebut dapat menghasilkan hasil yang menyesatkan. Oleh karena itu, representasi dan kualitas data adalah hal yang paling penting sebelum menjalankan analisis.

Persiapan data mencakup berbagai teknik yang berkaitan dengan analisis data mentah untuk menghasilkan data berkualitas. Teknik-teknik ini meliputi pengumpulan data, integrasi data, transformasi data, pembersihan data, reduksi data, dan diskritisasi data(Fernandes *et al.*, 2023). Persiapan data dapat lebih memakan waktu daripada penambangan data itu sendiri dan dapat menimbulkan tantangan yang serupa. Pentingnya persiapan data terletak pada fakta bahwa data dunia nyata adalah impure.

Teknik persiapan data dapat dikategorikan menjadi beberapa bagian, termasuk pembersihan data, transformasi data, integrasi data, normalisasi data, imputasi nilai yang hilang, dan identifikasi noise. Pembersihan data atau data cleansing melibatkan operasi yang memperbaiki data buruk, menyaring beberapa data yang tidak benar dari dataset, dan mengurangi detail yang tidak perlu dari data. Transformasi data melibatkan penghalusan, konstruksi fitur, agregasi atau ringkasan data, normalisasi, diskritisasi, dan generalisasi.

Persiapan data adalah langkah krusial dalam proses penambangan data yang tidak boleh diabaikan. Teknik-teknik persiapan data memastikan bahwa data yang digunakan dalam algoritma penambangan data adalah berkualitas tinggi, yang pada gilirannya meningkatkan keandalan dan akurasi hasil yang diperoleh. Dengan memahami dan menerapkan teknik-teknik persiapan data yang tepat, para peneliti dan praktisi dapat meningkatkan efektivitas proses penambangan data mereka.



Gambar 7.1. Bentuk-bentuk persiapan data ((García, Luengo and Herrera, 2015)

7.2.1 Data Cleaning

Data cleaning, atau pembersihan data, merupakan langkah penting dalam *preprocessing* data yang bertujuan untuk meningkatkan kualitas data sebelum proses data mining (DM) dilakukan. Proses ini melibatkan berbagai operasi yang memperbaiki data yang buruk, menyaring data yang salah dari set data, dan mengurangi detail yang tidak perlu dari data. Data cleaning adalah konsep umum yang mencakup atau tumpang tindih dengan teknik *preprocessing* data lainnya, termasuk penanganan data yang hilang dan identifikasi kebisingan, meskipun kedua kategori tersebut dipisahkan untuk analisis yang lebih mendalam(Zou, 2022).

Pentingnya data cleaning terletak pada fakta bahwa data yang terintegrasi ke dalam set data belum tentu bebas dari kesalahan. Integrasi dapat menghasilkan proporsi data yang kotor, yang mencakup data yang hilang, data yang salah, dan representasi non-standar dari data yang sama. Jika proporsi data yang kotor

tinggi, penerapan proses DM pasti akan menghasilkan model yang tidak dapat diandalkan. Sumber data yang kotor termasuk kesalahan entri data, kesalahan pembaruan data, kesalahan transmisi data, dan bahkan bug dalam sistem pemrosesan data.

Proses pembersihan data tidak hanya penting untuk menghilangkan atau memperbaiki data yang kotor, tetapi juga untuk memastikan bahwa set data yang konsisten dan hampir bebas dari kesalahan siap untuk algoritma DM. Beberapa algoritma dalam DM, seperti jaringan saraf tiruan (ANNs) atau metode berbasis jarak, bekerja jauh lebih baik dengan nilai atribut yang dinormalisasi. Selain itu, ada algoritma yang tidak dapat bekerja dengan atribut bernilai nominal atau mendapat manfaat dari transformasi halus dalam data. Teknik normalisasi data dan transformasi data telah dirancang untuk menyesuaikan set data dengan kebutuhan algoritma DM yang akan diterapkan setelahnya.

Secara keseluruhan, data cleaning merupakan langkah krusial yang tidak hanya mempersiapkan data untuk proses DM tetapi juga memastikan bahwa pengetahuan yang diekstrak oleh algoritma DM akurat dan dapat diandalkan. Mengabaikan langkah ini dapat mengakibatkan kesalahan interpretasi yang signifikan dari hasil DM, yang pada akhirnya dapat mempengaruhi keputusan yang dibuat berdasarkan analisis data tersebut.

7.2.2 Transformasi Data

Transformasi data merupakan langkah penting dalam *preprocessing* data yang bertujuan untuk mengkonversi atau mengkonsolidasikan data sehingga proses penambangan data (data mining, DM) dapat diterapkan dengan lebih efisien atau hasilnya menjadi lebih optimal. Proses ini mencakup berbagai sub-tugas seperti penghalusan (*smoothing*), konstruksi fitur, agregasi atau ringkasan data, normalisasi, diskritisasi, dan generalisasi. Meskipun transformasi data sering kali dianggap sebagai bagian dari keluarga teknik *preprocessing* data secara umum, beberapa tugas di dalamnya memerlukan pengawasan manusia dan sangat bergantung pada data, sehingga dianggap sebagai teknik transformasi data klasik(Aspin, 2022).

Transformasi data biasanya menggabungkan atribut-atribut mentah asli menggunakan berbagai rumus matematika yang berasal dari model bisnis atau rumus matematika murni. Tujuan dari transformasi ini adalah untuk menciptakan atribut baru yang sering disebut sebagai transformasi atribut atau set atribut. Dalam konteks penemuan ilmiah dan kontrol mesin, normalisasi saja mungkin tidak cukup untuk menyesuaikan data agar dapat meningkatkan kualitas model yang dihasilkan. Dalam kasus seperti ini, menggabungkan informasi yang terkandung dalam berbagai atribut bisa sangat bermanfaat. Salah satu metode sederhana yang dapat digunakan untuk tujuan ini adalah transformasi linier, yang didasarkan pada transformasi aljabar sederhana seperti penjumlahan, rata-rata, rotasi, translasi, dan sebagainya.

Selain itu, transformasi data juga dapat mencakup penggunaan transformasi min-max untuk menormalkan nilai-nilai numerik atribut ke dalam rentang yang ditentukan, sehingga semua atribut memiliki skala yang sama dan memberikan bobot yang sama pada semua atribut, yang sangat berguna dalam metode pembelajaran statistik. Transformasi peringkat merupakan transformasi sederhana lainnya yang menggantikan nilai atribut dengan peringkatnya, yang kemudian dapat diubah menjadi skor normal yang mewakili probabilitasnya dalam distribusi normal, mengungkapkan hubungan yang mungkin tersembunyi oleh distribusi atribut sebelumnya.

Dalam beberapa kasus, transformasi data dapat melibatkan penciptaan atribut baru yang tidak menghasilkan atribut tambahan tetapi mengubah distribusi nilai-nilai asli menjadi satu set nilai baru dengan properti yang diinginkan. Ketika set data sangat besar, teknik reduksi data dapat diterapkan untuk mengurangi ukuran set data sambil mencoba mempertahankan integritas dan informasi dari set data asli sebanyak mungkin, yang merupakan bagian dari proses transformasi data.

Secara keseluruhan, transformasi data memainkan peran kunci dalam mempersiapkan data untuk analisis DM dengan mengoptimalkan struktur dan distribusi data. Melalui berbagai teknik transformasi, data dapat disesuaikan untuk meningkatkan kinerja dan akurasi model DM, memungkinkan penemuan pola dan

wawasan yang lebih mendalam dari data. Transformasi data, dengan demikian, bukan hanya tentang mengubah data tetapi juga tentang meningkatkan nilai informasi yang dapat diekstraksi dari data tersebut.

7.2.3 Integrasi Data

Data Integration adalah proses penting dalam *preprocessing* data yang melibatkan penggabungan data dari berbagai sumber untuk menciptakan satu set data yang konsisten dan terpadu. Proses ini merupakan langkah awal yang krusial dalam analisis data dan penambangan data (DM), karena data yang diperoleh dari berbagai sumber sering kali berbeda dalam format, skema, dan konvensi penamaan, yang dapat menyebabkan redundansi dan inkonsistensi dalam set data yang dihasilkan(Schildhauer, 2018).

Salah satu tantangan utama dalam integrasi data adalah mencocokkan skema dari berbagai sumber. Perbedaan dalam nama atribut atau skema tabel dapat menghasilkan contoh-contoh yang tidak seragam yang perlu dikonsolidasikan. Selain itu, nilai atribut mungkin mewakili konsep yang sama tetapi dengan nama yang berbeda, menciptakan inkonsistensi dalam instansi yang diperoleh. Jika beberapa atribut dihitung dari atribut lain, set data akan menunjukkan ukuran besar tetapi informasi yang terkandung tidak akan meningkat secara proporsional. Oleh karena itu, mendeteksi dan mengeliminasi atribut yang redundan diperlukan untuk mengoptimalkan ukuran set data dan meningkatkan efisiensi proses DM selanjutnya.

Integrasi data juga melibatkan penggabungan data dari berbagai penyimpanan data. Proses ini harus dilakukan dengan hati-hati untuk menghindari redundansi dan inkonsistensi dalam set data yang dihasilkan. Operasi tipikal yang dilakukan dalam integrasi data termasuk identifikasi dan penyatuan variabel dan domain, analisis korelasi atribut, duplikasi tupel, dan deteksi konflik dalam nilai data dari sumber yang berbeda.

Pendekatan otomatis untuk integrasi data telah ditemukan dalam literatur, mulai dari teknik yang mencocokkan dan menemukan skema data, hingga prosedur otomatis yang menyelaraskan skema yang berbeda. Namun, masalah seperti

atribut yang redundan dan duplikasi tupel serta inkonsistensi tetap menjadi tantangan yang harus diatasi. Atribut dianggap redundan ketika dapat diturunkan dari atribut lain atau sekumpulan atribut. Inkonsistensi dalam dimensi atau nama atribut juga dapat menyebabkan redundansi.

Integrasi data yang tidak tepat dapat mengakibatkan penurunan akurasi dan kecepatan proses DM selanjutnya. Oleh karena itu, sangat penting untuk memastikan bahwa proses integrasi data dilakukan dengan benar untuk menghindari redundansi dan inkonsistensi yang akan segera muncul, sehingga menghasilkan set data yang bersih dan terpadu yang siap untuk analisis lebih lanjut.

Kesimpulannya, integrasi data adalah langkah kritis dalam *preprocessing* data yang memastikan bahwa informasi dari berbagai sumber dapat digabungkan menjadi satu set data yang kohesif dan bebas dari redundansi dan inkonsistensi. Melalui proses integrasi yang cermat, data yang dihasilkan menjadi lebih relevan dan dapat diandalkan untuk analisis DM, yang pada gilirannya meningkatkan akurasi dan efisiensi teknik DM yang digunakan.

7.2.4 Normalisasi Data

Data normalization merupakan langkah penting dalam *preprocessing* data yang bertujuan untuk meningkatkan kualitas dan efisiensi analisis data. Proses ini melibatkan penskalaan dan transformasi data ke skala umum tanpa mendistorsi perbedaan dalam rentang nilai atau kehilangan informasi. Normalisasi data sangat penting untuk algoritma yang sensitif terhadap skala data, seperti k-nearest neighbors (k-NN), jaringan saraf tiruan (ANNs), dan algoritma optimasi gradien turun, di mana skala variabel dapat berdampak signifikan terhadap kinerja model.

Tujuan utama dari normalisasi data adalah untuk menghilangkan data redundan dan memastikan konsistensi dalam dataset. Dalam data dunia nyata, variabel sering kali ada dalam skala dan unit yang berbeda, yang dapat menyebabkan bias dalam analisis dan menghasilkan hasil yang menyesatkan. Misalnya, dalam dataset yang berisi variabel pendapatan dan usia, nilai pendapatan mungkin berkisar dalam ribuan sementara nilai usia umumnya di bawah 100.

Tanpa normalisasi, model mungkin secara tidak proporsional memberikan bobot lebih pada pendapatan daripada usia, hanya karena skala nilai daripada kepentingan aktual mereka .

Teknik normalisasi juga meningkatkan efisiensi komputasi algoritma, mengurangi kompleksitas model, dan membantu mencapai konvergensi lebih cepat selama pelatihan. Selain itu, mereka berkontribusi pada interpretasi yang lebih sederhana dari bobot atau koefisien yang diberikan kepada fitur dalam model pembelajaran mesin, memudahkan untuk memahami pentingnya relatif setiap fitur.

Beberapa teknik dapat digunakan untuk normalisasi data, masing-masing dengan kasus penggunaan dan keuntungannya sendiri. Pilihan teknik tergantung pada sifat data dan kebutuhan analisis atau model yang digunakan(Surendra Kumar Reddy Koduru, 2022) .

1. Min-Max *Normalization*
   Min-max *normalization* adalah salah satu metode paling sederhana, yang mengubah rentang fitur menjadi skala dalam rentang [0, 1] atau [-1, 1]. Metode ini bermanfaat ketika minimum dan maksimum aktual dari data diketahui dan dapat dibatasi.
2. Z-Score *Normalization* (Standardisasi)
   Z-score *normalization*, atau standardisasi, mengubah fitur sehingga mereka memiliki rata-rata 0 dan standar deviasi 1. Metode ini sangat berguna ketika data mengikuti distribusi Gaussian dan kurang dipengaruhi oleh pencilan dibandingkan dengan normalisasi min-max.
3. *Decimal Scaling*
   Decimal scaling menormalisasi dengan memindahkan titik desimal dari nilai fitur. Jumlah tempat desimal yang dipindahkan tergantung pada nilai absolut maksimum dari fitur. Metode ini kurang umum tetapi dapat digunakan untuk normalisasi yang cepat dan sederhana tanpa memerlukan perhitungan ekstensif.

Meskipun normalisasi data bermanfaat, tidak tanpa tantangan. Harus dipilih dengan hati-hati teknik normalisasi yang

tepat berdasarkan distribusi data dan kebutuhan algoritma. Aplikasi normalisasi yang salah dapat menyebabkan kehilangan informasi, terutama jika data mengandung pencilan atau jika metode normalisasi tidak sesuai dengan sifat data.

Selain itu, normalisasi harus diterapkan secara konsisten pada data pelatihan dan pengujian untuk mencegah kebocoran data dan memastikan bahwa model berkinerja seperti yang diharapkan pada data yang tidak terlihat. Konsistensi ini penting untuk menjaga integritas dan keandalan analisis.

Data *normalization* adalah langkah kritis dalam *preprocessing* data, memastikan bahwa dataset dalam bentuk yang sesuai untuk analisis dan pemodelan. Dengan menstandarkan skala fitur, teknik normalisasi membantu meningkatkan akurasi model, efisiensi, dan interpretasi. Namun, pilihan metode normalisasi harus dibuat dengan bijak, mempertimbangkan karakteristik khusus data dan kebutuhan model analitis yang digunakan. Dengan pendekatan yang tepat terhadap normalisasi data, ilmuwan data dan analis dapat membuka hasil yang lebih akurat dan berwawasan dari upaya berbasis data mereka.

7.2.5 Imputasi Data yang Hilang

Dalam dunia analisis data, kita seringkali menghadapi masalah data yang hilang atau *missing data*. Data yang hilang ini dapat menyebabkan distorsi dalam analisis statistik dan pengambilan keputusan. Oleh karena itu, imputasi data yang hilang menjadi langkah penting dalam pra-pengolahan data untuk memastikan kualitas dan keakuratan hasil analisis. Imputasi data yang hilang adalah proses penggantian nilai data yang hilang dengan nilai pengganti yang diestimasi.

Data dapat hilang karena berbagai alasan, seperti kesalahan dalam pengumpulan data, kerusakan data, atau ketidaklengkapan respons dari responden. Secara umum, data yang hilang dapat dikategorikan menjadi tiga jenis, yaitu (Hameed and Ali, 2023):

1. Data Hilang Secara Acak (*Missing at Random*, MAR): Probabilitas data hilang berkaitan dengan informasi yang teramati.

2. Data Hilang Sepenuhnya Acak (*Missing Completely at Random*, MCAR): Probabilitas data hilang sama untuk semua observasi.
3. Data Hilang Tidak Acak (*Missing Not at Random*, MNAR): Probabilitas data hilang berkaitan dengan informasi yang tidak teramati.

Berbagai metode telah dikembangkan untuk mengatasi masalah data yang hilang, masing-masing dengan kelebihan dan keterbatasannya. Berikut adalah beberapa metode imputasi yang umum digunakan(Gomer and Yuan, 2023):

1. Imputasi Rata-rata: Mengganti nilai yang hilang dengan rata-rata dari kolom yang bersangkutan.
2. Imputasi Median: Menggunakan median dari kolom yang bersangkutan sebagai pengganti nilai yang hilang.
3. Imputasi Modus: Mengganti nilai yang hilang dengan modus atau nilai yang paling sering muncul.
4. Imputasi K-Nearest Neighbors (KNN): Menggunakan KNN untuk menemukan nilai yang paling mirip dan menggantikan nilai yang hilang dengan nilai tersebut.
5. Imputasi Regresi: Menggunakan model regresi yang dibangun dari data yang teramati untuk memprediksi nilai yang hilang.
6. Multiple Imputation: Menghasilkan beberapa set imputasi untuk data yang hilang dan menggabungkan hasilnya untuk mendapatkan estimasi yang lebih akurat.

Dengan kemajuan teknologi dan algoritma *machine learning*, pendekatan baru dalam imputasi data yang hilang telah dikembangkan. Metode-metode ini mencakup:

1. *Imputasi berbasis Fuzzy K-means*: Menggunakan logika fuzzy dalam proses k-means clustering untuk imputasi data.
2. *Support Vector Machine Imputation* (SVMI): Menggunakan SVM untuk memprediksi dan mengimputasi data yang hilang.

3. *Bayesian Principal Component Analysis (*BPCA): Menggunakan model probabilistik untuk estimasi nilai yang hilang.
4. *Event Covering* (EC): Metode yang berfokus pada pemodelan hubungan antar variabel untuk mengimputasi data yang hilang.

Imputasi data yang hilang sangat penting dalam analisis data untuk beberapa alasan. Pertama, itu memungkinkan peneliti untuk mempertahankan sampel data yang lebih besar, meningkatkan kekuatan statistik analisis. Kedua, dengan mengurangi bias yang disebabkan oleh data yang hilang, imputasi membantu dalam membuat estimasi yang lebih akurat dan keputusan yang lebih tepat. Ketiga, imputasi memungkinkan penggunaan metode analisis data yang memerlukan dataset lengkap.

Imputasi data yang hilang adalah langkah kritis dalam pra-pengolahan data yang memastikan integritas dan keakuratan analisis data. Dengan berbagai metode imputasi yang tersedia, penting untuk memilih metode yang paling sesuai dengan jenis data dan mekanisme kehilangan data. Pendekatan modern yang memanfaatkan algoritma machine learning menawarkan solusi yang lebih robust dan akurat untuk mengatasi masalah data yang hilang. Namun, penting juga untuk mempertimbangkan kompleksitas dan keterbatasan masing-masing metode dalam konteks data dan tujuan analisis yang spesifik.

7.2.6 *Noise Identification*

Dalam dunia pengolahan data dan pembelajaran mesin, *noise* atau kebisingan data merupakan salah satu tantangan utama yang harus dihadapi. *Noise* adalah informasi yang tidak diinginkan atau kesalahan yang terdapat dalam data, yang dapat mengganggu analisis dan model pembelajaran mesin, sehingga menghasilkan output atau prediksi yang tidak akurat. Identifikasi *noise* adalah langkah penting dalam proses pra-pengolahan data untuk memastikan kualitas data yang akan dianalisis atau digunakan dalam pembelajaran mesin.

Noise dapat disebabkan oleh berbagai faktor, termasuk kesalahan pengukuran, kesalahan transmisi data, atau kesalahan entri data. Dalam konteks pembelajaran mesin, *noise* dapat mempengaruhi performa model dengan cara mengurangi akurasi klasifikasi, memperpanjang waktu pembangunan model, dan mengurangi ukuran serta interpretasi dari model yang dibangun (Johnson and Khoshgoftaar, 2022).

Noise dalam data dapat dikategorikan menjadi dua jenis utama, yaitu class *noise* dan attribute *noise*. Class *noise* terjadi ketika label kelas dari contoh data salah, sedangkan attribute *noise* terjadi ketika nilai atribut dari contoh data tidak akurat. Kedua jenis *noise* ini memiliki dampak yang signifikan terhadap kualitas data dan efektivitas model pembelajaran mesin.

Identifikasi *noise* merupakan tugas yang kompleks dan memerlukan pendekatan khusus untuk mengidentifikasi dan memisahkan *noise* dari data yang valid. Beberapa metode telah dikembangkan untuk identifikasi *noise*, termasuk analisis statistik dan penggunaan algoritma khusus yang dirancang untuk mendeteksi ketidaksesuaian dalam data.

Setelah *noise* teridentifikasi, ada dua pendekatan utama yang dapat diambil untuk mengatasinya: memodifikasi dan membersihkan data, dan merancang algoritma pembelajaran mesin yang robust terhadap *noise*. Pendekatan pertama melibatkan teknik seperti filtering, di mana data yang diidentifikasi sebagai *noise* dihilangkan atau dikoreksi. Pendekatan kedua melibatkan pengembangan model pembelajaran mesin yang dapat menoleransi tingkat *noise* tertentu dalam data.

Penelitian telah menunjukkan bahwa penggunaan filter *noise* dapat meningkatkan hasil analisis data dan performa model pembelajaran mesin dengan mengurangi atau mengeliminasi pengaruh *noise*. Filter *noise* dirancang khusus untuk mengatasi class *noise* dan telah terbukti efektif dalam meningkatkan akurasi klasifikasi dalam kehadiran *noise*.

Identifikasi dan penanganan *noise* dalam data adalah langkah kritis dalam proses analisis data dan pembelajaran mesin. Dengan memahami sumber dan jenis *noise*, serta menerapkan metode yang tepat untuk identifikasi dan eliminasi *noise*, dapat

meningkatkan kualitas data dan efektivitas model pembelajaran mesin. Pendekatan yang komprehensif, yang mencakup baik pembersihan data maupun pengembangan algoritma yang robust terhadap *noise*, adalah kunci untuk mengatasi tantangan yang disebabkan oleh *noise* dalam data.

7.3 *Data Reduction* (Reduksi Data)

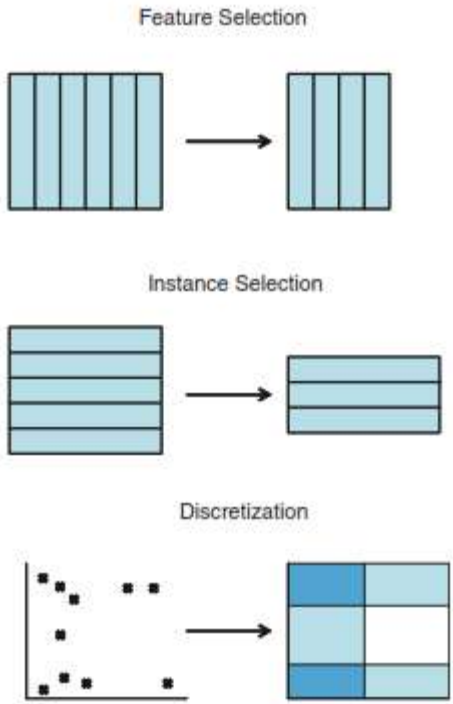
Dalam era big data, pengolahan dan analisis data menjadi tantangan tersendiri karena volume data yang sangat besar. Reduksi data menjadi salah satu strategi kunci untuk mengatasi tantangan ini, dengan tujuan untuk memperkecil volume data tanpa kehilangan informasi penting yang terkandung di dalamnya. Reduksi data tidak hanya memudahkan proses analisis tetapi juga meningkatkan efisiensi waktu dan sumber daya komputasi.

Reduksi data dapat dilakukan melalui berbagai teknik yang bertujuan untuk mengurangi dimensi data, baik dari sisi atribut maupun jumlah instansi. Teknik-teknik ini termasuk seleksi fitur (*Feature Selection*, FS), *Instance Selection* (*Instance Selection*, IS), dan diskritisasi .

1. Seleksi Fitur (FS): Teknik ini bertujuan untuk mengidentifikasi dan memilih subset atribut yang paling relevan dari dataset. Dengan mengeliminasi fitur yang tidak relevan atau redundan, FS membantu meningkatkan kecepatan dan efektivitas model pembelajaran mesin serta memudahkan interpretasi data.
2. *Instance Selection* (IS): Teknik ini melibatkan pemilihan subset instansi atau contoh dari dataset yang besar untuk menghasilkan representasi yang lebih kecil namun tetap mampu mencerminkan karakteristik keseluruhan data. IS berguna untuk mengurangi waktu pelatihan model tanpa mengorbankan akurasi yang signifikan.
3. Diskritisasi: Proses ini mengubah atribut numerik menjadi nominal dengan mengelompokkan nilai-nilai ke dalam beberapa kategori atau bin. Diskritisasi membantu mengurangi kompleksitas data dan mempermudah analisis terutama pada data yang memiliki distribusi yang luas.

Manfaat utama dari reduksi data adalah peningkatan efisiensi dalam pengolahan dan analisis data. Dengan mengurangi jumlah fitur dan instansi, reduksi data dapat mengurangi kompleksitas komputasi dan mempercepat proses pembelajaran mesin. Selain itu, reduksi data juga dapat meningkatkan kualitas model dengan mengurangi overfitting dan mempermudah interpretasi hasil analisis.

Reduksi data merupakan strategi penting dalam pengolahan data skala besar. Melalui teknik-teknik seperti seleksi fitur, *Instance Selection*, dan diskritisasi, reduksi data membantu mengatasi tantangan komputasi dan analisis yang dihadapi dalam era big data. Dengan demikian, reduksi data tidak hanya meningkatkan efisiensi dan efektivitas proses analisis tetapi juga memastikan integritas informasi yang diperoleh dari data tersebut.



Gambar 7.2. Bentuk-bentuk Reduksi data ((García, Luengo and Herrera, 2015)

7.3.1 Pemilihan Fitur (*Feature Selection*)

Dalam dunia data yang semakin berkembang, kita seringkali dihadapkan pada dataset yang sangat besar dengan banyak fitur yang tidak semua relevan untuk analisis atau model pembelajaran mesin yang kita ingin kembangkan. Di sinilah peran dari "Feature Selection" atau seleksi fitur menjadi sangat penting. Feature Selection adalah proses memilih subset fitur yang paling relevan dari dataset untuk digunakan dalam pembuatan model. Tujuannya adalah untuk memperbaiki performa model, mengurangi kompleksitas komputasi, dan membantu dalam interpretasi data yang lebih baik.

Feature Selection membantu dalam mengurangi dimensi data dengan menghilangkan fitur-fitur yang tidak penting atau redundan. Ini tidak hanya mempercepat proses pelatihan model tetapi juga meningkatkan akurasi model dengan mengurangi risiko overfitting. Selain itu, dengan mengurangi jumlah fitur, model menjadi lebih sederhana dan lebih mudah untuk diinterpretasikan, yang sangat penting dalam banyak aplikasi praktis.

Ada berbagai teknik dalam melakukan seleksi fitur, yang dapat dibagi menjadi tiga kategori utama: metode filter, metode wrapper, dan metode embedded (García, Luengo and Herrera, 2015).

1. Metode Filter: Metode ini mengevaluasi kepentingan fitur berdasarkan karakteristik statistik dari data yang tidak bergantung pada model yang digunakan. Contohnya termasuk penggunaan mutual information, chi-square test, dan *correlation coefficient scores*. Metode filter umumnya lebih cepat dan kurang kompleks secara komputasi dibandingkan dengan metode wrapper dan embedded.
2. Metode Wrapper: Metode ini menggunakan model prediktif sebagai bagian dari proses seleksi fitur, untuk menilai pentingnya atau kontribusi dari setiap fitur. Teknik ini melibatkan pencarian subset fitur yang optimal dengan mengevaluasi kombinasi fitur berdasarkan performa model. Contoh dari metode wrapper adalah *recursive feature elimination* (RFE).

3. Metode Embedded: Metode ini mengintegrasikan proses seleksi fitur sebagai bagian dari proses pelatihan model dan menggunakan kriteria pelatihan model untuk menentukan kepentingan fitur. Contohnya termasuk LASSO (*Least Absolute Shrinkage and Selection Operator*) untuk regresi, yang menambahkan penalty terhadap jumlah fitur untuk mengontrol kompleksitas model.

Salah satu tantangan utama dalam seleksi fitur adalah menemukan subset fitur yang optimal dari kombinasi fitur yang sangat besar, yang seringkali membutuhkan pencarian ekstensif dan bisa sangat memakan waktu. Selain itu, metode yang berbeda mungkin menghasilkan subset fitur yang berbeda, yang menimbulkan pertanyaan tentang stabilitas dan keandalan metode seleksi fitur.

Feature Selection adalah proses kritis dalam pengolahan data dan pembuatan model pembelajaran mesin yang efektif. Dengan memilih subset fitur yang paling relevan, kita dapat meningkatkan performa model, mengurangi waktu pelatihan, dan mempermudah interpretasi model. Meskipun ada berbagai teknik yang tersedia, pemilihan metode yang tepat bergantung pada konteks spesifik dari masalah yang dihadapi, serta sifat dari dataset yang digunakan. Memahami kelebihan dan keterbatasan dari masing-masing metode adalah kunci untuk melakukan seleksi fitur yang sukses.

7.3.2 Pemilihan Instansi (*Instance Selection*)

Dalam dunia pembelajaran mesin dan pengenalan pola, "*Instance Selection*" (IS) atau *Instance Selection* merupakan teknik reduksi data yang sangat penting. Teknik ini melibatkan pemilihan subset instansi yang relevan dari kumpulan data yang besar. Tujuan utamanya adalah untuk meningkatkan efisiensi algoritma data mining dengan mengurangi ukuran data, memfokuskan pada data yang relevan untuk tugas spesifik, dan meningkatkan kualitas data dengan menghilangkan instansi redundan dan berisik. Dengan terus bertambahnya ukuran database, IS menjadi semakin penting, dan baik peneliti maupun praktisi memberikan perhatian lebih terhadapnya. Strategi seleksi yang tepat dapat menghasilkan

peningkatan signifikan dalam performa data mining(Luengo *et al.*, 2020).

Instance Selection membantu dalam mengatasi tantangan yang disebabkan oleh dataset besar dalam pembelajaran mesin. Dengan memilih subset instansi yang paling relevan, IS dapat secara signifikan mengurangi waktu komputasi dan sumber daya yang diperlukan untuk pelatihan model. Selain itu, dengan mengeliminasi data yang redundan atau tidak relevan, IS dapat meningkatkan akurasi klasifikasi dan membuat model lebih mudah untuk diinterpretasikan.

Berbagai algoritma telah dikembangkan untuk *Instance Selection*, yang dapat diklasifikasikan menjadi beberapa metode, seperti decremental, batch, hybrid, dan wrapper. Beberapa contoh algoritma meliputi Edited Nearest Neighbor (ENN), Repeated-ENN (RENN), Multiedit, dan Modified Edited Nearest Neighbor (MENN). Algoritma-algoritma ini bertujuan untuk mengurangi ukuran set pelatihan sambil mempertahankan atau meningkatkan akurasi klasifikasi.

Selain itu, ada algoritma yang berfokus pada pendekatan kondensasi, yang bertujuan untuk mempertahankan instansi yang lebih dekat dengan batas keputusan, dan algoritma edisi, yang bertujuan untuk mengedit instansi berisik dan memperhalus batas keputusan. Beberapa algoritma populer dalam kategori ini termasuk CNN, TCNN, dan MCNN.

Salah satu tantangan utama dalam *Instance Selection* adalah menemukan subset instansi yang optimal dari kombinasi instansi yang sangat besar, yang seringkali membutuhkan pencarian ekstensif dan bisa sangat memakan waktu. Selain itu, metode yang berbeda mungkin menghasilkan subset instansi yang berbeda, yang menimbulkan pertanyaan tentang stabilitas dan keandalan metode *Instance Selection*.

Sebuah studi empiris yang dilakukan pada 42 metode seleksi prototipe (PS) menggunakan dataset kecil dan menengah mengidentifikasi metode yang berkinerja terbaik dalam hal tingkat reduksi, akurasi, dan efisiensi. Hasilnya menunjukkan bahwa metode decremental umumnya berkinerja lebih baik daripada metode incremental, dengan RNN dan MSS menonjol dalam

keluarga decremental. Metode hybrid seperti DROP3, CPruner, dan NRMCS juga menunjukkan hasil yang menjanjikan.

Instance Selection adalah komponen kritis dalam pengolahan data dan pembuatan model pembelajaran mesin yang efektif. Dengan memilih subset instansi yang paling relevan, kita dapat meningkatkan performa model, mengurangi waktu pelatihan, dan mempermudah interpretasi model. Meskipun ada berbagai teknik yang tersedia, pemilihan metode yang tepat bergantung pada konteks spesifik dari masalah yang dihadapi, serta sifat dari dataset yang digunakan. Memahami kelebihan dan keterbatasan dari masing-masing metode adalah kunci untuk melakukan *Instance Selection* yang sukses.

7.3.3 Diskritisasi

Diskritisasi merupakan proses konversi data kontinu menjadi data kategorikal atau diskrit. Dalam konteks pembelajaran mesin, teknik ini memegang peranan penting karena dapat mempermudah pemrosesan data dan meningkatkan performa model dengan mengurangi kompleksitas data yang diberikan. Proses diskritisasi sangat bergantung pada tugas Data Mining (DM) yang dihadapi, yang menjelaskan hubungannya dengan metode dan masalah lain dalam bidang ini (Chen and Anderson, 2023).

Dalam literatur, berbagai teknik diskritisasi telah diusulkan, mencerminkan pentingnya proses ini dalam mempersiapkan data untuk analisis pembelajaran mesin. Pendekatan yang berbeda telah dikembangkan, termasuk yang berbasis pada entropi informasi, uji statistik χ^2, kemungkinan, dan set kasar, antara lain. Metode-metode ini diklasifikasikan berdasarkan kriteria seperti univariat/multivariat, supervisi/tanpa supervisi, top-down/bottom-up, global/lokal, statis/dinamis, dan lebih lanjut, yang membantu dalam identifikasi diskritisasi terbaik untuk setiap situasi.

1. Taxonomi Diskritisasi
   Sebuah taxonomi diskritisasi yang komprehensif dan terbaru telah diusulkan, berdasarkan properti utama yang diamati dalam metode diskritisasi. Taxonomi ini membantu dalam mengkarakterisasi kelebihan dan kekurangan metode diskritisasi dari sudut pandang teoretis,

memudahkan pemilihan diskritisasi yang tepat. Teknik-teknik diskritisasi yang paling sering dibandingkan dalam studi ini termasuk EqualWidth, EqualFrequency, MDLP, ID3, ChiMerge, 1R, D2, dan Chi2, yang menunjukkan pentingnya mereka dalam penelitian.

2. Pendekatan Diskritisasi Baru
   Selain itu, penelitian terus berlanjut dalam mendefinisikan pendekatan diskritisasi baru seperti diskritisasi fuzzy, yang membagi atribut domain menjadi wilayah fuzzy dengan nilai keanggotaan, dan diskritisasi sensitif biaya, yang mempertimbangkan biaya kesalahan daripada hanya meminimalkan jumlah total kesalahan. Pendekatan semi-supervisi juga telah dieksplorasi, menunjukkan ekivalensinya dengan pendekatan supervisi dalam masalah klasifikasi semi-supervisi.
3. Studi Empiris dan Analisis Komparatif
   Studi ini melibatkan jumlah dataset yang lebih besar dibandingkan dengan pekerjaan sebelumnya, dan kesimpulan yang dicapai bersifat netral terhadap diskritisasi spesifik. Penelitian ini mengkonfirmasi peningkatan yang diusulkan dalam versi terbaru dari Chi2 dalam hal akurasi, menghilangkan kebutuhan pemilihan parameter oleh pengguna. Selain itu, analisis teoretis dan analitis intensif terkait Naïve Bayes telah dikembangkan, menghasilkan pengembangan PKID dan FFD berdasarkan kesimpulan tersebut.

Diskritisasi adalah proses kunci dalam pembelajaran mesin, memungkinkan konversi data kontinu menjadi format diskrit yang lebih mudah dikelola. Dengan berbagai teknik yang tersedia, pemilihan metode diskritisasi yang tepat sangat bergantung pada konteks spesifik dari setiap tugas DM. Taxonomi yang diusulkan dan analisis komparatif dari metode diskritisasi memberikan panduan berharga bagi peneliti dan praktisi dalam memilih diskritisasi yang paling sesuai untuk kebutuhan mereka. Melalui pendekatan baru dan studi empiris, bidang diskritisasi terus berkembang,

menawarkan solusi yang lebih efektif dan efisien untuk tantangan pemrosesan data dalam pembelajaran mesin.

Dalam dunia pembelajaran mesin, persiapan data merupakan langkah awal yang krusial sebelum memulai proses pembelajaran atau pengklasifikasian. Dua teknik penting dalam tahap ini adalah ekstraksi fitur (*feature extraction*) dan generasi instansi (*instance generation*). Kedua teknik ini memainkan peran penting dalam mengurangi dimensi data, menghilangkan redundansi, dan meningkatkan kualitas model yang dihasilkan.

7.3.4 Ekstraksi Fitur / Generasi Instansi (*Feature Extraction/Instance Generation*)

Ekstraksi fitur adalah proses transformasi data mentah menjadi set fitur yang lebih kecil dan lebih relevan yang dapat mewakili data asli dengan baik. Proses ini tidak hanya mengurangi jumlah data yang harus diproses oleh algoritma pembelajaran mesin tetapi juga dapat meningkatkan performa model dengan menghilangkan fitur-fitur yang tidak relevan atau redundan (Chang, 2023).

Dalam ekstraksi fitur, atribut-atribut yang tidak memberikan informasi signifikan untuk tugas pembelajaran mesin dapat dihilangkan. Selain itu, beberapa atribut dapat digabungkan atau dapat berkontribusi pada pembuatan atribut pengganti buatan. Proses ini sangat penting dalam pembelajaran terawasi karena dapat membantu model lebih baik dalam memahami batas keputusan.

Sementara itu, generasi instansi berkaitan dengan modifikasi atau penyesuaian contoh atau instansi dalam dataset. Teknik ini memungkinkan pembuatan atau penyesuaian contoh pengganti buatan yang dapat merepresentasikan batas keputusan dalam pembelajaran terawasi dengan lebih baik. Dengan demikian, generasi instansi dapat meningkatkan kualitas dataset pelatihan yang pada gilirannya meningkatkan akurasi model pembelajaran mesin.

Generasi instansi dapat dilakukan melalui berbagai cara, termasuk tetapi tidak terbatas pada, penghapusan instansi yang redundan atau konfliktif, serta penambahan instansi sintetis yang

dapat membantu model memahami pola dalam data dengan lebih baik.

Kedua teknik ini sangat penting dalam proses reduksi data, yang merupakan langkah penting dalam fase penemuan pengetahuan dan penambangan data. Reduksi data tidak hanya bertujuan untuk mengurangi kompleksitas komputasi tetapi juga untuk meningkatkan kualitas model yang dihasilkan. Dengan mengurangi dimensi data dan memperbaiki kualitas instansi, ekstraksi fitur dan generasi instansi secara langsung berkontribusi pada peningkatan performa model pembelajaran mesin.

Selain itu, dalam konteks data yang sangat besar atau kompleks, teknik-teknik ini menjadi sangat penting untuk memastikan bahwa model pembelajaran mesin dapat dilatih dengan efisien tanpa kehilangan informasi penting yang diperlukan untuk membuat prediksi yang akurat.

Ekstraksi fitur dan generasi instansi adalah dua teknik kunci dalam persiapan data untuk pembelajaran mesin. Keduanya berperan penting dalam mengurangi dimensi data, menghilangkan redundansi, dan meningkatkan kualitas dataset pelatihan. Melalui proses ini, model pembelajaran mesin dapat dilatih lebih efisien dan menghasilkan prediksi yang lebih akurat. Oleh karena itu, pemahaman yang mendalam tentang kedua teknik ini sangat penting bagi para peneliti dan praktisi dalam bidang pembelajaran mesin.

Dari pembahasan di atas, dapat disimpulkan bahwa teknik *preprocessing* data, termasuk ekstraksi fitur, generasi instansi, integrasi data, normalisasi, imputasi data yang hilang, dan identifikasi kebisingan, memainkan peran krusial dalam mempersiapkan data untuk analisis dan pemodelan yang efektif. Melalui proses-proses ini, data dapat diubah menjadi format yang lebih sesuai untuk analisis, dengan mengurangi dimensi, menghilangkan redundansi, dan meningkatkan kualitas data secara keseluruhan. Ini tidak hanya meningkatkan performa model pembelajaran mesin tetapi juga efisiensi dalam pelatihan model tersebut, yang pada akhirnya berkontribusi pada keakuratan dan keandalan hasil analisis.

Selanjutnya, reduksi data melalui seleksi fitur, *Instance Selection*, dan diskritisasi menunjukkan pentingnya memilih informasi yang paling relevan dari dataset besar untuk memastikan analisis yang efisien dan efektif. Proses *preprocessing* dan reduksi data ini membantu dalam mengatasi tantangan yang disebabkan oleh volume data yang besar dan kompleksitas data, memungkinkan peneliti dan praktisi untuk menghasilkan wawasan yang lebih mendalam dan membuat keputusan yang lebih tepat berdasarkan data. Dengan demikian, *preprocessing* dan reduksi data merupakan langkah-langkah penting yang tidak hanya memperkuat fondasi analisis data tetapi juga memperluas kemungkinan penemuan baru dalam bidang pembelajaran mesin dan analisis data.

DAFTAR PUSTAKA

Aspin, A. (2022) 'Data Transformation', *Pro Data Mashup for Power BI*, pp. 321–345. Available at: https://doi.org/10.1007/978-1-4842-8578-7_11.

Chang, X. (2023) 'Research on image features extraction based on machine learning algorithms', in H. Sheng and H. Dong (eds) *Eighth International Conference on Electronic Technology and Information Science (ICETIS 2023)*. SPIE, p. 75. Available at: https://doi.org/10.1117/12.2682449.

Chen, D. and Anderson, C.J. (2023) 'Categorical data analysis', in *International Encyclopedia of Education(Fourth Edition)*. Elsevier, pp. 575–582. Available at: https://doi.org/10.1016/B978-0-12-818630-5.10070-3.

Fernandes, A.A.A. *et al.* (2023) 'Data Preparation: A Technological Perspective and Review', *SN Computer Science*, 4(4), p. 425. Available at: https://doi.org/10.1007/s42979-023-01828-8.

García, S., Luengo, J. and Herrera, F. (2015) *Data Preprocessing in Data Mining*. Intelligen. Switzerland: Springer International Publishing. Available at: https://doi.org/10.1007/9783319102474.

Gomer, B. and Yuan, K.-H. (2023) 'Missing data analysis', in R.J. Tierney, F. Rizvi, and K.B.T.-I.E. of E. (Fourth E. Ercikan (eds). Oxford: Elsevier, pp. 805–818. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-818630-5.10090-9.

Hameed, W.M. and Ali, N.A. (2023) 'Missing value imputation Techniques: A Survey', *UHD Journal of Science and Technology*, 7(1), pp. 72–81. Available at: https://doi.org/10.21928/uhdjst.v7n1y2023.pp72-81.

Han, J., Kamber, M. and Pei, J. (2012) '3 - Data Preprocessing', in J. Han, M. Kamber, and J.B.T.-D.M. (Third E. Pei (eds) *The Morgan Kaufmann Series in Data Management Systems*. Boston: Morgan Kaufmann, pp. 83–124. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00003-4.

Jamshed, H. *et al.* (2019) 'Data Preprocessing: A preliminary step for web data mining', *3C Tecnología_Glosas de innovación aplicadas a la pyme*, pp. 206–221. Available at: https://doi.org/10.17993/3ctecno.2019.specialissue2.206-221.

Jassim, M.A. and Abdulwahid, S.N. (2021) 'Data Mining preparation: Process, Techniques and Major Issues in Data Analysis', *IOP Conference Series: Materials Science and Engineering*, 1090(1), p. 012053. Available at: https://doi.org/10.1088/1757-899X/1090/1/012053.

Johnson, J.M. and Khoshgoftaar, T.M. (2022) 'A Survey on Classifying Big Data with Label Noise', *Journal of Data and Information Quality*, 14(4), pp. 1–43. Available at: https://doi.org/10.1145/3492546.

Luengo, J. *et al.* (2020) 'Data Reduction for Big Data BT - Big Data Preprocessing: Enabling Smart Data', in J. Luengo et al. (eds). Cham: Springer International Publishing, pp. 81–99. Available at: https://doi.org/10.1007/978-3-030-39105-8_5.

Schildhauer, M. (2018) 'Data Integration: Principles and Practice BT - Ecological Informatics: Data Management and Knowledge Discovery', in F. Recknagel and W.K. Michener (eds). Cham: Springer International Publishing, pp. 129–157. Available at: https://doi.org/10.1007/978-3-319-59928-1_8.

Surendra Kumar Reddy Koduru (2022) 'A Comprehensive Analysis Of Normalization Approaches For Privacy Protection In Data Mining', *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 144–157. Available at: https://doi.org/10.32628/CSEIT1228529.

Zou, F. (2022) 'Research on data cleaning in big data environment', in *2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT)*. IEEE, pp. 145–148. Available at: https://doi.org/10.1109/3CBIT57391.2022.00037.

BAB 8
KONSEP DASAR DARI POLA UNTUK DATA MINING

Oleh Mutmainnah Muchtar

8.1 Pendahuluan

Dalam dunia yang semakin terhubung dan data yang semakin melimpah, pola menjadi kunci untuk membuka potensi besar yang tersimpan dalam setiap informasi yang kita miliki. Dalam data mining, konsep pola menjadi fondasi utama yang memungkinkan kita untuk menggali wawasan berharga dari lautan data yang tak terbatas. Sebagaimana sebuah peta memberikan arah di tengah kebingungan, demikian pula pola dalam data mining memberikan panduan untuk memahami dan menginterpretasikan data secara lebih mendalam (Jiawei, Micheline and Jian, 2011).

Data mining tidak hanya tentang mengekstraksi informasi dari kumpulan data besar. Namun juga menyelidiki, mengurai, dan mengungkap rahasia yang tersembunyi di dalamnya. Di bawah permukaan angka dan entitas, terdapat pola-pola yang memberikan wawasan yang berharga, memungkinkan kita untuk membuat prediksi, mengidentifikasi tren, dan mengambil keputusan yang didasarkan pada bukti yang kuat (Witten *et al.*, 2016).

Pola-pola yang tersembunyi dalam data menjadi kunci untuk memahami esensi dan mendapatkan nilai dari informasi yang tersimpan di dalamnya. Dalam bab ini, konsep dasar yang mendasari pola-pola dalam konteks data mining akan dieksplorasi. Pola tidak hanya sekedar representasi dari hubungan antar variabel atau entitas dalam dataset; mereka adalah cermin dari kompleksitas dunia nyata yang direfleksikan melalui titik-titik data (Tan, Steinbach and Kumar, 2006).

Ketika berbicara tentang pola dalam data mining, merujuk pada lebih dari sekadar kumpulan angka dan atribut. Pola-pola ini adalah pencerminan dari struktur, relasi, dan kecenderungan yang

mungkin tersembunyi di antara kerumitan data yang luas. Dari asosiasi sederhana hingga pola sekuensial yang kompleks, beragam bentuk pola yang dapat diidentifikasi dalam data mining akan dibahas. Melalui pemahaman yang mendalam tentang konsep dasar dari pola untuk data mining, pembaca akan mendapatkan pengetahuan yang diperlukan untuk menjelajahi dan memanfaatkan potensi tak terbatas yang terkandung dalam setiap titik data.

8.2 Pola Dalam Data Mining

8.2.1 Definisi Pola

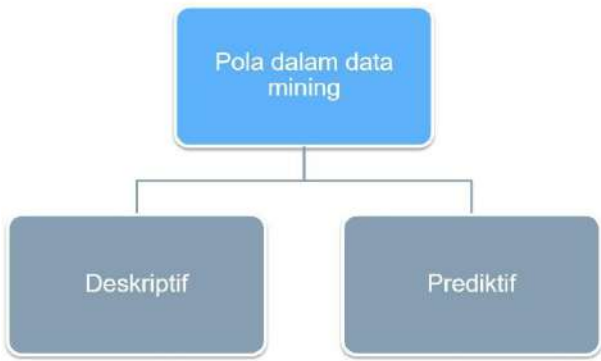
Pola dalam data mining merupakan konsep yang mendasar dan penting dalam penggalian informasi dari kumpulan data. Christopher Bishop, seorang ahli dalam bidang kecerdasan buatan, mendefinisikan pola sebagai "struktur tersembunyi dalam data yang dapat diidentifikasi oleh algoritma data mining." Hal ini menunjukkan bahwa pola bukanlah sekadar sekumpulan titik data, tetapi representasi dari struktur atau hubungan yang tersembunyi di antara titik-titik tersebut.

Pendekatan lain dalam mendefinisikan pola bersumber dari Jiawei Han dan Micheline Kamber, dalam bukunya "Data Mining: *Concepts and Techniques*", dimana mereka menjelaskan bahwa pola adalah "hubungan, *pattern* atau tren yang dapat ditemukan dalam data." Definisi ini menyoroti pentingnya mengidentifikasi hubungan dan tren yang ada di dalam data, yang dapat memberikan wawasan berharga bagi pengambil keputusan (Jiawei, Micheline and Jian, 2011).

Lebih lanjut, David Hand, seorang profesor terkemuka dalam bidang statistik, memberikan perspektif tambahan dengan mengatakan bahwa pola dalam data "mencerminkan perilaku atau kejadian yang dapat dikenali atau dipahami." Pernyataan ni menekankan bahwa pola tidak hanya sekadar fenomena statistik, tetapi juga memiliki implikasi dalam memahami perilaku manusia atau kejadian di dunia nyata.

Manfaat utama dari pola dalam data mining adalah memberikan wawasan yang berharga bagi pengambil keputusan. Dengan mengidentifikasi pola-pola yang ada dalam data, kita dapat memahami tren, meramalkan perilaku masa depan, mendeteksi

anomali, dan mengoptimalkan strategi bisnis(Shmueli *et al.*, 2019). Misalnya, dalam bisnis ritel, pola pembelian pelanggan dapat digunakan untuk menyusun strategi promosi yang lebih efektif atau mengoptimalkan stok barang. Untuk memaksimalkan manfaat dari pola-pola yang ada, penting untuk menggunakan alat dan teknik data mining yang tepat. Hal ini meliputi penggunaan berbagai metode analisis data seperti asosiasi, klastering, klasifikasi, dan regresi. Selain itu, penggunaan teknik visualisasi data juga dapat membantu dalam memahami dan mengkomunikasikan pola-pola yang ditemukan dengan lebih efektif.



Gambar 8.1. Pola dalam data mining

Pola dibagi menjadi dua jenis utama: pola deskriptif dan pola prediktif (Jiawei, Micheline and Jian, 2011). Pola deskriptif membantu dalam memahami apa yang telah terjadi dalam data tanpa membuat prediksi tentang data baru. Ini meliputi asosiasi antara item atau variabel, pola frekuensi yang sering muncul, dan klaster dari data dengan karakteristik yang serupa. Sementara itu, pola prediktif digunakan untuk membuat prediksi tentang data baru berdasarkan pola yang ditemukan dalam data historis. Ini termasuk klasifikasi data ke dalam kelas yang ditentukan, regresi untuk memprediksi nilai variabel kontinu, deteksi pola anomali, dan analisis evolusi untuk memahami perubahan pola dari waktu ke waktu. Pemahaman tentang jenis pola ini membantu dalam merencanakan dan menerapkan teknik data mining yang sesuai untuk mencapai tujuan analisis data yang diinginkan.

Dengan pemahaman yang kuat tentang pola dalam data mining, kita dapat menggali informasi berharga dan mendapatkan wawasan yang mendalam dari data yang ada. Pemahaman ini memberikan kesempatan untuk pengambilan keputusan yang lebih tepat dan strategi yang lebih efektif dalam berbagai bidang.

8.2.2 Pattern Recognition

Pattern recognition atau Pengenalan Pola adalah cabang ilmu yang berfokus pada identifikasi, klasifikasi, dan interpretasi pola atau struktur yang terdapat dalam data. Tujuannya adalah untuk mengenali pola-pola yang tersembunyi atau terulang dalam data, baik itu dalam bentuk gambar, suara, teks, atau data lainnya. Secara umum, pengenalan pola melibatkan proses pengenalan dan interpretasi pola yang ada dalam data dengan menggunakan metode komputasional atau algoritma tertentu. Proses ini sering kali melibatkan berbagai langkah, termasuk pra-pemrosesan data, ekstraksi fitur, pemilihan model, dan evaluasi kinerja (Duda, Hart and Stork, 2012; Muchtar and Cahyani, 2015).

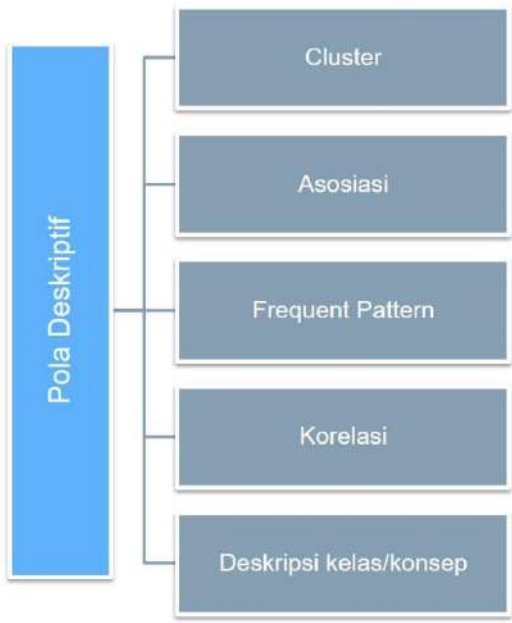
Contoh penggunaan pengenalan pola meliputi:

1. Pengenalan wajah dalam aplikasi pengenalan biometrik.
2. Klasifikasi dokumen teks berdasarkan isi atau topiknya.
3. Pengenalan suara untuk mengubah ucapan menjadi teks atau mengidentifikasi pembicara.
4. Pengenalan pola dalam citra medis untuk mendeteksi penyakit atau kondisi tertentu.
5. Klasifikasi pola cuaca untuk meramalkan kondisi cuaca di masa depan.

Pengenalan pola memiliki berbagai aplikasi di berbagai bidang, termasuk pengolahan citra, pengenalan karakter optik, pemrosesan bahasa alami, kecerdasan buatan, dan banyak lagi. Dengan kemajuan teknologi dan perkembangan metode pengenalan pola yang semakin canggih, aplikasi dan penemuan baru dalam bidang ini terus berkembang pesat.

8.3 Pola Deskriptif

Pola deskriptif atau *descriptive pattern* merupakan struktur, hubungan, atau tren yang ditemukan dalam data yang membantu dalam pemahaman tentang apa yang telah terjadi dalam data tanpa membuat prediksi tentang data baru. Menurut Jiawei Han dan Micheline Kamber dalam buku "Data Mining: Concepts and Techniques" mendefinisikan pola deskriptif sebagai "hubungan, pola atau tren yang dapat ditemukan dalam data(Jiawei, Micheline and Jian, 2011). Sementara itu, menurut Christopher Bishop, pola deskriptif adalah "struktur tersembunyi dalam data yang dapat diidentifikasi oleh algoritma pembelajaran mesin"(Bishop, 2006).



Gambar 8.2. Pola Deskriptif dan jenis-jenisnya

Beberapa jenis pola deskriptif meliputi:

1. Klaster: Pola *cluster* merupakan pola yang mengidentifikasi kelompok atau segmen data yang memiliki karakteristik atau properti yang serupa. Pola ini membantu dalam memahami struktur data dengan membagi data menjadi kelompok-

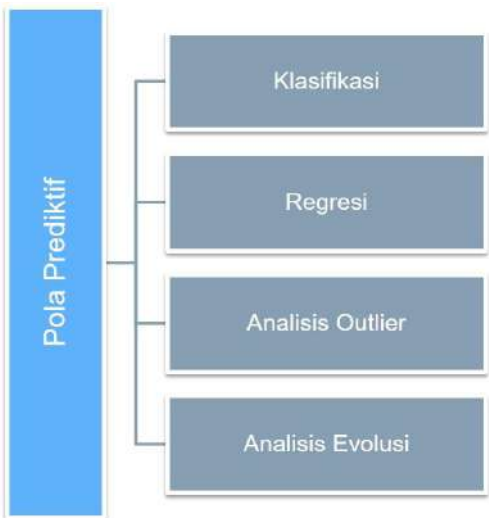
kelompok yang homogen. Contohnya adalah pengelompokan pelanggan berdasarkan pola pembelian yang serupa.

2. Asosiasi: Pola yang mengungkapkan hubungan antara *item* atau variabel dalam data, menemukan asosiasi antara *item-item* yang sering muncul bersama dalam suatu transaksi atau kejadian. Misalnya, hubungan antara produk yang sering dibeli bersamaan.
3. Frequent Patterns: Pola yang sering muncul atau berulang dalam kumpulan data. Pola-pola ini mencerminkan hubungan atau asosiasi antara *item* atau variabel dalam dataset. Dengan kata lain, frequent patterns mengidentifikasi kombinasi dari *item* atau variabel yang muncul bersama secara teratur dalam data. Contohnya, dalam bisnis ritel, frequent patterns dapat mencakup kombinasi produk yang sering dibeli bersama oleh pelanggan dalam satu transaksi.
4. *Correlations*: Pola korelasi adalah pola yang mengungkapkan hubungan statistik antara variabel atau *item* dalam data. Pola ini melibatkan analisis untuk menemukan korelasi positif, negatif, atau tidak ada korelasi antara atribut atau nilai yang terkait, atau antara dua set *item*. Contohnya adalah menemukan apakah ada hubungan antara penjualan produk tertentu dengan faktor eksternal seperti cuaca atau musim.
5. Desksripsi kelas/konsep: Pola yang membantu dalam mendeskripsikan kelas atau konsep tertentu dalam data. Proses ini melibatkan karakterisasi atau diskriminasi data untuk mendapatkan pemahaman yang lebih baik tentang karakteristik umum dari kelas atau konsep tersebut. Contohnya adalah menganalisis pola pembelian pelanggan "*big spenders*" untuk memahami karakteristik mereka yang membedakan dari pelanggan lainnya.

8.4 Pola Prediktif

Pola prediktif merupakan konsep dalam data mining yang bertujuan untuk membuat prediksi atau estimasi tentang data yang belum diketahui berdasarkan pola atau tren yang ditemukan dalam data historis. Para ahli memiliki definisi yang serupa tentang pola prediktif (Bishop, 2006; Jiawei, Micheline and Jian, 2011):

1. Jiawei Han dan Micheline Kamber, dalam buku "Data Mining: Concepts and Techniques", menggambarkan pola prediktif sebagai "memprediksi informasi yang belum diketahui atau yang akan terjadi di masa depan berdasarkan data yang ada."
2. Christopher Bishop, seorang ahli dalam kecerdasan buatan, menjelaskan bahwa pola prediktif adalah "struktur tersembunyi dalam data yang dapat digunakan untuk membuat prediksi tentang data baru."



Gambar 8.3. Pola Prediktif dan jenis-jenisnya

Jenis-jenis pola prediktif meliputi:

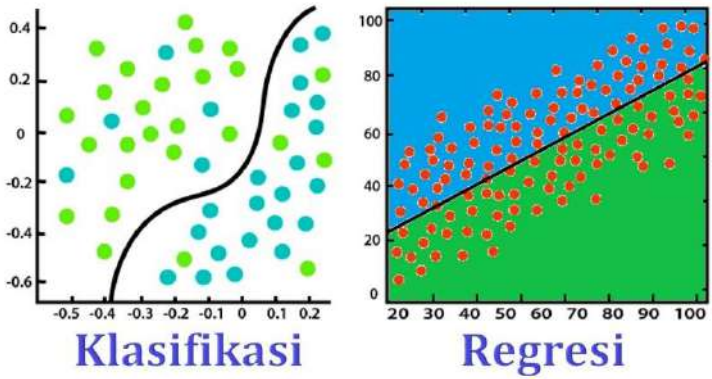
1. Klasifikasi: Pola dalam klasifikasi melibatkan pengelompokan data ke dalam kelas atau label yang telah ditentukan berdasarkan pola yang ditemukan dalam data pelatihan. Proses ini membantu dalam membuat prediksi tentang kelas atau label dari data yang belum diketahui. Misalnya, klasifikasi email sebagai spam atau bukan spam berdasarkan pada pola yang ditemukan dalam email yang sudah diketahui.
2. Regresi: Pola dalam regresi digunakan untuk memprediksi nilai variabel kontinu berdasarkan pola yang ditemukan dalam data. Ini membantu dalam memahami hubungan

antara variabel dan memprediksi nilai variabel target. Misalnya, memprediksi harga rumah berdasarkan fitur-fitur seperti luas tanah, jumlah kamar, dan lokasi geografis.

3. Analisis Outlier: Pola untuk *outlier analysis* digunakan dalam mendeteksi pola anomali atau data yang tidak biasa yang mungkin memerlukan perhatian khusus. *Outlier analysis* membantu dalam mengidentifikasi data yang tidak biasa atau tidak sesuai yang mungkin menandakan kecurangan, kesalahan, atau aktivitas ilegal. Misalnya, mendeteksi transaksi penjualan yang tidak biasa yang mungkin menandakan kecurangan atau aktivitas penipuan.
4. Analisis Evolusi: Pola untuk *evolution analysis* digunakan dalam menganalisis perubahan atau evolusi pola dari waktu ke waktu dalam data. *Evolution analysis* membantu dalam memahami tren atau perubahan dalam data seiring berjalannya waktu. Misalnya, memprediksi penjualan tahunan suatu produk berdasarkan tren penjualan dari beberapa tahun sebelumnya atau memantau perubahan dalam perilaku pelanggan dari waktu ke waktu.

Pola klasifikasi dan regresi seringkali dianggap sama karena keduanya berurusan dengan prediksi, namun keduanya berbeda dalam jenis variabel target yang diprediksi. Perbedaan utama antara pola klasifikasi dan regresi adalah jenis variabel target yang diprediksi: kategori untuk klasifikasi dan nilai numerik untuk regresi. Dengan kata lain, pola klasifikasi digunakan untuk memisahkan data menjadi kelas atau kategori yang berbeda(Muchtar and Muchtar, 2024), sedangkan pola regresi digunakan untuk memprediksi nilai numerik berkelanjutan(James *et al.*, 2013).

Gambar 8.4 adalah ilustrasi sederhana yang memperlihatkan perbedaan antara pola klasifikasi dan regresi. Dalam gambar tersebut, pola klasifikasi digunakan untuk memisahkan dua kelas data (biru dan hijau) menggunakan garis pemisah (garis lurus atau garis tidak lurus), sementara pola regresi digunakan untuk menemukan hubungan linier antara variabel x dan y (garis lurus).



Gambar 8.4. Contoh perbedaan pola klasifikasi dan regresi (Sumber : https://tutorialforbeginner.com/regression-vs-classification-in-machine-learning)

Dengan menggunakan pola prediktif, kita dapat membuat estimasi yang lebih baik tentang apa yang mungkin terjadi di masa depan berdasarkan pola dan tren yang ada dalam data historis, yang dapat digunakan untuk pengambilan keputusan yang lebih baik di berbagai bidang.

Tabel 8.1. Perbandingan secara umum antara pola deskriptif dan pola prediktif dalam data mining

Sisi perbandingan	Pola Deskriptif	Pola Prediktif
Tujuan Utama	Memahami struktur dan karakteristik data	Memprediksi atau menemukan pola di masa depan
Fokus	Menganalisis pola yang ada	Mencari pola yang belum diketahui
Jenis Analisis	Analisis deskriptif	Analisis prediktif
Hasil yang diberikan	Deskripsi atau ringkasan data	Prediksi atau estimasi
Metode yang Digunakan	Clustering, association, summarization	Classification, regression, clustering
Contoh	Mendeskripsikan profil pelanggan	Memprediksi apakah pelanggan akan

Sisi perbandingan	Pola Deskriptif	Pola Prediktif
	berdasarkan pembelian sebelumnya	membeli produk tertentu berdasarkan karakteristiknya
Pemanfaatan	Memahami tren pasar, mengidentifikasi pola pembelian, memahami perilaku pelanggan, dal lain sebagainya.	Membuat keputusan bisnis, peramalan penjualan, pemasaran yang terpersonalisasi, dan lain sebagainya.

Secara keseluruhan, pola deskriptif dalam data mining digunakan untuk merangkum dan menggambarkan pola yang ada dalam data, sementara pola prediktif digunakan untuk membuat prediksi tentang peristiwa di masa depan berdasarkan pola yang ditemukan dalam data historis. Pola deskriptif membantu kita memahami struktur dan karakteristik data, sementara pola prediktif memungkinkan kita untuk membuat keputusan berdasarkan pemahaman kita tentang data tersebut. Kedua teknik ini memiliki kelebihan dan aplikasi yang berbeda, dan pilihan antara keduanya tergantung pada tujuan analisis dan sifat data yang dihadapi. Dengan menggunakan kedua teknik ini secara bersamaan, kita dapat menggali wawasan yang lebih dalam dan membuat keputusan yang lebih baik dalam berbagai bidang, mulai dari bisnis hingga ilmu pengetahuan.

8.5 Representasi Pola dalam Berbagai Bentuk Data

Pola dalam data mining dapat muncul dalam berbagai bentuk data, mencakup data tabular, gambar, teks, suara, dan format data lainnya(Tan, Steinbach and Kumar, 2006). Pertama, dalam data tabular, pola sering tercermin melalui hubungan antar atribut atau variabel, seperti asosiasi antara item pembelian dalam transaksi ritel atau korelasi antara variabel dalam data keuangan. Ahli data mining, Jiawei Han, menggarisbawahi pentingnya pola dalam data tabular, mengatakan bahwa pola-pola ini merupakan "abstraksi atau representasi dari struktur yang signifikan, peningkatan, atau anomali yang secara konsisten muncul dalam data."

Gambar 8.5. Beberapa contoh representasi pola dalam berbagai data

Selanjutnya, dalam data gambar, pola dapat termanifestasi sebagai fitur visual atau struktur yang muncul secara berulang, seperti bentuk, warna, atau tekstur yang menggambarkan objek dalam gambar. Ahli lain, seperti Alex Graves, menekankan pentingnya penggunaan teknik deep learning dalam menemukan pola dalam data gambar yang kompleks. Selain itu, dalam data teks, pola dapat ditemukan dalam pola kata atau frasa yang sering muncul bersama, topik yang dibahas, atau sentimen yang diungkapkan. Penggunaan teknik pemrosesan bahasa alami dan analisis teks dapat membantu mengidentifikasi pola dalam data teks dengan lebih efektif.

Bahkan dalam data suara, pola dapat muncul sebagai pola gelombang suara atau karakteristik akustik tertentu yang mencerminkan informasi yang terkandung dalam suara tersebut. Teknik analisis sinyal suara dapat digunakan untuk mengekstraksi pola dari data suara dengan akurasi tinggi. Dengan demikian, representasi pola dalam berbagai bentuk data memungkinkan analisis yang holistik dan mendalam terhadap beragam informasi yang tersimpan dalam berbagai format data.

8.6 Tantangan dalam Menemukan Pola

Tantangan dalam menemukan pola dalam data mining meliputi beberapa aspek yang mendasar. Pertama, kompleksitas dan dimensi data yang besar mempersulit proses analisis, terutama ketika data memiliki beragam atribut dan variabel. Kualitas data yang rendah, seperti kekurangan data, ketidakakuratan, atau inkonsistensi, juga menjadi tantangan, karena dapat mengganggu kemampuan untuk menemukan pola yang valid dan bermakna. Selain itu, risiko overfitting juga menjadi masalah, di mana model

data mining menjadi terlalu rumit, sehingga menghasilkan pola yang tidak relevan atau tidak dapat diprediksi.

Tantangan lainnya meliputi keterbatasan sumber daya komputasi yang dapat menghambat proses analisis, interpretasi dan validasi yang benar dari pola-pola yang ditemukan, serta perhatian terhadap privasi dan keamanan data. Dengan mengatasi tantangan-tantangan ini, diharapkan pengembangan teknik dan metode data mining dapat memberikan hasil analisis yang lebih akurat dan bermakna, sambil memperhatikan etika dalam penggunaannya.

DAFTAR PUSTAKA

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer-Verlag.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012) *Pattern Classification*. John Wiley & Sons.

James, G. *et al.* (2013) *An Introduction to Statistical Learning With Applications in R*. Springer New York LLC.

Jiawei, H., Micheline, K. and Jian, P. (2011) *Data Mining: Concepts and Techniques*. Elsevier Science.

Muchtar, M. and Cahyani, L. (2015) 'Klasifikasi Citra Daun dengan Metode Gabor Co-Occurence', *ULTIMA Computing*, VII(2), pp. 39–47.

Muchtar, M. and Muchtar, R.A. (2024) 'Perbandingan Metode KNN dan SVM Dalam Klasifikasi Kematangan Buah Mangga Berdasarkan Citra HSV dan Fitur Statistik', *Jurnal Informatika dan Teknik Elektro Terapan*, 12(2), pp. 876–884.

Shmueli, G. *et al.* (2019) *Data mining for business analytics: concepts, techniques and applications in Python*. John Wiley & Sons.

Tan, P.-N., Steinbach, M. and Kumar, V. (2006) *Introduction to data mining*. Pearson Addison Wesley.

Witten, I.H. *et al.* (2016) *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science.

BAB 9
KONSEP DASAR DARI KLASIFIKASI DATA MINING

Oleh Wawan Firgiawan

9.1 Pendahuluan

Dalam era di mana data menjadi harta karun modern, kekuatan untuk menggali makna dari "gunung" data yang tak terbatas telah menjadi kunci bagi kemajuan kita dalam berbagai bidang. Di antara beragam teknik yang digunakan untuk menjelajahi kekayaan data ini, salah satu yang paling penting adalah klasifikasi dalam data mining.

Buku ini mengajak Anda untuk menjelajahi dunia klasifikasi dalam data mining, sebuah konsep yang menjadi tulang punggung bagi pemahaman yang mendalam tentang pola-pola yang tersembunyi dalam kumpulan data. Dengan menggali lebih dalam ke dalam konsep dasar dari klasifikasi, Anda akan memperoleh wawasan yang kuat tentang bagaimana algoritma ini bekerja, bagaimana mereka mengelompokkan dan memprediksi data, serta bagaimana mereka mampu memberikan wawasan yang berharga bagi berbagai aplikasi dunia nyata.

Dalam era di mana informasi menjadi hal yang paling berharga, kemampuan untuk memahami, mengelola, dan mengambil manfaat dari data telah menjadi aset yang sangat berharga bagi organisasi dan individu. Namun, dengan ledakan jumlah data yang tersedia, tantangan utama bukan lagi sebatas ketersediaan data, tetapi lebih kepada kemampuan untuk mengurai dan memahami informasi yang terkandung di dalamnya. Inilah tempat di mana klasifikasi dalam data mining menjadi sangat relevan.

Melalui teknik klasifikasi, kita dapat mengelompokkan data ke dalam kategori-kategori yang bermakna, mengidentifikasi pola yang tersembunyi, dan bahkan memprediksi perilaku atau hasil di

masa depan. Dari aplikasi sederhana seperti klasifikasi email sebagai spam atau bukan spam, hingga penemuan pola-pola kompleks dalam data medis untuk mendukung diagnosis penyakit, klasifikasi memiliki dampak yang luas dan penting dalam berbagai aspek kehidupan modern.

Melalui bab ini, Anda akan diajak untuk memahami konsep dasar klasifikasi, menjelajahi berbagai algoritma klasifikasi yang populer, mempelajari teknik evaluasi kinerja, dan menyaksikan bagaimana klasifikasi digunakan dalam berbagai aplikasi dunia nyata. Tujuan kami bukan hanya untuk memberikan pemahaman teoritis, tetapi juga untuk memberikan keterampilan praktis yang dapat Anda terapkan langsung dalam pekerjaan Anda atau proyek-proyek riset yang dilakukan.

Dengan semangat penasaran dan ketertarikan yang mendalam, mari kita mulai menjelajahi dunia klasifikasi dalam data mining. Mari kita bergerak maju untuk memahami dan menguasai konsep-konsep yang akan membuka pintu bagi pemahaman yang lebih dalam tentang dunia yang tersembunyi dalam data kita.

9.2 Pengenalan Klasifikasi dalam Data Mining

Dalam bab ini, kita akan memperkenalkan konsep klasifikasi dalam data mining. Kita akan membahas mengapa klasifikasi data sangat penting, bagaimana klasifikasi bekerja, dan bagaimana klasifikasi dapat memberikan wawasan yang berharga dalam analisis data.

9.2.1 Pengertian Dasar tentang Klasifikasi

Data mining adalah sebuah konsep yang merujuk pada proses penggalian informasi berharga dan pengetahuan yang tersembunyi di dalam tumpukan data yang terpusat atau dikenal dengan *database*. Proses ini melibatkan penerapan berbagai teknik seperti statistik, matematika, kecerdasan buatan, dan *machine learning* guna mengekstraksi serta mengidentifikasi pola atau hubungan yang signifikan dari sejumlah besar data yang tersedia (Turban, et al., 2005). Selain istilah "data mining", terdapat beberapa frasa lain yang memiliki makna serupa, seperti *Knowledge discovery in databases* (KDD), ekstraksi pengetahuan *(knowledge extraction)*,

analisis pola/data *(data/pattern analysis)*, kecerdasan bisnis *(business intelligence)*, serta arkeologi data dan penggalian data (Larose, 2014).

Dengan demikian, data mining menjadi suatu proses yang vital dalam pemrosesan data modern, memungkinkan dalam menggali informasi berharga yang dapat digunakan untuk pengambilan keputusan yang lebih baik dan efisien.

Untuk memperoleh informasi yang berharga dari kumpulan data, diperlukan penerapan teknik yang tepat untuk analisis data. Teknik-teknik tersebut mencakup metode statistik, seperti regresi yang memungkinkan untuk menemukan hubungan dan pola dalam data bahkan dalam jumlah yang banyak. Selain itu, teknik-teknik pengelompokan *(clustering)* dapat juga digunakan untuk mengelompokkan data menjadi kategori-kategori yang serupa berdasarkan karakteristik tertentu.

Tidak hanya itu, pendekatan *machine learning* juga penting dalam analisis data, di mana algoritma dipelajari dari data untuk membuat prediksi atau mengidentifikasi pola yang kompleks. Teknik-teknik seperti *decision trees, neural networks,* dan *support vector machines* sering digunakan dalam hal ini.

Klasifikasi merupakan suatu tugas dalam proses penambangan data yang melibatkan pemberian label kelas kepada setiap contoh dalam kumpulan data berdasarkan karakteristiknya. Tujuan utama dari klasifikasi adalah untuk mengembangkan model yang dapat memprediksi dengan akurat label kelas dari contoh baru berdasarkan karakteristiknya. Terdapat dua jenis utama klasifikasi: binary classification dan *multi-class classification* (Geeksforgeeks, 2023). Klasifikasi biner menerapkan pengelompokan contoh ke dalam dua kelas yang berbeda, misalnya "spam" atau "bukan spam", sementara klasifikasi kelas jamak melibatkan pengelompokan data atau informasi ke dalam lebih dari dua kelas.

Dalam konteks klasifikasi, atribut - juga dikenal sebagai fitur atau karakteristik - merujuk pada informasi yang menggambarkan setiap objek atau instance dalam dataset. Sebagai contoh, dalam klasifikasi hewan, atribut mungkin mencakup jenis makanan, jumlah kaki, atau warna bulu. Atribut ini memberikan konteks atau

ciri-ciri yang relevan yang membantu model klasifikasi dalam membedakan antara kategori atau kelas yang berbeda.

Di sisi lain, label kelas adalah kategori atau kelas yang ingin diprediksi oleh model klasifikasi untuk setiap instance data. Sebagai contoh, dalam dataset klasifikasi hewan, label kelas dapat mencakup kategori seperti "mamalia", "burung", atau "reptil". Informasi inilah yang dijadikan sebagai dasar dalam melakukan klasifikasi atau untuk mengembangkan model atau sistem yang mampu mempelajari pola dari atribut-atribut yang ada, dan kemudian menggunakan pola-pola ini untuk memprediksi label kelas yang benar untuk instance data baru yang belum diklasifikasikan sebelumnya.

Dengan memahami peran dan hubungan antara atribut dan label kelas dalam konteks klasifikasi data, kita dapat memahami bagaimana model klasifikasi bekerja dan bagaimana mereka dapat diterapkan untuk memecahkan berbagai masalah dalam berbagai domain, mulai dari kesehatan hingga keuangan.

9.2.2 Peran Klasifikasi dalam Data Mining

Klasifikasi memiliki peran penting dalam proses analisis data. Salah satu peran utamanya adalah mengorganisir data ke dalam kelompok-kelompok yang bermakna, yang memungkinkan kita untuk mengidentifikasi pola-pola yang relevan dan membuat keputusan yang tepat. Beberapa peran Klasifikasi dalam Data Mining.

1. Mengorganisir Data
   Klasifikasi membantu mengorganisir data dengan mengelompokkan objek atau *instance* ke dalam kategori-kategori yang berbeda berdasarkan atribut-atribut tertentu. Dengan cara ini, data yang kompleks dan beragam dapat diatur secara sistematis, sehingga memudahkan pemahaman dan analisis lebih lanjut (Olson & Shi, 2005).
2. Mengidentifikasi Pola
   Dengan menggunakan klasifikasi data, kita dapat mengidentifikasi pola-pola yang mungkin tersembunyi di dalamnya (Putri, et al., 2023). Sebagai contoh, dengan menganalisis pola pembelian sebelumnya, kita dapat

memprediksi perilaku pembelian masa depan pelanggan. Selain itu, klasifikasi juga memungkinkan kita untuk menemukan asosiasi antara fitur-fitur tertentu dalam data (Barkah, et al., 2020), sehingga memberikan wawasan yang berharga untuk pengambilan keputusan di berbagai bidang seperti pemasaran, keuangan, dan lainnya.

3. Membuat Keputusan yang Tepat
   Dengan memahami pola-pola yang teridentifikasi, kita dapat membuat keputusan yang lebih tepat dan terinformasi. Contohnya, dalam bisnis, klasifikasi dapat digunakan untuk memprediksi apakah pelanggan akan membeli produk tertentu, sehingga perusahaan dapat mengarahkan strategi pemasaran atau stok barang dengan lebih efisien (Dewi, et al., 2016).

Contoh tersebut hanya salah satu dari banyak aplikasi klasifikasi dalam berbagai domain. Dalam pembahasan berikutnya pada bab ini, kita akan menjelajahi lebih lanjut tentang teknik-teknik klasifikasi yang lebih canggih dan bagaimana mereka dapat diterapkan dalam skenario-skenario praktis lainnya.

9.3 Tantangan dan Kemajuan Terbaru dalam Klasifikasi

Dalam bab ini, kita akan membahas tantangan utama yang dihadapi dalam klasifikasi data mining, serta kemajuan terbaru dalam mengatasi tantangan tersebut. Beberapa tantangan dalam penyelesaian kasus dalam klasifikasi data mining adalah sebagai berikut.

1. *Overfitting* dan *Underfitting*
   Overfitting terjadi ketika model klasifikasi terlalu kompleks dan cenderung pada "menghafal" data latih dengan sangat baik, tetapi gagal untuk menggeneralisasi dengan baik pada data uji yang belum pernah dilihat sebelumnya. Di sisi lain, *underfitting* terjadi ketika model terlalu sederhana dan gagal menangkap pola-pola yang penting dalam data. Kedua masalah ini merupakan tantangan utama dalam klasifikasi, dan untuk menghindari terjadinya *overfitting* dan

underfitting pada proses klasifikasi data dalam data mining, ada beberapa teknik yang dapat diterapkan pada kasus tersebut seperti berikut:

a. Penggunaan dataset yang cukup besar
   Menggunakan dataset yang cukup besar dapat membantu mengurangi risiko *overfitting* dan *underfitting*. Dataset yang lebih besar memberikan lebih banyak variasi dan representasi data, yang memungkinkan model untuk belajar pola yang lebih umum dan mencegahnya menjadi terlalu spesifik terhadap data latih.
b. Pembagian data dengan benar
   Membagi dataset dengan benar menjadi data latih, data validasi, dan data uji sangat penting. Data latih digunakan untuk melatih model, data validasi digunakan untuk menyesuaikan parameter model dan memonitor kinerjanya selama proses pelatihan, sementara data uji digunakan untuk menguji kinerja model yang telah dilatih. Pembagian yang tepat membantu mencegah overfitting dengan memvalidasi kinerja model pada data yang tidak terlibat dalam proses pelatihan.
c. Penyetelan Hyperparameter dengan baik
d. Regularisasi
   Regularisasi adalah teknik yang digunakan untuk mencegah *overfitting* dengan menambahkan penalti pada fungsi kerugian yang digunakan dalam pelatihan model. Teknik regularisasi yang umum digunakan termasuk *L1 (Lasso)* dan *L2 (Ridge) regularization*. Regularisasi membantu membatasi kompleksitas model dan mencegahnya mempelajari pola yang terlalu spesifik pada data latih (Fei, et al., 2013).
e. Validasi Silang (Cross-Validation)
   Validasi silang adalah teknik yang digunakan untuk mengevaluasi kinerja model dengan membagi dataset menjadi beberapa *subset* yang saling tumpang tindih, dan melatih serta menguji model pada *subset-subset* tersebut secara bergantian (Baumann, 2003). Ini membantu menghasilkan estimasi yang lebih stabil dan konsisten

tentang kinerja model, serta mencegah *overfitting* karena model dievaluasi pada data yang tidak terlibat dalam proses pelatihan.

Dengan menerapkan kombinasi teknik-teknik ini secara bijaksana, dapat meminimalkan risiko overfitting dan underfitting dalam proses klasifikasi data mining, sehingga meningkatkan kinerja dan generalisasi model.

2. Data yang Tidak Seimbang *(Imbalanced Data)*

   Dalam banyak kasus, dataset klasifikasi tidak seimbang, artinya jumlah instance dalam setiap kelas tidak seimbang. Ini dapat menyebabkan model klasifikasi cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas, yang sering kali merupakan informasi yang lebih penting. Dalam bab ini, kita akan membahas strategi-strategi untuk menangani data yang tidak seimbang, seperti *oversampling*, *undersampling*, dan teknik-teknik *ensemble* (Shelke, et al., 2017).

3. Teknik dan Algoritma Terbaru dalam Klasifikasi

   Kemajuan terbaru dalam teknik dan algoritma klasifikasi telah membawa perubahan yang signifikan dalam bidang data mining dan analisis data. Berikut adalah beberapa perkembangan utama:

   a. Jaringan Saraf Tiruan *(Neural Networks)*
   b. Teknik Pembelajaran Mendalam *(Deep Learning)*

      Deep learning telah menjadi fokus utama dalam pengembangan algoritma klasifikasi baru. Model-model *deep learning* mampu menangani data dengan dimensi tinggi dan kompleksitas yang tinggi, dan mereka dapat belajar pola yang sangat abstrak dan non-linear dari data. Kemajuan dalam teknik-teknik seperti *transfer learning, reinforcement learning,* dan *self-supervised learning* telah membawa kemampuan adaptif yang luar biasa pada model klasifikasi.
   c. Pemrosesan Berbasis GPU

      Penggunaan pemrosesan berbasis GPU *(Graphics Processing Unit)* telah mempercepat pelatihan model-model klasifikasi, terutama dalam konteks deep learning.

GPU memungkinkan paralelisasi yang kuat dari operasi-operasi matriks yang kompleks, yang merupakan komponen utama dalam pelatihan model *neural networks.* Hal ini telah mengurangi waktu yang diperlukan untuk melatih model dari berminggu-minggu menjadi hanya beberapa jam atau bahkan menit.

d. Komputasi Awan
   Teknologi komputasi awan telah mengubah cara kita membangun, melatih, dan menerapkan model klasifikasi. Layanan-layanan *cloud computing* seperti *Amazon Web Services* (AWS), *Google Cloud Platform* (GCP), dan *Microsoft Azure* menyediakan infrastruktur yang skalabel dan elastis untuk pelatihan dan penerapan model-model klasifikasi dalam skala besar. Ini memungkinkan organisasi untuk dengan mudah mengelola sumber daya komputasi yang diperlukan tanpa harus menginvestasikan dalam infrastruktur fisik sendiri.

Dengan memahami tantangan utama dan kemajuan terbaru dalam klasifikasi, pembaca akan dapat mengembangkan model klasifikasi yang lebih efektif dan responsif terhadap perubahan dalam data dan lingkungan yang kompleks.

9.4 Algoritma Klasifikasi

Dalam bab ini, kita akan menjelaskan beberapa algoritma klasifikasi yang paling umum digunakan dalam data mining dan analisis data. Algoritma-algoritma ini telah terbukti efektif dalam berbagai situasi dan merupakan fondasi dari banyak aplikasi data mining modern.

1. Algoritma Naïve Bayes
   Naïve Bayes merupakan salah satu algoritma klasifikasi yang sederhana namun efektif yang didasarkan pada teorema Bayes dengan asumsi bahwa fitur-fitur dalam dataset adalah independen satu sama lain (Murphy, 2006). Metode Naïve Bayes telah menjadi pilihan yang populer karena kemudahan implementasinya dan kinerja yang baik dalam berbagai kasus penggunaan.

Berikut langkah-langkah umum untuk menerapkan algoritma Naïve Bayes dalam klasifikasi:

a. Persiapan data

   Pada proses persiapan data terdapat 3 halyang umum dilakukan, yaitu: (1) Pengumpulan data: kumpulkan dataset yang sesuai untuk tugas klasifikasi Anda, (2) Pembersihan data: lakukan pembersihan data seperti menghapus nilai yang hilang atau memperbaiki data yang tidak lengkap, dan terakhir (3) Pemisahan data: Pisahkan dataset menjadi data latih dan data uji. Biasanya, sebagian besar data digunakan untuk melatih model (misalnya, 70-80%), sementara sisanya digunakan untuk menguji kinerja model.

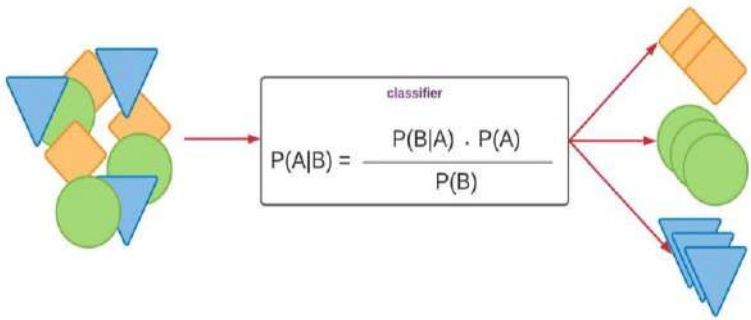
b. Ekstraksi fitur

   Dalam tahap ekstraksi fitur, langkah pertama adalah mengidentifikasi fitur-fitur yang relevan dalam dataset yang akan digunakan untuk klasifikasi. Ini melibatkan pemilihan atribut-atribut yang paling mempengaruhi target klasifikasi atau yang memiliki informasi yang paling relevan. Setelah fitur-fitur yang relevan diidentifikasi, langkah selanjutnya adalah mengubah data menjadi representasi numerik yang dapat diproses oleh algoritma Naïve Bayes. Proses ini dapat melibatkan berbagai teknik, tergantung pada jenis data yang digunakan. Misalnya, untuk data teks, teknik seperti TF-IDF *(Term Frequency-Inverse Document Frequency)* dapat digunakan untuk mengubah teks menjadi representasi numerik yang mempertimbangkan frekuensi kata dalam dokumen dan seberapa umum kata tersebut dalam seluruh korpus. Sementara itu, untuk data numerik, normalisasi sering digunakan untuk mengubah nilai-nilai atribut ke dalam rentang yang seragam, sehingga menghindari bobot yang tidak proporsional pada atribut dengan skala yang berbeda. Dengan melakukan ekstraksi fitur dengan cermat, kita dapat memastikan bahwa data yang diproses sesuai dengan format yang diperlukan oleh

algoritma Naïve Bayes untuk melakukan klasifikasi dengan akurat.

c. Pelatihan model

Dalam tahap pelatihan model, Anda perlu menginisialisasi model Naïve Bayes sesuai dengan jenis data yang Anda miliki, seperti Gaussian Naïve Bayes, Multinomial Naïve Bayes, atau Bernoulli Naïve Bayes. Setelah inisialisasi, maka selanjutnya dapat melatih model menggunakan data latih yang telah dipersiapkan sebelumnya. Proses pelatihan ini mencakup perhitungan probabilitas kelas dan probabilitas kondisional dari setiap atribut dalam setiap kelas (Jiang, 2007). Dengan demikian, model Naïve Bayes dapat "belajar" dari data latih untuk membuat prediksi yang akurat pada data baru. Berikut adalah ilustrasi dari model Naïve Bayes dalam *classifier* data (Ajaymehta, 2023).



Gambar 9.1. *Naïve Bayes Classifier*

Teorema Bayes menyatakan hubungan berikut antara variabel kelas tertentu y dan vektor fitur x_1 melalui x_n :

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P_{(y)}\ P_{(x_1,\ \ldots,x_n \mid y)}}{P_{(x_1,\ \ldots,x_n \mid y)}}$$

Gambar 9.2. *Teorema Bayes*

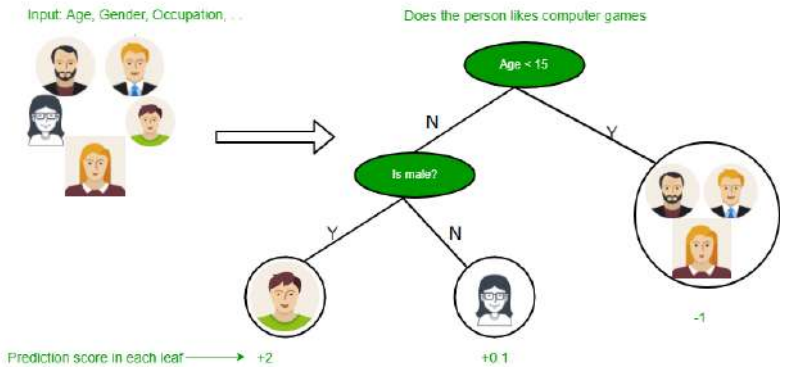
Keterangan:

1) $P(y \mid x_1, x_2, \ldots, x_n)$: adalah probabilitas kelas *y* diberikan vektor fitur x_1 hingga x_n *(probabilitas a posteriori).*
2) $P_{(x_1, x_2, \ldots, x_n \mid y)}$: adalah probabilitas vektor fitur x_1 hingga x_n diberikan kelas *y (likelihood).*
3) P(*y*): adalah probabilitas *a* priori dari kelas *y*.
4) $P_{(x_1, x_2, \ldots, x_n)}$: adalah probabilitas munculnya vektor fitur x_1 hingga x_n (pembagi normalisasi).

d. Evaluasi model
   Dalam evaluasi model, langkah pertama adalah menguji kinerja model menggunakan data uji yang telah dipisahkan sebelumnya, membandingkan kelas prediksi model dengan kelas sebenarnya dari data uji. Selanjutnya, evaluasi kinerja dilakukan menggunakan metrik seperti akurasi, presisi, *recall,* dan *F1-score* untuk menilai seberapa baik model Naïve Bayes dalam mengklasifikasikan data yang belum pernah dilihat sebelumnya.
e. Penyetelan hyperparameter (opsional)
   Jika diperlukan, lakukan penyetelan *hyperparameter* untuk meningkatkan kinerja model. Ini mungkin melibatkan mencoba berbagai nilai untuk parameter seperti *smoothing* parameter (untuk Multinomial Naïve Bayes) atau deviasi standar (untuk Gaussian Naïve Bayes).
f. Penerapan model
   Setelah model dilatih dan dievaluasi, model tersebut siap digunakan untuk mengklasifikasikan data baru yang diberikan. Gunakan model yang telah dilatih untuk melakukan prediksi pada data yang belum terlihat sebelumnya.

2. Algoritma Decision Trees
   Decision Trees atau pohon keputusan adalah algoritma klasifikasi yang intuitif dan mudah dipahami. Ia membagi data menjadi kelompok-kelompok yang lebih kecil dengan membuat serangkaian keputusan berbasis fitur-fitur dari

data. Decision Trees tidak hanya menghasilkan prediksi, tetapi juga memberikan wawasan tentang faktor-faktor yang mempengaruhi keputusan klasifikasi (Geeksforgeeks, 2024).



Gambar 9.3. Visualisasi Decision tree

Seperti yang Anda lihat dari gambar di atas, Pohon Keputusan atau *Decision tree* bekerja pada bentuk Jumlah Produk yang juga dikenal sebagai *Disjunctive Normal Form*. Pada gambar 3 di atas, kita memperkirakan penggunaan komputer dalam kehidupan sehari-hari masyarakat. Dalam Decision Tree, tantangan utama adalah melakukan identifikasi atribut untuk node akar di setiap level. Proses ini dikenal sebagai pemilihan atribut. Hal ini memiliki dua ukuran pemilihan atribut yang populer digunakan:

a) *Information gain*

   Saat menggunakan node pada pohon keputusan untuk mempartisi *instance* atau objek pelatihan menjadi *subset* yang lebih kecil, maka entropinya akan berubah. Sehingga perubahan informasi ini juga dikenal dengan ukuran perubahan entropi.

   1) Misalkan *S* adalah himpunan instance;
   2) *A* adalah atribut;
   3) S_v adalah himpunan bagian dari *S;*
   4) v mewakili nilai individual yang dapat diambil oleh atribut *A* dan Nilai *(A)* adalah himpunan semua nilai *A* yang mungkin, maka:

$$Gain(S, A) = Entropy(S) - \sum_{v}^{A} \frac{|S_v|}{|S|} \; x \; Entropy(S_v)$$

Gambar 9.4. Mencari nilai Gain

Entropy: adalah ukuran ketidakpastian suatu variabel acak, yang mencirikan ketidak murnian kumpulan contoh yang berubah-ubah. Semakin tinggi entropi, semakin banyak pula kandungan informasinya.
Misalkan *S* adalah himpunan kejadian, *A* adalah atribut, S_v adalah himpunan bagian dari *S* dengan *A = v*, dan Nilai *(A)* adalah himpunan seluruh nilai *A* yang mungkin, maka:

$$Gain(S, A) = Entropy(S) - \sum_{ve} Values(A) \frac{|S_v|}{|S|} \; x \; Entropy(S_v)$$

Gambar 9.5. Mencari nilai Gain

b) *Gini Index*

Gini Index adalah metrik yang digunakan untuk mengukur ketidakmurnian atau kekacauan dalam sebuah set data. Ini mewakili probabilitas salah mengklasifikasikan elemen secara acak dalam dataset tersebut. Dalam algoritma decision tree, seperti yang digunakan dalam tugas klasifikasi, atribut dengan nilai *Gini Index* yang lebih rendah lebih disukai karena menghasilkan pemisahan yang lebih baik dan, oleh karena itu, subset yang lebih murni (Geeksforgeeks, 2024).
Sklearn, sebuah perpustakaan pembelajaran mesin populer dalam bahasa pemrograman Python, mendukung kriteria "Gini" untuk menghitung Indeks Gini. Secara default, ia menggunakan nilai "gini" saat menghitung Indeks Gini.

$$Gini\ Index = 1 - \sum_{i=1}^{c} p_i^2$$

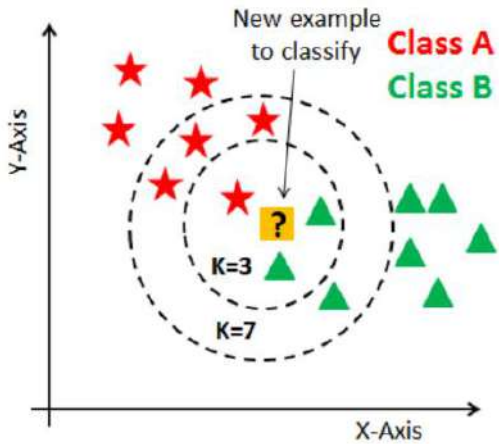
Gambar 9.6. Menghitung *Gini Index*

Gini Index adalah ukuran ketidaksetaraan atau ketidakmurnian distribusi, umumnya digunakan dalam pohon keputusan dan algoritma pembelajaran mesin lainnya. Rentang nilainya dari 0 hingga 0.5, di mana 0 menunjukkan himpunan yang murni (semua instansi berasal dari kelas yang sama), dan 0.5 menunjukkan himpunan yang maksimal tidak murni (instansi terdistribusi secara merata di seluruh kelas). Beberapa fitur dan karakteristik tambahan dari *Gini Index* adalah (Geeksforgeeks, 2024):

1) Diukur dengan menjumlahkan probabilitas kuadrat setiap hasil dalam distribusi dan mengurangkan hasilnya dari 1.
2) *Gini Index* yang lebih rendah menunjukkan distribusi yang lebih homogen atau murni, sementara Indeks Gini yang lebih tinggi menunjukkan distribusi yang lebih heterogen atau tidak murni.
3) Dalam pohon keputusan, *Gini Index* digunakan untuk mengevaluasi kualitas suatu pemisahan dengan mengukur perbedaan antara ketidakmurnian node induk dan ketidakmurnian terbobot dari node anak.
4) Dibandingkan dengan ukuran ketidakmurnian lain seperti entropi, *Gini Index* lebih cepat dalam perhitungan dan lebih sensitif terhadap perubahan dalam probabilitas kelas.
5) Salah satu kelemahan dari *Gini Index* adalah cenderung memilih pemisahan yang membuat node anak memiliki ukuran yang sama, bahkan jika itu tidak optimal untuk akurasi klasifikasi.
6) Dalam praktiknya, pilihan antara menggunakan *Gini Index* atau ukuran ketidakmurnian lainnya tergantung pada masalah dan dataset spesifik, dan seringkali memerlukan eksperimen dan penyetelan.

3. Algoritma K-Nearest Neighbors (KNN)
   Algoritma K-Nearest Neighbors (KNN) adalah salah satu algoritma klasifikasi yang sederhana namun kuat. Ia

memprediksi kelas dari instance baru dengan mempertimbangkan mayoritas kelas dari k tetangga terdekatnya dalam ruang fitur. Pendekatan ini membuat KNN cocok untuk data dengan struktur yang kompleks dan tidak linier, di mana pola-pola yang relevan mungkin tersebar di sepanjang berbagai dimensi dalam ruang fitur. Dengan prinsip sederhana ini, KNN telah terbukti efektif dalam banyak kasus, meskipun memerlukan pengaturan parameter k yang tepat untuk hasil yang optimal. Berikut adalah ilustrasi dari algoritma KNN.



Gambar 9.7. Ilustrasi Penentuan Kelas pada KNN (Shah, 2021)

Rumus umum yang digunakan dalam algoritma K-Nearest Neighbors (KNN) adalah sebagai berikut:

a) Menghitung jarak antara instance baru yang akan diprediksi dengan setiap instance dalam dataset latih menggunakan metrik jarak tertentu, seperti jarak Euclidean atau jarak Manhattan.
b) Mengidentifikasi k tetangga terdekat dari instance baru tersebut berdasarkan jarak yang dihitung.
c) Menentukan kelas mayoritas dari k tetangga terdekat tersebut.

   Secara matematis, jika x adalah instance baru yang akan diprediksi, x_i adalah instance dalam dataset latih, d adalah metrik jarak, dan K adalah jumlah tetangga yang

dipilih, maka langkah-langkah di atas dapat direpresentasikan sebagai berikut:

1) Menghitung jarak

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n}(xj - x_{ij})^2}$$

Gambar 9.8. Menghitung jarak

Di mana n adalah jumlah atribut atau dimensi fitur, dan x_j dan x_{ij} adalah nilai atribut ke-j dari instance baru x dan instance x_i dalam dataset latih, secara berturut-turut.

2) Mengidentifikasi k tetangga terdekat
Dengan menggunakan metrik jarak yang dihitung, kita mengidentifikasi k instance dengan jarak terdekat dengan instance baru x.

3) Menentukan kelas mayoritas
Setelah mengidentifikasi k tetangga terdekat, kita menentukan kelas mayoritas dari tetangga-tetangga ini. Prediksi kelas y untuk instance baru x diberikan oleh:

$$y = mode(y_1, y_1, \ldots, y_K)$$

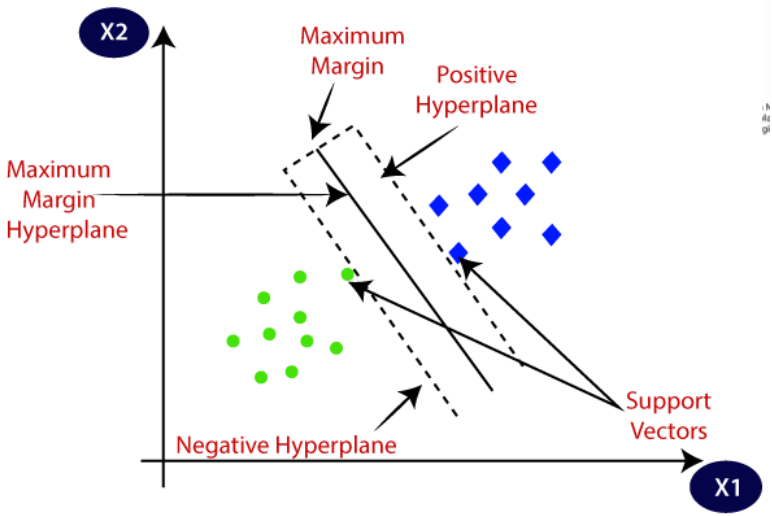
Gambar 9.9. Menghitung jarak

Di mana y_i adalah kelas dari tetangga ke- i, dan mode adalah fungsi yang mengembalikan nilai yang paling sering muncul dalam daftar nilai.

Ini adalah gambaran umum tentang langkah-langkah dan rumus dasar yang digunakan dalam algoritma KNN. Penting untuk dicatat bahwa KNN memerlukan pemilihan parameter K yang tepat dan metrik jarak yang sesuai dengan karakteristik dataset untuk hasil yang optimal.

4. Algoritma *Support Vector Machines* (SVM)
Support Vector Machines adalah algoritma klasifikasi yang kuat yang berusaha untuk menemukan *hyperplane*

garis/batas keputusan untuk memisahkan kelas dalam ruang berdimensi *n*. SVM dapat menangani data dengan jumlah fitur yang besar dan dapat diterapkan dalam berbagai kasus pengklasifikasi linier dan non-linier.



Gambar 9.10. Ilustrasi Penentuan Kelas pada SVM (Javatpoint, 2024)

Persamaan hyperplane linier dapat ditulis sebagai:

$$w^T x + b = 0$$

Gambar 9.11. Persamaan Hyperline linier

Vektor *W* mewakili vektor normal ke *hyperplane*. yaitu arah tegak lurus terhadap hyperplane. Parameter *b* dalam persamaan mewakili *offset* atau jarak hyperplane dari titik asal sepanjang vektor normal *w*.
Jarak antara titik data *x* ke-*i* dan batas keputusan dapat dihitung sebagai:

$$d_i = \frac{w^T x + b}{||w||}$$

Gambar 9.12. Persamaan *Hyperline linier*

dimana ||w|| mewakili norma *Euclidean* dari vektor bobot *w*. Norma Euclidean dari vektor normal *W.*
Adapun persamaan untuk Linear SVM classifier :

$$\bar{y} = \begin{cases} 1 : w^T x + b \geq 0 \\ 1 : w^T x + b < 0 \end{cases}$$

Gambar 9.13. Persamaan untuk Linear SVM classifier

Dimensi *hyperplane* bergantung pada fitur-fitur yang ada pada dataset. Secara umum jenis SVM dibagi menjadi 2 bagian, yaitu (Ghosh, et al., 2019):

1) *Linear SVM:* digunakan untuk data yang dapat dipisahkan secara linier. Ini berarti jika sebuah dataset dapat diklasifikasikan ke dalam dua kelas dengan menggunakan satu garis lurus, maka data tersebut disebut sebagai data yang dapat dipisahkan secara linier, dan pengklasifikasi yang digunakan disebut sebagai pengklasifikasi SVM Linier.
2) *Non-linear SVM:* digunakan untuk data yang tidak dapat dipisahkan secara linier. Ini berarti jika sebuah dataset tidak dapat diklasifikasikan menggunakan garis lurus, maka data tersebut disebut sebagai data non-linier, dan pengklasifikasi yang digunakan disebut sebagai pengklasifikasi SVM Non-linier.

9.5 Evaluasi Kinerja Klasifikasi

Dalam bagian ini, kita akan membahas cara-cara untuk mengevaluasi kinerja dari model klasifikasi. Evaluasi kinerja adalah langkah penting dalam memahami seberapa baik model kita bekerja dan seberapa dapat diandalkannya dalam menangani kasuk klasifikasi, prediksi, dan lain sebagainya. Berikut adalah teknik yang umum digunakan pada evaluasi kinerja metode klasifikasi.

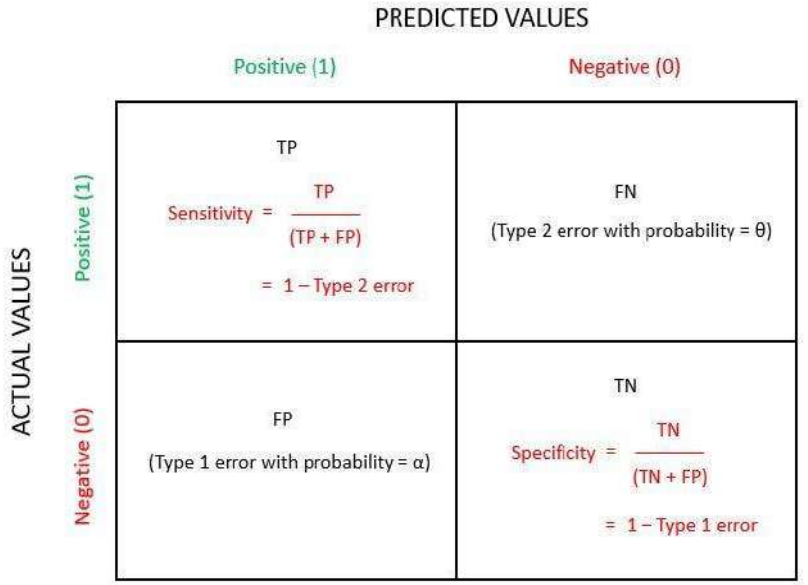
1. Metrik Evaluasi: *Akurasi, Presisi, Recall, F1-Score*
   Akurasi adalah metrik yang paling sederhana dan sering digunakan untuk mengevaluasi kinerja model. Namun, terkadang akurasi saja tidak memberikan gambaran yang

cukup tentang kinerja model, terutama jika dataset tidak seimbang. Oleh karena itu, kita juga perlu mempertimbangkan metrik-metrik lain seperti presisi, *recall, dan F1-Score*. Presisi mengukur proporsi instance positif yang benar-benar positif dari semua instance yang diklasifikasikan sebagai positif oleh model. *Recall,* di sisi lain, mengukur proporsi instance positif yang berhasil ditemukan oleh model dari semua instance positif yang sebenarnya. *F1-Score* adalah rata-rata harmonis antara presisi dan recall, yang memberikan nilai kinerja yang baik dalam kasus-kasus di mana kita peduli dengan keseimbangan antara presisi dan *recall*.

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

Gambar 9.14. Persamaan untuk Linear SVM classifier (Suresh, 2020)

Confusion matrix atau matriks kebingungan adalah alat yang digunakan dalam evaluasi kinerja model klasifikasi. Ini menggambarkan performa model dengan membandingkan nilai prediksi dengan nilai sebenarnya dari dataset.



Gambar 9.15. Menghitung *Confussion Matrix* (Suresh, 2020)

Matriks kebingungan biasanya berbentuk tabel dengan empat sel, masing-masing mewakili:

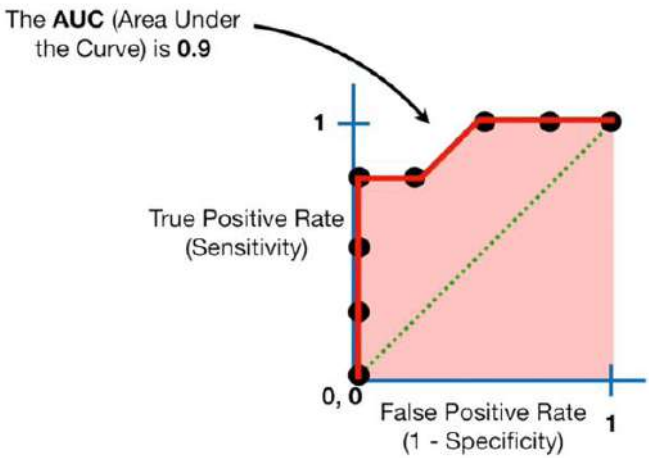
a. *True Positive* (TP): Prediksi yang benar untuk kelas positif.
b. *True Negative* (TN): Prediksi yang benar untuk kelas negatif.
c. *False Positive* (FP): Prediksi yang salah untuk kelas positif (disebut juga sebagai *Type I error*).
d. *False Negative* (FN): Prediksi yang salah untuk kelas negatif (disebut juga sebagai *Type II error*).

Dengan menggunakan matriks kebingungan, kita dapat menghitung metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk mengevaluasi kinerja model klasifikasi.

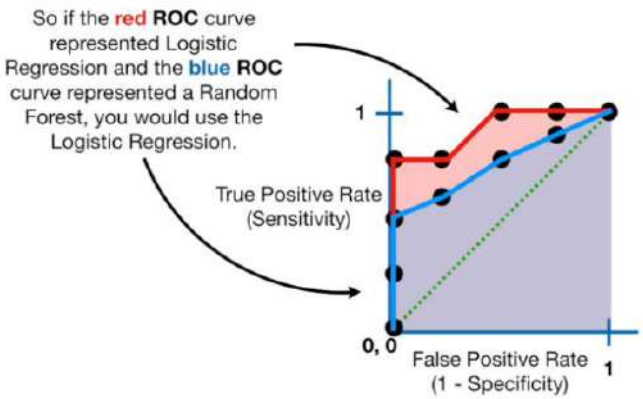
2. *Kurva ROC (Receiver Operating Characteristic)* dan *AUC (Area Under Curve)*

   Kurva ROC adalah grafik yang menampilkan hubungan antara *sensitivity (True Positive Rate)* dan *specificity (True*

Negative Rate) pada berbagai threshold pengklasifikasi. Ini memberikan gambaran tentang seberapa baik model dapat membedakan antara kelas positif dan negatif. AUC, yang merupakan area di bawah kurva ROC, adalah metrik yang sering digunakan untuk mengevaluasi secara keseluruhan kinerja model. Semakin tinggi nilai AUC, semakin baik modelnya dalam membedakan antara kedua kelas.



Gambar 9.16. Contoh Kurva AUC (Medium, 2019)



Gambar 9.17. Contoh Kurva ROC (Medium, 2019)

Sederhananya, Kurva ROC adalah grafik yang mengilustrasikan kinerja model klasifikasi pada berbagai tingkat ambang *(threshold)* (Zweig & Campbell, 1993). Ini menggambarkan perbandingan antara tingkat *true positive rate* (TPR) dan *false positive rate* (FPR) untuk berbagai nilai ambang. TPR adalah jumlah positif yang benar yang diprediksi oleh model dibagi dengan total jumlah positif yang sebenarnya, sedangkan FPR adalah jumlah negatif yang salah diprediksi sebagai positif dibagi dengan total jumlah negatif yang sebenarnya.

AUC, atau *Area Under the Curve*, adalah metrik numerik yang mengukur seberapa baik model dapat memisahkan kelas positif dan negatif (Purves, 1992). Nilai AUC berada dalam rentang 0 hingga 1, di mana nilai 1 menunjukkan kinerja yang sempurna (model sempurna memisahkan kelas dengan sempurna), dan nilai 0.5 menunjukkan kinerja yang sama dengan keberuntungan (model tidak memiliki kemampuan diskriminatif).

Kurva ROC dan AUC sering digunakan bersama-sama untuk mengevaluasi dan membandingkan kinerja model klasifikasi, terutama saat bekerja dengan data yang tidak seimbang *(imbalance)* atau ketika penting untuk memprioritaskan trade-off antara TPR dan FPR. Semakin besar nilai AUC, semakin baik model tersebut dalam memisahkan kelas positif dan negatif.

3. *Cross-Validation*

   Cross-validation adalah teknik yang penting untuk mengevaluasi kinerja model dengan menggunakan dataset yang terbatas (Baumann, 2003). Dalam *cross-validation,* dataset dibagi menjadi subset-subset yang saling tumpang tindih, dan model dilatih dan diuji pada subset-subset ini secara bergantian. Ini membantu menghasilkan estimasi yang lebih stabil dan konsisten tentang kinerja model, serta mengurangi risiko *overfitting* atau memilih model yang terlalu spesifik terhadap satu set data tertentu (Ghojogh & Crowley, 2019). Tujuannya adalah untuk memperkirakan

seberapa baik model akan berperforma pada data baru yang belum pernah dilihat sebelumnya.

Konsep dasar dari *cross-validation* adalah membagi dataset menjadi subset yang saling tumpang tindih *(folds),* di mana model dilatih pada beberapa fold dan diuji pada fold yang tersisa. Ada beberapa jenis *cross-validation*, yang paling umum adalah:

a. *K-Fold Cross-Validation:* Dataset dibagi menjadi *K-fold* yang sama besar. Model dilatih pada *K-1 fold* dan diuji pada fold yang tersisa. Proses ini diulangi *K* kali, dan hasil pengujian digabungkan.
b. *Stratified K-Fold Cross-Validation:* Mirip dengan *K-Fold,* tetapi memastikan bahwa setiap *fold* memiliki distribusi yang seimbang dari kelas target.
c. *Leave-One-Out Cross-Validation (LOOCV):* Setiap sampel digunakan sebagai data uji satu kali, sedangkan yang lainnya digunakan sebagai data latih. Ini cocok untuk dataset kecil, tetapi memerlukan waktu komputasi yang besar untuk dataset yang lebih besar.
d. *Time Series Cross-Validation:* Khusus untuk data *time series,* di mana pengujian dilakukan pada data yang lebih baru daripada data latih.

Dalam bagian ini, kita telah mendiskusikan penggunaan metrik evaluasi, kurva ROC, AUC, dan teknik *cross-validation* untuk mengevaluasi kinerja dari model klasifikasi. Dengan pemahaman yang baik tentang evaluasi kinerja, kita dapat membuat keputusan yang lebih baik tentang penggunaan dan penyesuaian model kita untuk kasus-kasus data mining yang berbeda.

9.6 Praktikum dan Studi Kasus

Pada bagian ini, kita akan melangkah lebih jauh dengan melihat penelitian terkait tentang implmentasi klasifikasi dalam data mining.

1. Implementasi Klasifikasi dengan Python

   Kita akan mempelajari langkah-langkah praktis untuk mengimplementasikan algoritma klasifikasi menggunakan

Python. Ini termasuk mempersiapkan data, membagi dataset menjadi data latih dan data uji, membangun model klasifikasi, mengevaluasi kinerjanya menggunakan metrik-metrik yang relevan, dan menyesuaikan model jika diperlukan. Kami juga akan melihat bagaimana menggunakan pustaka-pustaka populer seperti Scikit-learn untuk bahasa pemograman Python dalam proses implementasi ini.

2. Contoh Studi Kasus: Klasifikasi Sentimen pada Media Sosial
   Studi kasus pertama akan fokus pada klasifikasi sentimen pada media sosial. Kami akan menjelajahi bagaimana menggunakan teknik-teknik klasifikasi untuk menganalisis dan mengklasifikasikan sentimen dari postingan atau tweet pada platform media sosial. Ini adalah aplikasi yang penting dalam pemantauan merek, analisis opini publik, dan intelijen bisnis.
3. Contoh Studi Kasus: Klasifikasi Penyakit dengan Data Kesehatan
   Studi kasus kedua akan mempertimbangkan klasifikasi penyakit menggunakan data kesehatan. Kami akan mempelajari bagaimana menerapkan algoritma klasifikasi untuk memprediksi diagnosis penyakit berdasarkan gejala dan atribut kesehatan lainnya. Ini merupakan aplikasi yang sangat penting dalam bidang kesehatan, memungkinkan deteksi dini penyakit dan pengambilan keputusan yang lebih baik dalam perawatan pasien.
4. Contoh Studi Kasus: Klasifikasi Gambar
   Studi kasus ini akan fokus pada klasifikasi gambar, sebuah bidang yang semakin penting dalam era di mana penggunaan gambar semakin meluas dalam berbagai aplikasi, mulai dari pengenalan wajah hingga kendaraan otonom.
   Pada contoh ini, kita akan menjelajahi bagaimana menerapkan teknik-teknik klasifikasi untuk membangun model yang dapat mengenali dan mengklasifikasikan gambar ke dalam kategori-kategori yang sesuai. Contohnya, kita dapat membangun model untuk mengklasifikasikan gambar hewan menjadi kategori-kategori seperti kucing, anjing, atau

burung, atau mengklasifikasikan gambar manusia menjadi kategori-kategori berbeda berdasarkan emosi atau tindakan yang ditunjukkan.

Umunya pada pengimplementasian klasifikasi data dalam data mining, dilakukan dengan langkah-langkah sebagai berikut:

1. Pengumpulan dan Pra-pemrosesan Data: Mengumpulkan dataset gambar yang sesuai dengan tujuan klasifikasi kita. Pra-pemrosesan data mungkin melibatkan resizing gambar, normalisasi nilai piksel, dan augmentasi data untuk meningkatkan variasi dataset.
2. Pembagian Dataset: Memisahkan dataset menjadi subset latih dan uji untuk melatih dan menguji model.
3. Pembangunan Model Klasifikasi: Memilih arsitektur model yang sesuai, seperti *Convolutional Neural Networks* (CNNs), dan melatih model menggunakan dataset latih. Pengaturan hyperparameter dan optimasi model juga merupakan bagian penting dari proses ini.
4. Evaluasi Kinerja Model: Menggunakan subset uji untuk mengevaluasi kinerja model menggunakan metrik-metrik seperti akurasi, presisi, *recall, dan F1-Score.*
5. *Fine-tuning* dan Validasi Model: Jika diperlukan, melakukan *fine-tuning* pada model dan validasi menggunakan teknik seperti *cross-validation* untuk memastikan kinerja yang optimal.

Dengan mempelajari studi kasus ini, pembaca akan mendapatkan wawasan tentang bagaimana menerapkan teknik klasifikasi dalam konteks klasifikasi gambar, serta memahami tantangan dan strategi yang terlibat dalam membangun model klasifikasi yang efektif untuk tugas-tugas ini.

9.7 Penelitian Terkait dengan Klasifikasi Data Mining

Dalam bagian ini, kami akan membahas penelitian terkait yang telah dilakukan dalam domain klasifikasi data mining. Penelitian ini mencakup berbagai topik yang berkaitan dengan

pengembangan teknik, algoritma, dan aplikasi klasifikasi dalam berbagai konteks. Berikut adalah beberapa arah penelitian tentang klasifikasi pada data mining.

1. *A study on classification techniques in data mining*
   Hasil Penelitian: Penelitian ini mengulas pentingnya data mining sebagai proses untuk mengekstrak pengetahuan dari dataset besar, yang terdiri dari tiga komponen utama: Clustering atau Klasifikasi, Asosiasi Aturan, dan Analisis Urutan. Dalam konteks klasifikasi atau clustering, data dianalisis untuk menghasilkan aturan pengelompokan yang dapat digunakan untuk mengklasifikasikan data di masa depan. Data mining bertujuan untuk mengubah informasi dari dataset menjadi struktur yang dapat dimengerti melalui proses komputasi yang menggunakan metode dari kecerdasan buatan, pembelajaran mesin, statistik, dan sistem database. Terdapat enam kelas tugas umum dalam data mining, termasuk Deteksi Anomali, Pembelajaran Aturan Asosiasi, Clustering, Klasifikasi, Regresi, dan Ringkasan. Klasifikasi, sebagai teknik utama dalam data mining, digunakan untuk memprediksi keanggotaan kelompok untuk instance data dengan menggunakan teknik pembelajaran mesin. Studi ini menyajikan teknik klasifikasi dasar, seperti induksi pohon keputusan, jaringan Bayesian, dan klasifikasi tetangga terdekat, dengan tujuan memberikan tinjauan menyeluruh tentang berbagai teknik klasifikasi dalam data mining (Kesavaraj & Sukumaran, 2013).
2. *Performance Analysis of Data Mining Classification Techniques to Predict Diabetes*
   Hasil Penelitian: Studi ini menyoroti peran penting data mining dalam prediksi dan penanganan Diabetes Mellitus, yang merupakan tantangan kesehatan global yang meningkat pesat. Melalui penggunaan teknik ensemble seperti adaboost dan bagging, serta pohon keputusan J48 sebagai pembelajar dasar, penelitian ini bertujuan untuk meningkatkan akurasi prediksi diabetes. Dengan mengklasifikasikan pasien berdasarkan faktor risiko

diabetes menggunakan teknik data mining, penelitian ini dijalankan di tiga kelompok orang dewasa di Kanada. Hasil eksperimen menunjukkan bahwa metode ensemble adaboost memberikan kinerja yang lebih baik daripada metode bagging dan pohon keputusan J48 secara mandiri. Hal ini menunjukkan potensi besar teknik ensemble dalam meningkatkan akurasi sistem prediksi diabetes dan dapat menjadi landasan untuk pengembangan sistem pendukung keputusan klinis yang lebih efektif di masa depan (Perveen, et al., 2016).

3. *Application of data mining techniques in customer relationship management: A literature review and classification*

   Hasil Penelitian: Studi ini adalah tinjauan literatur pertama yang mengidentifikasi penerapan teknik data mining dalam CRM. Melalui analisis 900 artikel dari periode 2000–2006, 87 artikel terpilih dikategorikan dalam empat dimensi CRM dan tujuh fungsi data mining. Temuan menunjukkan fokus penelitian terbesar adalah pada retensi pelanggan, terutama dalam konteks pemasaran satu-satu dan program loyalitas. Model klasifikasi dan asosiasi menjadi yang paling umum digunakan dalam data mining untuk CRM. Analisis ini memberikan panduan bagi penelitian masa depan dan memfasilitasi peningkatan pengetahuan tentang aplikasi teknik data mining dalam CRM (Ngai, et al., 2009).

4. *Data mining in mining engineering: results of classification and clustering of shovels failures data*

   Hasil Penelitian: Penelitian ini menyoroti pentingnya perencanaan perawatan yang cerdas dan efisien dalam menghindari downtime yang mahal pada peralatan pertambangan. Dengan adanya kumpulan data besar tentang kinerja dan keandalan peralatan dari sistem monitoring kesehatan dan dispatch pabrikan, tambang modern memiliki kesempatan untuk melakukan perencanaan perawatan yang lebih efisien. Studi ini mengeksplorasi penerapan pendekatan klasifikasi dan klasterisasi untuk pengenalan pola dan prediksi kegagalan pada shovel tambang. Perilaku

kegagalan dari sepuluh shovel tambang selama satu tahun operasi diselidiki menggunakan teknik-teknik ini. Shovel diklasifikasikan ke dalam empat kelompok menggunakan algoritma klasterisasi k-means. Kegagalan di masa depan diprediksi menggunakan teknik klasifikasi *Support Vector Machine* (SVM). Data kegagalan historis dan waktu perbaikan digunakan untuk memprediksi jenis kegagalan berikutnya untuk semua shovel. Teknik SVM terbukti berhasil dengan akurasi prediksi lebih dari 75%. Ini merupakan upaya pertama (sejauh yang kami ketahui) dimana jenis kegagalan diprediksi berdasarkan data kegagalan/perbaikan historis untuk peralatan pertambangan. Pengelompokan shovel berdasarkan keandalannya dapat digunakan untuk alokasi peralatan dan perencanaan perawatan, yang tidak dapat dicapai dengan pemodelan keandalan tradisional. Penerapan teknik-teknik ini dengan sukses akan memberikan masukan berharga untuk pengambilan keputusan selama penjadwalan perawatan preventif (Dindarloo & Siami-Irdemoosa, 2017).

5. *Images Classification of Dogs and Cats using Fine-Tuned VGG Models*

   Hasil Penelitian: Studi ini menunjukkan bahwa menggunakan model VGG yang disesuaikan untuk mengenali berbagai ras anjing dan kucing (CDC) menghasilkan hasil yang cukup baik. Meskipun mengklasifikasikan gambar kucing dan anjing relatif mudah, mengklasifikasikan gambar berbagai ras dengan tingkat akurasi yang tinggi merupakan tantangan yang lebih besar. Namun, dengan menggunakan model VGG16 dan VGG19, kami mencapai tingkat akurasi yang cukup tinggi dalam pengujian. Akurasi pelatihan dan validasi di kedua model cukup tinggi, mencapai lebih dari 98%, yang menunjukkan bahwa model-model tersebut mampu mempelajari pola yang kompleks dalam gambar-gambar tersebut. Meskipun demikian, terdapat sedikit penurunan akurasi saat diuji, tetapi tetap mencapai tingkat yang memuaskan, dengan VGG19 memberikan hasil yang sedikit lebih baik dari VGG16. Ini menunjukkan bahwa

model VGG yang disesuaikan dapat menjadi pilihan yang baik untuk aplikasi pengklasifikasian gambar CDC (Mahardi, et al., 2020).

Dengan demikian, penelitian terkait dengan klasifikasi data mining mencakup spektrum luas topik-topik yang berkontribusi pada pengembangan teoritis dan praktis dalam bidang ini. Melalui eksplorasi penelitian ini, pemahaman tentang kemajuan terkini dan tren dalam klasifikasi data mining dapat diperluas, memberikan wawasan yang berharga untuk pengembangan model klasifikasi yang lebih baik di masa depan.

DAFTAR PUSTAKA

Ajaymehta, 2023. *Medium.* [Online] Available at: https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7. [Accessed 30 Maret 2024].

Barkah, N., Sutinah, E. & Agustina, N., 2020. Metode Asosiasi Data Mining Untuk Analisa Persediaan Fiber Optik Menggunakan Algoritma Apriori. *Jurnal Kajian Ilmiah,* 20(3), pp. 237-248.

Baumann, K., 2003. Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry,* 22(6), pp. 395-406.

Dewi, L., Hanifa, A., Setyaningrum & Eka, F., 2016. Penerapan Metode Asosiasi Menggunakan Algoritma Apriori Pada Aplikasi Analisa Pola Belanja Konsumen (Studi Kasus Toko Buku Gramedia Bintaro). *Jurnal Teknik Informatika,* 9(2).

Dindarloo, S. R. & Siami-Irdemoosa, E., 2017. Data mining in mining engineering: results of classification and clustering of shovels failures data. *International Journal of Mining, Reclamation and Environment,* 31(2), pp. 105-118.

Fei, W., Chawla, S. & Liu, W., 2013. *Tikhonov or lasso regularization: Which is better and when.* Herndon, VA, USA, IEEE.

Geeksforgeeks, 2023. *Basic Concept of Classification (Data Mining).* [Online]
Available at: https://www.geeksforgeeks.org/basic-concept-classification-data-mining. [Accessed 30 Maret 2024].

Geeksforgeeks, 2024. *Decision Tree in Machine Learning.* [Online] Available at: https://www.geeksforgeeks.org/decision-tree-introduction-example/. [Accessed 30 Maret 2024].

Ghojogh, B. & Crowley, M., 2019. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv,* 1905(12787).

Ghosh, S., Dasgupta, A. & Swetapadma, A., 2019. *A study on support vector machine based linear and non-linear pattern classification.* s.l., IEEE.

Javatpoint, 2024. *Support Vector Machine Algorithm.* [Online] Available at: https://www.javatpoint.com/machine-

learning-support-vector-machine-algorithm. [Accessed 30 Maret 2024].

Jiang, L., 2007. *Survey of improving naive bayes for classification.* Harbin, China, August 6-8, 2007: Springer Berlin Heidelberg.

Kesavaraj, G. & Sukumaran, S., 2013. *A study on classification techniques in data mining.* Tiruchengode, India, IEEE.

Larose, D. T., 2014. *Discovering Knowledge in Data : An Introduction to Data Mining.* Jakarta: John Wiley & Sons, Inc.

Mahardi, Wang, I.-H., Lee, K.-C. & Chang, S.-L., 2020. *Images Classification of Dogs and Cats using Fine-Tuned VGG Models.* Yunlin, Taiwan, IEEE.

Medium, 2019. *Memahami ROC dan AUC.* [Online] Available at: https://datasans.medium.com/memahami-roc-dan-auc-2e0e4f3638bf. [Accessed 30 Maret 2024].

Murphy, K. P., 2006. *Naive bayes classifiers.* s.l.:University of British Columbia.

Ngai, E., Xiu, L. & Chau, D., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications,* 36(2), pp. 2592-2602.

Olson, D. & Shi, Y., 2005. *Introduction to Business Data Mining.* Jakarta: McGraw-Hill Companies,Incorporated.

Perveen, S., Shahbaz, M., Guergachi, A. & Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science,* Volume 82, pp. 115-121.

Purves, R. D., 1992. Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). *Journal of pharmacokinetics and biopharmaceutics,* Volume 20, pp. 211-226.

Putri, H., Irianti, A. & Firgiawan, W., 2023. Pengelompokan Kecamatan Berdasarkan Produktifitas Hasil Perkebunan di Kabupaten Polewali Mandar Menggunakan Metode Fuzzy C-Means. *Journal of Computer and Information System (J-CIS),* 6(2), pp. 40-50.

Shah, R., 2021. *Introduction to k-Nearest Neighbors (kNN) Algorithm.* [Online] Available at: https://ai.plainenglish.io/introduction-

to-k-nearest-neighbors-knn-algorithm-e8617a448fa8. [Accessed 30 Maret 2024].

Shelke, M. S., R, P. & Deshmukh, V. K., 2017. A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering & Research (IJRTER),* 3(4), pp. 444-449.

Suresh, A., 2020. *What is a confusion matrix?.* [Online] Available at: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5. [Accessed 30 Maret 2024].

Turban, E., Prabantini, D., Aronson, J. E. & Ting-Peng, L., 2005. *Desision support systems and intelligent systems (diterjemahkan oleh Dwi Prabantini).* Yogyakarta: Andi.

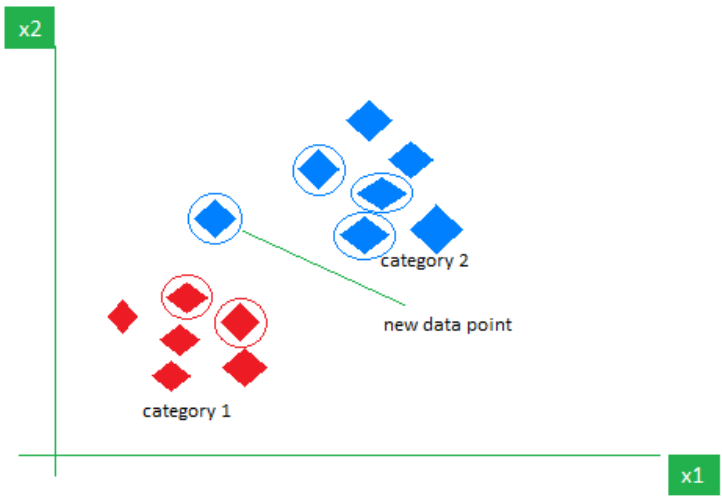
Zweig, M. H. & Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry,* 39(4), pp. 561-577.

BAB 10
PENDEKATAN KLASIFIKASI

Oleh Nurhikma Arifin

10.1 K-Nearest Neighbors

K-Nearest Neighbors (K-NN) merupakan salah satu algoritma klasifikasi yang digunakan dalam data mining untuk memprediksi kelas suatu data berdasarkan mayoritas kelas dari tetangga terdekatnya. Konsep dasar dari K-NN adalah bahwa data yang memiliki atribut atau fitur yang serupa cenderung berada dalam kelas yang sama. Algoritma K-NN bekerja dengan mengklasifikasikan objek berdasarkan jarak terdekatnya dari data pelatihan. Fungsinya adalah memberikan label pada data berdasarkan kumpulan data pembelajaran yang memiliki jarak paling dekat dengan objek yang akan diklasifikasikan. Jumlah tetangga (K) yang dipertimbangkan dalam pengklasifikasian suatu data menjadi faktor kunci dalam algoritma ini (Amna et al., 2023).



Gambar 10.1. Ilustrasi *K-Nearest Neighbors*
(Sumber : geeksforgeeks.org)

Proses KNN ini sangat sederhana dan mudah diterapkan, terutama dalam konteks klasifikasi. Selain itu, algoritma ini juga dapat digunakan untuk tujuan prediksi atau estimasi. Dalam hal ini, objek yang akan diklasifikasikan akan diberikan label atau kategori yang sama dengan mayoritas K tetangga terdekatnya.

Secara konseptual, K-NN memiliki kesamaan dengan metode *clustering*, di mana data yang baru akan dikelompokkan berdasarkan jaraknya terhadap data yang sudah ada atau tetangga terdekat. Meskipun K-NN dan metode *clustering* memiliki tujuan yang berbeda, yaitu klasifikasi dan pengelompokan, keduanya menggunakan pendekatan yang mirip dalam memanfaatkan informasi jarak antara objek-objek data.

Berikut Cara Kerja K-NN:

1. Menentukan Nilai K (jumlah tetangga terdekat). Jumlah K dapat dipilih berdasarkan analisis data dan kebutuhan spesifik tugas klasifikasi. Nilai K mengindikasikan berapa banyak tetangga terdekat yang akan diambil dalam pertimbangan untuk membuat prediksi. Pemilihan nilai K dapat dilakukan melalui validasi silang atau percobaan berulang.
2. Menghitung jarak antara data yang akan diprediksi dengan semua data dalam set pelatihan. Jarak ini dapat diukur dengan berbagai metrik, seperti *Euclidean distance* atau *Manhattan distance*. Rumus umum *euclidean distance* yang digunakan sebagai berikut (Helilintar et al., 2017) :

   $$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$$

   Keterangan:
   - x_1 = Sampel Data
   - x_2 = Data Uji / Testing
   - i = Variabel Data
   - d = Jarak
   - P = Dimensi Data
3. Menentukan tetangga terdekat berdasarkan jarak yang diukur sebelumnya.
5. Melakukan *Voting* mayoritas diantara tetangga terdekat.

6. Menentukan Prediksi label kelas atau nilai yang akan diprediksi berdasarkan hasil *voting* dari tetangga terdekat.
7. Mengevaluasi model K-NN dengan menggunakan metrik evaluasi yang sesuai.

Kelebihan algoritma K-NN:

1. Sederhana dan Mudah Dimengerti. Tidak memerlukan asumsi kompleks dan cocok untuk pemula dalam pemodelan data.
2. Fleksibel digunakan untuk berbagai masalah baik klasifikasi maupun regresi.
3. *Non-Parametrik* yang berarti tidak ada asumsi tertentu tentang distribusi data sehingga memungkinkan model untuk beradaptasi dengan data dengan baik.
4. Mampu Menangani Data Multikelas dan dapat bekerja dengan baik dalam kasus di mana hubungan antara fitur dan *output* tidak linear.
5. Pelatihan Cepat karena K-NN adalah algoritma "*lazy learning*", yang berarti tidak memerlukan fase pelatihan yang panjang. Modelnya cepat untuk diterapkan pada data baru.

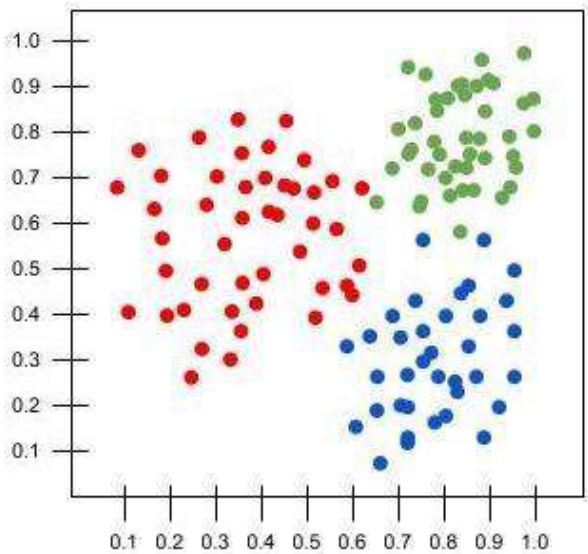
Kekurangan algoritma K-NN:

1. Sensitif terhadap *outlier* sehingga mempengaruhi hasil prediksi.
2. Mahal secara komputasional, terutama ketika *dataset* sangat besar. Perhitungan jarak harus dilakukan untuk setiap pasangan data poin.
3. Pentingnya Pemilihan Nilai K yang Tepat. Nilai K yang terlalu kecil dapat menyebabkan model menjadi rentan terhadap kebisingan, sementara nilai K yang terlalu besar dapat menghasilkan prediksi yang kurang sensitif.
4. Efek Dimensi Tinggi (*Curse of Dimensionality*) menyebabkan performa K-NN menurun di mana sebagian besar data poin menjadi semakin jauh satu sama lain dalam ruang fitur.

Selain kelebihan dan kekurangan dari algoritma K-NN yang perlu dipahami dengan cermat evaluasi model juga menjadi krusial. Metrik seperti akurasi dan presisi memberikan gambaran sejauh mana K-NN dapat mengklasifikasikan data dengan tepat. Evaluasi akan membantu penyetelan parameter, terutama nilai K, dan mengidentifikasi kelemahan model seperti sensitivitas terhadap *outlier*. Dengan evaluasi yang cermat, praktisi data mining dapat membuat keputusan yang terinformasi, memperbaiki model dan memastikan bahwa penerapan K-NN sesuai dengan tujuan analisis.

10.2 Naïve Bayes

Naïve Bayes merujuk pada metode klasifikasi yang bergantung pada konsep probabilitas dan statistik yang diilustrasikan oleh peneliti Inggris bernama Thomas Bayes (Sigid Widodo et al., 2023). Metode ini mengevaluasi probabilitas kejadian di masa depan berdasarkan pengalaman sebelumnya, yang dikenal dengan nama Teorema Bayes. *Naïve Bayes* menerapkanan *Teorema Bayes* dengan asumsi *independensi* antara fitur-fitur yang digunakan dalam klasifikasi. Hal ini menunjukkan algoritma *Naïve Bayes* mengasumsikan bahwa nilai-nilai fitur yang diamati dalam data pelatihan adalah *independen* satu sama lain ketika kelasnya diketahui (Sabry, 2023). Meskipun asumsi ini seringkali tidak sepenuhnya realistis dalam situasi dunia nyata, algoritma *Naïve Bayes* sering memberikan hasil yang baik, terutama dalam klasifikasi teks dan pengenalan pola (Andreas, 2016).



Gambar 10.2. Ilustrasi *Naive Bayes*
(Sumber : ml-science.com)

Dalam konteks klasifikasi, algoritma Naïve Bayes digunakan untuk memprediksi probabilitas kelas target (misalnya, apakah sebuah email adalah spam atau bukan) berdasarkan nilai-nilai fitur yang diamati dalam data pelatihan. Kemudian, saat diberikan contoh baru yang belum dilihat sebelumnya, algoritma menghitung probabilitasnya untuk setiap kelas dan memilih kelas dengan probabilitas tertinggi sebagai prediksi akhir. Algoritma *Naïve Bayes* adalah salah satu teknik klasifikasi yang paling umum digunakan dalam *machine learning* karena efisiensinya dan kemudahannya. Adapun persamaan algoritma *Naïve Bayes* sebagai berikut:

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{P(C_k).P(x_1|C_k).P(x_2|C_k)\dots P(x_n|C_k).}{P(x_1).P(x_2)\dots P(x_n)}$$

Keterangan:

- $P(C_k|x_1, x_2, \dots, x_n)$ adalah probabilitas kelas C_k terjasi jika fitur-fitur $x_1, x_2, \dots, x_n$ diamati
- $P(\ C_k)$ adalah probabilitas prior dari kelas C_k
- $P(x_i|C_k)$ adalah probabilitas bahwa fitur x_i diamati dalam kelas C_k

- $P(x_i)$ adalah probabilitas prior dari fitur x_i, yang dapat diabaikan dalam prediksi karena merupakan konstanta untuk setiap kelas

Berikut Cara Kerja *Naïve Bayes*:

1. Menghitung Probabilitas *Prior* Kelas $P(C_k)$: untuk setiap kelas C_k , hitung jumlah kemunculan kelas tersebut di dalam data pelatihan dan bagi dengan total *instance* untuk mendapatkan probabilitas *prior* $P(C_k)$.
2. Menghitung Probabilitas Fitur dalam Kelas $P(x_i|C_k)$: untuk setiap fitur x_i dalam setiap kelas C_k, hitung jumlah kemunculan fitur tersebut dalam kelas dan bagi dengan jumlah total instance di kelas tersebut untuk mendapatkan probabilitas $P(x_i|C_k)$.
3. Menghitung Probabilitas kelas setelah mengamati fitur $P(C_k|x_1, x_2, \dots, x_n)$: gunakan persamaan *Naïve Bayes* yang diberikan untuk menghitung probabilitas kelas C_k setelah mengamati fitur-fitur $x_1, x_2, \dots, x_n$. Kalikan probabilitas *prior* $P(C_k)$ dengan probabilitas kondisional setiap fitur $P(x_i|C_k)$ untuk semua fitur x_i yang diamati. Bagi dengan konstanta $P(x_1)xP(x_2)x \dots x\, P(x_n)$ yang merupakan faktor normalisasi yang sama untuk semua kelas.
4. Prediksi Kelas: Tentukan kelas dengan probabilitas terbesar sebagai prediksi. Ini bisa dihitung dengan membandingkan probabilitas kelas yang dihitung pada langkah sebelumnya.
5. Evaluasi Model: Evaluasi model menggunakan metrik yang sesuai seperti akurasi, presisi, *recall*, atau *F1-score* menggunakan data pengujian.
6. Penerapan pada data baru: Setelah model dilatih, model *Naïve Bayes* dapat digunakan untuk membuat prediksi pada data baru dengan menggunakan langkah-langkah di atas.

Kelebihan algoritma *Naïve Bayes* (Novia et al., 2020):

1. Dapat digunakan untuk data kuantitatif dan kualitatif.
2. Tidak memerlukan banyak data.
3. Tidak memerlukan pelatihan data yang rumit.

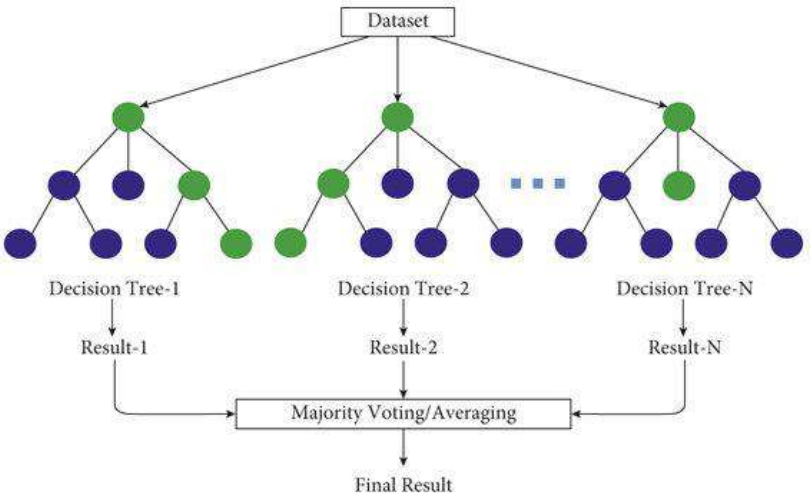
4. Jika ada nilai yang hilang, bisa diabaikan dalam perhitungan.
5. Perhitungannya cepat dan efisien.
6. Mudah dipahami dan diimplementasikan.
7. Dapat digunakan untuk masalah klasifikasi biner atau multikelas dan memungkinkan model untuk beradaptasi dengan data dengan baik.

Kekurangan algoritma *Naïve Bayes* (Novia et al., 2020):

1. Jika probabilitas kondisional mencapai nol, maka prediksi probabilitas juga akan menjadi nol.
2. Ketika setiap variabel independen diperkenalkan, ini dapat mengurangi akurasi karena biasanya ada korelasi di antara variabel tersebut.
3. Akurasi tidak dapat diukur hanya dengan satu probabilitas. Diperlukan bukti tambahan untuk mendukungnya.
4. Untuk membuat keputusan, pengetahuan sebelumnya atau pengalaman masa lalu diperlukan.

10.3 Random Forest

Random Forest adalah algoritma di bidang pembelajaran mesin yang termasuk dalam kategori *ensemble learning*. *Ensemble learning* adalah pendekatan yang menggabungkan beberapa model pembelajaran mesin sederhana untuk membentuk model yang lebih kuat. Dalam algoritma *Random Forest,* sebuah koleksi pohon keputusan dibangun, di mana setiap pohon dibuat menggunakan subset acak dari data pelatihan. Setiap pohon dalam *ensemble* melakukan prediksi secara *independen*, dan hasil prediksi dari semua pohon digabungkan untuk menghasilkan prediksi akhir (Supriadi and Andarsyah, 2023).



Gambar 10.3. Ilustrasi *Random Forest*
(Sumber : analyticsvidhya.com)

Berikut cara kerja *Random Forest*:

1. Pemilihan data secara acak: Dalam hal ini random forest menggunakan teknik *bagging* untuk memilih sebagian dari data secara acak untuk setiap pohon keputusan yang dibangun.
2. Membangun pohon keputusan: untuk setiap *subset* data yang di pilih dibangun pohon keputusan dengan cara berikut: Pilih sejumlah fitur secara acak dari total fitur yang tersedia. Selanjutnya pilih fitur terbaik untuk membagi data pada setiap node pohon keputusan, pemilihan fitur ini dapat menggunakan kriteria seperti entropi atau indeks gini dan lain-lain untuk mengukur ketidakmurnian atribut.
   Berikut ini rumus entropi dan indeks gini:
   Entropi di hitung dengan menggunakan rumus:

$$H(S) = \sum_{i-1}^{C} pi\ log_2(pi)$$

   Dimana *pi* adalah proporsi dari kelas *i* dalam node dan c adalah jumlah kelas
   Indeks gini di hitung dengan menggunakan rumus:

$$Gini(S) = 1 - \sum_{i-1}^{C} p_i^2$$

Dimana *pi* adalah proporsi dari kelas *i* dalam node dan c adalah jumlah kelas

3. Mengulangi langkah 1-2 sebanyak k kali untuk mendapatkan k pohon keputusan acak.
4. Prediksi dengan menggunakan ensemble: setelah semua pohon keputusan dibangun, prediksi dilakukan dengan cara melakukan *voting* (menggunakan mayoritas suara untuk klasifikasi dan rata-rata untuk regresi). berikut ini rumus untuk mayoritas suara:

$$c = argmax_c \sum_{i-1}^{N} I(ci = c)$$

Dimana, *I* adalah fungsi indikator yang bernilai 1 jika *ci* sama dengan *c* (prediksi pohon keputusan ke-i cocok dengan kelas *c*) dan 0 jika tidak. Rumus ini memberikan kelas dengan jumlah suara terbanyak dari semua pohon keputusan.

Random Forest memiliki fitur internal yang memperkirakan kesalahan generalisasi modelnya sendiri, yang disebut *out-of-bag* (OOB) *error estimate*. Saat membangun pohon keputusan, sebagian data asli (sekitar 2/3) digunakan dalam sampel bootstrap untuk membangun pohon, sedangkan sisanya (1/3) tidak digunakan dan disimpan untuk menguji kinerja pohon tersebut. OOB *error estimation* adalah rata-rata dari kesalahan prediksi untuk setiap kasus latihan yang tidak termasuk dalam sampel *bootstrap* pohon tersebut. Ketika Random Forest selesai dibangun, setiap kasus latihan melewati setiap pohon, dan matriks kedekatan setiap kasus dihitung berdasarkan pasangan kasus yang mencapai terminal *node* yang sama dalam pohon tersebut (Aziz, 2021).

Kelebihan algoritma *Random Forest* (Aziz, 2021):

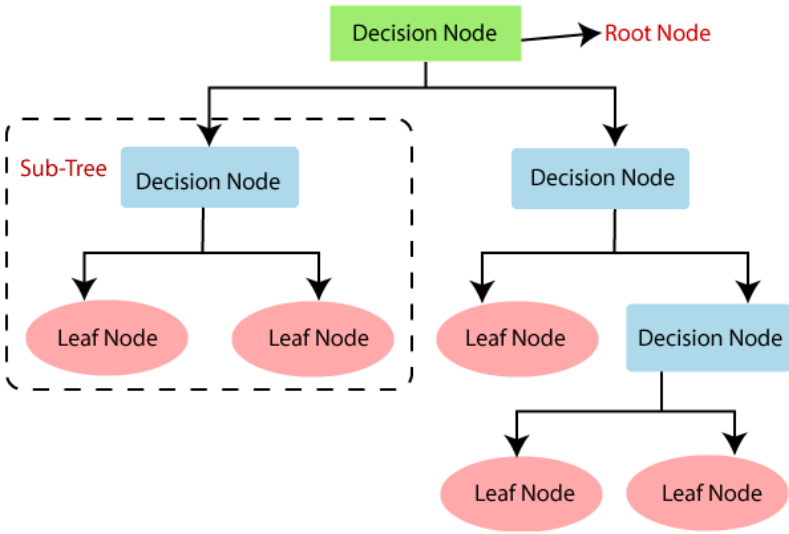
1. Kemampuannya menghasilkan *error* yang lebih rendah,
2. Memberikan hasil yang baik dalam klasifikasi,
3. Dapat mengatasi data latih dalam jumlah besar dengan efisien,
4. Dan merupakan metode yang efektif untuk mengestimasi data yang hilang.

Kekurangan algoritma *Random Forest* (Aziz, 2021):

1. Waktu pemrosesan lama karena menggunakan data yang banyak dan membangun model *tree* yang banyak pula untuk membentuk *random trees* karena menggunakan *single processor*.
2. Interpretasi yang sulit dan membutuhkan mode penyetelan yang tepat untuk data.
3. Ketika digunakan untuk regresi, mereka tidak dapat memprediksi di luar kisaran dalam data percobaan, hal ini di mungkinkan data terlalu cocok dengan kumpulan data pengganggu (*noisy*).

10.4 Decision Tree

Decision tree adalah sebuah struktur pohon yang digunakan dalam pembelajaran mesin untuk mengambil keputusan berdasarkan serangkaian aturan dan atribut (Putra et al., 2023). Pohon (*tree*) adalah sebuah struktur data yang terdiri dari simpul (*node*) dan rusuk (*edge*). Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar (*root/node*), simpul percabangan/internal (*branch/internal node*) dan simpul daun (*leaf node*) (Nasrullah, 2021). Setiap *node* dalam pohon mewakili keputusan atau pengujian terhadap suatu atribut, cabang-cabangnya merepresentasikan hasil dari pengujian tersebut, dan daun-daunnya merepresentasikan kelas atau nilai prediksi. *Decision tree* membagi dataset menjadi bagian-bagian yang lebih kecil dan lebih homogen berdasarkan atribut-atribut yang ada, sehingga memungkinkan untuk melakukan prediksi atau klasifikasi berdasarkan serangkaian aturan yang mudah dipahami. Algoritma Ini adalah salah satu algoritma pembelajaran mesin yang populer karena kemampuannya yang mudah diinterpretasikan oleh manusia dan dapat digunakan untuk klasifikasi serta regresi.



Gambar 10.4. Ilustrasi *Decision Tree*
(Sumber : addepto.com)

Berdasarkan tipe kategori data, *decision tree* dibagi menjadi dua jenis, yaitu *classification tree* dan *regression tree*. Metode *decision tree*, baik *classification tree* maupun *regression tree*, telah mengalami banyak variasi yang dikembangkan oleh peneliti sebelumnya. Beberapa contohnya termasuk algoritma C45, CART, CHAID, CRUISE, GUIDE, QUEST, ID3 dan M5. Menurut (Latifah and Wulandari, 2019).

Berikut Algoritma *decision tree* yang umum digunakan meliputi:

1. C4.5 atau C5.0: Algoritma ini merupakan pengembangan dari algoritma ID3 yang memungkinkan untuk menangani data yang memiliki atribut dengan nilai yang hilang. C4.5 menggunakan pendekatan top-down, memilih atribut terbaik pada setiap langkah untuk membagi *dataset* menjadi *subset* yang lebih kecil. Algoritma ini juga dapat menghasilkan pohon keputusan yang lebih kompleks dengan menggunakan teknik post-pruning untuk menghindari *overfitting*. C5.0 merupakan versi yang diperbarui dari C4.5 dengan peningkatan kinerja dan fitur tambahan.

2. CART (*Classification and Regression Trees*): Algoritma CART bekerja dengan membagi *dataset* menjadi *subset* yang semakin kecil berdasarkan aturan-aturan yang ditemukan dari atribut-atribut dalam data. Tujuan akhir dari CART adalah untuk membangun pohon keputusan yang optimal yang dapat mengklasifikasikan atau memprediksi target dengan akurasi yang tinggi.
3. CHAID (*Chi-square Automatic Interaction Detection*): Berbeda dengan beberapa algoritma pohon keputusan lainnya yang hanya menggunakan satu metrik untuk membagi dataset, CHAID menggunakan uji statistik *chi-square* untuk mengevaluasi hubungan antara variabel dependen dan independen. Algoritma ini secara otomatis mengidentifikasi interaksi antara variabel-variabel prediktor dan menggabungkan variabel yang memiliki hubungan yang signifikan dalam pembentukan pohon keputusan. Hasilnya adalah pohon keputusan yang lebih kompleks dan informatif, memungkinkan interpretasi yang lebih mendalam dari data.
4. ID3 (*Iterative Dichotomiser 3*): ID3 adalah salah satu algoritma yang paling awal dan populer untuk membangun pohon keputusan dalam pembelajaran mesin. Algoritma ID3 memilih atribut dengan gain informasi tertinggi untuk membagi *node* dalam pohon keputusan.

Kelebihan algoritma *Decision Tree* (Jollyta et al., 2023):

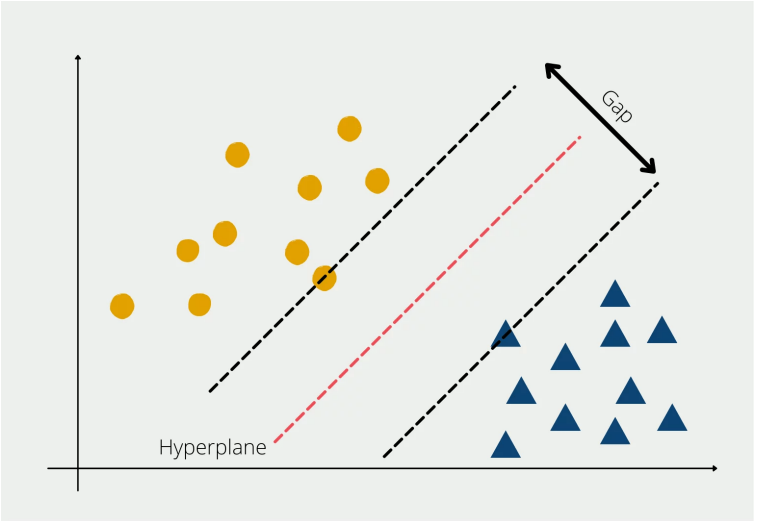
1. Mudah diinterpretasikan. Keputusan yang diambil berdasarkan atribut yang memiliki dua atau lebih pilihan keputusan dapat dengan mudah dipahami dan digunakan oleh pengguna karena dapat dilihat dari struktur pohon secara visual. Bentuk hierarki yang menjadi ciri *Decision Tree* memudahkan dalam menentukan atribut yang memiliki pengaruh paling besar.
2. Persiapan data mudah: Persiapan data untuk *Decision Tree* memerlukan upaya yang relatif kecil. *Decision Tree* mampu mengelola beragam jenis data, termasuk nilai diskrit dan kontinu. Nilai kontinu dapat diubah menjadi kategori menggunakan interval yang ditentukan. Selain itu, *Decision*

Tree dapat mengatasi masalah data yang hilang atau tidak lengkap.
3. Lebih fleksibel: *Decision Tree* dapat digunakan untuk klasifikasi dan regresi. Algoritma yang digunakan untuk klasifikasi dapat memilih atribut yang memiliki korelasi yang tinggi.

Kekurangan algoritma *Decision Tree* (Jollyta et al., 2023):
1. Rentan Terhadap *Overfitting*: Pohon keputusan yang rumit memiliki kecenderungan untuk mengalami overfitting dan tidak dapat memperluas pola dengan baik ke data baru. Masalah ini bisa diatasi dengan menggunakan metode pre-pruning atau post-pruning. Pre-pruning menghentikan pertumbuhan pohon saat data yang cukup tidak tersedia. Sementara itu, post-pruning menghapus sub-pohon yang tidak memiliki data yang memadai setelah pohon dibangun
2. *Estimator* dengan variasi tinggi: Fluktuasi kecil dalam data bisa menghasilkan variasi yang besar dalam pohon keputusan. *Bagging*, atau metode estimasi rata-rata, bisa digunakan untuk mengurangi variasi dari pohon keputusan.
3. Lebih Mahal: *Decision Tree* melakukan perhitungan pada semua atribut dan *instance* sampai menemukan aturan klasifikasi. Ini memerlukan biaya yang lebih tinggi untuk melatih data.

10.5 Support Vector Machine

SVM adalah metode klasifikasi yang mampu menangani data linear dan *non-linear*. Algoritma ini menggunakan pemetaan *non-linear* untuk mengubah data ke dimensi yang lebih tinggi, di mana ia mencari *hyperplane* pemisah linear optimal untuk memisahkan kelas-kelas data (Han et al., 2012). Algoritma SVM memiliki 4 *kernel function* yaitu kernel linear, *Radial Basic Function* (RBF), *sigmoid* dan *polynomial* (Areni et al., 2019).



Gambar 10.5. Ilustrasi *Support Vector Machine*
(Sumber : databasecamp.de)

Dalam algoritma SVM, terdapat beberapa konsep yang perlu dipahami:

1. Kernel: Algoritma SVM menggunakan konsep kernel untuk melakukan transformasi data ke dimensi yang lebih tinggi. Kernel adalah fungsi matematika yang memetakan data ke dimensi yang lebih tinggi tanpa harus secara eksplisit menghitung seluruh fitur dalam dimensi tinggi tersebut.
2. Metode Optimasi: SVM menggunakan metode optimasi untuk menggambarkan *hiperplane* dengan margin maksimum. Metode optimasi yang umum digunakan dalam algoritma SVM adalah metode Gradien Turunan atau *Stochastic Gradient Descent*.
3. *Support vector*: Dalam algoritma SVM, *support vector* merupakan data *training* yang berada tepat di sepanjang margin garis pemisah antara dua kelas.
4. Margin: Margin merupakan jarak antara garis pemisah atau *hyperplane* dengan *support vector* terdekat. Margin memiliki peran penting dalam algoritma SVM untuk mencari garis pemisah dengan margin maksimum yang dapat meminimalkan *error* klasifikasi.

Dalam proses pembelajaran algoritma SVM, terdapat tahapan-tahapan yang dilakukan (Moguerza and Muñoz, 2006):

1. Pemilihan Kernel: Tahap pertama dalam algoritma SVM adalah pemilihan kernel untuk mengubah data ke dimensi yang lebih tinggi sehingga dapat memisahkan kelas-kelas data *yang non-linear.*
2. Optimasi *Hyperplane*: Setelah data diubah ke dimensi yang lebih tinggi, algoritma SVM mengoptimalkan hyperplane dengan menggunakan metode optimasi seperti *gradient descent*. Metode optimasi ini bertujuan untuk mencapai *hyperplane* terbaik dengan margin maksimum antara kelas-kelas data. Selanjutnya, *support vector machine* akan mengklasifikasikan data baru dengan memproyeksikannya ke *hyperplane* dan menentukan kelasnya berdasarkan posisinya.
3. Pengaturan Parameter: Salah satu poin penting dalam algoritma SVM adalah pengaturan parameter seperti C dan gamma yang perlu diatur dengan tepat untuk mencapai hasil yang optimal. Setelah parameter diatur dengan tepat, algoritma SVM siap digunakan untuk melakukan klasifikasi data.

Kelebihan algoritma *Support Vector Machine* (Soofi and Awan, 2017):

1. Mampu menangani masalah klasifikasi data yang memiliki banyak fitur (*high-dimensional data*) dengan efisien.
2. Akurasi yang tinggi dalam klasifikasi data, terutama ketika jumlah sampel data relatif kecil.
3. Mampu mengatasi masalah *non-linearity* dalam data dengan menggunakan kernel *function*.
4. Mampu mengatasi masalah *overfitting* dengan menggunakan parameter yang tepat, seperti regularisasi.
5. Tahan terhadap adanya outlier dalam data, karena hanya *support vector* yang mempengaruhi posisi *hyperplane.*

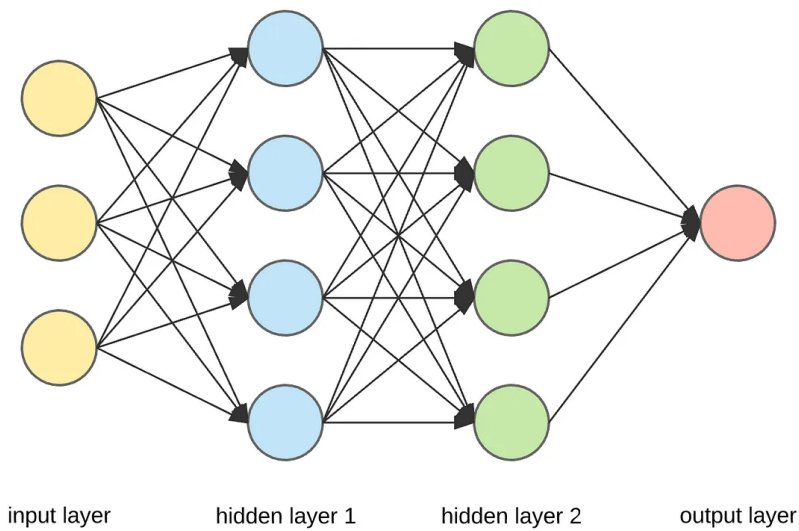
Kekurangan algoritma *Support Vector Machine* (K, 2023):

1. Memiliki kompleksitas komputasi yang tinggi ketika jumlah data sangat besar. Hal ini disebabkan oleh proses transformasi data ke dimensi yang lebih tinggi dan optimisasi *hyperplane*.

2. Sensitif terhadap pilihan kernel. Sebagai contoh, jika menggunakan kernel yang tidak sesuai dengan struktur data, performa SVM dapat menurun drastis. Dalam penggunaan SVM, perlu diperhatikan bahwa algoritma ini sensitif terhadap pemilihan parameter, seperti C dan gamma.
3. Tidak dapat memberikan penjelasan probabilistik untuk klasifikasi data.
4. Kinerja SVM menurun jika data memiliki banyak *noise* atau jika kelas target saling tumpang tindih.

10.6 Artifial Neural Network

Artificial Neural Network (ANN), atau dalam bahasa Indonesia dikenal sebagai Jaringan Saraf Tiruan, adalah model matematika yang terinspirasi oleh struktur dan fungsi jaringan saraf biologis dalam otak manusia (T. Sutojo et al., 2011). ANN digunakan dalam pembelajaran mesin untuk memodelkan hubungan kompleks antara *input* dan *output* dan dapat digunakan untuk tugas-tugas seperti klasifikasi, regresi, pengenalan pola, dan lainnya.



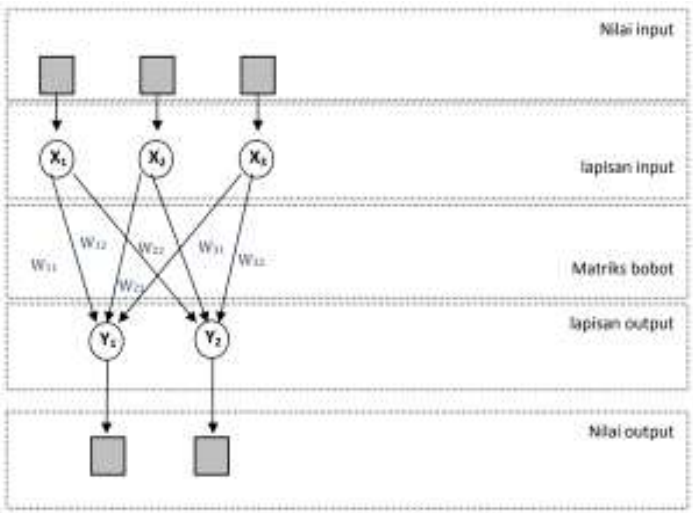
Gambar 10.6. Ilustrasi Support Vector Machine
(Sumber : towardsdatascience.com)

Terdapat 3 lapisan penyusun ANN bekerjasama dalam memproses informasi dan menghasilkan keluaran yang ditargetkan (Lesnussa et al., 2015):

1. Lapisan *input*, unit-unit di dalamnya disebut unit *input*. Unit *input* ini menerima pola data eksternal yang menggambarkan masalah yang hendak diselesaikan.
2. Lapisan tersembunyi, unit-unit di dalamnya disebut unit tersembunyi. Keluaran dari lapisan ini tidak dapat diamati langsung.
3. Lapisan *output*, di mana unit-unit di dalamnya disebut unit *output*. *Output* dari lapisan ini adalah solusi yang dihasilkan *Artificial Neural Network* untuk suatu masalah.

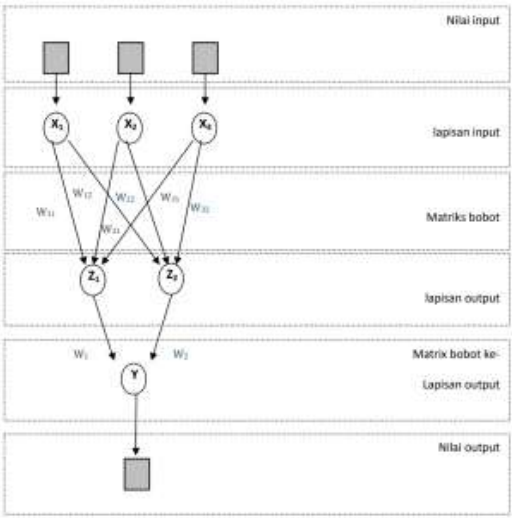
Terdapat 3 jenis arsitektur dalam *Artificial Neural Network* (ANN), (Hasan et al., 2019), yaitu:

1. Jaringan Lapisan Tunggal (*Single Layer Network*)
   Jaringan berlapis satu terdiri dari satu lapisan bobot koneksi. Lapisan ini terdiri dari unit *input* yang menerima sinyal dari luar dan unit *output* yang meiliki kemampuan untuk memeriksa tanggapan ANN.



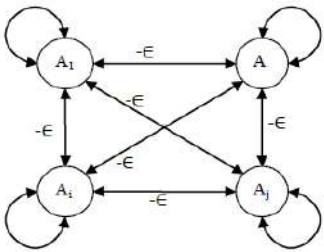
Gambar 10.7. Ilustrasi Jaringan Lapisan Tunggal
(Sumber : kantinit.com)

2. Jaringan Lapisan Jamak (*Multi Layer Network*)
   Jaringan berlapis banyak merupakan jaringan yang memiliki satu atau lebih lapisan tersembunyi. Meskipun lebih efektif dalam menyelesaikan masalah, jaringan ini lebih kompleks dalam proses pelatihannya.



Gambar 10.8. Ilustrasi Jaringan Lapisan Jamak
(Sumber : kantinit.com)

3. Jaringan Lapisan Kompetitif (*Competitive Layer Net*)
   Jaringan lapisan kompetitif merupakan jenis jaringan saraf tiruan yang terdiri dari satu lapisan *neuron* yang bersaing untuk menjadi aktivasi yang paling tinggi atau paling sesuai dengan *input* yang diberikan. Setiap *neuron* dalam lapisan ini berjuang untuk menjadi *neuron* pemenang yang menghasilkan *respons* terhadap pola *input* tertentu.



Gambar 10.9. Ilustrasi Jaringan Lapisan Kompetitif
(Sumber : media-rakyat45.blogspot.com)

Kelebihan algoritma *Artificial Neural Network* (Siregar and Octariadi, 2021):

1. Kemampuan beradaptasi dengan berbagai jenis data.
2. Kemampuan untuk mengenali dan mengklasifikasikan pola.
3. Cocok untuk mengolah *input* dan *output* yang kontinu atau berkelanjutan.
4. Kemudahan implementasi dalam berbagai aplikasi.

Kekurangan algoritma *Artificial Neural Network* (Siregar and Octariadi, 2021):

1. Proses menentukan parameter terbaik memerlukan percobaan berulang dan proses *trial* and *error*.
2. Proses iterasi memakan waktu yang relatif lama.
3. Tantangan dalam menginterpretasikan hasil yang dihasilkan oleh jaringan saraf tiruan.

DAFTAR PUSTAKA

Amna, A., S, W., Sudipa, I.G.I., E. Putra, T.A., Wahidin, A.J., Syukrilla, W.A., Anindya Khrisna, W., Heryana, N., Indriyani, T., Santoso, L.W., 2023. Data Mining, 1st ed. PT Global Eksekutif Teknologi Anggota IKAPI No. 033/SBA/2022, Jl. Pasir Sebelah No. 30 RT 002 RW 001 Kelurahan Pasie Nan Tigo Kecamatan Koto Tangah Padang Sumatera Barat.

Andreas C. Müller & Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.

Areni, I.S., Amirullah, I., Arifin, N., 2019. Klasifikasi Kematangan Stroberi Berbasis Segmentasi Warna dengan Metode HSV. J. Penelit. Enj. 23, 113–116. https://doi.org/10.25042/jpe.112019.03

Aziz, W.A., 2021. Implementasi Metode Random Forest Pada Klasifikasi Data Ulasan Konsumen Perusahaan (Studi Kasus : Aplikasi KAI Access). Universitas Islam Negeri Syarif Hidayatullah Jakarta, Jakarta.

Dertat, A., 2017. Applied Deep Learning - Part 1: Artificial Neural Networks, Toward Data Science. [Online] Available at: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6. [Accessed 30 Maret 2024].

Geeksforgeeks, 2024. K-Nearest Neighbor(KNN) Algorithm. [Online]
Available at: https://www.geeksforgeeks.org/k-nearest-neighbours//. [Accessed 30 Maret 2024].

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques, Third Edition. ed. Morgan Kaufmann Publishers.

Haponik, A., 2020. Decision Tree Machine Learning Model, addepto Warsaw - HQ. [Online] Available at: https://addepto.com/blog/decision-tree-machine-learning-model/.[Accessed 30 Maret 2024].

Hasan, N.F., Kusrini, K., Fatta, H.A., 2019. Analisis Arsitektur Jaringan Syaraf Tiruan Untuk Peramalan Penjualan Air Minum Dalam Kemasan. J. Rekayasa Teknol. Inf. JURTI 3, 1. https://doi.org/10.30872/jurti.v3i1.2290.

Helilintar, R., Ramadhani, R.A., Rochana, S., 2017. Data Mining K-Nears Neigbours. Fak. Tek. Univ. Nusant. PGRI Kediri Kampus II Mojoroto Gang No6 Kediri, 1 1, 1–52.

Jollyta, D., Prihandoko, P., Hajjah, A., Haerani, E., Siddik, M., 2023. Algoritma Klasifikasi untuk Pemula Solusi Python dan RapidMiner. Deepublish Digital, Yogyakarta.

K, Dhiraj., 2023. Top 4 advantages and disadvantages of Support Vector Machine or SVM. dhirajkumarblog.medium.com. Available at: https://dhirajkumarblog.medium.com/top-4-advantages-and disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 (Accessed: 22 March 2024).

Kantin IT. Belajar Jaringan Syaraf Tiruan. [Online] Available at: https://kantinit.com/kecerdasan-buatan/belajar-jaringan-syaraf-tiruan-jst-pengertian-arsitektur-cara-kerja-dan-jenis-jenisnya/#google_vignette.[Accessed 30 Maret 2024].

Latifah, R., Wulandari, E.S., 2019. Model Decision Tree untuk Prediksi Jadwal Kerja menggunakan Scikit-Learn. Semin. Nas. Sains Dan Teknol. 2019 Fak. Tek. Univ. Muhammadiyah Jkt. 1–6

Lesnussa, Y.A., Latuconsina, S., Persulessy, E.R., 2015. Aplikasi Jaringan Saraf Tiruan Backpropagation untuk Memprediksi Prestasi Siswa SMA (Studi kasus: Prediksi Prestasi Siswa SMAN 4 Ambon). J. Mat. Integratif 11.

Moguerza, J.M., Muñoz, A., 2006. Support Vector Machines with Applications. Stat. Sci. 21. https://doi.org/10.1214/088342306000000493.

Nasrullah, A.H., 2021. Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. J. Ilm. ILMU Komput. 7, 45–51. https://doi.org/10.35329/jiik.v7i2.203

Novia, E.A., Rahayu, W.I., Prianto, C., 2020. Sistem Perbandingan Algoritma K-Means dan Naïve Bayes untuk Memprediksi, 1st ed. Kreatif Industri Nusantara, Jl. Ligar Nyawang No. 2 Bandung.

Putra, R.F., Zebua, R.S.Y., Budiman, B., Rahayu, P.W., Bangsa, T.A., Zulfadhilah, M., Choirina, P., Wahyudi, F., Andiyan, Ar., 2023. DATA MINING : Algoritma dan Penerapannya, 1st ed. PT. Sonpedia Publishing Indonesia, Jambi.

Sabry, F., 2023. Naive Bayes Classifier : Fundamentals and Applications. One Billion Knowledgeable.

Sharma, A., 2024. Random Forest vs Decision Tree | Which Is Right for You?, Analytics Vidhya. [Online] Available at: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/.[Accessed 30 Maret 2024].

Sigid Widodo, A.Z.M., Pandu Kusuma, A., Dwi Puspitasari, W., 2023. Algoritma Naive Bayes Classifier (NBC) Pada Klasifikasi Tingkat Minat Barang Di Toko Violet Cell. JATI J. Mhs. Tek. Inform. 7, 87–94. https://doi.org/10.36040/jati.v7i1.5692

Siregar, A.C., Octariadi, B.C., 2021. Perbandingan Metode Jaringan Syaraf Tiruan Pada Klasifikasi Motif Kain Tenun Sambas. CYBERNETICS 4. https://doi.org/10.29406/cbn.v4i02.2489

Soofi, A.A. and Awan, A., 2017. Classification Techniques in Machine Learning: Applications and Issues', Journal of Basic & Applied Sciences. 13(1). pp. 459–465. https://pdfs.semanticscholar.org/2678/e213cec548d278879ceaf01582ee8913cc3f.pdf (Accessed: 22 March 2024).

Support Vector Machine (SVM) – easily explained!. 2021. Data Bas Camp. [Online] Available at: https://databasecamp.de/en/ml/svm-explained. [Accessed 30 Maret 2024].

Supriadi, M.R., Andarsyah, R., 2023. Deteksi Halaman Website Phishing Menggunakan Algoritma Machine Learning Gradient Boosting Classifier. Penerbit Bukupedia.

T. Sutojo, Edy Mulyanto, Vincent Suhartono, 2011. Kecerdasan Buatan. ANDI.

BAB 11
Konsep Cluster Analysis

Oleh Chairi Nur Insani

11.1 Konsep Dasar

Clustering bergantung pada konsep *similarity* antar objek data. Sehingga himpunan datanya tidak memiliki label kelas dan hanya dilihat dari kemiripan satu sama lain. *Clustering* dapat memiliki berbagai tujuan, tergantung pada kebutuhan bisnis atau ilmiah. *Clustering* melibatkan pengelompokan objek atau data ke dalam kelompok yang homogen berdasarkan kesamaan karakteristik.. *Clustering* data mining merupakan teknik yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki kesamaan tertentu. Berikut beberapa poin terkait *clustering* dalam konteks data mining.

1. Analisis Pola: Pengelompokan memungkinkan mengidentifikasi pola tersembunyi dalam data tanpa memerlukan label atau anotasi sebelumnya.
2. Segmentasi Pasar: Dalam penambangan data, *clustering* sering digunakan untuk segmentasi pasar. Ini termasuk mengelompokkan pelanggan atau konsumen berdasarkan perilaku, preferensi, atau demografinya.
3. Pengelompokan Dokumen: Dalam pemrosesan bahasa alami dan analisis teks, pengelompokan dapat digunakan untuk mengelompokkan dokumen berdasarkan topik, tema, atau pola dalam kalimatnya.
4. Analisis Biomedis: Dalam bidang biomedis, pengelompokan digunakan untuk mengelompokkan pasien berdasarkan respons mereka terhadap pengobatan, karakteristik genetik, atau faktor risiko lainnya.
5. Web Mining: *Clustering* juga digunakan dalam penambangan Web untuk mengelompokkan halaman Web berdasarkan topik, struktur, atau popularitas.

6. Pengelompokan Spasial: Dalam analisis spasial, klasterisasi dapat digunakan untuk mengidentifikasi pola spasial dalam data geografis, seperti pengelompokan wilayah berdasarkan kemiripan karakteristik atau tingkat kepadatan.
7. Deteksi Anomali: Selain mengelompokkan data yang mirip, klasterisasi juga dapat digunakan untuk mendeteksi anomali atau *outlier* dengan mengidentifikasi klaster yang tidak biasa atau tidak representatif.

Beberapa algoritma yang sering digunakan dalam data mining utuk membentuk kelompok bedasarkan struktur dalam data yaitu k-means dan *hierarchical clustering* (Sudipa *et al.*, 2024).

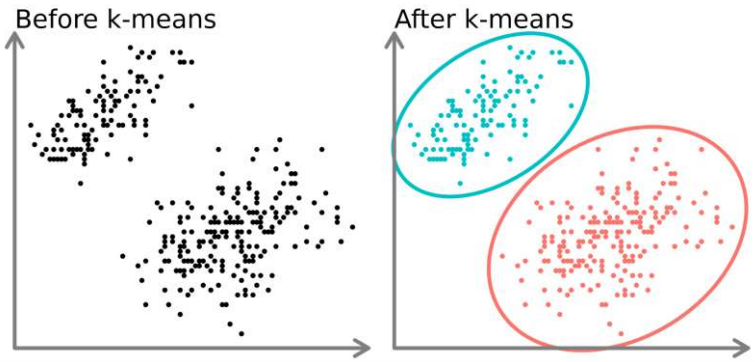
11.2 K-Means

K-Means memiliki kemampuan untuk mengelompokkan data ke dalam data *cluster*. Algoritma ini dapat menerima data tanpa label kategori. Algoritma pengelompokan K-Means juga merupakan teknik non-hierarki. Metode yang digunakan yaitu dengan mengelompokkan beberapa data ke dalam kelompok-kelompok yang menjelaskan bahwa data pada suatu kelompok mempunyai karakteristik yang sama dan berbeda dengan data pada kelompok lainnya. K-means terlebih dahulu menentukan jumlah cluster (K) dan kemudian secara iteratif memindahkan pusat (centroid) setiap cluster untuk meminimalkan *varians* dalam *cluster*.

Metode K-Means merupakan metode clustering yang paling sederhana dan umum. Hal ini dikarenakan K-Means dapat mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang cepat dan efisien. K-Means merupakan algoritma *clustering* yang mencakup algoritma k-Medoids berbasis objek serta metode partisi berdasarkan titik pusat (centroids). Algoritma ini pertama kali dikemukakan oleh MacQueen (1967) dan dikembangkan oleh Hartigan dan Wong pada tahun 1975(Suyanto, 2017).

Algoritma ini bekerja dengan cara membagi data ke dalam k klaster, di mana k adalah jumlah klaster yang ditentukan sebelumnya. Proses k-means dimulai dengan menentukan k titik pusat klaster acak yang disebut sebagai centroid. Setiap titik data

kemudian diatribusikan ke klaster yang memiliki centroid terdekat. Langkah ini diikuti dengan perhitungan ulang posisi centroid dengan menggunakan rata-rata dari semua titik data yang telah diatribusikan ke klaster tersebut.



Gambar 11.1. Ilustrasi K-Means
(Sumber : datacamp.com)

Proses iteratif k-means dilakukan sampai tidak ada lagi perubahan dalam atribusi titik data ke klaster atau sampai kriteria konvergensi lainnya terpenuhi. Salah satu keuntungan utama dari k-means adalah efisiensinya yang tinggi, terutama untuk data dengan jumlah besar. Namun, keberhasilan k-means sangat tergantung pada pemilihan jumlah klaster yang tepat dan inisialisasi centroid yang baik. Selain itu, k-means cenderung memberikan hasil yang baik untuk data yang memiliki bentuk klaster yang terdefinisi dengan baik, tetapi mungkin tidak efektif untuk data yang memiliki bentuk klaster yang kompleks atau tidak teratur(Wahono, 2023).

Tujuannya agar dapat membagi titik data M berdimensi N menjadi k cluster, dimana dilakukan proses clustering untuk meminimalkan jumlah kuadrat. Jarak antara data dengan masing-masing pusat cluster (berbasis centroid). Algoritma k-means pada implementasi memerlukan tiga parameter yang semuanya ditentukan oleh pengguna. Yaitu jumlah cluster k, inisialisasi cluster, dan jarak sistem. Biasanya, algoritma ini pada dasarnya mengelompokkan data hanya ke arah minimum lokal, sehingga k-

Means dijalankan secara independen dengan inisialisasi berbeda, menghasilkan cluster akhir yang berbeda(Jain, 2010).

Berikut adalah langkah-langkah kerja dari K-means(Solichin *et al.*, 2020) :

1. Penentuan jumlah *cluster* : menentukan berapa jumlah k *cluster* yang akan digunakan dari kelompok data.
2. Penentuan awal titik *centroid* (k) : penentuan secara acak *centroid* awal bedasarkan jumlah *cluster* dan jumlah data yang akan diproses.
3. Perhitungan jarak dari semua data dengan titik pusat *cluster* : ada bebrabagai metrik yang dapat digunakan untuk menghitung jarak data dengan titik *centroid* salah satu metrik yang umum digunakan adalah *euclidean distance*(Insani, Arifin and Rasyid, 2023). Rumus metode perhitungan jarak tersebut dapat dilihat pada persamaan berikut.

   $$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$$

   Keterangan:
   - x_1 = Sampel Data
   - x_2 = Data Uji / Testing
   - i = Variabel Data
   - d = Jarak
   - p = Dimensi Data
4. Pengelompokan data bedasarkan *cluster* terdekat : pemilihan *cluster* yang memiliki jarak terdekat dengan data. Kemudian mengelompokkan data kedalam *cluster* tersebut.
5. Perhitungan titik *centroid cluster* yang baru : Setelah pengelompokan data ke dalam *cluster*, dilakukan perhitungan titik pusat *cluster* yang baru dengan menghitung rata-rata jarak data dengan titik pusat *cluster*.
6. Membandingkan *cluster* baru dengan *cluster* awal : hal ini dilakukan apabila *cluster* baru memiliki nilai *centroid* yang berbeda dengan *cluster* awal, maka proses tahapannya harus diulangi dari tahapan nomer 5. Apabila *centroid* dari *cluster* baru memiliki nilai yang sama dengan/selisihnya

kurang dari 0.1 dengan *cluster* awal atau sebelumnya maka proses *clustering* dapat di hentikan.

11.3 Hierarchical Clustering

Pengelompokan *hierarchical clustering* bedasarkan bagan hirarji yang dimana terdapat gabungan antar dua grup terdekat pada tiap iterasi ataupun pada bagian set data ke dalam *cluster.*

Hierarchical clustering adalah salah satu teknik klasterisasi yang menghasilkan hierarki klaster dalam bentuk pohon atau dendrogram. Pendekatan ini tidak memerlukan jumlah klaster yang ditentukan sebelumnya, melainkan membangun struktur hierarki berdasarkan tingkat kemiripan antara titik data. Dalam metode agglomerative, setiap titik data awalnya dianggap sebagai klaster tunggal, lalu secara bertahap digabungkan berdasarkan kemiripan, sementara dalam metode divisive, semua titik data dianggap sebagai satu klaster yang kemudian dibagi menjadi klaster yang lebih kecil.

Salah satu keunggulan utama dari hierarchical clustering adalah kemampuannya untuk memberikan pemahaman yang lebih dalam tentang struktur data melalui dendrogram yang dihasilkan. Dendrogram ini menggambarkan bagaimana klaster berkelompok bersama berdasarkan tingkat kemiripan antara titik data, sehingga memungkinkan pengguna untuk memilih jumlah klaster yang sesuai dengan data atau memahami hubungan hierarkis antar klaster. Namun, kerugian dari pendekatan ini adalah biaya komputasi yang tinggi, terutama untuk data besar, serta sensitivitas terhadap inisialisasi dan metode pengukuran jarak yang dipilih.

Hierarchical clustering memiliki berbagai aplikasi dalam berbagai bidang, termasuk biologi, ilmu sosial, dan analisis citra. Dalam biologi, teknik ini digunakan untuk klasifikasi genetik, sementara dalam ilmu sosial, hierarchical clustering dapat digunakan untuk menganalisis pola hubungan antarindividu dalam jaringan sosial. Di bidang analisis citra, metode ini digunakan untuk segmentasi gambar, di mana objek yang mirip secara visual dikelompokkan bersama dalam klaster yang sesuai. Dengan pemahaman yang baik tentang konsep dan karakteristik hierarchical clustering, peneliti dan praktisi dapat mengaplikasikan teknik ini dengan efektif dalam analisis data mereka.

Metode *hierarchical clustering* memilikicara kerja dengan mengelompokan data ke dalam struktur pohon *cluster*. Metode *hierarchical clustering* dapat digolongkan ke dalam *agglomerative* dan *divisive hierarchical clustering*, tergantung pada dekomposisi dibentuk menurut cara *bottom-up* atau *top-down*.

1. *Agglomerative and Divisive Hierarchical Clustering*. Secara umum, ada dua tipe dari metode hierarchical clustering :
   a. *Agglomerative hierarchical clustering* : Strategi bottom-up dimulai dengan menempatkan setiap objek pada masing-masing *cluternya* yang kemudian *centroid* akan masuk pada *cluster* yang lebih besar dan lebih besar lagi, hingga semua data yang berada dalam titik *centroid* memiliki kondisi yang terpenuhi. Sebagian besar metode hierarchical clustering menerapkan kategori ini.
   b. *Divisive hierarchical clustering* : berbeda dengan agglomerative hierarchical clustering, strategi top-down dimulai dengan memasukan semua data ke dalam satu *cluster.* Kemudian membaginya kedalam *cluster* yang lebih kecil hingga tiap-tiap data membentuk *cluster* dalam *cluaster*nya sendiri sampai pada kondisi yang memenuhi.
2. BIRCH : *Balanced Iterative Reducing and Clustering.* BIRCH merupakan metode pengelompokan hirarki yang terintegrasi. Terdapat dua konsep pada metode ini yaitu *clustering feature* dan *clustering feature tree* (CF tree), kedua metode tersebut digunakan untuk menggambarkan ringkasan cluster. Pada fase-fase metodenya membantu proses pengelompokan data dengan kecepatan dan skalabilitas yang baik untuk set data besar. Metode ini juga efektif digunakan karena proses pengelompokan yang dinamik dengan model pengelompokan sedikit-sedikit data dan teratur pada tiap set data. Berikut beberapa tahapan metode ini :
   a. Tahapan awal : BIRCH membaca set data dalam membangun suatu inisial memori cf tree untuk mempertahankan struktur dari set data tersebut.

b. Tahapan Akhir : menggunakan pengelompokan yang selektif dalam menentukan *clustering* pada cf tree dengan skalabilitas linier dalam hal jumlah data dan kualitas pengelompokan data. Akan tetapi setiap cf tree hanya dapat menampung data yang terbatas maka metode ini jarang digunakan karena keterbatasan memori cf tree. Kelebihan lain yang dimiliki metode adalah konsep radius penentuan titik *centroid* yang dapat menentukan batas *cluster.*

3. *Cure : Clustering using representatives*, juga dikenal sebagai *Representative-based Clustering* (RBC), adalah pendekatan klasterisasi di mana representatif atau titik-titik pusat diidentifikasi terlebih dahulu, dan kemudian titik-titik data lainnya diatribusikan ke klaster terdekat berdasarkan jarak atau kesamaan. Pendekatan ini bertentangan dengan metode tradisional yang memulai klasterisasi dengan memilih titik data acak sebagai pusat klaster. Pendekatan RBC mencakup beberapa langkah. Pertama, representatif awal harus dipilih, sering kali dengan metode pemilihan titik-titik yang tepat seperti metode k-means. Kemudian, titik-titik data lainnya diatribusikan ke representatif terdekat, biasanya menggunakan pengukuran jarak seperti jarak Euclidean atau jarak Manhattan. Proses ini berlanjut dengan menghitung ulang posisi representatif berdasarkan anggota klaster yang baru diatribusikan, dan langkah ini diulang hingga konvergensi dicapai. Keuntungan dari pendekatan RBC termasuk efisiensi komputasi yang lebih tinggi dibandingkan dengan beberapa metode klasterisasi tradisional, karena hanya ada perhitungan jarak antara titik-titik data dan representatif, bukan antara semua pasangan titik data. Selain itu, dengan representatif yang dipilih secara cermat, pendekatan ini dapat menghasilkan klaster yang lebih stabil dan terdefinisi dengan baik. Namun, ada beberapa pertimbangan dalam menggunakan pendekatan RBC. Salah satunya adalah pemilihan representatif awal yang baik, karena titik-titik ini akan sangat memengaruhi hasil klasterisasi. Selain itu, kinerja algoritma dapat

dipengaruhi oleh jumlah dan distribusi representatif yang dipilih, dan proses klasterisasi mungkin lebih sensitif terhadap titik-titik outlier atau anomali dalam data. Dengan pemahaman yang baik tentang karakteristik dan implikasi dari pendekatan RBC, peneliti dapat menggunakannya secara efektif dalam analisis data mereka..

4. *Chameleon : A Hierarchical Clustering Algorithm Using Dynamic Modeling* (HCDM) adalah suatu metode klasterisasi hierarkis yang menggunakan pendekatan pemodelan dinamis untuk mengelompokkan data. Pendekatan ini menggabungkan prinsip klasterisasi hierarkis dengan teknik pemodelan dinamis untuk mengidentifikasi struktur klaster dalam data. Langkah pertama dalam HCDM adalah membangun representasi awal dari struktur hierarkis dengan menggunakan teknik klasterisasi tradisional, seperti agglomerative hierarchical clustering atau divisive hierarchical clustering. Kemudian, dengan memanfaatkan teknik pemodelan dinamis, struktur hierarkis ini diperbaiki atau disesuaikan secara iteratif untuk meningkatkan akurasi klasterisasi. Pemodelan dinamis dalam HCDM melibatkan proses iteratif di mana struktur klaster dianalisis dan disesuaikan berdasarkan hubungan antar titik data. Teknik ini dapat mencakup pemodelan probabilitas, penyesuaian jarak atau kemiripan antar titik data, atau penyesuaian representasi klaster. Proses ini terus berlanjut hingga konvergensi tercapai atau hingga kriteria berhenti lainnya terpenuhi. Keuntungan utama dari HCDM adalah kemampuannya untuk mengatasi beberapa kelemahan dari metode klasterisasi hierarkis tradisional, seperti sensitivitas terhadap inisialisasi dan pemilihan jarak yang tepat. Dengan memanfaatkan pendekatan pemodelan dinamis, HCDM dapat menyesuaikan struktur klaster secara lebih adaptif sesuai dengan karakteristik dan kompleksitas data. Meskipun demikian, ada beberapa tantangan dalam implementasi HCDM, termasuk kompleksitas komputasi yang lebih tinggi karena penggunaan teknik pemodelan dinamis, serta sensitivitas terhadap parameter dan

pengaturan algoritma. Namun, dengan pemahaman yang mendalam tentang prinsip-prinsip dan teknik yang terlibat, HCDM dapat menjadi alat yang kuat untuk analisis data dan penemuan pola.

DAFTAR PUSTAKA

Insani, C.N., Arifin, N. and Rasyid, M.R. (2023) 'Deteksi Gerakan Bahasa Isyarat Menggunakan Euclidean Distance', *Informatik : Jurnal Ilmu Komputer*, 19(1), pp. 99–106. Available at: https://doi.org/10.52958/iftk.v19i1.5658.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*, 31(8), pp. 651–666. Available at: https://doi.org/10.1016/j.patrec.2009.09.011.

Solichin *et al.* (2020) 'Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta', *Fountain of Informatics Journal*, 5(2), pp. 52–59.

Sudipa, I.G.I. *et al.* (2024) *Buku ajar data mining*.

Suyanto (2017) *Data mining untuk klasifikasi dan klasterisasi data*. 1st edn. Indonesia: Informatika.

Wahono, R.S. (2023) *Data Mining Data mining*, *Mining of Massive Datasets*. Available at: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part.

Han, J., Kamber, M., Pei, J.: Data Mining Concept and Techniques, 3rd ed. Morgan Kaufmann-Elsevier, Amsterdam (2012)

Jain. A.K (2009). Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 2009.

Tan, P.N., Steinbach, M., Kumar, V. (2006) Introduction to Data Mining.

Boston:Pearson Education

Han, Jiawei & Kamber, Micheline, Data Mining – Consepts and Techniques, Simon Fraser University, USA : Morgan Kaufmann, 2001

Shah, N. A., dan Paul, R. (2017). Survey on different grid based clustering algorithms of data mining. International Journal of Advance Research and Innovative Ideas in Education (IJARIIE), 3(1), 1118–1123.

Suman, dan Rani, P. (2017). A survey on sting and clique grid based clustering methods. International Journal of Advanced Research in Computer Science, 8(5), 1510–1512.

Liao, W.-k., Liu, Y., dan Choudhary, A. (2004). A grid-based clustering algorithm using adaptive mesh refinement. Dalam Appears in the 7th workshop on mining scientific and engineering datasets 2004 (hal. 1–9).

Parikh, M., dan Varma, T. (2014). Iog - an improved approach to find optimal grid size using grid clustering algorithm. IOSR Journal of Computer Engineering, 16(3), 114–118.

Datacamp, 2024. K-Means Clustering. [Online]
Available at: https://www. https://www.datacamp.com/tutorial/k-means-clustering-python//. [Accessed 20 April 2024].

BAB 12
JENIS-JENIS KNOWLEDGE

Oleh Frans Mikael Sinaga

12.1 Pendahuluan

Jenis-jenis *Knowledge* pada Data Mining menekankan peran krusial pengetahuan dalam perkembangan data mining. Dengan memandang data tidak hanya sebagai kumpulan fakta, buku ini bertujuan untuk menguraikan berbagai bentuk pengetahuan yang dapat diidentifikasi dan dimanfaatkan dalam ekosistem data mining. Fokusnya adalah pada pemahaman bahwa pengetahuan tidak hanya berupa informasi terorganisir, melainkan juga menjadi entitas strategis yang menjadi dasar untuk pengambilan keputusan informasional dan pembentukan wawasan mendalam. Melalui pembahasan konsep dasar, definisi, dan implementasi pengetahuan dalam berbagai aspek data mining, buku ini berupaya membuka perspektif baru terhadap cara data mining dapat diinterpretasikan dan dimanfaatkan sebagai sumber daya yang memiliki potensi besar (Ghavami, P. 2019).

12.2 Definisi dan Konsep Dasar

Berikut ini dijelaskan mengenai definisi data mining, konsep dasar, pengetahuan eksplisit dan implisit (Shmueli, G. et al. 2019).

1. Definisi

   Data Mining merupakan sebuah proses analisis data yang dilakukan secara otomatis atau semi-otomatis dengan tujuan mengekstraksi pola yang memiliki nilai atau informasi yang belum terungkap dari sebuah kumpulan data. Fokus utama dari kegiatan data mining adalah mengidentifikasi keterkaitan atau pola yang tersembunyi dalam dataset, dengan harapan dapat memberikan wawasan baru atau keuntungan strategis yang bermanfaat**.**

2. Konsep Dasar
   Klasifikasi Pengetahuan pada Data Mining mengacu pada pemahaman yang diperoleh melalui analisis data menggunakan metode data mining. Pengetahuan ini dapat dikategorikan menjadi dua tipe utama:
3. Pengetahuan Eksplisit
   a. Pengetahuan yang Eksplisit adalah pengetahuan yang dapat diungkapkan atau dinyatakan secara jelas.
   b. Contoh dari jenis pengetahuan ini mencakup aturan bisnis, fakta, atau informasi terstruktur lainnya.d
   c. Pengetahuan eksplisit dapat dengan mudah direpresentasikan dalam bentuk aturan atau basis data.
4. Pengetahuan Implisit
   a. Pengetahuan yang Implisit mengacu pada pemahaman atau keahlian yang tidak selalu diungkapkan secara eksplisit.
   b. Melibatkan unsur-unsur seperti intuisi, wawasan, atau pola yang mungkin sulit dijelaskan dengan kata-kata.
   c. Pemahaman semacam itu dapat muncul melalui analisis pola atau tren yang ditemukan dalam data.

Dalam ranah data mengidentifikasi jenis-jenis pengetahuan tersebut. Proses ini memberikan peluang bagi pengguna untuk menjelajahi struktur dalam data, mengenali hubungan antar variabel, dan menghasilkan pengetahuan yang dapat digunakan untuk meningkatkan keputusan atau merancang strategi bisnis yang lebih efektif. Mining, berbagai teknik seperti clustering, klasifikasi, asosiasi, dan regresi dapat digunakan untuk.

Dengan memeriksa konsep dasar ini, pengguna data mining dapat meningkatkan pemahaman mereka tentang peran pengetahuan dalam mengelola serta memanfaatkan informasi yang terdapat dalam dataset besar. Ini membuka peluang untuk mengembangkan solusi dan strategi yang lebih efektif**.**

12.3 Klasifikasi Knowledge dalam Data Mining

Jenis pengetahuan berdasarkan tingkatannya dibagi menjadi tiga bagian, yaitu pengetahuan rendah, pengetahuan menengah, pengetahua tinggi (Kantardzic, M. 2019).

1. Pengetahuan Rendah
   Pengetahuan pada tingkat rendah membicarakan aspek pengetahuan yang bersifat dasar, mungkin terkait dengan informasi umum atau aturan sederhana. Bagian ini akan menguraikan metode di mana data mining dapat mengidentifikasi dan mengkategorikan pengetahuan rendah, serta bagaimana penerapannya dapat diimplementasikan.
2. Pengetahuan Menengah
   Pengetahuan pada tingkat menengah melibatkan pemahaman yang lebih kompleks, mencakup hubungan yang lebih dalam antara variabel atau pola yang lebih rumit. Dengan memanfaatkan algoritma klasifikasi yang lebih canggih atau analisis regresi, bagian ini akan menginvestigasi bagaimana data mining mampu mengenali serta memahami pengetahuan pada tingkat menengah. Studi kasus akan memberikan contoh aplikasi dalam konteks situasi nyata.
3. Pengetahuan Tinggi
   Pengetahuan pada tingkat tinggi merupakan bentuk pengetahuan yang paling kompleks dan mendalam, kemungkinan melibatkan pemahaman yang mendalam tentang bagaimana variabel berinteraksi dan berdampak pada suatu kejadian. Teknik seperti ensemble learning atau advanced neural networks dapat berperan dalam mengungkapkan pengetahuan pada tingkat tinggi ini. Melalui studi kasus, akan terlihat bagaimana pengetahuan ini dapat memberikan wawasan yang signifikan untuk mendukung pengambilan keputusan strategis.
   Dengan pemahaman yang lebih mendalam tentang klasifikasi pengetahuan dalam data mining, diharapkan pembaca mampu mengungkap potensi maksimal dari informasi yang ada dalam dataset besar dan menyusun

solusi serta strategi yang lebih efektif, terutama dalam beragam konteks aplikatif**.**

12.4 Pengetahuan Implisit vs. Pengetahuan Eksplisit

Bagian ini akan menguraikan perbedaan antara pengetahuan implisit dan eksplisit, serta bagaimana keduanya memiliki peran penting dalam konteks data mining. Kedua jenis pengetahuan ini membentuk dasar pemahaman yang mendalam tentang informasi yang terdapat dalam dataset, memberikan kemampuan kepada pengguna data mining untuk menggali wawasan yang memiliki nilai signifikan (Tan, P. N. 2020)

1. Pengetahuan Eksplisit
   a. Definisi dan Sifat: Pengetahuan eksplisit adalah informasi yang dapat dijelaskan dengan jelas dan telah didokumentasikan secara komprehensif. Hal ini mencakup aturan bisnis, fakta, dan data terstruktur lainnya yang dapat diwakili dengan spesifik dalam suatu format tertentu.
   b. Pengenalan Melalui Data Mining: Pada bagian ini, akan dibahas berbagai teknik dan algoritma data mining yang dapat diterapkan untuk mengenali, mengekstrak, dan memahami pengetahuan eksplisit. Contoh penerapan dalam berbagai sektor akan diuraikan guna memberikan pemahaman yang lebih baik terhadap konsep tersebut.
2. Pengetahuan Implisit
   a. Definisi dan Sifat: Pengetahuan implisit mencakup pemahaman, intuisi, atau wawasan yang tidak selalu dapat dijelaskan secara tegas. Hal ini kemungkinan terkait dengan pola yang sulit
      diungkapkan secara verbal atau pengetahuan yang melekat dalam pengalaman.
   b. Penarikan Informasi dengan Data Mining: Pada bagian ini, akan dibahas bagaimana data mining mampu mengeksplorasi pengetahuan implisit melalui teknik seperti clustering, analisis sentimen, atau analisis pola

yang kompleks. Studi kasus akan memberikan contoh nyata untuk menjelaskan penerapan konsep ini.

3. Integrasi Pengetahuan Eksplisit dan Implisit
    a. Manfaat Gabungan: Fokus dari pembahasan ini akan ditempatkan pada signifikansi mengintegrasikan keduanya. Bagaimana pengetahuan eksplisit dan implisit dapat digabungkan secara saling melengkapi, menghasilkan pemahaman data yang lebih menyeluruh dan mendalam.
    b. Contoh Penerapan Gabungan: Dengan merujuk pada situasi nyata, akan diuraikan bagaimana organisasi atau individu berhasil menggabungkan pengetahuan eksplisit dan implisit untuk meningkatkan proses pengambilan keputusan atau merancang strategi inovatif.
4. Tantangan dan Peluang
    a. Hambatan dalam Pengelolaan Pengetahuan: Pada bagian ini, akan diuraikan beberapa hambatan yang mungkin timbul dalam mengelola keduanya, beserta dengan bagaimana data mining dapat berperan dalam mengatasi rintangan tersebut.
    b. Potensi Inovasi: Mendeskripsikan potensi inovatif yang muncul ketika kedua jenis pengetahuan, baik eksplisit maupun implisit, dikelola dengan efektif. Hal ini memberikan dasar bagi pembaca untuk merancang pendekatan yang lebih proaktif dalam memanfaatkan informasi.

Bagian ini akan merangkum pentingnya memahami dan mengelola kedua jenis pengetahuan ini dalam konteks data mining. Penggunaan efektif dari pengetahuan eksplisit dan implisit dapat menjadi kunci untuk mendapatkan wawasan yang mendalam dan berharga dari dataset besar.

12.5 Penemuan Pengetahuan dalam Basis Data / *Knowledge Discovery in Databases* (KDD)

Bagian ini akan memperluas perspektif tentang konsep *Knowledge Discovery in Databases* (KDD) dan bagaimana hal ini

menjadi bagian yang tak terpisahkan dalam pemahaman jenis-jenis pengetahuan dalam data mining. KDD merupakan pendekatan sistematis untuk mengekstraksi pengetahuan yang bernilai dan sebelumnya tidak diketahui dari basis data besar (Zhang, D. 2021).

1. Pengantar KDD
   Definisi dan Lingkup: *Knowledge Discovery in Databases* (KDD) dapat dijelaskan sebagai suatu pendekatan terorganisir untuk mengekstraksi pengetahuan yang bernilai dari dataset besar yang sebelumnya tidak diketahui. Lingkup KDD mencakup serangkaian langkah dalam proses, dimulai dari pemilihan data hingga penafsiran hasil.
2. Langkah-langkah Proses KDD
   a. Pemilihan Data (*Selection)*: Memilih data awal yang relevan dan memiliki potensi untuk menghasilkan informasi yang bernilai.
   b. Pemurnian Data (*Pre-processing*): Proses membersihkan dan mentransformasi data dengan tujuan meningkatkan kualitasnya.
   c. Transformasi Data (*Transformation*): Merubah data ke dalam format yang lebih cocok untuk keperluan analisis.
   d. Data Mining: Penggunaan teknik-teknik seperti clustering, klasifikasi, dan asosiasi guna mengidentifikasi pola dan relasi.
   e. Evaluasi: Mengukur efektivitas model atau pola yang telah diidentifikasi.
   f. Interpretasi: Mentranslasikan hasil data mining menjadi pengetahuan yang dapat dipahami dan diterapkan.
3. Peran KDD dalam Pemahaman Jenis-jenis Pengetahuan
   a. Pengetahuan Eksplisit: Cara KDD mengenali dan mengekstrak pengetahuan eksplisit dari data yang terstruktur, seperti aturan bisnis atau fakta yang terdokumentasi.
   b. Pengetahuan Implisit: Tentang bagaimana KDD dapat mendukung penemuan pengetahuan implisit melalui analisis pola atau tren dalam data yang mungkin sulit dijelaskan secara verbal.

4. Implementasi KDD dalam Kasus Nyata
   Studi Kasus: Penggambaran kasus konkret yang menjelaskan penerapan konsep KDD dalam berbagai konteks, mulai dari sektor keuangan hingga kesehatan.
5. Tantangan dan Inovasi dalam KDD
   a. Hambatan: Mengenali rintangan yang mungkin timbul selama proses KDD, seperti kompleksitas data atau perluasan cakupan.
   b. Inovasi: Membahas perkembangan terbaru dalam KDD dan dampak teknologi terkini dalam membentuk lingkungan penemuan pengetahuan.

Dengan menyelami Knowledge Discovery in Databases (KDD) secara menyeluruh, diharapkan pembaca dapat memperoleh pemahaman yang kuat tentang peran inti KDD dalam lingkungan data mining serta bagaimana konsep ini membentuk fondasi untuk memahami jenis-jenis pengetahuan secara lebih mendalam (Ma, J., & Sun, D. 2022).

12.6 Teknik Data Mining untuk Menggali Pengetahuan

Bagian ini akan secara rinci menjelaskan pemanfaatan variasi teknik data mining, seperti klasterisasi, pengejaran aturan asosiasi, dan teknik lainnya, dalam usaha mengekstrak pengetahuan yang bernilai. Keterlibatan studi kasus dan contoh nyata diharapkan memberikan pemahaman yang jelas dan aplikatif (Jamsa, K. 2021).

1. *Clustering* (Klasterisasi)
   a. Definisi dan Penerapan: Penjelasan mendalam tentang apa itu klasterisasi dalam konteks data mining dan bagaimana teknik ini digunakan untuk mengelompokkan data yang serupa.
   b. Manfaat: Menjelaskan keuntungan yang diperoleh dari penerapan klasterisasi dalam mengidentifikasi pola yang tersembunyi dalam suatu kumpulan data.
2. *Association Rule Mining* (Aturan Asosiasi)
   a. Konsep dan Implementasi: Menguasai dasar-dasar pengejaran aturan asosiasi dan bagaimana teknik ini

diterapkan untuk mengenali keterkaitan antar variabel dalam suatu dataset.

b. Penerapan Praktis: Menghadirkan contoh kasus atau situasi nyata yang memberikan gambaran konkret tentang bagaimana aturan asosiasi diterapkan dalam konteks dunia nyata.

3. Studi Kasus dan Contoh Nyata
   a. Penerapan dalam Lingkungan Bisnis: Memberikan ilustrasi kasus dari perusahaan atau organisasi yang berhasil menerapkan teknik data mining untuk menghasilkan wawasan strategis.
   b. Pengaruh Positif: Menjelaskan konsekuensi positif yang dihasilkan dari penerapan teknik data mining, meliputi peningkatan efisiensi, pengambilan keputusan yang lebih optimal, atau terciptanya inovasi.
4. Tantangan dan Solusi
   a. Hambatan dalam Implementasi: Mengenali tantangan yang dapat timbul selama penerapan teknik data mining dan strategi yang dapat diadopsi oleh organisasi untuk mengatasi hambatan tersebut.
   b. Solusi yang Efektif: Memaparkan alternatif solusi yang efektif dalam mengatasi kendala yang mungkin muncul selama tahap implementasi.
5. Pertimbangan Etis
   Keamanan dan Privasi: Menyelidiki aspek etis yang terkait dengan penerapan teknik data mining, termasuk upaya untuk melindungi keamanan dan privasi data.

Melalui penelusuran rinci terhadap teknik data mining dan penerapannya dalam kasus-kasus studi, diharapkan pembaca dapat memperoleh pemahaman yang menyeluruh dan praktis tentang bagaimana teknologi ini dapat dimanfaatkan untuk mengeksplorasi pengetahuan berharga dalam berbagai situasi.

12.7 Pengetahuan dalam Prediksi dan Klasifikasi

Bagian ini akan mengulas peranan pengetahuan dalam aspek prediksi dan klasifikasi, dengan penekanan pada bagaimana wawasan yang diperoleh dari teknik data mining dapat digunakan untuk meramalkan hasil atau mengelompokkan data. Melibatkan contoh aplikasi di berbagai sektor industri diharapkan akan memberikan gambaran praktis tentang bagaimana pengetahuan dapat diaplikasikan dalam skenario dunia nyata (Gupta et al. 2021).

1. Penggunaan Pengetahuan dalam Prediksi
    a. Definisi Prediksi: Menguraikan konsep prediksi dalam kerangka data mining serta bagaimana pengetahuan dapat menjadi faktor krusial untuk meramalkan hasil berdasarkan pola atau tren yang teridentifikasi dalam data.
    b. Penerapan Metode Prediktif: Mendiskusikan variasi teknik prediksi, termasuk regresi, analisis deret waktu, atau pembelajaran mesin, dan bagaimana pengetahuan dapat memperbaiki keakuratan prediksi.
2. Peran Pengetahuan dalam Klasifikasi
    a. Klasifikasi dalam Data Mining: Menjelaskan secara terinci konsep klasifikasi dan cara pengetahuan dapat mendukung pengelompokan data ke dalam kategori atau kelas yang relevan.
    b. Metode Klasifikasi yang Umum: Menginformasikan tentang teknik klasifikasi seperti pohon keputusan, jaringan saraf tiruan, dan mesin vektor dukungan, dengan fokus pada cara pengetahuan dapat memperbaiki hasil klasifikasi.
3. Contoh Aplikasi dalam Berbagai Industri
    a. Bidang Keuangan: Bagaimana pengetahuan dimanfaatkan untuk meramalkan arah tren pasar, menilai risiko investasi, atau mendeteksi kecenderungan perubahan dalam kondisi ekonomi.
    b. Kesehatan: Ilustrasi pemanfaatan pengetahuan dalam mengelompokkan penyakit, meramalkan penyebaran penyakit, atau mengidentifikasi faktor risiko.

c. Pemasaran: Cara pengetahuan mendukung meramalkan perilaku konsumen atau mengelompokkan target pasar untuk merancang strategi pemasaran yang lebih efektif.

4. Keuntungan dan Tantangan
   a. Manfaat Penerapan Pengetahuan: Menjelaskan bagaimana pemanfaatan pengetahuan dalam proses prediksi dan klasifikasi dapat menghasilkan keuntungan, seperti pengambilan keputusan yang lebih tepat atau pengenalan peluang baru.
   b. Tantangan yang Potensial Timbul: Mengeksplorasi rintangan atau isu etis yang mungkin timbul dalam menggunakan pengetahuan untuk prediksi dan klasifikasi.
5. Etika dalam Penggunaan Pengetahuan
   Pertimbangan Etis: Fokus pada pertimbangan etis terkait pemanfaatan pengetahuan dalam prediksi dan klasifikasi, mencakup aspek privasi dan kejelasan.

Dengan penjelasan yang komprehensif ini, diharapkan pembaca memperoleh pemahaman yang mendalam tentang kontribusi pengetahuan dalam meramalkan hasil dan mengelompokkan data, beserta penerapannya di berbagai sektor industri.

12.8 Kesimpulan

1. Definisi dan Konsep Dasar
   Jenis-jenis Knowledge pada Data Mining merujuk pada pemahaman yang mendalam terhadap berbagai bentuk pengetahuan yang ditemukan melalui analisis data dengan memanfaatkan teknik data mining. Sentral dalam perhatian ini adalah pemahaman bahwa pengetahuan tidak terbatas pada informasi terorganisir semata, melainkan juga menjadi suatu entitas strategis yang menjadi dasar bagi pengambilan keputusan dan pembentukan wawasan yang mendalam.
2. Klasifikasi Knowledge dalam Data Mining
   Klasifikasi Knowledge dalam Data Mining membahas bagaimana pengetahuan dapat dikelompokkan berdasarkan tingkat kompleksitas dan pemahaman. Terdapat kategori

pengetahuan rendah, menengah, dan tinggi, masing-masing membawa implikasi unik dalam proses data mining.

3. Pengetahuan Implisit vs. Pengetahuan Eksplisit
   Dalam kerangka Data Mining, diuraikan perbedaan antara pengetahuan eksplisit dan implisit. Pengetahuan eksplisit adalah informasi yang dapat dijelaskan secara tegas, sedangkan pengetahuan implisit melibatkan pemahaman atau keahlian yang mungkin sulit diungkapkan secara verbal.
4. Penemuan Pengetahuan dalam Basis Data (Knowledge Discovery in Databases - KDD)
   Knowledge Discovery in Databases (KDD) merupakan suatu pendekatan sistematis dalam mengekstraksi pengetahuan yang berharga dari dataset besar yang sebelumnya tidak diketahui. Tahapan proses KDD melibatkan pemilihan data, pembersihan, transformasi, data mining, evaluasi, dan interpretasi.
5. Teknik Data Mining untuk Menggali Pengetahuan
   Penerapan teknik data mining, seperti clustering, association rule mining, dan teknik lainnya, diuraikan untuk mengekstrak pengetahuan yang berharga dari dataset. Melibatkan studi kasus dan contoh spesifik memberikan pemahaman praktis tentang bagaimana teknik-teknik ini diimplementasikan.
6. Pengetahuan dalam Prediksi dan Klasifikasi
   Pengetahuan dimanfaatkan secara efektif dalam prediksi dan klasifikasi dengan menggunakan teknik seperti regresi, analisis time series, dan machine learning. Contoh aplikasi dalam sejumlah industri, termasuk keuangan, kesehatan, dan pemasaran, memberikan gambaran nyata tentang penerapan pengetahuan dalam berbagai konteks dunia nyata.

DAFTAR PUSTAKA

Ghavami, P. (2019). Big Data Analytics Methods. Walter de Gruyter GmbH & Co KG.

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). Data Mining for Business Analytics. John Wiley & Sons.

Ma, J., & Sun, D. (2022). Data Mining And Machine Learning Applications. John Wiley & Sons.

Gupta, B.B., Perakovi, D., Ahmed and Gupta, D. (2021). Data Mining Approaches for Big Data and Sentiment Analysis in Social Media. IGI Global.

Zhang, D. (2021). Fundamentals of Image Data Mining. Springer Nature.

Jamsa, K. (2021). Introduction to data mining and analytics : with machine learning in R and Python. Burlington, Ma Jones Et Bartlett Learning.

Pang-Ning Tan (2020). Introduction to data mining : international edition. Harlow: Pearson Education.

Mehmed Kantardzic (2019). DATA MINING : concepts, models, methods, and algorithms.

BAB 13
FASE PADA KNOWLEDGE MANAGEMENT LIFE CYCLE

Oleh Nurul Afifah Arifuddin

13.1 Pendahuluan

Pada era informasi saat ini, organisasi menghadapi tantangan dalam mengelola jumlah data yang terus meningkat. Terlebih lagi, data tersebut cenderung kompleks dan memiliki potensi untuk menghasilkan nilai yang signifikan jika dikelola dengan baik. Oleh karena itu, pengelolaan pengetahuan (*knowledge management*) menjadi krusial untuk membantu organisasi dalam memahami, menyimpan, dan menggunakan pengetahuan yang terkandung dalam data.

Pentingnya pengelolaan pengetahuan semakin diperkuat dalam konteks data mining, di mana organisasi dapat menggunakan teknik dan algoritma untuk mengeksplorasi pola, tren, dan pengetahuan yang tersembunyi dalam data besar. Sebagai contoh, perusahaan yang mampu mengoptimalkan *Knowledge Management Life Cycle* (KMLC) dalam konteks data mining dapat memperoleh wawasan mendalam dan mengambil keputusan yang lebih tepat.

Bab ini bertujuan untuk memberikan wawasan mendalam tentang konsep Knowledge Management Life Cycle (KMLC) dan bagaimana penerapannya dalam konteks data mining dapat memberikan nilai tambah signifikan bagi organisasi. Dengan memahami tahapan-tahapan KMLC dan cara integrasinya dengan teknik data mining, diharapkan pembaca dapat mengembangkan pemahaman yang lebih baik tentang pengelolaan pengetahuan dalam konteks yang dinamis dan kompleks ini.

Dalam perkembangan bab ini, kita akan membahas secara rinci setiap tahap KMLC dan bagaimana data mining dapat diterapkan secara efektif dalam setiap tahapan tersebut. Selain itu, kita akan menjelaskan tantangan yang mungkin dihadapi dalam

implementasi KMLC dengan data mining serta strategi untuk mengatasi tantangan tersebut. Studi kasus praktis akan mendukung pemahaman konsep dan memberikan gambaran nyata tentang penerapan KMLC dalam lingkungan bisnis sehari-hari.

Dengan demikian, bab ini diharapkan dapat menjadi panduan yang komprehensif bagi pembaca untuk mengimplementasikan KMLC dalam konteks data mining serta memastikan pengelolaan pengetahuan yang efektif dan berkelanjutan.

13.2 Konsep Dasar *Knowledge Management Life Cycle*

Pada era di mana informasi menjadi salah satu aset paling berharga bagi organisasi, Konsep Dasar *Knowledge Management Life Cycle* (KMLC) telah menjadi landasan strategis dalam upaya mengelola pengetahuan. KMLC mencakup serangkaian tahapan sistematis, yang dalam konteks data mining, menjadi semakin krusial. Tahapan-tahapan ini, yang melibatkan pengumpulan, penyimpanan, distribusi, dan pemanfaatan pengetahuan, membentuk fondasi yang kokoh untuk memahami, mengelola, dan memanfaatkan informasi dalam lingkungan organisasi yang dinamis. Dalam bab ini, kita akan mengeksplorasi definisi KMLC, membahas peran penting data mining dalam setiap tahapan, dan menyelidiki bagaimana integrasi kedua konsep ini dapat membawa dampak positif pada kemampuan organisasi untuk membuat keputusan yang informasional dan proaktif.

13.2.1 Definisi KMLC

Knowledge Management Life Cycle (KMLC) merupakan suatu rangkaian tahapan yang sistematis dalam pengelolaan pengetahuan dari awal hingga akhir. KMLC mencakup proses pengumpulan, penyimpanan, distribusi, dan pemanfaatan pengetahuan dalam suatu organisasi (Bernard and Tichkiewitch, 2010). Dalam konteks data mining, KMLC menjadi kritis karena melibatkan ekstraksi pengetahuan dari data yang kompleks.

Beberapa definisi penting dalam konteks KMLC adalah:

1. Menentukan sumber-sumber data yang relevan untuk organisasi.
2. Menggunakan teknik data mining untuk mengidentifikasi pola dan tren dalam data.
3. Menerapkan strategi penyimpanan pengetahuan yang sesuai untuk memastikan aksesibilitas dan keberlanjutan.
4. Membangun sistem distribusi pengetahuan yang efisien.
5. Mengoptimalkan pemanfaatan pengetahuan dalam pengambilan keputusan organisasi.

13.2.2 Tahapan KMLC

1. Pengumpulan Pengetahuan (*Knowledge Acquisition*)
   Tahapan pertama KMLC yaitu Pengumpulan Pengetahuan, memainkan peran sentral dalam menghasilkan informasi yang bernilai dari data yang ada. Dalam konteks data mining, pendekatan ini melibatkan identifikasi sumber data yang relevan dan penerapan berbagai teknik analisis data. Algoritma klasifikasi membantu dalam mengelompokkan data, sedangkan algoritma *clustering* membantu mengidentifikasi pola yang dapat memberikan wawasan tentang hubungan antar elemen data. Selain itu, teknik regresi digunakan untuk memahami hubungan kausal antar variabel. Dengan memahami dan menerapkan teknik ini, organisasi dapat menggali pengetahuan yang mendalam dari data serta membuka pintu untuk pengambilan keputusan yang lebih baik. (Fayyad, Piatetsky-Shapiro and Smyth, 1996)
2. Penyimpanan Pengetahuan (*Knowledge Storage*)
   Setelah pengetahuan berhasil dikumpulkan, langkah berikutnya dalam KMLC adalah penyimpanan pengetahuan. Ini melibatkan pemilihan sistem penyimpanan yang tepat untuk memastikan pengetahuan dapat diakses dan dikelola secara efisien. Hasil analisis data dapat disimpan dalam basis data atau *data warehouse*. Keputusan dalam penyimpanan data ini penting untuk memastikan keberlanjutan dan aksesibilitas pengetahuan. Penyimpanan pengetahuan (data) yang efektif, memungkinkan organisasi untuk merinci temuan dari analisis

data mining dan membuatnya siap untuk distribusi dan pemanfaatan. (Dyché, 2000)

3. Distribusi Pengetahuan (*Knowledge Distribution*)
   Distribusi pengetahuan adalah tahap penting untuk memastikan bahwa wawasan yang ditemukan dapat diakses oleh pemangku kepentingan yang relevan dalam organisasi. Infrastruktur distribusi yang baik dapat mempercepat proses pengambilan keputusan. Teknologi seperti portal pengetahuan, sistem manajemen dokumen, dan intranet membantu menyampaikan informasi secara efisien. Dalam hal ini, data mining juga dapat digunakan untuk memahami kebutuhan pengguna dan mempersonalisasi pengalaman distribusi pengetahuan. Dengan membangun sistem distribusi yang adaptif, organisasi dapat memastikan bahwa pengetahuan yang dihasilkan dapat mencapai pemangku kepentingan yang tepat pada waktu yang tepat. (White and chaffey, 2010)
4. Pemanfaatan Pengetahuan (*Knowledge Utilization*)
   Tahap terakhir KMLC, Pemanfaatan Pengetahuan, mengarah pada integrasi hasil data mining dalam pengambilan keputusan organisasi. Hasil analisis data mining diaplikasikan dalam strategi bisnis, membantu meningkatkan efektivitas operasional dan mendukung pengambilan keputusan yang berbasis bukti. Dalam tahapan ini, organisasi memastikan bahwa pengetahuan yang dihasilkan dari data mining benar-benar memberikan nilai tambah dan relevan untuk mencapai tujuan bisnis.

Dengan memahami setiap tahapan KMLC, organisasi dapat mencapai siklus pengelolaan pengetahuan yang holistik dan berkelanjutan. Integrasi data mining dalam tahap pengumpulan pengetahuan membuka peluang baru untuk mendapatkan wawasan mendalam dari data yang ada. (Sharda *et al.*, 2014)

13.3 Integrasi Data Mining dalam KMLC

Integrasi *Knowledge Management Life Cycle* (KMLC) dengan teknik data mining membentuk fondasi penting dalam mengelola dan memanfaatkan pengetahuan organisasi secara efektif. Dalam

bab ini, kita akan mengeksplorasi bagaimana data mining memperkaya setiap tahapan KMLC, dari pengumpulan hingga pemanfaatan pengetahuan. Fokusnya akan tertuju pada peran utama data mining dalam mendukung proses pengambilan keputusan dan meningkatkan efisiensi pengelolaan pengetahuan organisasi.

13.3.1 Peran Data Mining dalam Pengumpulan Pengetahuan

Data Mining memegang peran sentral dalam tahapan pertama *Knowledge Management Life Cycle* (KMLC), yakni Pengumpulan Pengetahuan. Melibatkan proses identifikasi dan pengumpulan data yang relevan, teknik data mining membuka pintu untuk pemahaman yang mendalam dari data organisasi. Algoritma klasifikasi, salah satu alat data mining, memberikan kemampuan untuk mengelompokkan data ke dalam kategori yang jelas berdasarkan atribut tertentu. Misalnya, dalam sebuah perusahaan *e-commerce*, klasifikasi dapat membantu mengelompokkan pelanggan ke dalam segmen berdasarkan perilaku pembelian mereka.

Selain itu, teknik *clustering* dalam data mining memungkinkan organisasi mengidentifikasi pola dan hubungan dalam data yang tidak memiliki kategori sebelumnya. Misalnya, dalam industri kesehatan, clustering dapat membantu mengelompokkan pasien berdasarkan gejala, memungkinkan identifikasi pola penyakit yang mungkin terlewatkan secara manual.

Selanjutnya, teknik regresi membantu memahami hubungan kausal antar variabel. Contohnya dalam konteks keuangan, regresi dapat membantu memahami faktor-faktor yang mempengaruhi kinerja saham.

Integrasi data mining dalam tahap Pengumpulan Pengetahuan juga membawa manfaat dalam mengidentifikasi anomali atau pencilan dalam data. Dengan mendeteksi pola yang tidak biasa, organisasi dapat merespons secara proaktif terhadap situasi yang tidak terduga. Sebagai contoh, dalam industri manufaktur, data mining dapat digunakan untuk mengidentifikasi proses produksi yang tidak efisien atau bahan baku yang memiliki tingkat cacat tinggi.

Pentingnya data mining dalam tahap ini juga terlihat dalam konteks eksplorasi dan ekstraksi wawasan dari data yang kompleks. Algoritma data mining mampu mengidentifikasi tren dan pola yang mungkin tidak terlihat secara langsung. Dalam perusahaan media, misalnya, data mining dapat digunakan untuk menganalisis perilaku pemirsa dan mengidentifikasi tren konten yang paling populer.

Dengan adanya data mining, organisasi dapat memaksimalkan kegunaan data mereka. Teknik-teknik data mining tidak hanya memproses data secara cepat tetapi juga dapat menangkap hubungan yang rumit diantara variabel. Oleh karena itu, data mining menjadi alat yang sangat efektif untuk menggali potensi pengetahuan yang terkandung dalam data besar.

Namun, tantangan yang dihadapi dalam peran data mining ini mencakup kompleksitas data yang terus berkembang, kebutuhan akan sumber daya komputasi yang besar, dan kebutuhan untuk melindungi privasi dan keamanan data. Organisasi perlu mempertimbangkan metode pengelolaan data yang efektif dan solusi keamanan yang andal untuk memaksimalkan manfaat dari peran data mining dalam pengumpulan pengetahuan.

Penerapan data mining dalam tahap ini juga memerlukan pemahaman yang mendalam tentang bisnis dan kebutuhan spesifik organisasi. Diperlukan kolaborasi erat antara tim analis data dan pemangku kepentingan bisnis untuk memastikan bahwa analisis data mining relevan dan bermanfaat untuk pencapaian tujuan organisasi.

Kesimpulannya, peran data mining dalam Pengumpulan Pengetahuan sangatlah signifikan. Dengan kemampuannya untuk mengidentifikasi pola, tren, dan hubungan dalam data, data mining dapat membantu organisasi memahami informasi yang terkandung dalam data besar. Melalui penerapan teknik-teknik data mining, organisasi dapat membuka potensi wawasan baru, mendukung pengambilan keputusan yang informasional, dan memacu inovasi dalam pemanfaatan pengetahuan. (Sharda *et al.*, 2014) (Fayyad, Piatetsky-Shapiro and Smyth, 1996)

13.3.2 Penyimpanan Pengetahuan yang efisien melalui data mining

Penyimpanan pengetahuan, tahap kedua dalam *Knowledge Management Life Cycle* (KMLC), membutuhkan strategi efisien untuk memastikan pengetahuan yang dikumpulkan dapat diakses dan dikelola dengan optimal. Dalam konteks ini, data mining memainkan peran penting dalam memilih dan mengoptimalkan sistem penyimpanan yang sesuai dengan karakteristik pengetahuan yang dihasilkan.

Secara teknis, data mining membantu organisasi untuk menentukan jenis sistem penyimpanan yang paling mendukung pengetahuan yang dihasilkan. Analisis data mining dapat memberikan wawasan mendalam tentang struktur dan pola pengetahuan, memungkinkan organisasi untuk memilih antara penyimpanan berbasis basis data, *data warehouse*, atau solusi manajemen pengetahuan yang lebih kompleks.

Algoritma *clustering* dalam data mining dapat membantu dalam mengelompokkan pengetahuan yang mirip atau terkait ke dalam unit yang terorganisir. Dengan menggunakan teknik ini, organisasi dapat memahami bagaimana pengetahuan dapat diorganisir secara efisien dalam penyimpanan, meminimalkan redundansi dan meningkatkan aksesibilitas.

Selanjutnya, teknik regresi dapat digunakan untuk merinci hubungan antar variabel dalam pengetahuan yang disimpan. Ini memungkinkan organisasi untuk menentukan sejauh mana variabel dapat berdampak pada pengetahuan dan bagaimana variabel tersebut dapat dioptimalkan dalam konteks penyimpanan.

Selain itu, data mining membantu dalam identifikasi dan pemisahan pengetahuan yang kritis dari yang kurang penting. Dengan analisis yang cermat, organisasi dapat menentukan pengetahuan yang perlu disimpan dalam penyimpanan inti, sementara pengetahuan yang kurang kritis dapat diarahkan ke penyimpanan yang lebih ekonomis.

Keamanan data juga menjadi aspek krusial dalam penyimpanan pengetahuan. Data mining dapat digunakan untuk mengidentifikasi potensi risiko keamanan dan mendukung implementasi langkah-langkah keamanan yang sesuai. Misalnya,

teknik deteksi anomali dapat membantu mendeteksi perilaku yang mencurigakan atau serangan siber yang potensial.

Penting untuk dicatat bahwa penyimpanan pengetahuan yang efisien melibatkan pemahaman mendalam tentang kebutuhan bisnis dan pengguna akhir. Data mining memfasilitasi pengumpulan umpan balik dari pemangku kepentingan dan analisis kebutuhan pengguna, memastikan bahwa sistem penyimpanan dikonfigurasi sesuai dengan persyaratan bisnis yang sebenarnya.

Tantangan dalam penyimpanan pengetahuan melalui data mining termasuk manajemen volume data yang terus berkembang, kebutuhan akan teknologi penyimpanan yang skalabel, dan implementasi strategi penyimpanan yang mempertimbangkan perkembangan bisnis di masa depan.

Adapun solusi teknis untuk meningkatkan efisiensi penyimpanan melalui data mining termasuk penggunaan teknologi kompresi data, distribusi data, dan otomatisasi manajemen siklus hidup data. Analisis data mining juga memungkinkan organisasi untuk merancang strategi pemulihan bencana yang lebih efektif.

Penyimpanan pengetahuan yang efisien melalui data mining melibatkan pemilihan dan pengelolaan sistem penyimpanan yang sesuai. Dengan memanfaatkan teknik clustering, regresi, dan identifikasi risiko keamanan, organisasi dapat meningkatkan aksesibilitas, keamanan, dan efisiensi penyimpanan pengetahuan mereka. Data mining menjadi kunci untuk mengoptimalkan penyimpanan pengetahuan dalam konteks *Knowledge Management Life Cycle.* (Han, Kamber and Pei, 2011)

13.3.3 Distribusi Pengetahuan yang Dipercepat

Distribusi pengetahuan merupakan tahapan penting dalam *Knowledge Management Life Cycle* (KMLC), di mana informasi yang dikumpulkan dan disimpan diarahkan ke pemangku kepentingan yang tepat. Dalam konteks distribusi pengetahuan yang dipercepat, data mining memegang peran strategis dalam memahami kebutuhan pemangku kepentingan dan meningkatkan efisiensi proses distribusi.

Teknik data mining dapat digunakan untuk memahami preferensi dan kebutuhan pemangku kepentingan secara lebih

mendalam. Analisis klaster atau pengelompokan data membantu dalam mengidentifikasi pola dan karakteristik kelompok pemangku kepentingan yang serupa. Sebagai contoh, disektor e-commerce, data mining dapat mengidentifikasi kelompok pelanggan dengan preferensi produk serupa, memudahkan penyusunan strategi distribusi yang lebih tepat.

Dalam meningkatkan kecepatan distribusi pengetahuan, personalisasi pengalaman pengguna menjadi fokus penting. Data mining memungkinkan organisasi untuk memahami preferensi individu pemangku kepentingan dan menyajikan pengetahuan dengan cara yang paling relevan dan menarik bagi setiap pengguna. Algoritma rekomendasi yang berbasis data mining, seperti Collaborative Filtering, dapat mempercepat distribusi dengan menyajikan konten yang sesuai dengan minat dan kebutuhan masing-masing pengguna.

Teknik data mining juga dapat digunakan untuk memprediksi tren dan kebutuhan masa depan, memungkinkan organisasi untuk mempersiapkan distribusi pengetahuan dengan lebih proaktif. Algoritma time series forecasting dalam data mining dapat membantu memahami pola perilaku pemangku kepentingan dari waktu ke waktu, memberikan wawasan yang berharga untuk merancang strategi distribusi jangka panjang.

Dalam situasi di mana waktu distribusi kritis, teknik data mining seperti Association Rules dapat membantu memahami hubungan antar elemen pengetahuan. Dengan memahami korelasi antara informasi, organisasi dapat menyusun paket pengetahuan yang lebih terkait secara kontekstual dan meningkatkan efisiensi distribusi.

Data mining juga dapat memainkan peran kunci dalam pemilihan platform distribusi yang paling efisien. Dengan menganalisis preferensi dan kebiasaan pemangku kepentingan, organisasi dapat memilih platform yang sesuai, seperti portal pengetahuan, sistem manajemen dokumen, atau aplikasi seluler, untuk mendukung proses distribusi. (Aggarwal, 2015)

Namun, perlu diperhatikan bahwa distribusi pengetahuan yang dipercepat juga melibatkan tantangan dalam menjaga keamanan dan keakuratan informasi yang didistribusikan. Data

mining dapat membantu dalam deteksi dini potensi risiko atau informasi yang tidak akurat, memberikan lapisan tambahan perlindungan dalam proses distribusi.

Pentingnya data mining dalam distribusi pengetahuan yang dipercepat juga mencakup pemahaman mengenai keberlanjutan strategi distribusi. Melalui analisis data mining, organisasi dapat mendapatkan umpan balik langsung dari pemangku kepentingan terkait efektivitas dan kepuasan distribusi, memungkinkan mereka untuk melakukan perubahan atau penyesuaian yang diperlukan.

Dalam menghadapi tantangan integrasi data yang semakin besar dan kompleks, kebutuhan akan algoritma dan teknik data mining yang efisien semakin penting. Pemilihan model analisis yang sesuai dan pemrosesan data yang cepat menjadi kunci dalam mendukung distribusi pengetahuan yang dipercepat.

Dalam kesimpulan, distribusi pengetahuan yang dipercepat memanfaatkan data mining sebagai sarana untuk memahami kebutuhan, preferensi, dan perilaku pemangku kepentingan. Dengan memanfaatkan teknik data mining, organisasi dapat mempercepat proses distribusi, meningkatkan personalisasi, dan menjawab kebutuhan pemangku kepentingan dengan lebih tepat waktu dan efisien dalam konteks *Knowledge Management Life Cycle.*

13.3.4 Meningkatkan Pengambilan Keputusan Melalui Data Mining

Meningkatkan pengambilan keputusan merupakan esensi dari keberhasilan sebuah organisasi. Tahapan terakhir *Knowledge Management Life Cycle* (KMLC), Pemanfaatan Pengetahuan, menjadi landasan bagi pengambilan keputusan yang strategis. Dalam konteks ini, data mining memegang peran sentral dalam memberikan wawasan dan analisis mendalam yang mendukung pengambilan keputusan yang lebih baik.

Pertama-tama, data mining membantu organisasi untuk mengidentifikasi pola dan tren yang mungkin tidak terlihat secara manual. Algoritma klasifikasi dan clustering digunakan untuk mengelompokkan dan mengidentifikasi hubungan antar data, memberikan dasar yang kuat untuk pemahaman kontekstual yang lebih mendalam. Sebagai contoh, di sektor keuangan, data mining

dapat membantu mengidentifikasi pola perilaku pasar yang mempengaruhi keputusan investasi.

Teknik regresi dalam data mining membantu organisasi untuk memahami hubungan kausal antar variabel. Ini memungkinkan pembuatan model matematis yang dapat memprediksi dampak perubahan dalam satu variabel terhadap variabel lainnya. Dalam industri manufaktur, misalnya, regresi dapat digunakan untuk meramalkan dampak perubahan dalam proses produksi terhadap kualitas produk.

Selanjutnya, data mining membantu dalam identifikasi faktor kritis yang mempengaruhi hasil keputusan. Algoritma Association Rules dan Sequential Pattern Mining membantu dalam mengungkapkan keterkaitan antar variabel, memberikan pemahaman lebih lanjut tentang faktor-faktor yang dapat memengaruhi hasil keputusan. Dalam sektor e-commerce, ini dapat digunakan untuk memahami faktor-faktor yang memotivasi konsumen untuk melakukan pembelian.

Meningkatkan pengambilan keputusan juga melibatkan prediksi berdasarkan data historis. Time series forecasting dalam data mining membantu organisasi untuk meramalkan tren masa depan berdasarkan data masa lalu. Misalnya, dalam sektor kesehatan, data mining dapat digunakan untuk meramalkan tingkat penyebaran penyakit berdasarkan pola historis dan faktor-faktor tertentu.

Dalam aspek personalisasi pengambilan keputusan, data mining memungkinkan pengembangan model rekomendasi yang disesuaikan dengan preferensi individu. Collaborative Filtering dan Content-Based Filtering adalah teknik data mining yang digunakan untuk menyajikan rekomendasi yang relevan dan disesuaikan dengan kebutuhan pengguna. Hal ini dapat diterapkan dalam platform e-learning untuk memberikan materi pembelajaran yang sesuai dengan minat dan kebutuhan siswa.

Pentingnya data mining dalam meningkatkan pengambilan keputusan juga terlihat dalam pengoptimalkan strategi bisnis. Melalui analisis data mining, organisasi dapat memahami dinamika pasar, kecenderungan konsumen, dan perubahan dalam lingkungan bisnis. Sebagai contoh, perusahaan ritel dapat menggunakan data

mining untuk mengoptimalkan strategi pemasaran dan penempatan produk di toko berdasarkan pola pembelian pelanggan.

Data mining juga membantu dalam merinci keputusan berdasarkan berbagai skenario dan kondisi. Dengan memanfaatkan teknik simulasi dan skenario analisis, organisasi dapat mengevaluasi dampak potensial dari berbagai keputusan yang mungkin diambil. Ini memungkinkan pengambilan keputusan yang lebih informasional dan berbasis bukti. (Witten, Frank and Hall, 2011a)

Tantangan dalam meningkatkan pengambilan keputusan melalui data mining mencakup kebutuhan untuk mengelola kompleksitas data yang terus berkembang, memastikan keakuratan model prediktif, dan mengatasi bias dalam data. Organisasi perlu menjaga keseimbangan antara kompleksitas teknis dan pemahaman bisnis yang mendalam.

Dalam menghadapi tantangan ini, organisasi perlu mengembangkan keahlian analisis data mining, memastikan ketersediaan sumber daya komputasi yang memadai, dan fokus pada kualitas data yang akurat. Dengan demikian, organisasi dapat memaksimalkan potensi data mining dalam memberikan wawasan yang diperlukan untuk mendukung pengambilan keputusan yang lebih baik. Dengan menerapkan teknik dan algoritma data mining yang tepat, organisasi dapat meminimalkan risiko dan mencapai keberhasilan dalam pengambilan keputusan strategis.

13.4 Tantangan dan Strategi Implementasi

Dalam Bab 13.4 ini, kita akan menjelajahi dinamika menarik antara implementasi *Knowledge Management Life Cycle* (KMLC) dan pemanfaatan data mining. Mengintegrasikan KMLC dengan data mining menawarkan potensi besar untuk meningkatkan pengelolaan pengetahuan dan pengambilan keputusan di organisasi. Namun, perjalanan ini tidak terlepas dari tantangan-tantangan kritis yang perlu dihadapi. Dari keamanan data hingga pengelolaan sumber daya dan resistensi organisasional, bab ini secara rinci menggali tantangan implementasi yang mungkin muncul. Selanjutnya, kita akan merinci strategi-strategi yang dapat diterapkan untuk mengatasi tantangan tersebut, memastikan bahwa

integrasi KMLC dengan data mining memberikan dampak positif dan berkelanjutan dalam konteks pengelolaan pengetahuan.

13.4.1 Tantangan dalam Implementasi KMLC dengan Data Mining

Implementasi *Knowledge Management Life Cycle* (KMLC) dengan memanfaatkan data mining dapat memberikan manfaat besar, tetapi juga dihadapkan pada sejumlah tantangan yang perlu diatasi dengan cermat.

1. Keamanan Data
   Salah satu tantangan utama dalam integrasi KMLC dengan data mining adalah masalah keamanan data. Dalam pengumpulan dan penyimpanan pengetahuan, perlindungan terhadap data menjadi kritis. Data mining melibatkan akses dan analisis data yang luas, sehingga perlindungan privasi dan keamanan menjadi perhatian utama. Dalam konteks ini, organisasi perlu mengembangkan kebijakan keamanan data yang ketat dan menerapkan teknologi enkripsi yang kuat untuk melindungi informasi yang sensitif.
2. Sumber Daya
   Integrasi KMLC dengan data mining membutuhkan sumber daya yang signifikan, termasuk perangkat keras yang kuat, perangkat lunak analisis data, dan tim ahli data mining. Tantangan terkait sumber daya termasuk biaya implementasi, kebutuhan akan kompetensi analis data yang tinggi, dan keberlanjutan pemeliharaan infrastruktur. Dalam mengatasi tantangan ini, organisasi perlu mengembangkan model biaya yang berkelanjutan dan melibatkan pelatihan karyawan untuk meningkatkan kapabilitas analisis data internal.
3. Resistensi Organisasional:
   Perubahan selalu dihadapi dengan resistensi, dan implementasi KMLC dengan data mining tidak terkecuali. Organisasi sering mengalami ketidaknyamanan atau kekhawatiran terkait dengan perubahan proses kerja yang sudah mapan. Menyelaraskan kebijakan perusahaan, memberikan klarifikasi terkait manfaat dan tujuan, serta

mengidentifikasi pemenang dalam organisasi dapat membantu mengurangi resistensi tersebut. (Han, Kamber and Pei, 2011)(Witten, Frank and Hall, 2011b)

13.4.2 Strategi Mengatasi Tantangan Implementasi

1. Pelatihan Karyawan:
   Strategi pertama dalam mengatasi tantangan implementasi adalah memberikan pelatihan yang memadai kepada karyawan. Karyawan perlu memahami konsep KMLC dan bagaimana data mining dapat meningkatkan proses pengelolaan pengetahuan. Pelatihan ini dapat mencakup penggunaan alat data mining, interpretasi hasil analisis, dan praktik terbaik dalam pengambilan keputusan berbasis data.
2. Kebijakan Keamanan Data yang Jelas
   Membentuk kebijakan keamanan data yang jelas dan ketat menjadi langkah kunci dalam mengatasi tantangan keamanan. Organisasi perlu mengidentifikasi data apa yang dapat diakses oleh siapa, serta menerapkan kontrol akses dan enkripsi data yang sesuai. Pemahaman dan kesadaran akan kebijakan keamanan ini perlu ditanamkan dalam seluruh organisasi.
3. Integrasi Sistem yang Teliti
   Strategi penting lainnya adalah memastikan integrasi yang teliti antara sistem KMLC dan alat data mining. Ini melibatkan pemilihan perangkat lunak yang kompatibel, pengaturan antarmuka yang efisien, dan pemantauan kinerja sistem secara terus-menerus. Integrasi yang baik memastikan efisiensi operasional dan pengambilan keputusan yang optimal.
4. Partisipasi Aktif Pemangku Kepentingan:
   Melibatkan pemangku kepentingan secara aktif dalam proses implementasi dapat membantu mengatasi resistensi organisasional. Komunikasi terbuka, sesi kolaborasi, dan pengumpulan umpan balik secara berkala dapat menciptakan pemahaman bersama dan meredakan kekhawatiran.

Dengan mengakui dan mengatasi tantangan ini melalui strategi yang tepat, organisasi dapat memaksimalkan potensi integrasi KMLC dengan data mining, mencapai pengelolaan pengetahuan yang lebih efektif, dan meningkatkan kemampuan pengambilan keputusan mereka.

13.5 Studi Kasus

Dalam bab ini, kami akan mengeksplorasi sebuah studi kasus yang mengilustrasikan keberhasilan integrasi *Knowledge Management Life Cycle (KMLC)* dengan teknologi data mining dalam konteks organisasi yang mencari untuk mengoptimalkan pengelolaan pengetahuan mereka.

1. Latar Belakang Organisasi:
   Sebuah perusahaan teknologi global dengan kehadiran di berbagai sektor, mulai dari manufaktur hingga layanan pelanggan. Organisasi ini menyadari potensi besar yang dimiliki oleh data yang mereka kumpulkan dan menyimpan, dan ingin memanfaatkannya dengan lebih efektif untuk meningkatkan pengambilan keputusan dan inovasi di seluruh lapisan perusahaan.
2. Tantangan yang Diidentifikasi:
   Organisasi ini menghadapi sejumlah tantangan dalam pengelolaan pengetahuan mereka, termasuk kesulitan dalam mengekstrak wawasan yang bernilai dari volume data yang terus berkembang, dan kurangnya visibilitas terhadap pengetahuan yang ada di seluruh departemen. Mereka menyadari perlunya pendekatan terstruktur untuk mengelola pengetahuan dan memutuskan untuk mengimplementasikan KMLC.
3. Implementasi KMLC dengan Data Mining:
   Organisasi memutuskan untuk mengadopsi pendekatan KMLC dengan integrasi data mining untuk memecahkan tantangan yang dihadapi. Mereka memulai dengan tahap Pengumpulan Pengetahuan, menggunakan algoritma data mining untuk mengidentifikasi pola dan tren yang tersembunyi dalam data mereka. Ini mencakup analisis klasifikasi dan clustering untuk memahami relasi antar variabel dan mengelompokkan pengetahuan yang serupa.

Dalam tahap Penyimpanan Pengetahuan, organisasi memilih sistem penyimpanan yang didukung oleh teknologi data mining untuk memastikan penyimpanan yang efisien dan aksesibilitas optimal. Mereka menggunakan analisis regresi untuk memahami sejauh mana variabel pengetahuan memengaruhi keputusan dan untuk merinci hubungan kausal yang mungkin tidak terlihat sebelumnya.

Dalam Distribusi Pengetahuan, data mining digunakan untuk mempercepat proses distribusi dengan mengidentifikasi preferensi dan kebutuhan pemangku kepentingan secara lebih akurat. Algoritma rekomendasi dipakai untuk menyajikan pengetahuan yang relevan dan menyesuaikan pengalaman pengguna berdasarkan pola perilaku individu.

4. Hasil dan Manfaat:
   Melalui integrasi KMLC dengan data mining, organisasi berhasil mengoptimalkan pengelolaan pengetahuan mereka. Mereka melihat peningkatan dalam pengambilan keputusan berbasis data, pengurangan waktu distribusi pengetahuan, dan peningkatan kolaborasi antar departemen. Selain itu, organisasi mencapai efisiensi operasional yang lebih baik dan meningkatkan daya saing mereka di pasar.
5. Tantangan yang Diatasi:
   Meskipun berhasil, implementasi ini tidak lepas dari tantangan, seperti kebutuhan untuk merancang kebijakan keamanan data yang kuat dan mengatasi resistensi organisasional terhadap perubahan. Organisasi dengan cermat mengevaluasi dan menyesuaikan strategi mereka untuk mengatasi tantangan ini selama fase implementasi.

Studi kasus ini menggambarkan bagaimana integrasi KMLC dengan data mining dapat memberikan dampak positif yang signifikan terhadap pengelolaan pengetahuan suatu organisasi, menyoroti keberhasilan, tantangan, dan manfaat yang dapat diperoleh melalui pendekatan ini.

13.6 Kesimpulan

Bab 13 menjelaskan implementasi *Knowledge Management Life Cycle* (KMLC) dengan pendekatan data mining, membawa pembaca melalui serangkaian tahap penting dalam mengelola pengetahuan di lingkungan organisasi. Dengan fokus pada Pengumpulan Pengetahuan, Penyimpanan Pengetahuan, Distribusi Pengetahuan, dan Pemanfaatan Pengetahuan, bab ini memberikan wawasan mendalam tentang bagaimana data mining dapat memperkuat setiap tahap KMLC.

Tantangan yang mungkin muncul selama implementasi, seperti keamanan data, sumber daya, dan resistensi organisasional, diperinci bersama dengan strategi untuk mengatasi tantangan tersebut. Pelatihan karyawan, kebijakan keamanan data yang kuat, dan integrasi sistem menjadi kunci sukses dalam menghadapi kompleksitas integrasi KMLC dengan data mining.

Studi kasus yang bersifat umum memberikan ilustrasi konkret tentang bagaimana integrasi KMLC dengan data mining dapat memberikan manfaat nyata bagi sebuah perusahaan teknologi e-commerce. Peningkatan pengalaman pelanggan, pengambilan keputusan yang lebih baik, dan efisiensi operasional adalah hasil langsung dari penerapan pendekatan ini.

Dengan demikian, bab ini menyajikan kerangka kerja yang kokoh untuk organisasi yang ingin meningkatkan manajemen pengetahuan mereka melalui pemanfaatan *Knowledge Management Life Cycle* yang terintegrasi dengan teknologi data mining. Dalam era di mana data menjadi aset berharga, integrasi ini menjadi kunci untuk mencapai keunggulan kompetitif dan inovasi berkelanjutan.

DAFTAR PUSTAKA

Aggarwal, C.C. (2015) *Data Mining*. Cham: Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-14142-8.

Bernard, A. and Tichkiewitch, S. (2010) *Methods and Tools for Effective Knowledge Life-Cycle-Management*. Springer Berlin Heidelberg. Available at: https://books.google.co.id/books?id=-xW7cQAACAAJ.

Dyché, J. (2000) *E-Data: Turning Data Into Information with Data Warehousing*. Addison-Wesley (Addison-Wesley information technology series). Available at: https://books.google.co.id/books?id=FnbTB9dRJpUC.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) *From Data Mining to Knowledge Discovery in Databases) (© AAAI)*. Available at: www.ffly.com/.

Han, J., Kamber, M. and Pei, J. (2011) *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Sharda, R. *et al.* (2014) *Business Intelligence and Analytics: Systems for Decision Support*. Pearson (Always learning). Available at: https://books.google.co.id/books?id=FLYDnwEACAAJ.

White, G. and chaffey, D. (2010) *Business Information Management: Improving performance using information systems*.

Witten, I.H., Frank, E. and Hall, M.A. (2011a) 'Chapter 1 - What's It All About?', in I.H. Witten, E. Frank, and M.A. Hall (eds) *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Boston: Morgan Kaufmann, pp. 3–38. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-374856-0.00001-8.

Witten, I.H., Frank, E. and Hall, M.A. (2011b) 'Chapter 4 - Algorithms: The Basic Methods', in I.H. Witten, E. Frank, and M.A. Hall (eds) *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Boston: Morgan Kaufmann, pp. 85–145. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-374856-0.00004-3.

BAB 14
MODEL KNOWLEDGE MANAGEMENT

Oleh Yulita Salim

14.1 Pendahuluan

Knowledge Managemen (Manajemen Pengetahuan) merupakan sebuah sistem atau proses yang terarah yang melibatkan proses identifikasi, organisasi, penyimpanan, dan berbagi pengetahuan dalam organisasi. Tujuan Model Manajemen Pengetahuan adalah agar kemampuan organisasi dapat ditingkatkan dalam hal pengelolaan, penyimpanan, dan pemanfaatan pengetahuan agar dapat digunakan secara lebih efektif dan efisien. Pada dasarnya Manajemen Pengetahuan dapat dibagi kedalam dua kelompok yaitu:

1. *Tacit Knowledge*
   Pengetahuan Tacit merupakan jenis pengetahuan yang tidak dapat dikodifikasi atau diartikulasi secara eksplisit. Pengetahuan ini melekat pada diri setiap individu yang terbentuk dari pengalaman hidup, budaya dan interaksi sosial yang terjadi dalam kehidupannya sehari-hari. Dibalik setiap keahlian yang dimiliki seseorang, terdapat pengetahuan tersembunyi yang tidak dapat diungkapkan. Pengetahuan seperti ini bersemayam dalam diri dan menjadi keahlian yang seringkali tidak disadari tetapi menjadi kemampuan dasar yang dapat membantu seseorang dalam menyelesaikan masalah secara efektif.
   Jenis pengetahuan Tacit sangat diperlukan dalam melakukan diagnosa berbagai masalah dalam organisasi. Pengetahuan ini dapat menjadi sumber keunggulan tak berwujud yang sangat penting sehingga perlu dikelola dengan baik melalui berbagai cara seperti pelatihan, mentoring atau kolaborasi.
2. *Explicit Knowledge*

Pengetahuan eksplisit adalah jenis pengetahuan yang dapat diakses, dipelajari, dan disebarluaskan. Pengetahuan ini dapat dikodifikasi dan ditransfer dalam bentuk formal dan sistematis. Berbagai jenis pengetahuan eksplisit tersebar dalam bentuk tertulis seperti laporan, dokumen, buku panduan, ataupun database yang dapat dipahami oleh siapa saja yang membacanya. Selain itu, pengetahuan ini juga dapat tertuang dalam bentuk digital seperti artikel online, video tutorial atau aplikasi pembelajaran.
Seorang ahli yang memiliki pengalaman dan keahlian tertentu dapat menuangkan pengetahuan tacit-nya kedalam bentuk pengetahuan eksplisit sehingga setiap orang dapat mengetahui dan memanfaatkan informasi tersebut untuk digunakan bersama-sama dalam pengembangan organisasi. Pengetahuan eksplisit sangat dibutuhkan untuk membantu organisasi untuk membangun dan membagikan pengetahuan secara sistematis agar menjadi pondasi yang membantu pertumbuhan organisasi. Dengan mengelola pengetahuan eksplisit secara efektif maka organisasi dapat meningkatkan efisiensi, konsistensi dan daya saing pasar.

Selain kedua jenis pengetahuan diatas, perlu diketahui bahwa setiap pengetahuan juga dapat digunakan secara bersama-sama dalam suatu kelompok. Pengetahuan jenis seperti ini dapat disebut *Sharing Knowledge*. Proses penciptaan pengetahuan adalah proses spiral yang merupakan interaksi antara *Tacit Knowledge* dan *Explisit Knowledge*. Praktik berbagi pengetahuan dapat diwujudkan dalam berbagai bentuk seperti diskusi, mentoring, atau percakapan santai yang mengalir dari seseorang kepada orang lain yang memungkinkan adanya transfer keahlian, wawasan, maupun pengalaman setiap individu dalam suatu komunitas atau kelompok.
Berbagi pengetahuan dapat mendorong terciptanya lingkungan yang terbuka dan saling percaya sehingga menimbulkan adanya harmonisasi kolaborasi yang menjadikan sebuah organisasi menjadi lebih kuat, adaptif, dan inovatif.

14.2 Kerangka Manajemen Pengetahuan

Untuk mengelola dan meningkatkan pengetahuan dalam sebuah organisasi diperlukan sebuah framework yang dapat membantu agar konsep pengetahuan dalam sebuah organisasi dapat diidentifikasi, diciptakan, disimpan, dan disebarkan untuk meningkatkan kinerja organisasi.

Manajemen Pengetahuan terlaksana melalui Sistem Pengelolaan Pengetahuan atau *Knowledge Management System* (KMS). Kerangka manajemen pengetahuan terdiri atas: Manusia (*People*), Proses (*Process*), dan Teknologi (*Technology*). Ketiga komponen saling terkait dan harus diintegrasikan secara efektif untuk mencapai manajemen pengetahuan yang sukses.

14.2.1 Manusia (*People*)

Manusia merupakan faktor paling penting dalam manajemen pengetahuan. Mereka adalah pemilik, pencipta, dan pengguna pengetahuan dalam organisasi. Aspek manusia dalam manajemen pengetahuan meliputi:

1. Budaya organisasi yang mendukung berbagi pengetahuan, kolaborasi, dan pembelajaran berkelanjutan.
2. Kompetensi dan keterampilan individu dalam mengelola dan memanfaatkan pengetahuan.
3. Peran dan tanggung jawab yang terkait dengan manajemen pengetahuan, seperti manajer pengetahuan, subjek matter expert, dan tim manajemen pengetahuan.
4. Insentif dan penghargaan untuk mendorong kontribusi dan berbagi pengetahuan.

14.2.2 Proses (*Processes*)

Proses manajemen pengetahuan mencakup serangkaian aktivitas yang sistematis untuk mengelola siklus hidup pengetahuan dalam organisasi, meliputi:

1. Identifikasi dan akuisisi pengetahuan dari sumber internal dan eksternal.
2. Penciptaan pengetahuan baru melalui penelitian, pengembangan, dan inovasi.

3. Penyimpanan dan organisasi pengetahuan dalam repositori atau sistem manajemen pengetahuan.
4. Berbagi dan diseminasi pengetahuan melalui berbagai mekanisme, seperti komunitas praktik, pelatihan, dan kolaborasi.
5. Pemanfaatan pengetahuan untuk pengambilan keputusan, pemecahan masalah, dan inovasi produk atau layanan.
6. Evaluasi dan perbaikan berkelanjutan untuk memastikan efektivitas manajemen pengetahuan.

14.2.3 Teknologi (*Technology*)

Teknologi berperan sebagai enabler atau pendukung dalam manajemen pengetahuan. Teknologi memfasilitasi proses pengelolaan pengetahuan dan memberikan infrastruktur yang diperlukan, seperti:

1. Sistem manajemen dokumen dan repositori pengetahuan untuk menyimpan dan mengorganisir pengetahuan eksplisit.
2. Portal pengetahuan dan mesin pencarian untuk mengakses dan menemukan pengetahuan dengan mudah.
3. Alat kolaborasi dan komunikasi, seperti forum diskusi, chat, video konferensi, untuk berbagi pengetahuan dan berinteraksi.
4. Teknologi pemetaan pengetahuan dan taksonomi untuk mengorganisir dan mengkategorikan pengetahuan.
5. Alat analitik dan pelaporan untuk mengukur dan mengevaluasi kinerja manajemen pengetahuan.

Ketiga komponen ini harus diintegrasikan secara holistik dalam kerangka manajemen pengetahuan. Manusia adalah inti dari manajemen pengetahuan, sementara proses menyediakan metodologi sistematis, dan teknologi menjadi pendukung yang memfasilitasi proses tersebut. Keseimbangan dan harmoni antara ketiga komponen ini akan menghasilkan manajemen pengetahuan yang efektif dan memberikan nilai bagi organisasi.
Model yang digambarkan oleh Nonaka dan Takeuchi (1995) tentang proses konvensi pengetahuan yaitu: sosialisasi, eksternalisasi,

kombinasi dan internalisasi. Setiap proses konvensi yang terjadi dapat melibatkan satu bentuk pengetahuan ke dalam bentuk pengetahuan yang lain (*tacit* atau *explicit*).

14.3 Klasifikasi Manajemen Pengetahuan

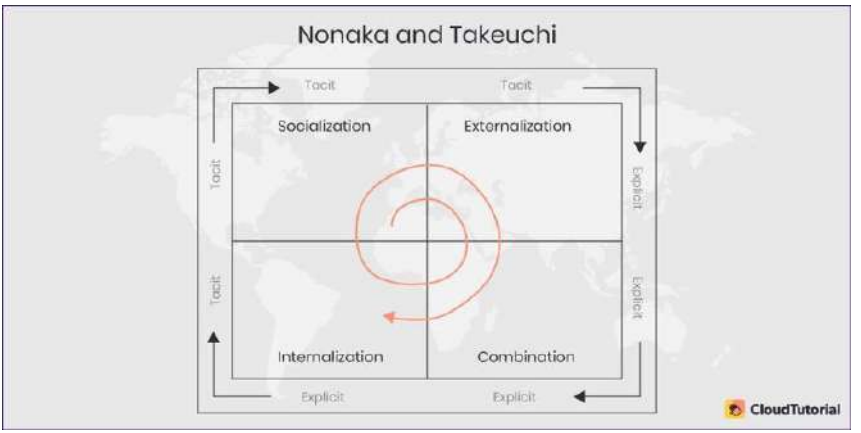
Manajemen Pengetahuan diklasifikasikan dalam beberapa tipe sebagai berikut:

1. Mengumpulkan dan menggunakan ulang pengetahuan terstruktur.
2. Mengumpulkan dan berbagi pelajaran yang sudah di pelajari (lessons learned) berdasarkan praktek yang sudah dilakukan.
3. Mengidentifikasi sumber dan jaringan kepakaran.
4. Membuat struktur dan memetakan pengetahuan yang diperlukan untuk meningkatkan performansi.
5. Mengukur dan mengelola nilai ekonomis dari pengetahuan
6. Menyusun dan menyebarkan pengetahuan dari sumber eksternal

Beberapa model *knowledge management* yang populer:

1. Model SECI (*Socialization, Externalization, Combination, Internalization*)
   Model SECI dikembangkan oleh Ikujiro Nonaka dan Hirotaka Takeuchi. Model SECI menggambarkan proses konversi pengetahuan tacit (tersembunyi) menjadi pengetahuan eksplisit (tertulis) dan sebaliknya. Proses tersebut terdiri dari:
   a. Sosialisasi: Pertukaran pengetahuan tacit antar individu melalui interaksi langsung, observasi, dan praktik bersama.
   b. Eksternalisasi: Mengonversikan pengetahuan tacit menjadi pengetahuan eksplisit dalam bentuk dokumen, laporan, atau database.
   c. Kombinasi: Menggabungkan dan mengintegrasikan berbagai sumber pengetahuan eksplisit untuk menciptakan pengetahuan baru.

d. Internalisasi: Mengonversi pengetahuan eksplisit menjadi pengetahuan tacit individu melalui pembelajaran, simulasi, atau praktik langsung.



Gambar 14.1. Model Nonaka dan Takeuchi

2. Model Bukowitz dan Williams
   Model ini berfokus pada proses penciptaan, penyebaran, dan pemanfaatan pengetahuan dalam organisasi. Terdiri dari dua komponen utama:
   a. Proses Taktis: Meliputi aktivitas mendapatkan, menggunakan, mempelajari, berbagi, mengukur, dan memelihara pengetahuan.
   b. Proses Strategis: Mencakup strategi, kepemimpinan, dan aspek budaya yang mendukung manajemen pengetahuan secara berkelanjutan.
3. Model Wiig
   Model ini mengkategorikan pengetahuan menjadi pengetahuan faktual, konseptual, prosedural, dan metakognitif. Fokus utamanya adalah pada pemeliharaan dan pemanfaatan pengetahuan melalui proses penciptaan, manifestasi, penggunaan, dan transfer pengetahuan.

Gambar 14.2. Model WIIGS KM

Model Wiig merupakan salah satu model manajemen pengetahuan yang dikembangkan oleh Karl M. Wiig, seorang pakar di bidang manajemen pengetahuan. Model ini berfokus pada pemeliharaan dan pemanfaatan pengetahuan melalui serangkaian proses yang berkesinambungan.

Dalam Model Wiig, pengetahuan dikategorikan menjadi empat jenis:

a. Pengetahuan Faktual (*Factual Knowledge*)
   Mencakup pengetahuan tentang fakta-fakta, data, dan informasi yang objektif.
b. Pengetahuan Konseptual (*Conceptual Knowledge*)
   Mencakup pengetahuan tentang konsep, teori, model, dan kerangka kerja.
c. Pengetahuan Prosedural (*Procedural Knowledge*)
   Mencakup pengetahuan tentang cara melakukan sesuatu, seperti prosedur, proses, teknik, dan metode.
d. Pengetahuan Metakognitif (*Metacognitive Knowledge*)
   Mencakup pengetahuan tentang pemahaman diri sendiri, seperti kekuatan, kelemahan, preferensi, dan gaya belajar.

Model Wiig mengusulkan empat proses inti, yaitu:

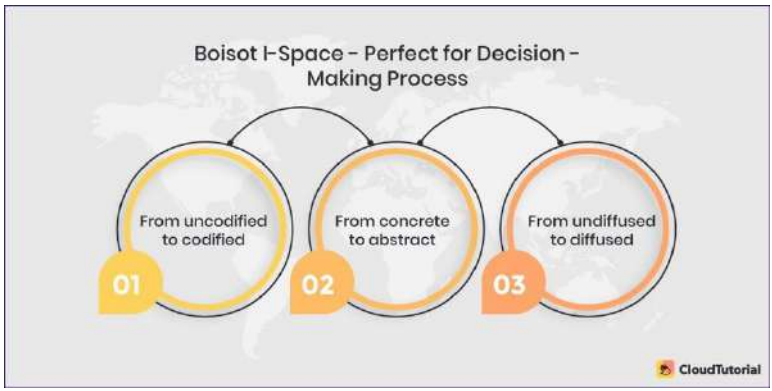
a. Penciptaan Pengetahuan (*Knowledge Creation*)
   Mencakup aktivitas seperti penelitian dan pengembangan, inovasi, pembelajaran dari pengalaman, dan akuisisi pengetahuan dari sumber eksternal.
b. Manifestasi Pengetahuan (*Knowledge Manifestation*)
   Mencakup proses mentransformasikan pengetahuan tacit menjadi pengetahuan eksplisit, seperti mendokumentasikan, mengkodifikasikan, dan menyimpan pengetahuan dalam repositori.
c. Penggunaan Pengetahuan (*Knowledge Use*)
   Mencakup pemanfaatan pengetahuan dalam pengambilan keputusan, pemecahan masalah, perencanaan, dan implementasi aktivitas organisasi.
d. Transfer Pengetahuan (*Knowledge Transfer*)
   Mencakup proses berbagi dan menyebarkan pengetahuan di seluruh organisasi melalui pelatihan, mentoring, komunitas praktik, dan sistem komunikasi.

Model Wiig menekankan pentingnya mengelola keempat jenis pengetahuan tersebut melalui empat proses inti secara berkesinambungan. Hal ini bertujuan untuk memastikan bahwa pengetahuan dimanfaatkan secara optimal dalam mendukung tujuan dan strategi organisasi.

Salah satu keunggulan Model Wiig adalah penekanannya pada pengetahuan metakognitif, yang sering diabaikan dalam model manajemen pengetahuan lainnya. Pengetahuan ini membantu individu dan organisasi memahami kekuatan, kelemahan, dan preferensi mereka dalam mengelola pengetahuan.

Meskipun Model Wiig cukup komprehensif, ia juga memiliki beberapa keterbatasan, seperti tidak secara eksplisit menyoroti aspek budaya dan kepemimpinan dalam manajemen pengetahuan. Namun, secara keseluruhan, Model Wiig memberikan kerangka kerja yang berguna untuk memahami dan mengelola pengetahuan dalam organisasi.

4. Model Boisot
   Model Boisot menganalisis pengetahuan berdasarkan dua dimensi: kodifikasi (tingkat pengetahuan yang dapat dikodifikasikan) dan difusi (tingkat penyebaran pengetahuan). Model ini memberikan panduan untuk mengelola aliran pengetahuan dalam organisasi, termasuk kapan pengetahuan harus dikodifikasikan atau disebarkan.



Gambar 14.3. Model Boisot I-Space

5. Model Standar Manajemen Pengetahuan (KM Standards Model)
   Model ini dikembangkan oleh British Standards Institution (BSI) dan menyediakan kerangka kerja yang komprehensif untuk implementasi manajemen pengetahuan. Terdiri dari beberapa komponen utama, seperti kepemimpinan, budaya, proses, teknologi, dan pengukuran kinerja.

Pemilihan model manajemen pengetahuan yang tepat bergantung pada kebutuhan, budaya, dan tujuan organisasi. Model-model ini memberikan panduan dan kerangka kerja untuk mengelola aset pengetahuan secara efektif, memfasilitasi penciptaan, penyebaran, dan pemanfaatan pengetahuan dalam organisasi.

DAFTAR PUSTAKA

Sri Suwarsi. 2012. Model Implementasi Knowledge Management Dalam Menciptakan Inovasi Pada Industri Kreatif Di Kota Bandung

Budiastuti, Dyah. *Model Knowledge Management di Perguruan Tinggi*. *Binus Business Review*, vol. 3, no. 1, 2012, pp. 52-60.

K. Dalkir. Knowledge Management in Theory and Practice

R. Nurcahyo. D. I. Sensuse. 2019. Knowledge Management System Dengan Seci Model Sebagai Media Knowledge Sharing Pada Proses Pengembangan Perangkat Lunak Di Pusat Komputer Universitas Tarumanagara. Jurnal Teknologi Terpadu. Vol. 5, No. 2.

https://www.thecloudtutorial.com/knowledge-management-models/

BAB 15
METODE PENGUKURAN KINERJA KNOWLEDGE MANAGEMENT DALAM ORGANISASI

Oleh A. Amirul Asnan Cirua

15.1 Knowledge Management

Knowledge Management (KM) berkembang seiring dengan kemajuan komputer, jaringan, dan sistem manajemen data. Berbagi dan berkolaborasi di antara ribuan orang yang tersebar di seluruh dunia bergantung pada teknologi koneksi dan penyimpanan konten yang terorganisir, dan banyak proyek pengetahuan berfokus pada pembangunan sistem untuk menghubungkan orang-orang dan menangkap pengetahuan. Pada dasarnya, penerapan *Knowledge Management* (KM) adalah strategi untuk mencapai visi dan misi suatu organisasi. Namun, terkadang penerapan KM tidak berjalan dengan mudah atau sukses. Salah satu faktor yang dapat menyebabkan organisasi gagal menerapkan KM adalah kurangnya kesiapan organisasi dari segi manusia, proses, dan teknologi. (Ibrahim *et al.*, 2021). Penelitian Penerapan strategi KM yang baik secara efektif dan menjadi perusahaan berbasis pengetahuan dipandang sebagai syarat wajib bagi keberhasilan organisasi saat mereka memasuki era ekonomi pengetahuan. Namun, metrik standar diperlukan untuk mengukur pengetahuan dan untuk sepenuhnya meyakinkan manajemen dan pemangku kepentingan mengenai nilai inisiatif KM (Bose, 2004).

Tujuan dalam mengelola dan memanfaatkan pengetahuan perusahaan adalah untuk memaksimalkan keuntungan bagi organisasi. Ini menunjukkan kemampuan untuk mengukur investasi pokok dan hasil investasi tersebut secara berkala. Namun, mengukur manfaat bisnis KM saat ini cukup sulit. Pengukuran adalah aspek KM dan upaya transfer praktik terbaik yang paling

sedikit dikembangkan karena sulitnya mengukur sesuatu yang tidak dapat dilihat, misalnya pengetahuan. Sistem KM harus menunjukkan nilainya. Tanpa keberhasilan yang terukur, antusiasme dan dukungan terhadap KM tidak mungkin berlanjut. Secara khusus, tanpa keberhasilan yang terukur, manajer tidak akan mampu mengetahui cara kerja sistem KM organisasi, di mana hambatannya, dan bagaimana sistem dapat ditingkatkan. Oleh karena itu, manajer harus mampu mengukur dampak upaya KM terhadap kinerja organisasinya (Claire McInerney, 2002).

Dalam melakukan pengukuran maka suatu metrik yang terstandarisasi diperlukan untuk mengukurnya. Pengembangan metrik KM telah dimulai dalam beberapa tahun terakhir dan metrik ini diterapkan oleh beberapa organisasi, namun penelitian lebih lanjut diperlukan untuk mendefinisikan ukuran-ukuran ini dengan lebih baik dan menjadikannya universal (Bose, 2004).

Knowledge Management System (KMS) memfasilitasi KM dengan menentukan alur pengetahuan dari orang yang tahu ke orang yang perlu tahu di seluruh organisasi sambil pengetahuan berkembang dan berkembang selama proses tersebut. KMS terdiri dari berbagai alat dan teknologi.

15.1.1 Prinsi Dasar Knowledge Management

Prinsip dasar dalam manajemen pengetahuan adalah sebagai berikut :

1. **Akuisisi Pengetahuan,** Ini adalah proses memperoleh atau mengembangkan aset intelektual, seperti pemahaman pribadi, keahlian, pengalaman, dan hubungan antar data. Proses ini melibatkan pendataan menyimpan ke dalam basis data pengetahuan organisasi atau gudang pengetahuan..
2. **Berbagi Pengetahuan,** yaitu proses memberikan pengetahuan kepada orang-orang yang membutuhkan dalam organisasi penggunanya. Proses ini dapat terjadi secara tradisional melalui kolokium dan diskusi atau lebih modern melalui teknologi.
3. **Memanfaatkan Pengetahuan,** yaitu bagaimana organisasi menggunakan pengetahuan. Termasuk di dalamnya adalah penerapan dalam pembentukan panduan-panduan kerja

yang didasarkan pada pengetahuan dan pengalaman masa lalu. Aktivitas pengembangan dan pemutakhiran lebih lanjut dari pengetahuan yang didapatkan. (Ministry of National Development Planning/Bappenas, 2022).

15.1.2 Siklus Proses Knowledge Management

Pada proses KM, terdapat 6 siklus yang berperan dalam pembentukannya, yaitu sebagai berikut :

1. ***Create Knowledge,*** Pengetahuan tersebut berasal dari pengalaman dan keterampilan karyawan. Mengembangkan pengetahuan baru atau menemukan cara baru untuk melakukan sesuatu disebut pengetahuan. Jika pengetahuan tersebut tidak berada dalam organisasi, pengetahuan eksternal berpeluang dibawa masuk, seperti terjadinya transfer teknologi yang terjadi dari laboratorium penelitian ke dalam organisasi bisnis.
2. ***Capture Knowledge,*** Pengetahuan yang diciptakan perlu disimpan dalam bentuk mentahnya ke dalam sebuah database. Kebanyakan organisasi menggunakan berbagai jenis repositori pengetahuan untuk menangkap pengetahuan (Bose, 2004).
3. ***Refine Knowledge,*** Pengetahuan baru harus ditempatkan dalam konteks yang dapat ditindaklanjuti. Wawasan manusia atau pengetahuan tacit diterima dan disempurnakan bersama dengan pengetahuan eksplisit (Herschel, And and Steiger, 2001).
4. ***Store Knowledge****,* Kodifikasi pengetahuan diam-diam dan eksplisit membantu membuat pengetahuan tersebut dapat dimengerti dan dapat digunakan dikemudian hari.
5. ***Manage Knowledge,*** pengetahuan harus selalu diperbarui. Hal ini ditinjau untuk memverifikasi bahwa itu relevan dan akurat. Sebagian besar perusahaan Fortune mempunyai departemen yang terdefinisi dengan baik yang benar-benar menjaga agar pengetahuan tetap terkini.
6. ***Disseminate Knowledge,*** Pengetahuan harus selalu tersedia bagi siapa pun yang membutuhkannya baik kapan pun dan dimana pun. Teknologi baru seperti *groupware*,

Internet/intranet dan teknologi DSS lainnya membantu penyebaran pengetahuan.

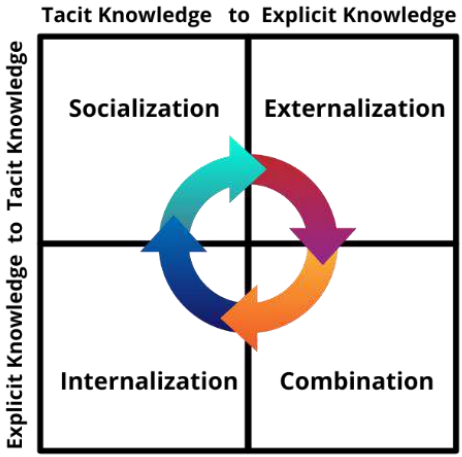
15.1.3 Mekanisme *Knowledge Management*

Manajemen pengetahuan didefinisikan sebagai melakukan apa yang diperlukan untuk mendapatkan hasil maksimal dari sumber daya pengetahuan. Secara umum, manajemen pengetahuan berfokus pada pengorganisasian dan penyediaan pengetahuan yang diperlukan, dimanapun dan kapanpun dibutuhkan (Pelealu, 2022).

Dalam hubungan sosial dengan orang lain, pengetahuan juga muncul dan berkembang. Pengetahuan eksplisit dan pengetahuan tacit adalah dua jenis pengetahuan yang ada pada orang. Pengetahuan eksplisit adalah jenis pengetahuan yang telah disimpan pada media penyimpanan.

1. ***Tacit Knowledge*** merupakan pengetahuan yang dimiliki seseorang dan sangat sulit untuk diformalisasikan. Tacit knowledge ini terbukti dalam tindakan dan pengalaman seseorang, termasuk dalam nilai emosional dan idealismenya.
2. ***Explicit Knowledge*** dapat berupa kata-kata, yang kemudian dapat dikalkulasi dan dibagikan dalam bentuk data, spesifikasi produk, atau prinsip universal. Tidak sama dengan pengetahuan tersirat, yang tidak dapat dibagi menjadi jenis pengetahuan seperti pengalaman, keterampilan, dan pemahaman. Pengetahuan eksplisit ini selalu siap untuk diberikan secara formal dan sistematik kepada orang lain.

Teori yang mengusung mengenai proses atau mekanisme manajemen pengetahuan yang berisi empat proses secara umum, yaitu:



Gambar 15.1. Proses Manajemen Pengetahuan

1. ***Socialization*,** yaitu kombinasi pengetahuan tacit seseorang yang biasanya diperoleh melalui kegiatan bersama. seperti proses magang karyawan baru, transfer ide atau foto.
2. ***Externalization*,** yaitu mengubah pengetahuan yang tidak jelas menjadi yang jelas sehingga orang lain dapat lebih mudah memahaminya.
3. ***Internalization*,** yaitu mengubah pengetahuan/informasi yang jelas menjadi pengetahuan/informasi yang tidak jelas bagi seseorang.
4. ***Combination*,** yaitu pengetahuan yang jelas yang baru ditemukan melalui kombinasi. Proses ini dapat mencakup komunikasi, integrasi, dan sistemisasi berbagai sumber pengetahuan.

15.1.4 Indikator *Knowledge Management System*

Menurut (Pelealu, 2022) indikator *knowledge management* adalah :

Tabel 15.1. Tingkat Kesiapan *Knowledge Manajement*

Variable	Dimensi	Indikator
Knowledge Management	Persepsi Manajemen Pengetahuan	1. Pemahaman tentang manajemen pengetahuan 2. Komitmen terhadap manajemen pengetahuan 3. Persepsi terhadap pekerjaan manajemen pengetahuan
	Hubungan antara inovasi manajemen pengetahuan	1. Implementasi rencana manajemen pengetahuan 2. Kematangan dan pengendalian rencana pengelolaan pengetahuan
	Organisasi Pengetahuan Strategis	1. Menyimpan pengetahuan 2. Sosialisasi pengetahuan 3. Penggunaan teknologi informasi dan komunikasi (TIK) 4. Organisasi pengetahuan strategis 5. Kualitas informasi.

15.1.5 Tingkat Kesiapan *Knowledge Management*

Menurut Rao (2005) dalam (Herkanawati, 2015) telah mengidentifikasi tingkat kesiapan menjadi 5 (lima) level seperti terlihat pada tabel 15.2 di bawah ini:

Tabel 15.2. Tingkat Kesiapan *Knowledge Manajement*

Level	Nama Level	Konversi Nilai (%)	Karakteristik
1	*Not Ready*	0 – 20	Belum adanya pemahaman mengenai KM *(Knowledge Management)* Belum adanya pemahaman mengenai visi, misi dari KM Tidak menggambarkan fenomena atau permasalahan KM
2	*Preliminary (exploring knowledge management)*	21 - 40	Organisasi sudah mengenal pentingnya kegiatan KM Proses dalam organisasi sudah menggambarkan kegiatan KM Sudah terdapat individu yang menggalakkan *Knowledge Management System*
3	*Ready (accepted)*	41 – 60	Sudah stabil dan individu dalam organisasi sudah mempraktekkan aktifitas yang efektif untuk mendukung KM Kegiatan KM sudah dilakukan setiap waktu di setiap kegiatan pekerjaan Kegiatan KM sudah dapat ditemukan pada setiap individu Sudah ada sistem pendokumentasian

Level	Nama Level	Konversi Nilai (%)	Karakteristik
4	*Receptive (advocating and measuring)*	61 – 80	Sudah adanya efisiensi dari KM Kegiatan-kegiatan yang ada pada level dilanjutkan dan sudah dihasilkan suatu standard dan aturan
5	*Optimal (institutionalized knowledge management)*	81 - 100	Organisasi telah memiliki kemampuan untuk beradaptasi dan fleksibel terhadap syarat-syarat yang ditentukan untuk mencapai *KM Readiness*

Zaidah (2010) dalam (Herkanawati, 2015) membuat formula untuk mengkonversi nilai tingkat kesiapan. Zaidiah (2010) menyatakan bahwa untuk menentukan tingkat kesiapan organisasi untuk menerapkan KM, persentase kesiapan manajemen pengetahuan dapat dihitung dengan membagi jumlah angka atau skor setiap indikator faktor keberhasilan manajemen (KMCSF) dibagi total keseluruhan bobot maksimal, atau menggunakan formula 1.

$$P : \frac{Sn}{Sm}\ x\ 100\% \qquad (1)$$

Dimana :
P : persentase tingkat kesiapan.
Sn : jumlah skor x bobot.
Sm : total bobot x skor maksimal.

15.2 Pengukuran Kinerja

Salah satu pandangan menyatakan bahwa untuk menjamin kinerja organisasi secara keseluruhan, organisasi perlu mengelola dan mengukur sumber daya teknologi, manusia, dan keuangannya. Mengukur kinerja ini memvalidasi apakah upaya tersebut bermanfaat atau tidak, Sudut pandang ini menyatakan bahwa

proyek KM mempunyai hasil yang berwujud dan tidak berwujud yang pertama cocok untuk pengukuran yang dapat diukur, sedangkan yang terakhir menghasilkan hasil yang berharga yang tidak dapat diukur dengan mudah sehingga membuatnya sangat sulit untuk diukur (Tseng and Lee, 2009). Oleh karena itu, sangat diperlukan adanya apa yang disebut indikator kinerja, yang belum tentu menunjukkan peningkatan kinerja organisasi secara keseluruhan, namun menunjukkan apakah aktivitas pengetahuan meningkat atau tidak. Hal ini menentukan status proyek dan apakah proyek tersebut telah mencapai tingkat kepuasan atau memerlukan tindakan perbaikan.

15.2.1 Pengukuran Kinerja KM berbasis Strategi

Menurut APQC *(American Productivity and Quality Center)* mengidentifikasi enam strategi KM yang muncul dalam studi organisasi yang dianggap mencerminkan sifat dan kekuatan berbeda dari organisasi yang terlibat (Shannak, 2009):

1. Strategi Pengetahuan sebagai Strategi Bisnis yaitu Pendekatan KM yang komprehensif dan mencakup seluruh perusahaan, yang sering kali menjadikan pengetahuan sebagai produknya.
2. Strategi Pengelolaan Aset Intelektual yaitu strategi yang berfokus pada aset yang sudah ada dalam perusahaan yang dapat dieksploitasi secara lebih penuh atau ditingkatkan.
3. Strategi Tanggung Jawab Aset Pengetahuan Pribadi yaitu strategi yang mendorong dan mendukung setiap karyawan untuk mengembangkan keterampilan dan pengetahuan mereka serta berbagi pengetahuan satu sama lain.
4. Strategi Penciptaan Pengetahuan yaitu strategi yang menekankan inovasi dan penciptaan pengetahuan baru melalui R&D. Diadopsi oleh para pemimpin pasar yang menentukan arah masa depan sektor mereka.
5. Strategi Transfer Pengetahuan yaitu strategi yang mentransfer pengetahuan dan praktik terbaik untuk meningkatkan kualitas dan efisiensi operasional.

6. Strategi Pengetahuan yang Berfokus pada Pelanggan dengan tujuan untuk memahami pelanggan dan kebutuhan mereka sehingga memberikan apa yang mereka inginkan.

15.2.2 Indikator Pengukuran Kinerja

Secara umum, menentukan indikator mana yang paling berguna ketika mengukur kinerja suatu proyek sebagian besar berkaitan dengan penilaian. Beberapa indikator kinerja yang digunakan oleh perusahaan-perusahaan besar internasional ternama yang telah mengembangkan sistem KM secara internal (Shannak, 2009).

1. **Ericsson,** indikator kinerja yang diidentifikasi terdapat enam indikator sebagai berikut :
   a. Jumlah praktik terbaik yang teridentifikasi.
   b. Jumlah kontribusi.
   c. Jumlah kontribusi dalam database penggunaan kembali terkait dengan jumlah total proyek yang diselesaikan di unit pasar.
   d. Jumlah kontribusi yang terbukti menghasilkan bisnis baru/berulang.
   e. Jumlah kontribusi yang dinilai dapat digunakan kembali.
   f. Jumlah individu dalam masyarakat dan aktivitas masing-masing individu.
2. **Xerox,** indikator ini memiliki tujuan utama dari sistem KM adalah untuk mempertahankan keunggulan kompetitif dengan menggunakan pengetahuan yang dikumpulkan dari seluruh karyawan, arsip paten dan proses serta semua dokumen yang disimpan dalam berbagai format di semua lokasi. Berikut indikator kinerja yang teridentifikasi:
   a. Akses terhadap pengetahuan.
   b. Susunan ilmu pengetahuan.
   c. Jumlah panggilan ke administrasi.
   d. Dukungan dari manajemen.
   e. Kegunaan.

3. **Siemens,** para karyawan harus berbagi pengetahuan mereka melalui komunikasi informal tatap muka. Berikut indikator kinerja yang teridentifikasi:
   a. Keberhasilan pelanggan.
   b. Kepuasan karyawan.
   c. Indikator insentif.
   d. Indeks inovasi.
   e. Jumlah entri dalam database.
   f. Jumlah berbagi pengetahuan dalam setahun.
   g. Jumlah pesanan.
   h. Kualitas informasi.
   i. Penggunaan kembali pengetahuan.
   j. Kegunaan ilmunya.

DAFTAR PUSTAKA

Bose, R. (2004) 'Knowledge management metrics', *Industrial Management and Data Systems*, 104(6), pp. 457–468. doi: 10.1108/02635570410543771.

Claire McInerney (2002) 'Bulletin of the American Society for Information Science and Technology 1999', pp. 1–6.

Herkanawati, D. (2015) 'Measurement Readiness Level of Knowledge Management Balitbang HR Ministry of Communication and Informatics', *JURNAL IPTEKKOM : Journal of Science & Information Technology*, 17(2), p. 189.

Herschel, R. T., And, H. N. and Steiger, D. (2001) 'Tacit to explicit knowledge conversion', *Cognitive Processing*, 18(4), pp. 461–477. doi: 10.1007/s10339-017-0825-6.

Ibrahim, S. S. *et al.* (2021) 'Pengukuran Kesiapan Penerapan Knowledge Management di Institusi Pendidikan Tinggi', *Jambura Journal of Informatics*, 3(2), pp. 87–96. doi: 10.37905/jji.v3i2.11797.

Ministry of National Development Planning/Bappenas (2022) 'Guideline for Knowledge Management of the Ministry of National Development Planning/Bappenas', p. 51.

Pelealu, D. R. (2022) 'The Effect of Knowledge Management System and Knowledge Sharing on Employee Performance and Loyalty', *Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE)*, 5(1), pp. 371–389. doi: 10.31538/iijse.v5i1.2162.

Shannak, R. O. (2009) 'Measuring knowledge management performance', *European Journal of Scientific Research*, 35(2), pp. 242–253.

Tseng, Y. F. and Lee, T. Z. (2009) 'Comparing appropriate decision support of human resource practices on organizational performance with DEA/AHP model', *Expert Systems with Applications*, 36(3 PART 2), pp. 6548–6558. doi: 10.1016/j.eswa.2008.07.066.

BIODATA PENULIS

Aisyah Mutia Dawis, S.Kom., M.Kom.
Dosen Program Studi Sarjana Sistem dan Teknologi Informasi
Fakultas Sains dan Teknologi Universitas 'Aisyiyah Surakarta

Penulis lahir di Kota Surakarta. Ia Lulus S1 pada tahun 2013 hingga mendapat gelar Sarjana Komputer di Universitas Muhammadiyah Surakarta serta mendapatkan penghargaan sebagai lulusan terbaik Se Fakultas Komunikasi dan Informatika, kemudian ia melanjutkan studi S2, lulus pada tahun 2019 hingga mendapat gelar Magister Komputer di Universitas Amikom Yogyakarta. Saat ini ia tercatat sebagai dosen tetap di Program Studi Sarjana Sistem dan Teknologi Informasi di Universitas 'Aisyiyah Surakarta. Selain mengajar ia aktif dalam kegiatan tridharma lainnya diantaranya ialah penelitian dan pengabdian kepada masyarakat. Kegiatan penelitian internal dan eksternal pernah dilakukannya. Beberapa penelitian berjudul : *" Virtual Reality Tour Sebagai Media Informasi Pengenalan Gedung Kampus 2 Universitas' Aisyiyah Surakarta"* dan *"Utilization of Virtual Reality Technology in Knowing the Symptoms of Acrophobia and Nyctophobia"*. Ia sering menjadi *invited speaker* diantaranya menjadi tenaga pengajar BPPTIK Kementerian Kominfo. Serta sebagai bentuk pengabdian kepada masyarakat, ia pun pernah terlibat aktif sebagai trainer bimbingan teknis bantuan TIK SMP, SMA, Mahasiswa, Tenaga Pendidik, Guru dan Dosen dari tahun 2010 hingga saat ini.

BIODATA PENULIS

Dr. Irmawati, S.Kom., MMSI.
Dosen Program Studi Sistem Informasi
Fakultas Teknik dan Informatika
Universitas Multimedia Nusantara

Penulis adalah dosen tetap pada Program Studi Sistem Informasi Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Menyelesaikan pendidikan S1 dan S2 pada Program Studi Sistem Informasi Universitas Gunadarma dan S3 pada departemen Teknik Elektro Universitas Indonesia. Penulis menekuni bidang keilmuan meliputi *Artificial Intelligence*, *Image Processing*, *Machine Learning*, dan *Deep Learning*.

BIODATA PENULIS

Hermila A. S.SI., M.Pd.
Dosen Program Studi Pendidikan Teknologi Informasi
Fakultas Teknik Universitas Negeri Gorontalo

Penulis lahir di Polewali Mandar tanggal 25 Juli 1993. Penulis adalah dosen tetap pada Program Studi Pendidikan Teknologi Informasi Fakultas Teknik, Universitas Negeri Gorontalo. Menyelesaikan pendidikan S1 pada Jurusan Sistem Informasi STMIK Profesional Makassar dan melanjutkan S2 pada Jurusan Pendidikan Teknologi Kejuruan Kekhususan Pendidikan Teknik Informatika dan Komputer Universita Negeri Makassar. Penulis sekarang mulai menekuni bidang Menulis.

BIODATA PENULIS

Nahya Nur, S.T., M.Kom.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Polewali Mandar tanggal 05 November 1991. Penulis merupakan dosen tetap pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Program Studi Teknik Informatika di Universitas Hasanuddin dan melanjutkan S2 pada program studi Informatika di Institut Teknologi Sepuluh Nopember dengan bidang konsentrasi komputasi cerdas dan visi. Penulis menekuni bidang Menulis.

BIODATA PENULIS

Dr. Eng. Sulfayanti, S.Si., M.T.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Mamuju tanggal 17 Maret 1989. Penulis adalah dosen tetap pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Matematika dan melanjutkan S2 pada Jurusan Teknik Elektro peminatan Teknik Informatika Universitas Hasanuddin. Penulis menyelesaikan studi S3 pada *Graduate School of Computer Science and Systems Engineering*, Okayama Prefectural University.

BIODATA PENULIS

Siti Aulia Rachmini, S.T., M.T.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Jakarta tanggal 06 Juli 1982. Penulis adalah dosen tetap pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika Universitas Trisakti Jakarta dan melanjutkan S2 pada Jurusan Teknik Informatika Universitas Hasanuddin Makassar. Penulis menekuni bidang Menulis.

BIODATA PENULIS

Farid Wajidi, S.Kom., M.T.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Lapeo, Polewali Mandar tanggal 18 April 1989. Penulis adalah dosen tetap pada Program Studi Informatika, Fakultas Teknik Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika di STMIK Dipanegara dan melanjutkan S2 pada Jurusan Teknik Informatika di Universitas Hasanuddin.

BIODATA PENULIS

Mutmainnah Muchtar, S.T., M.Kom.
Dosen Program Studi Ilmu Komputer
Fakultas Teknologi Informasi, Universitas Sembilanbelas November Kolaka

Penulis lahir di Kendari, pada tanggal 12 Januari 1991 dan merupakan anak ke-dua dari empat bersaudara. Penulis merupakan dosen di Program Studi Ilmu Komputer Fakultas Teknologi Informasi, Universitas Sembilanbelas November Kolaka, Sulawesi Tenggara. Penulis menyelesaikan pendidikan program Sarjana (S1) di Universitas Halu Oleo Kendari Program Studi Teknik Informatika pada tahun 2012 dan menyelesaikan program Pasca Sarjana (S2) di Institut Teknologi Sepuluh Nopember Surabaya Program Studi Teknik Informatika pada tahun 2015. Hingga saat ini penulis telah menerbitkan beberapa artikel ilmiah nasional maupun internasional, dengan fokus bidang penelitian yaitu *Digital Image Processing*, *Computer Vision*, Data Mining dan Kecerdasan Buatan.

BIODATA PENULIS

Wawan Firgiawan, S.T., M.Kom.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Majene 04 November 1996. Penulis adalah dosen pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Awal karir akademis saya dimulai Ketika menyelesaikan gelar sarjana (S1) di Jurusan Teknik Informatika Universitas Sulawesi Barat, Majene. Kemudian melanjutkan pendidikan magister (S2) di bidang yang sama di Universitas Hasanuddin, Makassar. Dalam perjalanan pendidikannya, saya menemukan minat yang mendalam dalam pada bidang penulisan dan riset seperti *Artificial Intelligence, Evolutionary Algorithm, Decision Support System, Algorithm and Computation*.

Kontak: Alamat Jl. Poros Majene-Mamuju, Desa Tallu Banua, Kecamatan Sendana, kabupaten Majene, Sulawesi Barat. Untuk kontak lebih lanjut, Anda bisa menghubungi saya melalui email wawanfirgiawan@unsulbar.ac.id. Saya juga memiliki akun media sosial Instagram dan Facebook dengan username @wawanfirgiawan.

BIODATA PENULIS

Nurhikma Arifin, S.Kom., M.T.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Majene tanggal 25 April 1993. Penulis adalah dosen tetap pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika di Universitas Islam Negeri Alauddin Makassar dan melanjutkan S2 pada Jurusan Teknik Informatika di Universitas Hasanuddin. Selama masa studi tersebut, penulis mendalami bidang-bidang khusus dalam dunia informatika, mengejar pemahaman mendalam tentang aspek-aspek teknis dan konseptual yang mendukung perkembangan ilmu pengetahuan dan teknologi.

Selain menekuni bidang akademis, penulis juga memiliki minat dan keahlian dalam menulis. Kemampuan menulis tidak hanya terfokus pada karya ilmiah, tetapi juga melibatkan kreativitas dalam menyampaikan ide dan informasi. Penulis memandang bahwa kemampuan menulis merupakan aspek yang krusial dalam berbagi pengetahuan dan pengalaman kepada mahasiswa, serta sebagai sarana untuk menyampaikan konsep-konsep kompleks secara jelas dan komprehensif.

Sebagai seorang akademisi yang berkembang, penulis berkomitmen untuk terus berkontribusi dalam dunia pendidikan dan penelitian, serta memperluas jangkauan pengetahuan di bidang

Informatika. Penulis berharap dapat terus memberikan kontribusi positif bagi mahasiswa dan perkembangan ilmu pengetahuan di Indonesia.

BIODATA PENULIS

Chairi Nur Insani, S.Kom., M.T.
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Wonomulyo pada tanggal 27 Juli 1994. Penulis adalah dosen tetap pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika di STMIK Dipanegara Makassar yang telah berubah menjadi Universitas Dipa Makassar dan melanjutkan S2 pada Jurusan Teknik Informatika di Universitas Hasanuddin. Penulis menekuni bidang Menulis. Selama masa studi tersebut, penulis mendalami bidang-bidang khusus dalam dunia informatika, mempelajari pemahaman mendalam tentang aspek-aspek teknis dan konseptual yang mendukung perkembangan ilmu pengetahuan dan teknologi.

Kemampuan menulis sebagai seorang akademisi tidak hanya berfokus pada karya ilmiah penelitian, tetapi juga bisa menjadi aktor utama dalam praktik kreativitas untul menyampaikan ide dan informasi. Penulis memandang bahwa kemampuan menulis merupakan aspek yang krusial dalam berbagi pengetahuan dan pengalaman kepada mahasiswa, serta sebagai sarana untuk menyampaikan konsep-konsep kompleks secara jelas dan komprehensif.

Sebagai seorang akademisi yang berkembang, penulis berkomitmen untuk terus berkontribusi dalam dunia pendidikan dan penelitian, serta memperluas jangkauan pengetahuan di bidang

Informatika. Penulis berharap dapat terus memberikan kontribusi positif bagi mahasiswa dan perkembangan ilmu pengetahuan di Indonesia.

BIODATA PENULIS

Frans Mikael Sinaga, S.Kom., M.Kom.
Dosen Program Studi Teknik Informatika
Fakultas Informatika

Seorang penulis dan dosen tetap di Program Studi Teknik Informatika, Fakultas Informatika, Universitas Mikroskil Medan, lahir di Penggalangan pada 24 Oktober 1993, Sumatera Utara. Ia merupakan anak ketiga dari empat bersaudara dari pasangan Bapak Waristo Sinaga dan Ibu Linda Sitanggang. Penulis menyelesaikan pendidikan program Sarjana (S1) di STMIK-STIE Mikroskil Medan, Program Studi Teknik Informatika, dan melanjutkan program Pasca Sarjana (S2) di STMIK-STIE Mikroskil Medan, Program Studi Teknologi Informasi. Sejak tahun 2018, penulis telah aktif mengajar di Universitas Mikroskil dan terus berkontribusi hingga saat ini. Ia mengajar berbagai mata kuliah, antara lain Pemrograman Komputer menggunakan Python, Komunikasi Data dan Jaringan Komputer, Analisis dan Perancangan Sistem, Computer Vision, Organisasi dan Arsitektur Komputer. Selain kegiatan mengajar, penulis juga aktif dalam memberikan webinar dan seminar terkait Keamanan Jaringan dan Pemrograman Komputer. Buku berjudul " Jenis-jenis Knowledge" ini merupakan karya kedua yang ditulis oleh penulis.

BIODATA PENULIS

Nurul Afifah Arifuddin, S.Pd., M.T.
Dosen Program Studi S1-Informatika
Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta

Penulis memulai perjalanan hidupnya di kota Sinjai pada tanggal 7 November 1993. Penulis menapaki langkah pertamanya dalam dunia akademis dengan menyelesaikan pendidikan sarjana di Universitas Negeri Makassar, khususnya dalam bidang Pendidikan Teknik Informatika dan Komputer. Menyadari betapa pentingnya pengembangan diri, Penulis kemudian melanjutkan pendidikan ke jenjang magister di Universitas Hasanuddin, konsentrasi yang diambil adalah Informatika. Keputusannya untuk terus mengejar ilmu di bidang IT memberinya bekal yang kuat dalam berkontribusi sebagai dosen di Jurusan S1-Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta.

Namun, kehidupan Penulis tidak hanya terbatas pada dunia akademis. Di luar ruang kuliah, ia menemukan kegembiraan dan inspirasi dalam berbagai aktivitas. Hobi utamanya adalah berkeliling menikmati keanekaragaman budaya dan keindahan alam. Selain itu, di momen ketika mood menulis mengalir, Nurul Afifah menikmati menyelami dunia literasi melalui membaca novel dan menuangkan pemikirannya dalam bentuk tulisan.

BIODATA PENULIS

Yulita Salim
Dosen Program Studi Teknik Informatika
Fakultas Ilmu Komputer Universitas Muslim Indonesia Makassar

Penulis lahir di Batusitanduk, Palopo pada tanggal 22 Juli 1981. Penulis adalah Dosen Tetap Yayasan UMI pada Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Muslim Indonesia Makassar. Menyelesaikan pendidikan S1 pada Program Studi Teknik Informatika FIKOM UMI, S2 pada Jurusan Teknik Elektro Konsentrasi Informatics and Computer UNHAS dan melanjutkan studi S3 pada Program Teknologi Informasi di MIIT Universiti Kuala Lumpur, Malaysia.

Penulis dapat dihubungi melalui email: yulita.salim@umi.ac.id

BIODATA PENULIS

A. Amirul Asnan Cirua, S.T., M.Kom
Dosen Program Studi Informatika
Fakultas Teknik Universitas Sulawesi Barat

Penulis lahir di Wonomulyo tanggal 02 April 1998. Penulis adalah dosen pada Program Studi Informatika Fakultas Teknik, Universitas Sulawesi Barat. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika di Universitas Sulawesi Barat dan melanjutkan S2 pada Jurusan Teknik Informatika di Universitas Hasanuddin. Sebagai seorang dosen informatika, selain menempuh pendidikan formal, penulis juga banyak mengikuti kegiatan pengembangan dan peningkatan diri sebagai seorang dosen terkhusus pada bidang pengajaran, penelitian, dan pengabdian. Penulis juga aktif melakukan penelitan dan menulis jurnal baik dalam skala nasional maupun internasiona. Penulis juga tergabung dalam komunitas pengembang IT mandardeveloper dan merupakan pembina divisi IoT di Informatics Study Club.