Pengenalan Data Mining

Muttaqin ■ Wahyu Wijaya Widiyanto ■ Muhammad Munsarif
Green Ferry Mandias ■ Stenly Richard Pungus ■ Agung Widarman
Wiranti Kusuma Hapsari ■ Siska Aprilia Hardiyanti
Aslam Fatkhudin ■ Pasnur ■ Eva Firdayanti Bisono
Mochammad Anshori ■ Suryani ■ Nurirwan Saputra

Pengenalan Data Mining

UU 28 tahun 2014 tentang Hak Cipta

Fungsi dan sifat hak cipta Pasal 4

Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Perlindungan Pasal 26

Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

a. penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan informasi aktual;
b. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
c. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
d. penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

Sanksi Pelanggaran Pasal 113

1. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp1.000.000.000,00 (satu miliar rupiah).

Pengenalan Data Mining

Muttaqin, Wahyu Wijaya Widiyanto, Muhammad Munsarif
Green Ferry Mandias, Stenly Richard Pungus, Agung Widarman
Wiranti Kusuma Hapsari, Siska Aprilia Hardiyanti
Aslam Fatkhudin, Pasnur, Eva Firdayanti Bisono
Mochammad Anshori, Suryani, Nurirwan Saputra

Penerbit Yayasan Kita Menulis

Pengenalan Data Mining

Penulis:
Muttaqin, Wahyu Wijaya Widiyanto, Muhammad Munsarif
Green Ferry Mandias, Stenly Richard Pungus, Agung Widarman
Wiranti Kusuma Hapsari, Siska Aprilia Hardiyanti
Aslam Fatkhudin, Pasnur, Eva Firdayanti Bisono
Mochammad Anshori, Suryani, Nurirwan Saputra

Editor: Ronal Watrianthos & Janner Simarmata
Desain Sampul: Devy Dian Pratama, S.Kom.

Penerbit
Yayasan Kita Menulis
Web: kitamenulis.id
e-mail: press@kitamenulis.id
WA: 0821-6453-7176
IKAPI: 044/SUT/2021

Muttaqin., dkk.
Pengenalan Data Mining
Yayasan Kita Menulis, 2023
xiv; 190 hlm; 16 x 23 cm
ISBN: 978-623-342-791-3
Cetakan 1, April 2023
I. Pengenalan Data Mining
II. Yayasan Kita Menulis

Katalog Dalam Terbitan

Kata Pengantar

Data mining adalah salah satu bidang ilmu yang berkembang pesat dan penting dalam dunia bisnis dan industri. Melalui teknik-teknik analisis data yang canggih, data mining memungkinkan kita untuk mengidentifikasi pola-pola yang tersembunyi dalam data, sehingga dapat memberikan wawasan dan pemahaman yang lebih dalam mengenai suatu fenomena atau masalah.

Buku ini, berjudul "Pengenalan Data Mining", dirancang sebagai panduan praktis bagi siapa saja yang ingin mempelajari konsep-konsep dasar data mining dan menerapkannya dalam praktik. Buku ini mencakup berbagai topik penting dalam data mining, mulai dari pendahuluan, pengumpulan data, preprocessing data, eksplorasi data, hingga teknik-teknik modeling, evaluasi model, dan aplikasi data mining dalam berbagai bidang.

Setiap bab dalam buku ini dirancang untuk memberikan pemahaman yang mendalam mengenai konsep-konsep dasar, teknik-teknik, dan aplikasi data mining yang relevan. Selain itu, setiap bab juga dilengkapi dengan contoh kasus dan latihan-latihan praktis, sehingga pembaca dapat memperoleh pengalaman langsung dalam menerapkan teknik-teknik data mining dalam praktik.

Secara lengkap buku ini membahas :
Bab 1 Pendahuluan
Bab 2 Pengumpulan Data
Bab 3 Pre Processing
Bab 4 Eksplorasi Data
Bab 5 Pemodelan Data
Bab 6 Evaluasi Model
Bab 7 Pengklasifikasian
Bab 8 Regresi
Bab 9 Clustering

Bab 10 Association Rule
Bab 11 Time Series Analysis
Bab 12 Text Mining
Bab 13 Data Mining Dalam Big Data
Bab 14 Aplikasi Data Mining

Buku ini diharapkan dapat menjadi sumber belajar yang bermanfaat bagi mahasiswa, praktisi bisnis, peneliti, atau siapa saja yang tertarik dalam bidang data mining. Kami berharap bahwa buku ini dapat membantu pembaca dalam mengembangkan keterampilan dan pemahaman yang dibutuhkan dalam memanfaatkan potensi data mining untuk meningkatkan efisiensi dan efektivitas dalam berbagai kegiatan bisnis dan industri. Selamat membaca!

Salam,
Penulis.

Daftar Isi

Daftar Gambar

Daftar Tabel

Bab 1

Pendahuluan

1.1 Definisi Data Mining

Data mining adalah proses ekstraksi informasi yang bermanfaat dari data besar dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Informasi yang dihasilkan dari proses data mining dapat digunakan untuk membantu pengambilan keputusan dan prediksi di berbagai bidang, seperti bisnis, ilmu pengetahuan, teknologi, kesehatan, dan lain-lain. Definisi data mining telah berkembang seiring dengan perkembangan teknologi dan penelitian.

Berikut adalah beberapa definisi data mining yang terbaru:

1. Menurut Witten, Frank, Hall (2016) data mining sebagai proses pengembangan model yang berguna untuk menggali pengetahuan dari data. Mereka menyatakan bahwa model yang dihasilkan dari proses Data Mining dapat digunakan untuk memprediksi nilai atau kelas dari data baru.
2. Menurut Tan, Steinbach, dan Kumar (2018) data mining sebagai proses menemukan pola-pola menarik dari data besar dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Mereka menyatakan bahwa pola-pola yang ditemukan dapat

digunakan untuk membuat prediksi, memperbaiki kinerja sistem, atau mendukung pengambilan keputusan.

3. Menurut Han dan Kamber (2019) mendefinisikan data mining sebagai proses menemukan pola, hubungan, dan anomali yang menarik dari data besar dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Mereka menyatakan bahwa hasil dari proses data mining dapat digunakan untuk memprediksi, menjelaskan, atau mengoptimalkan fenomena yang diamati.
4. Menurut Kononenko dan Kukar (2007) mendefinisikan data mining sebagai proses ekstraksi pengetahuan yang berharga dari data dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Mereka menyatakan bahwa pengetahuan yang dihasilkan dapat digunakan untuk meningkatkan kinerja sistem, memperbaiki prediksi, dan mengoptimalkan pengambilan keputusan.

Keempat definisi di atas memiliki kesamaan dalam hal menekankan pentingnya proses ekstraksi informasi yang bermanfaat dari data besar dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Namun, terdapat perbedaan dalam hal penekanan pada pengembangan model, menemukan pola-pola menarik, menemukan pola, hubungan, dan anomali yang menarik, atau ekstraksi pengetahuan yang berharga.

Definisi data mining yang luas dan sering digunakan saat ini adalah definisi CRISP-DM (Cross Industry Standard Process for Data Mining). Definisi ini mencakup enam tahapan dalam proses data mining, yaitu:

1. understanding the business problem;
2. understanding the data;
3. data preparation;
4. modeling;
5. evaluation;
6. deployment.

Definisi CRISP-DM menyatakan bahwa data mining adalah proses yang terstruktur dan berulang-ulang untuk menemukan informasi yang bermanfaat dari data besar. Proses ini melibatkan tahapan-tahapan yang terdefinisi dengan

jelas, mulai dari memahami masalah bisnis hingga mengimplementasikan hasil Data Mining ke dalam sistem yang sudah ada.

Selain definisi-definisi di atas, terdapat juga beberapa definisi lain yang lebih spesifik, seperti:

1. Text mining adalah proses ekstraksi informasi bermanfaat dari teks dengan menggunakan teknik-teknik matematika, statistika, dan kecerdasan buatan. Informasi yang dihasilkan dari proses *text mining* dapat berupa konsep, topik, atau sentimen yang terkandung dalam teks (Grossman and Frieder, 2016).
2. Web mining adalah proses ekstraksi informasi bermanfaat dari data yang dihasilkan oleh aplikasi web, seperti mesin pencari, situs jejaring sosial, dan situs e-Commerce. Informasi yang dihasilkan dari proses web mining dapat berupa preferensi pengguna, tren pasar, atau profil konsumen (Zheng and Lytras, 2018).
3. Social media mining adalah proses ekstraksi informasi bermanfaat dari data yang dihasilkan oleh situs jejaring sosial, seperti Facebook, Twitter, dan Instagram. Informasi yang dihasilkan dari proses *social media mining* dapat berupa preferensi pengguna, opini publik, atau kecenderungan sosial (Zafarani, Abbasi and Liu, 2014).

Definisi-definisi ini menekankan pada jenis data tertentu yang diekstraksi dengan menggunakan teknik-teknik data mining yang khusus. Namun, kesamaannya adalah bahwa proses ekstraksi informasi yang bermanfaat dari data besar tetap menjadi fokus utama dari data mining (Larose, 2006).

Keuntungan Data Mining

Terdapat banyak keuntungan yang bisa didapatkan dari penerapan Data Mining dalam berbagai bidang, antara lain:

1. Meningkatkan efisiensi bisnis
   Data mining dapat membantu mengidentifikasi pola dan trend dalam data yang besar dan kompleks, sehingga membantu organisasi untuk membuat keputusan yang lebih tepat dan akurat. Misalnya, Data Mining dapat membantu organisasi dalam menemukan faktor-faktor yang memengaruhi penjualan, sehingga organisasi dapat membuat

keputusan yang lebih baik dalam hal manajemen persediaan dan pemasaran.

2. Mengidentifikasi peluang bisnis baru
   Data mining dapat membantu organisasi dalam mengidentifikasi peluang bisnis baru yang mungkin belum terpikirkan sebelumnya. Misalnya, dengan menganalisis data konsumen, organisasi dapat menemukan segmen pasar yang masih belum tersentuh dan memutuskan untuk mengembangkan produk atau layanan yang baru untuk segmen pasar tersebut.
3. Menemukan faktor penyebab masalah
   Data mining dapat membantu organisasi dalam menemukan faktor penyebab masalah yang mungkin sulit untuk diidentifikasi dengan metode konvensional. Misalnya, dengan menganalisis data gangguan pada mesin-mesin, organisasi dapat menemukan faktor-faktor yang menyebabkan mesin rusak, sehingga dapat mengambil tindakan yang tepat untuk mencegah terjadinya gangguan pada masa yang akan datang.
4. Meningkatkan keamanan dan deteksi penipuan
   Data mining dapat membantu organisasi dalam meningkatkan keamanan dan deteksi penipuan. Misalnya, dengan menganalisis pola transaksi pada kartu kredit, organisasi dapat menemukan transaksi yang mencurigakan dan mengambil tindakan yang tepat untuk mencegah terjadinya penipuan.
5. Meningkatkan efektivitas pelayanan pelanggan
   Data mining dapat membantu organisasi dalam meningkatkan efektivitas pelayanan pelanggan. Misalnya, dengan menganalisis data konsumen, organisasi dapat menemukan pola-pola perilaku konsumen dan kebutuhan pelanggan, sehingga dapat memberikan pelayanan yang lebih baik dan menyesuaikan produk atau layanan sesuai dengan kebutuhan pelanggan.

Keterbatasan Data Mining

Tentunya, seperti halnya teknologi atau metode lainnya, data mining memiliki beberapa keterbatasan yang perlu diperhatikan, antara lain:

1. Kesalahan dalam data
   Data mining sangat tergantung pada kualitas data yang digunakan. Jika data yang digunakan memiliki kesalahan atau kekurangan, maka hasil dari data mining dapat menjadi tidak akurat atau bahkan salah.
2. Ketergantungan pada algoritma dan model yang digunakan
   Hasil dari data mining sangat tergantung pada algoritma dan model yang digunakan. Jika algoritma atau model yang digunakan tidak sesuai dengan data atau tidak diatur dengan benar, maka hasil dari data mining dapat menjadi tidak akurat atau bahkan salah.
3. Keterbatasan dalam data yang tidak terstruktur
   Data mining lebih mudah dilakukan pada data yang terstruktur, seperti data numerik atau data yang tersimpan dalam database relasional. Namun, ketika data tidak terstruktur, seperti data teks atau gambar, maka akan lebih sulit untuk dilakukan data mining.
4. Keterbatasan dalam interpretasi hasil
   Hasil dari data mining dapat menjadi sangat kompleks dan sulit untuk diinterpretasikan oleh orang yang tidak memiliki latar belakang teknis atau statistik. Oleh karena itu, perlu ada upaya untuk membuat hasil yang lebih mudah dipahami dan disajikan dalam bentuk yang dapat dimengerti oleh orang yang tidak memiliki latar belakang teknis atau statistik.
5. Masalah privasi dan etika
   Dalam penggunaan data mining, ada masalah privasi dan etika yang perlu diperhatikan. Dalam beberapa kasus, penggunaan data dapat merusak privasi orang lain atau mengungkapkan informasi yang seharusnya tidak diberikan kepada orang lain.

Contoh Penerapan Data Mining

Data Mining telah banyak digunakan dalam berbagai bidang, termasuk bisnis, pemerintah, kesehatan, dan lain-lain. Berikut adalah beberapa contoh penerapan Data Mining dalam berbagai bidang:

1. Bisnis dalam bisnis
   Data mining dapat digunakan untuk menganalisis data pelanggan dan memprediksi perilaku pelanggan, memprediksi penjualan dan permintaan, menemukan pola pembelian, dan mengoptimalkan pengiriman dan persediaan. Data mining juga dapat digunakan untuk mengoptimalkan kampanye iklan, memilih lokasi toko, dan menemukan pola fraud.
2. Kesehatan dalam bidang kesehatan
   Data mining dapat digunakan untuk menganalisis data pasien dan memprediksi risiko penyakit, menemukan pola penyakit, memprediksi hasil tes dan pengobatan, dan mengoptimalkan penggunaan sumber daya kesehatan.
3. Pemerintah dalam bidang pemerintah
   Data mining dapat digunakan untuk mengidentifikasi pola kriminalitas, memprediksi kejadian kejahatan, menemukan pola perpajakan, dan mengoptimalkan alokasi sumber daya pemerintah.
4. Sumber daya alam dalam bidang sumber daya alam
   Data mining dapat digunakan untuk memprediksi potensi mineral, menemukan pola penggunaan lahan, dan memprediksi kejadian alam seperti gempa bumi dan banjir.

Tujuan Data Mining

Tujuan data mining adalah untuk menemukan pola atau informasi yang berguna dari kumpulan data yang besar dan kompleks. Data mining digunakan untuk mengidentifikasi pola dalam data yang sebelumnya tidak diketahui dan dapat digunakan untuk membantu dalam pengambilan keputusan, prediksi, dan perencanaan. Tujuan utama data mining adalah untuk menghasilkan informasi yang berguna dari data yang tersedia, sehingga dapat digunakan untuk meningkatkan efisiensi dan efektivitas dari sebuah organisasi atau proses.

Berikut adalah beberapa tujuan utama dari data mining:

1. Mengidentifikasi pola dan hubungan
   Tujuan utama data mining adalah untuk mengidentifikasi pola dan hubungan dalam data yang sebelumnya tidak diketahui. Dengan mengidentifikasi pola dan hubungan ini, organisasi dapat mengambil tindakan yang tepat dan efektif.
2. Memprediksi perilaku dan tren
   Data mining juga digunakan untuk memprediksi perilaku dan tren di masa depan berdasarkan data historis. Dengan memprediksi perilaku dan tren ini, organisasi dapat mengambil tindakan yang tepat dan mengambil keuntungan dari peluang yang muncul.
3. Menemukan informasi yang berguna
   Data mining juga digunakan untuk menemukan informasi yang berguna dari data yang sebelumnya tidak terlihat. Informasi ini dapat digunakan untuk memperbaiki proses bisnis, meningkatkan keefektifan organisasi, atau meningkatkan layanan yang diberikan.
4. Mendukung pengambilan keputusan
   Data mining juga dapat digunakan untuk mendukung pengambilan keputusan. Dengan mengumpulkan dan menganalisis data, organisasi dapat membuat keputusan yang lebih baik dan berdasarkan fakta.
5. Meningkatkan efisiensi dan efektivitas
   Tujuan utama Data Mining adalah untuk meningkatkan efisiensi dan efektivitas sebuah organisasi atau proses. Dengan mengumpulkan dan menganalisis data, organisasi dapat mengidentifikasi area yang memerlukan perbaikan dan mengambil tindakan untuk meningkatkan efisiensi dan efektivitas (Aggarwal, 2015). Serta untuk mengidentifikasi pola dan hubungan dalam data yang sebelumnya tidak diketahui, memprediksi perilaku dan tren, menemukan informasi yang berguna, mendukung pengambilan keputusan, dan meningkatkan efisiensi dan efektivitas.

Jenis-Jenis Data Mining

Data mining adalah proses menggali informasi yang berharga dari kumpulan data besar dan kompleks. Dalam proses data mining, berbagai teknik dan metode digunakan untuk menemukan pola dan hubungan dalam data yang sebelumnya tidak diketahui.

Jenis-jenis data mining mencakup penggalian asosiasi, klasifikasi, klastering, regresi, dan analisis anomali:

1. Penggalian Asosiasi - Teknik data mining yang digunakan untuk menemukan korelasi antara item dalam sebuah kumpulan data. Contoh penggalian asosiasi adalah menemukan bahwa konsumen yang membeli roti sering kali juga membeli mentega. Penggalian asosiasi biasanya digunakan dalam bidang e-Commerce dan pemasaran.
2. Klasifikasi - Teknik data mining yang digunakan untuk memprediksi kelas atau label dari sebuah objek berdasarkan karakteristiknya. Contoh klasifikasi adalah memprediksi apakah sebuah email adalah spam atau tidak spam berdasarkan kata-kata yang digunakan dalam email tersebut. Klasifikasi biasanya digunakan dalam bidang keamanan informasi, pemasaran, dan kesehatan.
3. Klastering - Teknik data mining yang digunakan untuk mengelompokkan objek berdasarkan karakteristik atau fitur yang sama. Contoh klastering adalah mengelompokkan konsumen berdasarkan preferensi dan kebiasaan belanja mereka. Klastering biasanya digunakan dalam bidang pemasaran, sosiologi, dan biologi.
4. Regresi - Teknik data mining yang digunakan untuk memprediksi nilai numerik dari sebuah variabel berdasarkan variabel lainnya. Contoh regresi adalah memprediksi harga rumah berdasarkan ukuran, jumlah kamar tidur, dan lokasi rumah tersebut. Regresi biasanya digunakan dalam bidang bisnis, ekonomi, dan keuangan.
5. Analisis Anomali - Teknik data mining yang digunakan untuk mengidentifikasi data yang berbeda atau tidak normal. Contoh analisis anomali adalah mendeteksi transaksi yang mencurigakan pada kartu kredit atau memantau kinerja sistem komputer untuk

mengidentifikasi masalah. Analisis anomali biasanya digunakan dalam bidang keamanan informasi dan kesehatan.

1.2 Proses Data Mining

Proses data mining adalah serangkaian langkah atau tahapan dalam menggali pengetahuan atau informasi dari data yang besar dan kompleks. Proses ini meliputi tahapan mulai dari pemilihan data, pre-prosesing data, pemodelan, dan evaluasi hasil. Proses data mining yang efektif membutuhkan perencanaan yang matang serta pemilihan teknik dan algoritma yang tepat untuk setiap tahapan.

Dalam artikel ini, akan dibahas tentang proses data mining secara detail, serta beberapa teknik yang umum digunakan dalam setiap tahapan prosesnya.

Tahapan Pemilihan Data

Tahapan pemilihan data merupakan langkah awal dalam proses Data Mining. Pemilihan data yang tepat akan memastikan bahwa hasil dari proses Data Mining lebih akurat dan relevan. Beberapa kriteria dalam pemilihan data antara lain: relevansi data, ketersediaan data, jumlah data yang memadai, dan kualitas data. Setelah pemilihan data dilakukan, langkah selanjutnya adalah melakukan pre-prosesing data.

Tahapan Pre-Prosesing Data

Pre-prosesing data adalah tahapan penting dalam proses Data Mining. Tujuan dari tahapan ini adalah untuk membersihkan data dari kesalahan, menghilangkan data yang tidak relevan, serta membuat data siap untuk proses selanjutnya.

Beberapa teknik dalam pre-prosesing data antara lain:

1. Data cleaning - Proses pembersihan data dari noise, kesalahan, dan ketidakakuratan.
2. Data integration - Proses penggabungan data dari berbagai sumber yang berbeda.
3. Data transformation - Proses transformasi data ke dalam format yang lebih tepat dan terstandarisasi.

4. Data reduction - Proses pengurangan jumlah data dengan menggunakan teknik sampling atau filtering.

Setelah pre-prosesing data dilakukan, langkah selanjutnya adalah memodelkan data.

Tahapan Pemodelan Data

Pemodelan data adalah tahapan dalam proses Data Mining yang menggunakan teknik dan algoritma untuk menghasilkan model atau pola dari data yang telah dipilih dan di preproses. Tujuan dari tahapan ini adalah untuk mengidentifikasi pola yang tersembunyi dalam data, dan membangun model yang dapat digunakan untuk membuat prediksi atau mengambil keputusan. B

Beberapa teknik dalam pemodelan data antara lain:

1. Asosiasi - Teknik yang digunakan untuk mengidentifikasi hubungan antara item-item dalam data.
2. Klasifikasi - Teknik yang digunakan untuk membangun model yang dapat mengklasifikasikan data ke dalam kategori tertentu.
3. Klastering - Teknik yang digunakan untuk mengelompokkan data yang serupa berdasarkan kesamaan karakteristik.
4. Regresi - Teknik yang digunakan untuk memprediksi nilai variabel yang kontinu berdasarkan variabel independen.
5. Analisis anomali - Teknik yang digunakan untuk mengidentifikasi data yang berbeda dari pola umum dalam data.

Setelah tahapan pemodelan selesai dilakukan, langkah selanjutnya adalah mengevaluasi hasil dari proses data mining.

Tahapan Evaluasi Hasil

Tahapan evaluasi hasil adalah tahapan dalam proses Data Mining yang dilakukan untuk mengevaluasi kualitas dan keakuratan model yang telah dibangun. Beberapa teknik dalam tahapan evaluasi hasil antara lain:

1. Confusion Matrix - Teknik yang digunakan untuk mengukur kinerja model dengan membandingkan prediksi dengan hasil aktual.

2. Validasi silang - Teknik yang digunakan untuk memvalidasi kinerja model dengan membagi data ke dalam subset yang berbeda untuk pelatihan dan pengujian.
3. Kurva ROC - Teknik yang digunakan untuk mengukur kinerja model pada berbagai *threshold* atau batas ambang.

Setelah tahapan evaluasi hasil selesai dilakukan, langkah terakhir adalah menerapkan hasil dari proses data mining.

Tahapan Aplikasi Hasil

Tahapan aplikasi hasil adalah tahapan terakhir dalam proses Data Mining, yang melibatkan penggunaan hasil yang telah ditemukan untuk membuat keputusan atau tindakan yang dapat meningkatkan efisiensi dan efektivitas bisnis atau organisasi.

Beberapa contoh aplikasi hasil data mining antara lain:

1. Prediksi - Menggunakan model yang dibangun untuk membuat prediksi atau perkiraan tentang suatu peristiwa atau kondisi.
2. Segmentasi - Menggunakan hasil klastering untuk mengidentifikasi kelompok pelanggan atau pasar yang berbeda.
3. Identifikasi anomali - Menggunakan hasil analisis anomali untuk mengidentifikasi kejadian yang tidak biasa atau tidak diharapkan.
4. Pemilihan fitur - Menggunakan hasil pemodelan untuk memilih fitur atau atribut yang paling relevan dan memberikan kontribusi terbesar terhadap prediksi.

Bab 2

Pengumpulan Data

2.1 Pendahuluan

Pengumpulan data bisa dikatakan sebagai prosedur mengumpulkan, mengukur, dan menganalisis wawasan yang akurat untuk penelitian menggunakan teknik standar yang divalidasi. Seorang peneliti dapat mengevaluasi hipotesis penelitian mereka berdasarkan data yang dikumpulkan. Dalam kebanyakan kasus, pengumpulan data adalah langkah utama dan paling penting untuk penelitian, terlepas dari bidang penelitian. Pendekatan pengumpulan data berbeda untuk berbagai bidang studi, tergantung pada informasi yang diperlukan.

Pengumpulan data adalah proses mengumpulkan dan mengukur informasi tentang variabel-variabel penelitian yang ditargetkan dalam suatu sistem yang mapan, yang kemudian memungkinkan seseorang untuk menjawab pertanyaan yang relevan dan mengevaluasi hasil. Pengumpulan data adalah komponen penelitian di semua bidang studi termasuk ilmu fisik dan sosial, humaniora, dan bisnis. Meskipun metode bervariasi berdasarkan disiplin ilmu, penekanannya adalah memastikan bahwa pengumpulan data dilakukan secara akurat dan jujur.

Metode pengumpulan data adalah teknik maupun cara yang dilakukan untuk mengumpulkan data. di mana metode menunjuk pada suatu cara sehingga bisa

diperlihatkan penggunaannya melalui angket penelitian, wawancara, pengamatan, tes, dokumentasi, dan sebagainya.

Metode pengumpulan data dapat dibedakan menjadi 2, yaitu:

Pengumpulan Data Sekunder

Data sekunder adalah jenis data yang telah diterbitkan dalam buku, surat kabar, majalah, jurnal, portal online, dan lain-lain. Ada banyak data yang tersedia dari sumber-sumber tersebut terkait bidang penelitian Anda, terlepas dari sifat bidang penelitian.

Oleh karena itu, penerapan seperangkat kriteria yang tepat untuk memilih data sekunder yang akan digunakan dalam penelitian ini memainkan peran penting dalam hal meningkatkan tingkat validitas dan reliabilitas penelitian. Kriteria ini termasuk (tetapi tidak terbatas pada) tanggal publikasi, kredensial penulis, keandalan sumber, kualitas diskusi, kedalaman analisis, tingkat kontribusi teks untuk pengembangan bidang penelitian dan lain-lain.

Pengumpulan Data Primer

Pengumpulan Data Primer merupakan metode pengumpulan data primer dapat dibagi menjadi dua kelompok besar yaitu teknik dalam metode penelitian kuantitatif dan penelitian kualitatif. Metode penelitian kuantitatif menggambarkan dan mengukur tingkat kejadian berdasarkan angka dan perhitungan. Selain itu, pertanyaan "berapa banyak?" dan "seberapa sering?" sering diajukan dalam studi kuantitatif. Dengan demikian, metode pengumpulan data kuantitatif didasarkan pada angka dan perhitungan matematis.

Penelitian kuantitatif dapat digambarkan sebagai penelitian yang melibatkan pengumpulan data numerik dan menunjukkan pandangan hubungan antara teori dan penelitian yang bersifat deduktif, kecenderungan untuk pendekatan ilmu pengetahuan alam, dan memiliki konsepsi objektivis tentang realitas sosial.

Metode pengumpulan data kualitatif bersifat eksploratif dan terutama berkaitan untuk mendapatkan wawasan dan pemahaman tentang alasan dan motivasi yang mendasarinya. Metode pengumpulan data kualitatif muncul setelah diketahui bahwa metode pengumpulan data kuantitatif tradisional tidak dapat mengekspresikan perasaan dan emosi manusia.

Tujuan dari semua pengumpulan data adalah membuktikan kualitas kegunaan pengumpulan data adalah untuk menangkap bukti kualitas yang memungkinkan analisis mengarah pada perumusan jawaban yang meyakinkan dan kredibel untuk pertanyaan yang diajukan.

Tentu saja terlepas dari bidang studi atau preferensi untuk mendefinisikan data (kuantitatif atau kualitatif), pengumpulan data yang akurat sangat penting untuk menjaga integritas penelitian. Di mana pemilihan instrumen pengumpulan data yang sesuai (yang ada, dimodifikasi, atau yang baru dikembangkan) dan instruksi yang dijelaskan dengan jelas untuk penggunaan yang benar mengurangi kemungkinan kesalahan.

Metode pengumpulan data yang dianggap sebagai teknik atau cara yang dilakukan oleh peneliti untuk mengumpulkan data. Pengumpulan data dilakukan untuk mendapatkan informasi topik penelitian yang diperlukan dalam rangka mencapai tujuan penelitian.

Rencana pengumpulan data adalah kunci penting untuk mengembangkan studi yang baik. Rencana menunjukkan bagaimana anda akan mengakses dan mengumpulkan informasi dari partisipan anda. Rencana pengumpulan data yang jelas pada tahap proposal dapat mengurangi stres dan memastikan bahwa para peneliti di masa depan dapat mereplikasi studi anda. Berikut pengumpulan data di data mining.

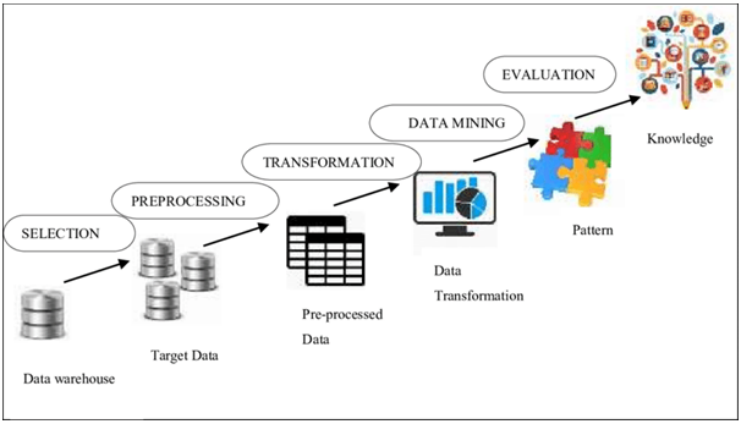
2.2 Proses Pengumpulan Data KDD

Bagaimana proses data diambil? Sebenarnya, pengumpulan data dilakukan melalui proses *Knowledge Discovery In Databases* atau KDD. Proses tersebut diawali dari data mentahan, dan terus dilakukan hingga berakhir pada informasi atau pengetahuan yang sudah diolah (Rukmana and Ramdani, 2018).

Knowledge Discovery in Database Process (KDD) adalah salah satu metode yang bisa digunakan dalam melakukan data mining. Fayyed et al. (1996) mendefinisikan KDD sebagai proses dari menggunakan metode data mining untuk mencari informasi-informasi yang berharga, pola yang ada di dalam data, yang melibatkan algoritma untuk mengidentifikasi pola pada data. Dunham (2003) meringkas proses KDD dari berbagai step, yaitu: seleksi data,

pra-proses data, transformasi data, data mining, dan yang terakhir interpretasi dan evaluasi (Yu et al., 2018).

Berikut adalah ilustrasi serta penjelasan mengenai proses KDD secara detail terlihat pada gambar 2.1 berikut:



Gambar 2.1: Proses KDD

Berdasarkan gambar di atas, masing-masing penjelasannya adalah sebagai berikut:

1. Data Cleansing
   Proses di mana data diolah lalu dipilih data yang dianggap bisa dipakai, di mana pada proses ini:
   a. Pemrosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan *noise* dilakukan.
   b. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD.
   c. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.
   d. Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain (eksternal).

2. Data Integration
   Proses menggabungkan data yang dianggap berulang akan digabungkan menjadi satu.
3. Selection
   Proses seleksi atau pemilihan data yang dianggap relevan terhadap analisis, di mana:
   a. Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, di mana penemuan (discovery) akan dilakukan.
   b. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.
4. Data Transformation
   Proses transformasi data terpilih ke dalam bentuk mining prosedur, seperti:
   a. Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada *goal* yang ingin dicapai.
   b. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.
5. Data Mining
   Proses di mana dilakukan beragam teknik untuk mengekstrak pola-pola potensial menghasilkan data yang berguna, di mana:
   a. Pemilihan tugas data mining, pemilihan *goal* dari proses KDD misalnya klasifikasi, regresi, clustering, dll.
   b. Pemilihan algoritma data mining untuk pencarian (searching).
   c. Proses data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

d. Teknik, metode, atau algoritma dalam data mining sangat bervariasi.
e. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

6. Pattern Evolution
   Proses di mana pola-pola yang telah diidentifikasi berdasarkan *measure* yang diberikan, antara lain:
   a. Penerjemahan pola-pola yang dihasilkan dari data mining.
   b. Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.
   c. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.
7. Knowledge Presentation
   Proses paling akhir dari proses KDD, data-data yang sudah diproses divisualisasikan agar lebih mudah dipahami oleh pengguna dan diharapkan bisa diambil tindakan berdasarkan analisis.

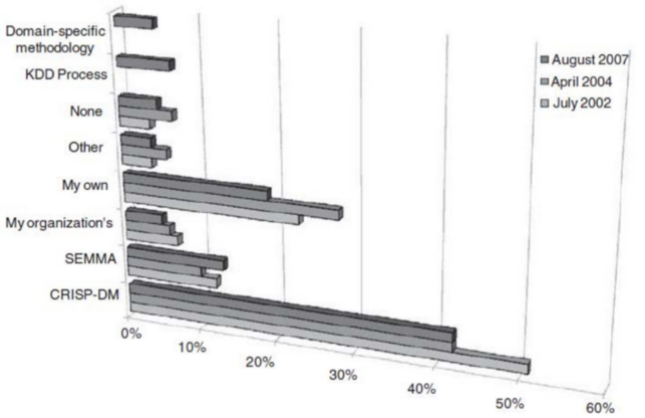
2.3 Proses Pengumpulan Data Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah salah satu model proses data mining (data mining framework) yang awalnya (1996) dibangun oleh 5 perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA.

Framework ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan *methodology standard non-proprietary* bagi data mining. Versi pertama dari metodologi ini dipresentasikan pada 4th CRISP-DM SIG Workshop di Brussels pada bulan Maret 1999 (Mahalakshmi,

Sridevi and Rajaram, 2016); dan langkah-langkah proses data mining berdasarkan model ini dipublikasikan pada tahun berikutnya.

Antara tahun 2006 dan 2008 terbentuklah grup CRISP-DM 2.0 SIG yang berkeinginan untuk mengupdate CRISP-DM proses model. Namun produk akhir dari inisiatif ini tidak diketahui. Banyak hasil penelitian yang mengungkapkan bahwa CRISP-DM adalah data mining model yang masih digunakan secara luas di kalangan industri, sebahagian dikarenakan keunggulannya dalam menyelesaikan banyak persoalan dalam proyek-proyek data mining.



Gambar 2.2: Survei Penggunaan Metodologi Data Mining (Mariscal, Marban, and Fernandez 2010)

Mariscal, Marba, dan Fernandez menyatakan CRISP-DM sebagai *de facto* menjadi standar untuk pengembangan proyek data mining dan *knowledge discovery* karena paling banyak digunakan dalam pengembangan data mining. Hal tersebut dapat terlihat dari survei yang ditunjukkan pada Gambar 2.2 yang dilakukan terhadap penggunaan metodologi dalam proyek data mining (Navisa, Hakim and Nabilah, 2021).

Hasil survei "Penggunaan Metodologi dalam Proyek Data Mining", memperlihatkan pengguna CRISP-DM di tahun 2002 mencapai 51%, kemudian menurun menuju 41% di tahun 2004. Meskipun persentase penggunaan CRISP-DM menurun 10%, jumlah pengguna metodologi ini masih terbilang lebih banyak daripada pengguna metodologi lain.

Model proses CRISP-DM memberikan gambaran tentang siklus hidup proyek data mining. CRISP-DM memiliki 6 tahapan yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation*, dan *Deployment* seperti ditunjukkan pada Gambar 2.2 (Rafi Muttaqin et al., 2022).

Gambar 2.3: Tahapan Dalam CRISP-DM

Masing-masing tahapan tersebut dijelaskan sebagai berikut:

Business Understanding

Ini adalah tahap pertama dalam CRISP-DM dan termasuk bagian yang cukup vital. Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokkan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun.

Kegiatan yang dilakukan antara lain menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

Data Understanding

Secara garis besar untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan (summary) dan mengidentifikasi potensi masalah dalam data. Tahap ini juga harus dilakukan secara cermat dan tidak

terburu-buru, seperti pada visualisasi data, yang terkadang *insight*-nya sangat sulit didapat jika dihubungkan dengan *summary* datanya.

Jika ada masalah pada tahap ini yang belum terjawab, maka akan mengganggu pada tahap modeling. Ringkasan atau summary dari data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan, atau mengungkapkan penyimpangan tak terduga yang perlu ditangani pada tahap selanjutnya, yaitu *data preparation*. Masalah dalam data biasanya seperti nilai-nilai yang hilang, outlier, berdistribusi spike, berdistribusi bimodal harus diidentifikasi dan diukur sehingga dapat diperbaiki dalam *data preparation*.

Data Preparation

Secara garis besar untuk memperbaiki masalah dalam data, kemudian membuat variabel *derived*. Tahap ini jelas membutuhkan pemikiran yang cukup matang dan usaha yang cukup tinggi untuk memastikan data tepat untuk algoritma yang digunakan. Bukan berarti saat *Data Preparation* pertama kali di mana masalah-masalah pada data sudah diselesaikan, data sudah dapat digunakan hingga tahap terakhir.

Tahap ini merupakan tahap yang sering ditinjau kembali saat menemukan masalah pada saat pembangunan model. Sehingga dilakukan iterasi sampai menemukan hal yang cocok dengan data. Tahap sampling dapat dilakukan di sini dan data secara umum dibagi menjadi dua, data training dan data testing.

Kegiatan yang dilakukan antara lain memilih kasus dan parameter yang akan dianalisis (Select Data), melakukan transformasi terhadap parameter tertentu (Transformation), dan melakukan pembersihan data agar data siap untuk tahap modeling (Cleaning).

Modeling

Secara garis besar untuk membuat model prediktif atau deskriptif. Pada tahap ini dilakukan metode statistika dan Machine Learning untuk penentuan terhadap teknik data mining, alat bantu data mining, dan algoritma data mining yang akan diterapkan.

Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap *data preparation*.

Beberapa modeling yang biasa dilakukan adalah *classification, scoring, ranking, clustering, finding relation*, dan *characterization*.

Evaluation

Melakukan interpretasi terhadap hasil dari data mining yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

Deployment

Tahap *deployment* atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk *deployment* dimulai selama *business understanding* dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional.

Pada akhirnya, rencana *sistem deployment* mengakui bahwa tidak ada model yang statis. Model tersebut dibangun dari data yang diwakili data pada waktu tertentu, sehingga perubahan waktu dapat menyebabkan berubahnya karakteristik data. Model pun harus dipantau dan mungkin diganti dengan model yang sudah diperbaiki (Singgalen, 2022).

Bab 3

Pre Processing

3.1 Pendahuluan

Pemilihan fitur yang nantinya akan di proses dalam metode data mining adalah sebuah fase terpenting dalam proses awal data mining. Di mana fitur tersebut merupakan fitur yang sudah diolah sebelumnya. Akurasi kinerja yang baik dilakukan dengan menerapkan aplikasi-aplikasi pemodelan fitur yang sudah saling dikombinasikan. Tentunya ini membutuhkan biaya yang tidak sedikit karena adanya komputasi yang kompleks dan mahal dalam pengadaannya. Tetapi ternyata dalam perkembangannya fitur yang banyak itu tidak memberikan kepastian bahwa kinerja sistem akan menjadi lebih baik dan maksimal.

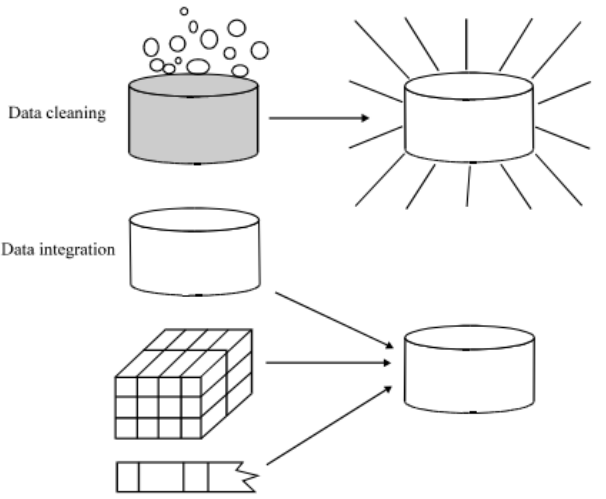
Penggunaan dua fitur yang memiliki diskriminan dua kelas ternyata lebih baik daripada jumlah fitur yang banyak tetapi tidak memiliki diskriminasi dua kelas. Bisa disimpulkan bahwa fitur yang digunakan haruslah fitur yang punya korelasi untuk mendiskriminasi kelas-kelas yang akan diproses. Perlu di perhatikan juga sifat generalisasi yang dibangun oleh klasifikator. Sifat generalisasi klasifikator yang dihasilkan akan semakin bagus jika rasio jumlah data latih terhadap jumlah parameter bebasnya tinggi.

Poin penting dalam pemilihan fitur adalah jika terdapat banyak kandidat fitur yang akan digunakan, maka cara pemilihan fitur yang paling penting di antara

kandidat-kandidat tersebut sehingga berkurang jumlahnya tetapi juga memberikan diskriminasi kelas yang baik juga. Jika kemudian fitur yang dipilih hanya memiliki diskriminasi yang kecil maka kinerja desain klasifikator yang dibentuk otomatis menjadi buruk. Sebaliknya jika fitur yang dipilih penuh dengan informasi maka desain klasifikator yang dibentuk sederhana.

Dari hal diatas maka saat pemilihan fitur harus diusahakan mengarah jarak perbedaan antara kelas yang besar dan variasi dalam kelas kecil. Artinya adalah bahwa fitur harus memiliki nilai beda yang jauh dalam kelas yang berbeda tetapi dekat nilainya dengan kelas yang sama. Maka dari itu fitur harus diuji satu demi satu, dan menghapus fitur yang punya kemampuan diskriminasi kecil (Prasetyo, 2014).

Solusi terbaik yaitu dengan menguji dua fitur atau lebih secara bersama sama karena bisa saja ada fitur yang punya korelasi lemah dan kuat. Fitur-fitur yang diuji tersebut tidak hanya dalam dimensi saat itu kadang ditemukan juga kasus data yang non linear tidak bisa ditemukan diskriminasinya hingga kemudian dilakukan transformasi fitur dari dimensi yang lama ke dimensi yang baru yang dimungkinkan akan mendapatkan kinerja yang lebih tinggi di mana bisa mentransformasi yang asalnya data non linear menjadi linear seperti pada gambar 3.1.



Gambar 3.1: Preprocessing Data (Suad A. Alasadi and Wesam S. Bhaya, 2017)

3.2 Fase Pre Processing Data

Fase pre processing data adalah hal paling penting dalam proses penambangan data (Charu C. Anggarwal, 2015). Namun, jarang dieksplorasi lebih jauh karena sebagian besar fokusnya adalah pada aspek analitis dari penambangan data.

Fase ini dimulai dengan pengumpulan data, dan itu terdiri dari langkah-langkah berikut:

Ekstraksi Fitur

Seorang analis mungkin dihadapkan dengan sejumlah besar dokumen mentah, log sistem, atau transaksi komersial dengan sedikit panduan tentang bagaimana data mentah ini harus diubah menjadi fitur database yang bermakna untuk diproses. Fase ini sangat bergantung pada analis untuk dapat mengabstraksi fitur-fitur yang ada dan yang paling relevan dengan aplikasi tertentu.

Misalnya, dalam aplikasi untuk mendeteksi penipuan kartu kredit, jumlah muatan, frekuensi pengulangan. Oleh karena itu, mengekstraksi fitur yang tepat sering kali merupakan keterampilan yang membutuhkan pemahaman tentang domain aplikasi spesifik yang ada.

Pembersihan Data

Data yang diekstraksi mungkin memiliki entri yang salah atau hilang. Karena itu, beberapa catatan mungkin perlu dihapus, atau entri yang hilang mungkin perlu diestimasi. Inkonsistensi mungkin perlu dihilangkan.

Pemilihan dan Transformasi Fitur

Ketika data berdimensi sangat tinggi, banyak algoritma penambangan data tidak bekerja secara efektif. Selain itu, banyak yang berdimensi tinggi fitur kurang baik sehingga menambah kesalahan pada proses penambangan data. Karena itu, berbagai metode digunakan untuk menghapus fitur yang tidak relevan atau mengubah serangkaian fitur saat ini ke ruang data baru yang lebih dapat menerima analisis.

Aspek lain yang terkait adalah transformasi data, di mana satu kumpulan data dengan kumpulan tertentu atribut dapat diubah menjadi kumpulan data dengan kumpulan atribut lainnya sama atau berbeda jenis. Sebagai contoh, sebuah atribut, seperti usia, dapat dipartisi ke dalam rentang untuk membuat nilai diskrit untuk kenyamanan analitis.

Proses pembersihan data membutuhkan metode statistik yang biasa digunakan untuk menghilangkan estimasi data. Selain itu, entri data yang salah sering dihapus untuk memastikan lebih banyak hasil penambangan yang akurat. Pemilihan fitur dan transformasi tidak boleh dianggap sebagai bagian dari pre processing data karena fase pemilihan fitur sering kali sangat bergantung pada analitik spesifik masalah yang sedang dipecahkan.

Dalam beberapa kasus, proses pemilihan fitur bahkan dapat terintegrasi dengan erat dengan algoritma atau metodologi tertentu yang digunakan, dalam bentuk pembungkus model atau model tertanam. Namun demikian, fase pemilihan fitur biasanya dilakukan sebelum menerapkan algoritma spesifik yang ada.

Tantangan utama dalam proses ini adalah bahwa setiap aplikasi penambangan data itu unik, dan karenanya sulit untuk membuat teknik umum dan dapat digunakan kembali di berbagai aplikasi. Namun demikian, banyak formulasi data mining berulang kali digunakan dalam konteks aplikasi yang berbeda.

Ini sesuai dengan "masalah super" utama atau blok bangunan dari proses penambangan data. Itu tergantung pada keterampilan dan pengalaman analis untuk menentukan bagaimana ini berbeda formulasi dapat digunakan dalam konteks aplikasi penambangan data tertentu (Joyce Jackson, 2002).

3.3 Basic Tipe Data

Salah satu aspek yang menarik dari proses data mining adalah banyaknya variasi tipe data yang tersedia untuk analisis. Ada dua jenis data yang luas, dengan kompleksitas yang berbeda-beda, untuk proses penambangan data:

Data Berorientasi Non-Ketergantungan

Ini biasanya mengacu pada tipe data sederhana seperti multidimensi data atau data teks. Tipe data ini adalah yang paling sederhana dan paling umum dihadapi. Dalam kasus ini, rekaman data tidak memiliki dependensi yang ditentukan antara item data atau atribut.

Contohnya adalah sekumpulan demografis catatan tentang individu yang berisi usia, jenis kelamin, dan kode pos mereka.

Tabel 3.1: Kumpulan Data Set (Prasetyo, 2014)

Nama	Umur	Jenis kelamin	Alamat	Kode pos
Dian nawangsih	41	Perempuan	Jl anjani 20 semarang	50199
Deni ardianto	43	Laki laki	Yogyakarta	34178
Yuni asmarani	42	Perempuan	Tebet jakarta selatan	80803
Sri rusniati	40	Perempuan	Cepu jawa tengah	43128

Data Berorientasi Ketergantungan

Dalam kasus ini, hubungan implisit atau eksplisit mungkin ada antar item data. Misalnya, kumpulan data jejaring sosial berisi kumpulan simpul (item data) yang dihubungkan bersama oleh sekumpulan sisi (relasi). Di sisi lain, deret waktu berisi dependensi implisit.

Misalnya, dua berturut-turut nilai yang dikumpulkan dari sensor cenderung terkait satu sama lain. Oleh karena itu, atribut waktu secara implisit menentukan ketergantungan antara pembacaan yang berurutan.

Data Biner

Data biner dapat dianggap sebagai kasus khusus dari data kategorikal multidimensi atau data kuantitatif multidimensi. Ini adalah kasus khusus dari kategori multidimensi data, di mana setiap atribut kategori dapat mengambil salah satu dari paling banyak dua nilai diskrit. Ini juga merupakan kasus khusus dari data kuantitatif multidimensi karena ada pemesanan antara kedua nilai tersebut.

Selain itu, data biner juga merupakan representasi dari *setwise* data, di mana setiap atribut diperlakukan sebagai indikator elemen set. Nilai 1 menunjukkan bahwa elemen harus disertakan dalam set.

Data Teks

Data teks dapat dilihat sebagai string, atau sebagai data multidimensi, tergantung caranya mereka terwakili. Dalam bentuk mentahnya, dokumen teks sesuai dengan string. Ini adalah sebuah tipe data berorientasi ketergantungan. Setiap string adalah a urutan karakter (atau kata-kata) yang sesuai dengan dokumen.

Namun, dokumen teks jarang direpresentasikan sebagai string. Ini karena sulit untuk langsung menggunakan pemesanan antara kata-kata dengan cara yang efisien untuk aplikasi skala besar, dan keuntungan tambahan memanfaatkan urutan sering terbatas dalam domain teks.

Dalam praktiknya, representasi vektor-ruang digunakan, di mana frekuensi kata-kata masuk dokumen digunakan untuk analisis. Kata-kata juga terkadang disebut sebagai istilah. Dengan demikian, urutan kata yang tepat hilang dalam representasi ini. Frekuensi ini biasanya dinormalisasi dengan statistik seperti panjang dokumen, atau frekuensi kata-kata individu dalam koleksi.

3.4 Pembuangan Outlier dan Normalisasi Data

Titik yang letaknya sangat jauh dari rata-rata variabel random yang biasanya berkorelasi dengan titik tersebut disebut dengan *noise* atau *outlier*. Pengukuran dilakukan pada ambang batas yang diberikan, umumnya nilai berapa kalinya standar deviasi. Untuk variabel random terdistribusi normal, jarak dua kali standar deviasi akan menjangkau 95 persen dari titik dan jarak tiga kali standar deviasi dapat menjangkau 99 persen dari titik. Titik dengan nilai yang sangat berbeda dengan nilai rata-rata maka akan menghasilkan eror yang besar saat pelatihan dan kemungkinan berefek buruk.

Efek ini bisa jadi lebih buruk saat *outlier* merupakan hasil pengukuran *noise*, Jumlah *outlier* biasanya sedikit dan *outlier* ini biasanya dibuang dari data yang diproses. Jika *outlier* ini bukan masalah dan dihasilkan dari distribusi dengan perhitungan yang panjang maka desainer harus memilih fungsi biaya dalam klasifikator yang sangat tidak sensitif terhadap kemungkinan *outlier*. Misalnya kriteria *least square* sangat sensitif terhadap keberadaan *outlier*.

Metode pendekatan seperti statistik, pemeriksaan kerapatan, K Nearest Neighbor adalah metode-metode yang digunakan untuk mendeteksi keberadaan *outlier*. Pada beberapa kasus, *outlier* tidak selalu data dengan perilaku menyimpang yang pada akhirnya harus dibuang tapi kadang malah data yang memang dicari karena perilakunya yang menyimpang itu.

Normalisasi data dalam praktiknya designer sering dihadapkan pada fitur dengan nilai yang terletak dalam jangkauan nilai berbeda. Akibatnya, fitur dengan nilai atau jangkauan yang besar mempunyai pengaruh yang lebih besar dalam fungsi biaya daripada fitur dengan nilai kecil atau jangkauan kecil.

Untuk mendapatkan jangkauan yang sama maka fitur teknik normalisasi dilakukan. Tanpa dilakukan normalisasi bisa jadi fitur X yang akan

mendominasi fungsi biaya pada klasifikator. Setelah dinormalisasi semua fitur akan berada dalam jangkauan yang sama sehingga proporsi pengaruh pada fungsi biaya dalam klasifikator menjadi seimbang.

Cara yang sederhana dan banyak digunakan adalah normalisasi linier untuk cara yang pertama masing-masing fitur dihitung nilai *mean* dan *varian* maka untuk N data yang ada dalam fitur ke –k akan didapatkan. Hasil normalisasi dengan cara tersebut didapatkan fitur yang mempunyai sifat *zero – mean* dan unit *variance*. Teknik linier yang lain adalah dengan menykalakan jangkauan setiap fitur dalam jangkauan.

Selain teknik linear, teknik non linear juga dapat digunakan dalam kasus di mana data bahkan tidak terdistribusi di sekitar rata-rata. Untuk melakukan normalisasi non linier bisa digunakan fungsi non linier seperti logaritma atau sigmoid untuk memetakan dalam interval yang ditentukan. Penskalaan *soft max* yang populer digunakan.

Data Yang Salah

Dalam praktiknya, data tertentu biasanya selalu ada nilai salah atau kosong pada satu atau lebih fitur dari satu atau lebih dari data vektor dalam data keseluruhan. Nilai yang salah ini bisa seharusnya bernilai angka tapi bernilai karakter, atau nilai yang disimpan berada di luar jangkauan nilai yang seharusnya dimasukkan.

Masalah seperti ini bisa terjadi karena banyak penyebab seperti input dari user yang dilakukan sembarangan, data yang didapat dari formulir kuesioner yang biasanya juga tidak diisikan secara lengkap oleh responden, basis data, dan antarmuka aplikasi yang tidak taat integritas data, alat ukur yang sudah tidak standar sehingga memberikan hasil yang salah dan sebagainya.

Berikut beberapa pilihan yang dapat digunakan untuk memberikan perlakuan pada data yang salah:

1. Membuang semua fitur dari vektor (1 vektor berisi beberapa fitur termasuk fitur yang nilainya salah) Pendekatan seperti ini bisa digunakan ketika jumlah vektor (data) yang mempunyai nilai yang salah jumlahnya sedikit dibandingkan dengan vektor lain yang nilai fiturnya ada. Jika tidak seperti itu masalahnya, maka pembuangan vektor akan berpengaruh pada sifat alami dari masalah.

2. Untuk fitur ke–i, hitung rata-rata berdasarkan nilai yang tersedia untuk fitur tersebut. Kemudian hasilnya digunakan untuk mengganti nilai fitur yang salah pada setiap vektor.
3. Vektor yang mempunyai fitur dengan nilai yang salah tidak dibuang. Yang harus dilakukan pada pendekatan ini adalah untuk semua pasangan fitur antara dua vektor *xi* dan *yi* didefinisikan dengan persamaan berikut ini:

$$b_i = \begin{cases} 0, \text{jika kedua nilai fitur xi dan yi ada} \\ 1, \text{jika kedua nilai fitur xi dan yi tidak ada} \end{cases}$$
$$i = 1,2.....,r$$

Pengamatan setiap fitur yang dibangkitkan secara independen dengan menguji kemampuan diskriminasinya pada masalah yang harus diselesaikan merupakan rangkaian proses sederhana untuk memilih fitur. Walaupun pengamatan setiap fitur dinilai kurang maksimal tetapi cara ini mampu membuang fitur yang jelek dan menguji fitur-fitur yang dipilih dengan metode berikutnya. Saat kita mendapatkan fitur yang besar diskriminasinya maka komputasi menjadi lebih murah ketika proses algoritma.

3.5 Tipe-Tipe Fitur

Elemen pada tiap vektor mengandung fitur yang beraneka ragam, misalnya pada data yang mengandung informasi tentang bentuk fisik manusia di mana ada berat badan yang nilainya bersifat kuantitatif. Fitur ini adalah fitur numerik yang bisa dibandingkan satu sama lain. Sedangkan fitur warna kulit manusia mempunyai nilai yang bersifat kualitatif karena itu tidak bisa dibandingkan. Fitur dibagi menjadi dua kategori yaitu kategorikal atau kualitatif dan tipe numerik yaitu kuantitatif.

Berikut ini adalah empat sifat penting dalam suatu fitur yaitu antara lain:

1. distinctness, meliputi sama dengan (=) dan tidak sama dengan (≠);
2. order, meliputi lebih kecil (<), lebih kecil atau sama dengan (≤), lebih besar (>), lebih besar atau sama dengan (≥);
3. addition, meliputi penjumlahan (+) dan pengurangan (-);
4. multiplication, meliputi perkalian (*) dan pembagian (/).

Dengan melihat sifat-sifat diatas maka diturunkan empat tipe fitur yaitu nominal, ordinal, interval dan rasio. Tipe fitur kategorikal atau kualitatif terdiri atas nominal dan ordinal. Nilai fitur yang bertipe nominal memberi nilai berupa nama, dengan menggunakan nama ini maka sebuah fitur bisa membedakan diri dengan vektor yang lain (=, ≠) contohnya nomor kartu tanda penduduk, nomor surat izin mengemudi, nomor induk mahasiswa dll.

Sedangkan nilai fitur yang bertipe ordinal mempunyai nilai berupa nama yang mempunyai informasi yang terurut (<, ≤,>,≥) misalnya suhu yang meliputi (dingin, normal dan panas) atau predikat kelulusan (cum laude, sangat memuaskan, memuaskan)

Selanjutnya adalah tipe numerik atau kuantitatif, dibagi menjadi dua yaitu fitur interval dan fitur rasio. Tipe fitur interval di mana nilai fitur ada perbedaan di antara dua nilai yang mengandung arti contoh tanggal, suhu (dalam Celsius atau Fahrenheit) selanjutnya tipe rasio yaitu nilai fitur di mana ada perbedaan di antara dua nilai dan rasio dua nilai yang mempunyai makna berarti (*,/) contohnya tinggi, panjang, rata rata, dll

Fitur nominal dan ordinal yang masuk kategorikal dan nilainya kualitatif contohnya kode pos atau nomor KTP. Nilai ini adalah nilai simbolik sehingga tidak bisa dilakukan operasi aritmetika seperti halnya nilai numerik. Sedangkan fitur interval dan rasio adalah jenis numerik yang nilainya kuantitatif. Di mana nilai ini bisa dilakukan operasi aritmetika dan bisa dipresentasikan dengan nilai integer atau kontinu.

Fitur itu sendiri masih bisa dibagi lagi menjadi diskrit dan kontinu berdasarkan angka nilainya. Jika memiliki nilai dalam jumlah himpunan yang terbatas maka fitur tersebut dapat bernilai diskrit. Fitur ini ditemukan pada fitur kategorikal yang hanya memiliki beberapa nilai variasi, contohnya suhu. Suhu hanya ada tiga pilihan normal, panas dan dingin.

Contoh lain yang bisa kita ambil adalah jenis kelamin manusia, yang hanya terdiri atas dua pilihan saja pria atau wanita. Nilai ini sering direpresentasikan dengan nilai biner misalnya ya/tidak, benar /salah, pria/wanita dll.

Sementara itu fitur yang bernilai kontinu akan memiliki jangkauan real. Variabel panjang, tinggi atau berat akan memakai nilai real, tetapi tetap menggunakan ukuran presisi jumlah angka di belakang koma.

Bab 4

Eksplorasi Data

4.1 Pendahuluan

Ada beberapa definisi eksplorasi data dalam data mining. Menurut Han dan Kamber (2018), eksplorasi data adalah proses untuk memperoleh pemahaman yang lebih dalam tentang data dengan menggali pengetahuan yang tersembunyi dalam data secara interaktif dan visual. Hal ini dilakukan untuk menemukan pola yang menarik, anomali, dan informasi yang berguna dalam data.

Menurut Fayyad, Piatetsky-Shapiro, and Smyth (2014), eksplorasi data adalah proses interaktif dan visual untuk menemukan pola dan informasi penting dalam data yang belum diketahui sebelumnya. Tujuan utama dari eksplorasi data adalah untuk menemukan pola dan hubungan antar variabel dalam data.

Menurut Witten and Frank (2016), eksplorasi data adalah proses untuk memperoleh pemahaman tentang data yang tidak diketahui sebelumnya. Tujuan utama dari eksplorasi data adalah untuk menemukan pola dan informasi yang berguna dalam data, dan membantu dalam mempersiapkan data untuk analisis lanjutan.

Menurut Tan, Steinbach, and Kumar (2018), eksplorasi data adalah proses untuk menemukan pola, struktur, dan hubungan dalam data dengan menggunakan teknik analisis statistik dan visualisasi data. Eksplorasi data

bertujuan untuk menghasilkan wawasan dan pengetahuan baru dari data. Menurut Caruana et al. (2015), eksplorasi data adalah proses untuk mengidentifikasi pola dan hubungan dalam data menggunakan teknik visualisasi data dan analisis statistik. Tujuan utama dari eksplorasi data adalah untuk menemukan informasi penting dalam data dan membantu dalam mengembangkan model prediksi yang akurat.

Secara umum, eksplorasi data dalam data mining adalah proses interaktif dan visual untuk menemukan pola dan informasi yang berguna dalam data yang belum diketahui sebelumnya. Tujuan utama dari eksplorasi data adalah untuk menghasilkan wawasan dan pengetahuan baru dari data, dan membantu dalam mempersiapkan data untuk analisis lanjutan atau pengembangan model prediksi yang akurat.

Tujuan dan Manfaat Eksplorasi Data

Tujuan utama eksplorasi data dalam data mining adalah untuk menemukan pola dan hubungan yang tersembunyi dalam data yang dapat membantu dalam pengambilan keputusan yang lebih baik dan efektif. Dalam eksplorasi data, data mining dilakukan dengan menggunakan teknik dan algoritma untuk mengidentifikasi pola dan hubungan dalam data, serta untuk memahami karakteristik dan tren yang muncul dalam data (Han dan Kamber, 2018).

Selain itu, eksplorasi data juga dapat membantu dalam memvalidasi hipotesis dan mengidentifikasi aspek penting yang terkait dengan suatu fenomena atau masalah yang sedang diteliti. Dalam hal ini, eksplorasi data dapat digunakan dalam berbagai bidang, termasuk bisnis, ilmu pengetahuan, kedokteran, dan lain sebagainya. Dengan menggali wawasan baru dari data, eksplorasi data dapat memberikan manfaat besar dalam meningkatkan kualitas dan efektivitas pengambilan keputusan.

Manfaat dari eksplorasi data antara lain:

1. Mengidentifikasi pola dan tren
   Eksplorasi data dapat membantu kita mengidentifikasi pola dan tren dalam data yang mungkin tidak terlihat pada pandangan pertama. Hal ini dapat membantu kita memahami lebih baik tentang data dan membuat keputusan yang lebih baik berdasarkan informasi yang didapat.

2. Mengurangi kesalahan
   Eksplorasi data dapat membantu mengurangi kesalahan dalam data dengan memeriksa nilai yang hilang atau data yang tidak valid, sehingga meningkatkan kualitas data dan akurasi analisis.
3. Mendukung pengambilan keputusan
   Eksplorasi data dapat membantu dalam pengambilan keputusan dengan memberikan informasi yang lebih jelas tentang data yang dianalisis. Hal ini dapat membantu kita membuat keputusan yang lebih baik dan lebih tepat.
4. Meningkatkan efisiensi dan efektivitas
   Eksplorasi data dapat membantu meningkatkan efisiensi dan efektivitas dengan mengidentifikasi area yang memerlukan perhatian khusus, sehingga kita dapat fokus pada area tersebut dan menghasilkan hasil yang lebih baik.
5. Menemukan peluang bisnis
   Eksplorasi data dapat membantu menemukan peluang bisnis baru dengan mengidentifikasi tren pasar, kebiasaan konsumen, atau perilaku pelanggan yang dapat digunakan untuk mengembangkan produk atau layanan baru.
6. Meningkatkan pengembangan produk
   Eksplorasi data dapat membantu meningkatkan pengembangan produk dengan mengidentifikasi fitur atau fungsi produk yang paling disukai oleh pelanggan. Informasi ini dapat digunakan untuk membuat produk yang lebih menarik bagi pelanggan dan meningkatkan kepuasan pelanggan.
7. Meningkatkan pelayanan pelanggan
   Eksplorasi data dapat membantu meningkatkan pelayanan pelanggan dengan mengidentifikasi masalah yang paling sering dikeluhkan oleh pelanggan. Informasi ini dapat digunakan untuk mengembangkan strategi untuk meningkatkan layanan pelanggan dan mengurangi tingkat keluhan pelanggan.

8. Mengoptimalkan operasi bisnis
   Eksplorasi data dapat membantu mengoptimalkan operasi bisnis dengan mengidentifikasi area di mana biaya dapat dikurangi atau efisiensi dapat ditingkatkan. Hal ini dapat membantu meningkatkan keuntungan bisnis dan membuat operasi bisnis lebih efisien.
9. Meningkatkan prediksi
   Eksplorasi data dapat membantu meningkatkan prediksi dengan mengidentifikasi pola dan tren dalam data historis. Hal ini dapat digunakan untuk membuat prediksi yang lebih akurat tentang tren masa depan dan membantu membuat keputusan yang lebih baik.
10. Mendukung penelitian dan pengembangan
    Eksplorasi data dapat digunakan untuk mendukung penelitian dan pengembangan dengan membantu mengidentifikasi area yang memerlukan penelitian lebih lanjut atau pengembangan produk baru. Hal ini dapat membantu meningkatkan inovasi dan memperkuat posisi perusahaan di pasar.

Tantangan Eksplorasi Data

Eksplorasi data adalah proses yang kompleks dan rumit yang melibatkan analisis dan visualisasi data untuk menemukan pola dan informasi yang berguna. Namun, eksplorasi data juga dapat menghadapi berbagai tantangan yang dapat mempengaruhi kualitas analisis dan kesimpulan yang dihasilkan (Witten dan Frank, 2016).

Salah satu tantangan utama dalam eksplorasi data adalah keterbatasan data. Data yang tidak lengkap atau tidak terstruktur dapat mempersulit proses eksplorasi data. Selain itu, kompleksitas data yang tinggi juga dapat menjadi tantangan karena data dapat terdiri dari banyak variabel yang saling terkait dan sulit untuk diinterpretasikan secara visual.

Selain itu, eksplorasi data juga dapat menghadapi tantangan dalam mengelola dan memproses data dalam skala yang besar. Proses eksplorasi data dapat memakan waktu dan sumber daya yang banyak, terutama jika data yang digunakan sangat besar. Oleh karena itu, eksplorasi data memerlukan teknik dan alat yang tepat untuk mempermudah proses analisis dan meminimalkan kesalahan dalam mengambil kesimpulan (Peng et al, 2021).

Berikut beberapa tantangan dalam eksplorasi data:

1. Kualitas data yang buruk
   Kualitas data yang buruk seperti nilai yang hilang, duplikat, atau tidak valid dapat menyulitkan eksplorasi data. Kualitas data yang buruk dapat menghasilkan hasil analisis yang tidak akurat atau tidak dapat diandalkan.
2. Keterbatasan teknologi dan sumber daya
   Eksplorasi data dapat memerlukan sumber daya yang besar seperti perangkat keras, perangkat lunak, dan tenaga ahli untuk menghasilkan hasil yang efektif dan efisien. Keterbatasan teknologi dan sumber daya dapat menjadi hambatan bagi eksplorasi data yang efektif.
3. Data yang tidak terstruktur
   Data yang tidak terstruktur seperti gambar, video, atau teks dapat menjadi sulit untuk dianalisis dan diinterpretasikan. Eksplorasi data pada data yang tidak terstruktur memerlukan teknik khusus dan alat yang sesuai.
4. Keragaman data
   Keragaman data dapat menjadi tantangan dalam eksplorasi data karena data yang berbeda memerlukan pendekatan analisis yang berbeda. Keragaman data dapat memerlukan kombinasi teknik analisis yang berbeda untuk menghasilkan hasil yang akurat dan dapat diandalkan.
5. Privasi dan keamanan data
   Privasi dan keamanan data dapat menjadi tantangan dalam eksplorasi data karena perlindungan data sensitif menjadi penting. Hal ini dapat membatasi akses data dan menghalangi eksplorasi data yang efektif.
6. Penemuan pola acak
   Eksplorasi data dapat menghasilkan banyak pola acak yang tidak bermakna atau tidak relevan. Pola acak ini dapat membingungkan dan memakan waktu dalam proses eksplorasi data.

7. Kesulitan dalam menginterpretasikan hasil
   Hasil eksplorasi data dapat sulit diinterpretasikan dan dipahami oleh pengguna yang tidak ahli dalam analisis data. Hal ini dapat menghasilkan hasil yang tidak efektif atau bahkan keliru.
8. Pilihan metode analisis
   Pilihan metode analisis yang tepat dapat menjadi tantangan dalam eksplorasi data. Metode analisis yang salah dapat menghasilkan hasil yang tidak akurat atau tidak relevan.

4.2 Persiapan Data

Pengumpulan data untuk eksplorasi data adalah proses mengumpulkan data yang akan digunakan untuk analisis lebih lanjut. Data dapat diperoleh dari berbagai sumber seperti survei, basis data, atau bahkan internet. Sebelum pengumpulan data dilakukan, perlu dilakukan perencanaan terlebih dahulu untuk menentukan jenis data yang akan dikumpulkan, metode pengumpulan yang akan digunakan, serta ukuran sampel yang dibutuhkan.

Setelah data terkumpul, langkah selanjutnya adalah memeriksa dan membersihkan data dari nilai yang hilang atau anomali sebelum melakukan analisis lebih lanjut. Proses pengumpulan data yang baik dan lengkap akan memastikan bahwa analisis eksplorasi data yang dilakukan berjalan dengan efektif dan memberikan hasil yang akurat. Selanjutnya persiapan data.

Persiapan data dalam eksplorasi data adalah proses mempersiapkan data mentah untuk analisis. Hal ini penting untuk memastikan data siap untuk dianalisis dan menghasilkan hasil yang akurat dan dapat diandalkan (Xu et al, 2018).

Beberapa langkah dalam persiapan data dalam eksplorasi data antara lain:

1. Pembersihan data
   Langkah pertama dalam persiapan data adalah membersihkan data dari nilai yang hilang, duplikat, atau tidak valid. Data yang tidak bersih dapat menghasilkan hasil analisis yang tidak akurat atau tidak dapat diandalkan.

2. Integrasi data
   Jika data berasal dari berbagai sumber, langkah selanjutnya adalah mengintegrasikan data. Integrasi data memastikan bahwa semua data tersedia dan terintegrasi dengan baik dalam satu set data yang dapat dianalisis.
3. Transformasi data
   Proses transformasi data melibatkan mengubah data mentah menjadi bentuk yang dapat dianalisis dengan mudah. Contoh transformasi data adalah mengubah data dalam format teks ke dalam bentuk angka, normalisasi data, dan penghapusan atribut yang tidak relevan.
4. Pemilihan fitur
   Pemilihan fitur melibatkan memilih atribut yang relevan dari data untuk dianalisis. Hal ini membantu memfokuskan eksplorasi data pada fitur yang paling relevan untuk tujuan analisis.
5. Pemilihan teknik analisis
   Pemilihan teknik analisis melibatkan memilih teknik analisis yang paling sesuai untuk data dan tujuan analisis. Contoh teknik analisis termasuk regresi, klasifikasi, dan klastering.
6. Pembagian data
   Pembagian data melibatkan membagi data menjadi subset yang lebih kecil untuk tujuan analisis. Hal ini memungkinkan analisis data yang lebih efisien dan efektif.
7. Pemfilteran data
   Pemfilteran data melibatkan memfilter data untuk fokus pada subset data yang relevan untuk tujuan analisis. Hal ini membantu meminimalkan pengaruh data yang tidak relevan pada hasil analisis.
8. Pemeriksaan kualitas data
   Pemeriksaan kualitas data melibatkan memastikan bahwa data yang digunakan dalam eksplorasi data berkualitas tinggi dan dapat diandalkan. Hal ini membantu meminimalkan kesalahan dan membuat hasil analisis lebih akurat dan dapat diandalkan.

4.3 Teknik Eksplorasi Data

Teknik eksplorasi data (data exploration) merupakan proses penemuan atau eksplorasi data yang berguna dalam mengungkap pola-pola yang terdapat dalam data mentah. Tujuan dari eksplorasi data adalah untuk mengidentifikasi pola atau hubungan yang tidak diketahui sebelumnya, memahami sifat data, dan membuat hipotesis yang dapat diuji dengan teknik analisis yang lebih lanjut.

Berikut adalah pembahasan tentang beberapa teknik eksplorasi data yang digunakan dalam analisis data:

Visualisasi Data

Visualisasi data merupakan salah satu teknik eksplorasi data yang paling umum digunakan dalam analisis data. Teknik ini melibatkan representasi visual data dalam bentuk grafik atau diagram. Visualisasi data dapat membantu pengguna untuk memahami data dengan lebih baik, mengungkap pola-pola yang tidak terlihat dalam tabel atau angka, dan memudahkan dalam pengambilan keputusan.

Contoh visualisasi data adalah diagram batang, diagram garis, diagram lingkaran, dan diagram scatter. Teknik ini dapat membantu pengguna untuk memahami hubungan antara variabel, melihat perubahan dalam data, dan mengidentifikasi pola atau anomali dalam data.

Regresi

Regresi adalah teknik eksplorasi data yang digunakan untuk memahami hubungan antara satu atau lebih variabel independen dan variabel dependen. Regresi dapat membantu pengguna untuk memprediksi nilai variabel dependen berdasarkan variabel independen. Regresi dapat digunakan untuk mengidentifikasi pola dalam data dan memahami bagaimana perubahan dalam satu variabel akan mempengaruhi variabel lain.

Klastering

Klastering adalah teknik eksplorasi data yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang serupa. Teknik ini dapat membantu pengguna untuk mengidentifikasi pola dalam data dan memahami hubungan antara variabel.

Klastering dapat digunakan untuk memahami perilaku pelanggan, mengelompokkan produk berdasarkan sifat atau karakteristik, dan mengelompokkan pasien berdasarkan gejala atau penyakit.

Klasifikasi

Klasifikasi adalah teknik eksplorasi data yang digunakan untuk mengklasifikasikan data ke dalam kategori yang diberikan. Teknik ini dapat membantu pengguna untuk memahami hubungan antara variabel dan mengidentifikasi pola dalam data. Klasifikasi dapat digunakan untuk mengklasifikasikan pelanggan berdasarkan preferensi, mengklasifikasikan produk berdasarkan sifat atau karakteristik, dan mengklasifikasikan pasien berdasarkan gejala atau penyakit.

Association Rule Mining

Association Rule Mining adalah teknik eksplorasi data yang digunakan untuk menemukan hubungan yang tersembunyi antara item atau produk yang dibeli bersamaan. Teknik ini dapat membantu pengguna untuk memahami hubungan antara variabel dan mengidentifikasi pola dalam data. Association Rule Mining dapat digunakan untuk meningkatkan penjualan dan mengoptimalkan persediaan produk.

Pemilihan metode eksplorasi data yang tepat juga penting untuk mengatasi tantangan dalam eksplorasi data. Pemilihan metode harus didasarkan pada tujuan dan karakteristik data yang akan dianalisis. Beberapa teknik eksplorasi data dapat digunakan bersamaan untuk menghasilkan hasil yang lebih akurat dan dapat dipercaya.

Secara keseluruhan, teknik eksplorasi data sangat penting dalam analisis data karena dapat membantu pengguna untuk mengungkap pola-pola yang terdapat dalam data, memahami sifat data, dan membuat hipotesis yang dapat diuji dengan teknik analisis yang lebih lanjut. Dalam melakukan eksplorasi data, diperlukan persiapan data yang baik dan pemilihan metode eksplorasi data yang tepat untuk mengatasi tantangan dalam eksplorasi data.

Dengan menggunakan teknik eksplorasi data yang tepat, pengguna dapat menghasilkan hasil yang akurat dan dapat dipercaya untuk pengambilan keputusan yang lebih baik.

4.4 Validasi Hasil Eksplorasi Data

Validasi hasil eksplorasi data adalah proses penting dalam analisis data yang memastikan bahwa hasil yang dihasilkan dari eksplorasi data adalah akurat dan dapat diandalkan. Hal ini dilakukan untuk menghindari kesalahan interpretasi atau kesimpulan yang salah dari data. Validasi hasil eksplorasi data adalah proses yang digunakan untuk memeriksa apakah hasil eksplorasi data yang dihasilkan secara kualitatif dan kuantitatif benar-benar mencerminkan data yang asli atau tidak.

Ada beberapa teknik validasi data yang dapat digunakan seperti:

1. Cross-validation adalah teknik validasi data yang membagi data menjadi beberapa bagian dan menggunakan beberapa bagian data sebagai data pelatihan dan sisa bagian sebagai data pengujian. Ini digunakan untuk menghindari overfitting pada data, di mana model terlalu cocok dengan data pelatihan tetapi tidak dapat digunakan untuk memprediksi data yang belum dilihat sebelumnya.
2. Hold-out validation adalah teknik validasi data yang membagi data menjadi dua bagian yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model dan data pengujian digunakan untuk menguji model. Teknik ini digunakan ketika jumlah data yang tersedia terbatas.
3. Bootstrapping adalah teknik validasi data yang menghasilkan banyak sampel data acak dari data yang tersedia. Setiap sampel digunakan untuk melatih model dan kemudian menguji model pada data yang belum dilihat sebelumnya. Teknik ini digunakan ketika jumlah data yang tersedia sangat terbatas dan tidak dapat dibagi menjadi data pelatihan dan data pengujian.

Validasi hasil eksplorasi data sangat penting dalam analisis data karena dapat memastikan bahwa hasil yang dihasilkan dari eksplorasi data benar-benar mencerminkan data yang asli. Hal ini memungkinkan pengambilan keputusan yang lebih akurat dan dapat diandalkan. Selain itu, validasi hasil eksplorasi data dapat membantu menghindari kesalahan interpretasi atau kesimpulan yang salah dari data, yang dapat memiliki konsekuensi serius.

Dalam kesimpulan, validasi hasil eksplorasi data adalah proses penting dalam analisis data yang memastikan bahwa hasil yang dihasilkan dari eksplorasi data adalah akurat dan dapat diandalkan. Hal ini dilakukan untuk menghindari kesalahan interpretasi atau kesimpulan yang salah dari data.

Ada beberapa teknik validasi data yang dapat digunakan seperti *cross-validation, hold-out validation*, dan *bootstrapping*. Validasi hasil eksplorasi data sangat penting dalam analisis data karena dapat memastikan pengambilan keputusan yang lebih akurat dan dapat diandalkan.

4.5 Aplikasi Eksplorasi Data

Eksplorasi data adalah proses yang digunakan untuk memahami dan menganalisis data secara lebih dalam. Dalam era digital yang semakin maju seperti sekarang, eksplorasi data menjadi semakin penting untuk membantu mengambil keputusan bisnis yang tepat dan efektif (Sajedi dan Ghazi, 2020).

Berikut adalah beberapa aplikasi eksplorasi data yang umum digunakan:

1. Bisnis dan Pemasaran - Eksplorasi data digunakan dalam bisnis dan pemasaran untuk mengidentifikasi tren dan pola dalam data pelanggan dan penjualan. Ini membantu perusahaan untuk menentukan strategi pemasaran dan pengambilan keputusan yang lebih akurat.
2. Kesehatan - Eksplorasi data digunakan dalam bidang kesehatan untuk mengidentifikasi pola dan trend dalam data pasien, dan untuk memperkirakan risiko penyakit dan mencari solusi yang tepat. Hal ini dapat membantu dalam diagnosis dan pengobatan penyakit serta pengembangan obat baru.
3. Keuangan - Eksplorasi data digunakan dalam bidang keuangan untuk mengidentifikasi tren dan pola dalam data pasar dan investasi. Ini membantu investor dan manajer keuangan untuk membuat keputusan investasi yang lebih tepat.
4. Sumber Daya Manusia - Eksplorasi data digunakan dalam sumber daya manusia untuk mengidentifikasi pola dan trend dalam data karyawan dan performa kerja. Ini membantu manajer untuk

menentukan strategi rekrutmen dan penempatan karyawan yang lebih tepat.

5. Pendidikan - Eksplorasi data digunakan dalam pendidikan untuk mengidentifikasi pola dan trend dalam data siswa dan hasil tes. Ini membantu guru dan dosen untuk menentukan metode pengajaran yang lebih efektif dan membuat keputusan yang lebih tepat tentang pendidikan.
6. Transportasi - Eksplorasi data digunakan dalam transportasi untuk mengidentifikasi pola dan trend dalam data lalu lintas dan mobilitas. Ini membantu manajer transportasi untuk membuat keputusan tentang perencanaan jalan dan jalur transportasi yang lebih tepat.

Dalam kesimpulan, eksplorasi data merupakan proses penting dalam dunia digital saat ini. Aplikasi eksplorasi data digunakan di berbagai bidang seperti bisnis, kesehatan, keuangan, sumber daya manusia, pendidikan, dan transportasi untuk mengidentifikasi pola dan trend dalam data dan membantu dalam pengambilan keputusan yang lebih akurat.

Bab 5

Pemodelan Data

5.1 Pendahuluan

Pemodelan data merupakan proses penting dalam dunia teknologi informasi dan sistem informasi. Pemodelan data membantu mendefinisikan, menganalisis, dan mengorganisir kebutuhan data yang diperlukan untuk mendukung proses bisnis dalam suatu organisasi. Dalam konteks ini, pemodelan data bertujuan untuk menciptakan struktur yang jelas dan koheren dari elemen data serta hubungan antara elemen-elemen tersebut. Proses ini memungkinkan pengembang, analis, dan pemangku kepentingan lainnya untuk memahami bagaimana data diorganisir dan bagaimana sistem informasi akan mengolah dan mengelola data tersebut.

Pemodelan data mencakup beberapa tahapan, mulai dari pemodelan konseptual hingga pemodelan fisik. Pemodelan konseptual melibatkan representasi visual dari entitas bisnis dan hubungan antara entitas tersebut, sering kali menggunakan teknik seperti *Entity-Relationship Model* (ER Model) atau *Unified Modeling Language* (UML). Pemodelan konseptual membantu menyederhanakan proses perancangan sistem dan memastikan bahwa kebutuhan bisnis dipahami dan ditangani dengan benar.

Pemodelan data logika merupakan tahap berikutnya, di mana model konseptual diterjemahkan menjadi struktur data yang lebih spesifik dan teknis,

seperti relational model atau model data berbasis objek. Pada tahap ini, model data disesuaikan dengan teknologi basis data yang akan digunakan, dan keputusan mengenai optimalisasi, normalisasi, dan kunci primer dan asing dibuat.

Terakhir, pemodelan data fisik melibatkan perancangan struktur penyimpanan data pada tingkat perangkat keras. Ini mencakup keputusan mengenai penyimpanan, indeks, dan strategi optimasi kinerja dan kapasitas yang diperlukan untuk menjalankan sistem secara efisien.

Secara keseluruhan, pemodelan data merupakan proses krusial yang memungkinkan organisasi untuk mengelola dan mengolah data mereka dengan lebih efektif. Dengan pemodelan data yang baik, organisasi dapat mengoptimalkan penggunaan data, meningkatkan pengambilan keputusan, dan menyediakan layanan yang lebih baik kepada pelanggan dan pengguna.

Definisi dan Tujuan pemodelan data

Pemodelan data adalah proses sistematis yang digunakan untuk mendefinisikan, menggambarkan, dan mengorganisir elemen data dalam suatu sistem informasi atau basis data. Proses ini melibatkan pembuatan model yang merepresentasikan struktur dan hubungan antara elemen data, serta aturan dan batasan yang diterapkan pada data tersebut. Model data yang dihasilkan memudahkan pemahaman tentang bagaimana data diorganisir, disimpan, dan diakses dalam sistem (Hoberman, 2009; Hernandez, 2003).

Beberapa tujuan utama dari pemodelan data meliputi:

1. Mendefinisikan struktur data
   Pemodelan data membantu mengidentifikasi elemen data yang diperlukan oleh sistem, seperti entitas, atribut, dan hubungan. Dengan menetapkan struktur yang jelas dan konsisten, model data memastikan bahwa semua elemen data relevan diidentifikasi dan dikelola dengan baik (Simsion & Witt, 2005).
2. Mendukung proses bisnis
   Pemodelan data memungkinkan organisasi untuk mendefinisikan dan menganalisis kebutuhan data yang diperlukan untuk mendukung proses bisnis mereka. Dengan memahami kebutuhan data, organisasi dapat mengembangkan sistem informasi yang lebih efisien dan efektif (Hoberman, 2009).

3. Mempermudah komunikasi antara pemangku kepentingan
   Model data yang jelas dan koheren memudahkan komunikasi antara anggota tim yang terlibat dalam pengembangan sistem, seperti analis bisnis, pengembang, dan manajer proyek. Model data juga memudahkan komunikasi dengan pemangku kepentingan lainnya, seperti pengguna akhir dan manajemen (Hernandez, 2003).
4. Meningkatkan kualitas dan integritas data
   Pemodelan data yang baik memastikan bahwa data disimpan dan diakses dengan cara yang konsisten dan terstruktur. Hal ini meningkatkan kualitas dan integritas data, memungkinkan organisasi untuk membuat keputusan yang lebih baik berdasarkan informasi yang akurat dan andal (Simsion & Witt, 2005).
5. Memudahkan pengembangan dan pemeliharaan sistem
   Model data yang baik memudahkan pengembangan sistem baru dan perubahan pada sistem yang ada. Dengan pemodelan data yang baik, organisasi dapat dengan mudah mengidentifikasi area yang memerlukan perubahan, mengurangi kompleksitas dalam pengembangan, dan memastikan bahwa perubahan tersebut tidak akan mengakibatkan masalah dalam sistem secara keseluruhan (Hoberman, 2009).

Sejarah dan Perkembangan Pemodelan Data

Pemodelan data telah berkembang seiring dengan evolusi teknologi dan kebutuhan bisnis. Awalnya, pemodelan data terfokus pada sistem basis data hierarkis dan jaringan, yang kemudian berkembang menjadi model relasional yang diperkenalkan oleh E.F. Codd pada tahun 1970.

Seiring waktu, berbagai teknik pemodelan data seperti *Entity-Relationship Model, Object-Oriented Model*, dan *Semantic Model* telah dikembangkan untuk mengatasi kebutuhan yang berbeda.

Jenis-Jenis Pemodelan Data

Ada tiga jenis utama pemodelan data yang mencerminkan tingkatan abstraksi yang berbeda:

1. Pemodelan data konseptual, pemodelan data logika, dan pemodelan data fisik. Pemodelan data konseptual bertujuan untuk merepresentasikan struktur data tingkat tinggi dan hubungan antar entitas.
2. Pemodelan data logika berfokus pada struktur data lebih rinci dan cara data diorganisir dalam sistem.
3. Pemodelan data fisik menggambarkan bagaimana data disimpan dan diakses dalam sistem komputer.

Pemodelan data berperan penting dalam pengembangan sistem informasi, karena membantu tim pengembang dalam merancang dan mengelola struktur data yang efisien dan skalabel. Pemodelan data memastikan bahwa data disimpan dan diakses dengan cara yang optimal dan sesuai dengan kebutuhan bisnis.

Selain itu, pemodelan data memungkinkan tim pengembang untuk mengidentifikasi dan memecahkan masalah dalam desain sistem sebelum implementasi, mengurangi biaya dan waktu pengembangan.

Proses Pemodelan Data

Proses pemodelan data melibatkan beberapa langkah, termasuk pengumpulan kebutuhan, analisis data, perancangan model data, validasi dan verifikasi model, dan implementasi dalam sistem. Pengumpulan kebutuhan melibatkan interaksi dengan pemangku kepentingan untuk memahami kebutuhan bisnis dan teknis.

Analisis data melibatkan evaluasi struktur data yang ada dan identifikasi hubungan antar entitas. Perancangan model data melibatkan penggunaan teknik pemodelan yang sesuai untuk menggambarkan struktur data. Validasi dan verifikasi model memastikan bahwa model data sesuai dengan kebutuhan bisnis dan teknis. Terakhir, model data diimplementasikan dalam sistem informasi untuk mendukung operasi bisnis.

5.2 Pemodelan Data Konseptual

Pemodelan data konseptual adalah sebuah proses yang digunakan untuk merepresentasikan struktur data secara abstrak dengan mengidentifikasi dan mendefinisikan objek, atribut, relasi, dan aturan yang terlibat dalam suatu domain bisnis atau sistem informasi. Pemodelan data konseptual biasanya dilakukan pada tahap awal dalam pengembangan sistem informasi untuk memahami kebutuhan bisnis dan merancang model data yang akan digunakan dalam sistem tersebut.

Menurut Hoffer, George, dan Valacich (2021), pemodelan data konseptual adalah "proses mendefinisikan entitas dan hubungan antar entitas, serta atribut yang terkait dalam sebuah sistem informasi." Pemodelan data konseptual dapat digunakan sebagai dasar untuk pengembangan database, desain sistem informasi, dan pengembangan aplikasi. Model ini biasanya diwakili dalam bentuk diagram *Entity-Relationship* (ER) yang menunjukkan entitas, atribut, dan hubungan antar entitas.

Pendekatan Top-Down dan Bottom-Up

Pendekatan top-down dalam pemodelan data konseptual dimulai dengan pemahaman menyeluruh tentang domain bisnis dan kemudian merinci struktur data yang diperlukan. Sebaliknya, pendekatan bottom-up dimulai dengan mengidentifikasi elemen data yang ada dan kemudian menggabungkannya menjadi struktur data yang lebih tinggi. Kedua pendekatan ini dapat digunakan secara bersamaan untuk menciptakan model data konseptual yang efisien dan efektif.

Pendekatan top-down dan bottom-up adalah dua pendekatan yang umum digunakan dalam pemodelan data. Pendekatan top-down adalah pendekatan yang dimulai dari pemahaman umum tentang organisasi dan memecahnya menjadi bagian yang lebih kecil dan lebih spesifik. Sedangkan pendekatan bottom-up dimulai dari informasi detail atau data yang ada dan membangun pemahaman yang lebih besar dari situ.

Menurut Shelly et al. (2021), pendekatan top-down sering kali digunakan dalam fase awal pemodelan data ketika informasi umum tentang organisasi dan bisnis belum sepenuhnya diketahui. Pendekatan ini membantu mengidentifikasi bagian-bagian penting dari organisasi yang kemudian dapat dipecah menjadi bagian yang lebih kecil dan lebih spesifik. Pendekatan top-

down membantu menentukan model data konseptual yang lebih abstrak dan umum sebelum memperinci informasi lebih lanjut.

Di sisi lain, pendekatan bottom-up sering kali digunakan ketika informasi yang spesifik tentang organisasi dan bisnis sudah diketahui. Pendekatan ini memungkinkan pemodelan data dilakukan dari bawah ke atas dengan membangun model data secara bertahap. Pendekatan bottom-up membantu memastikan bahwa model data memperhitungkan informasi detail yang relevan dan penting.

Entity Relationship Model (ER Model)

ER Model adalah teknik pemodelan data konseptual yang digunakan untuk menggambarkan struktur data dalam domain bisnis. Menurut Connolly dan Begg (2021), entitas dalam ER model merepresentasikan objek-objek atau konsep-konsep dalam dunia nyata, seperti orang, tempat, atau barang. Atribut merepresentasikan informasi khusus tentang entitas, seperti nama orang atau harga barang. Hubungan antar entitas menggambarkan bagaimana entitas tersebut saling berhubungan dalam sistem informasi, seperti hubungan antara pelanggan dan pesanan.

ER model biasanya diwakili dalam bentuk diagram ER yang menunjukkan entitas, atribut, dan hubungan antar entitas. Diagram ER membantu untuk memvisualisasikan struktur data dalam sistem informasi dan memastikan bahwa data tersebut diorganisir dan diintegrasikan dengan baik.

Entitas, Atribut, dan Hubungan

Entitas, atribut, dan hubungan adalah konsep utama dalam pemodelan data. Entitas merepresentasikan objek atau konsep dalam dunia nyata, seperti orang, tempat, atau barang. Atribut merepresentasikan informasi khusus tentang entitas, seperti nama orang atau harga barang. Hubungan menggambarkan bagaimana entitas saling berhubungan dalam sistem informasi.

Menurut Elmasri dan Navathe (2016), entitas dalam pemodelan data dapat didefinisikan sebagai "objek yang memiliki keberadaan yang terpisah dan dapat dibedakan dari objek lain." Atribut merepresentasikan "karakteristik khusus dari entitas", seperti nama, alamat, atau nomor identitas. Hubungan menggambarkan "keterkaitan antara entitas", seperti relasi antara pelanggan dan pesanan dalam sebuah sistem informasi.

Entitas, atribut, dan hubungan biasanya diwakili dalam diagram *Entity-Relationship* (ER), yang memungkinkan pemahaman yang lebih mudah

tentang struktur data dalam sebuah sistem informasi. Diagram ER juga membantu memastikan bahwa data diorganisir dan diintegrasikan dengan baik.

Kardinalitas dan Partisipasi

Cardinality dan partisipasi adalah konsep penting dalam pemodelan data yang membantu menggambarkan hubungan antar entitas dalam sebuah sistem informasi. Cardinality menggambarkan jumlah entitas dalam setiap hubungan, sedangkan partisipasi menggambarkan apakah entitas dalam sebuah hubungan harus terlibat atau tidak.

Menurut Connolly dan Begg (2021), cardinality dapat didefinisikan sebagai "jumlah entitas yang dapat terlibat dalam setiap sisi hubungan". Misalnya, pada hubungan "satu ke banyak", satu entitas di satu sisi hubungan dapat terhubung dengan banyak entitas di sisi lain hubungan. Pada hubungan "satu ke satu", satu entitas di setiap sisi hubungan terhubung dengan satu entitas di sisi lain hubungan.

Partisipasi, di sisi lain, menggambarkan apakah entitas dalam sebuah hubungan harus terlibat atau tidak. Partisipasi dapat didefinisikan sebagai "wajib atau opsionalnya partisipasi entitas dalam sebuah hubungan" (Connolly & Begg, 2021). Misalnya, pada hubungan "satu ke banyak", partisipasi pelanggan dalam pesanan mungkin wajib, artinya setiap pesanan harus memiliki pelanggan yang terkait dengannya.

Pemahaman tentang cardinality dan partisipasi penting dalam pemodelan data karena dapat memengaruhi struktur database dan operasi pengolahan data dalam sistem informasi.

Notasi Chen dan Notasi Crow’s Foot

Notasi Chen dan Notasi Crow's Foot adalah dua jenis notasi yang umum digunakan dalam pemodelan data. Notasi ini digunakan untuk merepresentasikan entitas, atribut, dan hubungan dalam sebuah sistem informasi.

Menurut Elmasri dan Navathe (2016), Notasi Chen menggunakan simbol kotak untuk merepresentasikan entitas, sedangkan atribut dinyatakan dalam bentuk lingkaran dan hubungan dinyatakan dengan menggunakan garis. Notasi Crow's Foot, di sisi lain, menggunakan simbol kaki burung (crow's foot) untuk merepresentasikan kardinalitas dan partisipasi dalam sebuah hubungan.

Notasi Chen dan Crow's Foot memiliki kelebihan dan kekurangan masing-masing. Notasi Chen cenderung lebih mudah dipahami dan lebih sering digunakan untuk pemodelan data konseptual, sedangkan Notasi Crow's Foot lebih cocok untuk pemodelan data logis dan fisik karena lebih detail dalam merepresentasikan kardinalitas dan partisipasi.

Kedua notasi ini memiliki tujuan yang sama, yaitu membantu dalam pemahaman tentang struktur data dalam sebuah sistem informasi dan memastikan bahwa data tersebut diorganisir dan diintegrasikan dengan baik.

Unified Modelling Language (UML)

Unified Modeling Language (UML) adalah bahasa pemodelan yang digunakan untuk menggambarkan dan merancang sistem berbasis objek. UML menggunakan konsep-konsep seperti entitas, atribut, dan hubungan untuk merepresentasikan objek dan hubungannya dalam sebuah sistem informasi.

Menurut Booch et al. (2005), UML adalah "bahasa pemodelan standar yang digunakan untuk menggambarkan sistem berbasis objek dari segi konsep, struktur, dan perilaku." UML menyediakan berbagai jenis diagram, seperti diagram kelas, diagram objek, dan diagram aktivitas, yang digunakan untuk memodelkan struktur dan perilaku sistem secara terpisah.

UML juga memungkinkan pengembang untuk memodelkan sistem dalam berbagai tingkat abstraksi, dari tingkat konseptual hingga tingkat implementasi. Hal ini memudahkan dalam pengembangan dan dokumentasi sistem, serta memudahkan dalam memastikan bahwa data diorganisir dan diintegrasikan dengan baik.

Teknik Validasi dan Evaluasi Model Konseptual

Teknik validasi dan evaluasi model konseptual adalah langkah penting dalam pemodelan data. Validasi bertujuan untuk memastikan bahwa model konseptual memenuhi persyaratan dan kebutuhan bisnis, sedangkan evaluasi bertujuan untuk mengukur kualitas model konseptual yang dibuat.

Menurut Hoffer, George, dan Valacich (2021), teknik validasi dapat mencakup wawancara dengan pemangku kepentingan bisnis, peninjauan ulang model dengan ahli bisnis, dan analisis kasus penggunaan (use case) untuk menguji model dalam situasi nyata. Teknik evaluasi meliputi pengujian fungsional dan pengujian kualitas data untuk memastikan bahwa model konseptual dapat memenuhi kebutuhan bisnis dan menghasilkan data yang akurat.

Teknik validasi dan evaluasi model konseptual sangat penting untuk memastikan bahwa model yang dibuat dapat diimplementasikan dengan baik dalam sistem informasi dan dapat memberikan manfaat yang diharapkan. Tanpa validasi dan evaluasi yang tepat, model konseptual dapat menghasilkan data yang tidak akurat atau tidak memenuhi kebutuhan bisnis.

5.3 Pemodelan Data Logika

Pemodelan data logika adalah proses memodelkan data dalam bentuk yang cocok untuk diproses oleh mesin. Model data logika mencakup struktur data, tipe data, dan aturan yang digunakan untuk memproses data dalam sistem informasi. Pemodelan data logika biasanya digunakan dalam pemrograman, basis data, dan pengembangan aplikasi.

Pemodelan data logika penting dalam pengembangan aplikasi dan basis data karena memungkinkan pengembang untuk memproses data dengan lebih efektif dan efisien. Model data logika memastikan bahwa data diorganisir dan diintegrasikan dengan baik, serta memastikan bahwa data yang dihasilkan oleh sistem informasi akurat dan terpercaya.

Relational Model

Model Relational yang dikenalkan oleh E.F. Codd pada tahun 1970 adalah model data logika yang didasarkan pada teori himpunan matematika dan aljabar relasional. Model ini memodelkan data sebagai kumpulan tabel yang saling terkait dan diorganisir dalam bentuk relasi.

Menurut Codd (1970), model Relational memungkinkan pengguna untuk mengakses data dengan cara yang sederhana dan intuitif melalui operasi relasional, seperti SELECT, JOIN, dan UNION. Model ini memastikan bahwa data diorganisir dengan baik dan konsisten, sehingga memudahkan dalam pengolahan dan pengambilan informasi.

Model Relational juga memungkinkan pengguna untuk menghindari redundansi data dan memastikan integritas data dengan mematuhi kaidah normalisasi. Kaidah normalisasi memastikan bahwa setiap tabel memenuhi syarat dan tidak mengandung informasi yang berlebihan atau berulang.

Menurut Date (2004), relational model didasarkan pada tiga konsep utama: entitas, atribut, dan hubungan antara entitas. Entitas merepresentasikan objek

atau konsep dalam dunia nyata, sedangkan atribut merepresentasikan informasi khusus tentang entitas. Hubungan menggambarkan keterkaitan antara entitas dalam sistem informasi.

Dalam relational model, data diorganisir dalam bentuk tabel (relations) yang memiliki baris (tuple) dan kolom (attribute). Setiap tabel merepresentasikan suatu entitas atau hubungan antara entitas, dan setiap baris dalam tabel merepresentasikan satu entitas atau satu instansi dari hubungan antara entitas.

Relational model juga memungkinkan pengguna untuk menggabungkan tabel menggunakan operasi join, serta mengatur data menggunakan perintah seperti SELECT, INSERT, UPDATE, dan DELETE. Hal ini memudahkan dalam pengolahan data dan menghasilkan informasi yang akurat dan konsisten.

Penerapan model Relational telah membawa perubahan besar dalam pemrosesan data dan pengembangan basis data. Model ini masih menjadi standar industri dan menjadi dasar bagi banyak sistem basis data yang digunakan saat ini.

Kunci Primer (Primary Key) dan Kunci Asing (Foreign Key)

Primary Key dan *Foreign Key* adalah konsep penting dalam pemodelan data logika. Primary Key adalah atribut atau kombinasi atribut yang dapat diidentifikasi sebagai kunci unik untuk setiap baris dalam sebuah tabel. Sedangkan Foreign Key adalah atribut atau kombinasi atribut yang mereferensikan Primary Key dari tabel lain.

Foreign Key, di sisi lain, digunakan untuk membangun hubungan antara tabel dalam basis data relasional. Setiap tabel yang mengandung Foreign Key mereferensikan Primary Key dari tabel lain. Hal ini memungkinkan pengguna untuk menggabungkan informasi dari beberapa tabel dan menghasilkan informasi yang lebih kompleks dan terintegrasi.

Normalisasi dan Bentuk Normal

Normalization adalah proses memastikan bahwa sebuah basis data relasional memenuhi kaidah normalisasi, yaitu proses pemisahan informasi dalam tabel menjadi beberapa tabel yang lebih kecil dan terkait satu sama lain dengan cara yang jelas dan terstruktur. Normal Form adalah aturan atau standar yang digunakan untuk mengukur tingkat normalisasi dari sebuah basis data.

Menurut Date (2004), Normal Form dibagi menjadi beberapa tingkatan, yaitu:

1. First Normal Form (1NF) - Memastikan bahwa setiap atribut dalam sebuah tabel hanya memiliki satu nilai.
2. Second Normal Form (2NF) - Memastikan bahwa setiap non-kunci atribut dalam sebuah tabel bergantung pada seluruh Primary Key dan bukan hanya sebagian saja.
3. Third Normal Form (3NF) - Memastikan bahwa setiap non-kunci atribut dalam sebuah tabel tidak memiliki ketergantungan transitif pada Primary Key.
4. Boyce-Codd Normal Form (BCNF) - Memastikan bahwa setiap atribut dalam sebuah tabel bergantung secara fungsional pada Primary Key.

Normal Form digunakan untuk memastikan bahwa sebuah basis data relasional memenuhi persyaratan tertentu, seperti kebutuhan untuk menghindari redundansi data, menghindari anomali data, dan memastikan bahwa setiap tabel diorganisir dengan baik dan konsisten.

Dengan menerapkan Normal Form, pengguna dapat memastikan bahwa data diorganisir dengan baik dan dapat diakses dengan mudah, serta dapat memastikan bahwa data yang dihasilkan oleh sistem informasi akurat dan konsisten.

5.4 Pemodelan Data Hierarkis

Pemodelan data hierarkis adalah salah satu pendekatan dalam pemodelan data yang digunakan untuk merepresentasikan data dalam bentuk pohon atau struktur berhierarki. Model ini memodelkan data sebagai satu set record atau file yang disusun dalam bentuk hierarki, di mana setiap record memiliki satu atau beberapa record anak.

Pemodelan data hierarkis sering digunakan dalam pengolahan data yang berkaitan dengan ilmu pengetahuan, seperti biologi dan zoologi, di mana hierarki digunakan untuk merepresentasikan klasifikasi dari organisme hidup.

Namun, kelemahan dari pemodelan data hierarkis adalah kurang fleksibel dalam mengakses data dan sulit dalam menangani data yang tidak terstruktur. Oleh karena itu, model ini lebih cocok untuk aplikasi bisnis dengan struktur data yang stabil dan terstruktur.

5.5 Pemodelan Data Jaringan dan Semantik

Pemodelan data jaringan adalah salah satu pendekatan dalam pemodelan data yang digunakan untuk merepresentasikan data dalam bentuk grafik yang terdiri dari kumpulan *node* dan *edge*. Model ini memodelkan data sebagai kumpulan record yang terkait satu sama lain melalui hubungan yang kompleks dan tidak terbatas. Setiap record dalam model jaringan memiliki satu atau lebih set elemen induk atau "owner" dan dapat memiliki satu atau lebih member elemen "member". Setiap set elemen dapat memiliki banyak member elemen dan dapat terhubung dengan set elemen lain dalam model.

Pemodelan data jaringan sering digunakan dalam pengolahan data yang melibatkan hubungan yang kompleks dan banyak-to-many, seperti dalam aplikasi bisnis yang melibatkan kategori produk, pelanggan, dan transaksi. Model ini memungkinkan pengguna untuk mengakses dan memanipulasi data dengan cara yang fleksibel dan dapat diadaptasi dengan cepat terhadap perubahan dalam kebutuhan bisnis.

Namun, kelemahan dari pemodelan data jaringan adalah kompleksitas dalam perancangan dan pengolahan data yang memerlukan keterampilan khusus dalam pemrograman dan analisis data

Pemodelan Data Semantik

Pemodelan data semantik adalah salah satu pendekatan dalam pemodelan data yang digunakan untuk merepresentasikan data dengan memperhatikan makna atau arti dari data tersebut. Model ini memodelkan data dalam bentuk yang dapat dimengerti oleh manusia dan mesin, sehingga memudahkan dalam pemrosesan dan pengambilan informasi.

Menurut Berners-Lee, Hendler, dan Lassila (2001), dalam pemodelan data semantik, data disimpan dalam bentuk grafik (graph) yang terdiri dari *node* dan

edge. Node merepresentasikan entitas dalam sistem informasi, sedangkan edge merepresentasikan hubungan antara entitas tersebut.

Pemodelan data semantik memungkinkan pengguna untuk menambahkan metadata atau informasi tentang makna atau arti dari entitas dan hubungan dalam sistem informasi. Metadata dapat digunakan untuk meningkatkan pemahaman tentang data dan memudahkan dalam integrasi data dari berbagai sumber yang berbeda.

Pemodelan data semantik juga memungkinkan pengguna untuk menggunakan bahasa yang lebih dekat dengan bahasa manusia dalam memproses data, seperti SPARQL (SPARQL Protocol and RDF Query Language), yang merupakan bahasa query untuk mencari informasi dalam data semantik.

Pemodelan data semantik sering digunakan dalam aplikasi yang memerlukan integrasi data dari berbagai sumber, seperti aplikasi web, e-Commerce, dan sistem informasi bisnis. Model ini memungkinkan pengguna untuk mengakses dan memanipulasi data dengan cara yang lebih intuitif dan efektif.

5.6 Pemodelan Data Objek dan Data Fisik

Pemodelan data objek adalah salah satu pendekatan dalam pemodelan data yang digunakan untuk merepresentasikan data dalam bentuk objek yang terdiri dari atribut dan metode. Model ini memodelkan data seperti objek dalam dunia nyata, sehingga lebih mudah dipahami dan diakses oleh pengguna.

Setiap objek dalam model objek memiliki atribut yang merepresentasikan data yang disimpan dalam objek tersebut. Setiap atribut memiliki tipe data dan nilai yang terkait dengan data yang disimpan. Metode dalam model objek merepresentasikan perilaku objek, seperti operasi atau fungsi yang dapat dilakukan pada objek tersebut.

Pemodelan data objek memungkinkan pengguna untuk mengakses dan memanipulasi data dalam bentuk yang lebih mudah dipahami dan intuitif. Model ini juga memungkinkan pengguna untuk mengorganisir data dalam struktur yang kompleks dan berlapis-lapis, sehingga lebih mudah dalam memahami hubungan antara objek.

Pemodelan data objek sering digunakan dalam pengembangan aplikasi yang kompleks, seperti game, simulasi, dan sistem informasi geografis. Model ini memungkinkan pengguna untuk mengembangkan aplikasi yang lebih fleksibel dan efektif, serta dapat diadaptasi dengan cepat terhadap perubahan dalam kebutuhan bisnis.

Pemodelan data fisik adalah proses desain bagaimana data akan disimpan, diakses, dan dikelola dalam sistem informasi. Pemodelan data fisik melibatkan pemilihan teknologi penyimpanan, pengaturan struktur data, dan strategi untuk mengoptimalkan kinerja dan keamanan sistem. Pendekatan pemodelan data fisik melibatkan konversi dari model data logika ke struktur data fisik yang sesuai dengan karakteristik dan batasan teknologi penyimpanan yang digunakan.

Pemodelan data fisik adalah salah satu pendekatan dalam pemodelan data yang digunakan untuk merepresentasikan struktur penyimpanan data fisik pada disk atau media penyimpanan lainnya. Model ini menggambarkan bagaimana data disimpan dalam basis data dan struktur fisik yang dibutuhkan untuk menyimpan data tersebut.

Pemodelan data fisik memperhatikan faktor-faktor teknis dalam penyimpanan data, seperti ukuran blok, indeks, dan algoritma penyimpanan. Model ini memastikan bahwa data tersimpan dengan efisien dan dapat diakses dengan cepat dan akurat.

Pemodelan data fisik sering digunakan dalam perancangan basis data dan pengembangan aplikasi yang memerlukan kinerja penyimpanan data yang tinggi, seperti sistem transaksi online dan sistem pengolahan data besar. Model ini memungkinkan pengguna untuk memperhitungkan faktor teknis dalam perancangan basis data dan memastikan bahwa data tersimpan dengan efisien dan dapat diakses dengan cepat.

Penyimpanan dan Indeks Data

Penyimpanan data adalah proses menyimpan dan mengatur data dalam media penyimpanan, seperti hard disk, solid-state drive, atau sistem penyimpanan terdistribusi. Penyimpanan data melibatkan pengaturan struktur data dalam format yang efisien untuk akses dan manipulasi data. Indeks data adalah struktur tambahan yang dibuat untuk mempercepat pencarian dan akses data dalam sistem. Indeks dapat berupa struktur seperti B-tree, hash, atau bitmap, tergantung pada karakteristik data dan pola akses yang diharapkan.

Optimalisasi Kinerja dan Kapasitas

Optimalisasi kinerja dan kapasitas melibatkan pengaturan dan penyesuaian sistem penyimpanan data untuk memaksimalkan kecepatan, efisiensi, dan skalabilitas. Strategi optimalisasi melibatkan pemilihan teknologi penyimpanan yang sesuai, penggunaan indeks dan partisi, serta pengaturan teknik kompresi dan caching. Optimalisasi kinerja dan kapasitas memastikan bahwa sistem informasi dapat menangani beban kerja yang diharapkan dan meningkatkan kualitas layanan bagi pengguna

Keamanan dan Integritas Data

Keamanan dan integritas data adalah aspek penting dalam pemodelan data fisik. Keamanan data melibatkan perlindungan data dari akses yang tidak sah, manipulasi, atau penghancuran. Strategi keamanan melibatkan penggunaan enkripsi, autentikasi, dan kontrol akses.

Integritas data adalah keandalan dan konsistensi data dalam sistem. Integritas data dijamin melalui penggunaan batasan, pemicuan, dan prosedur yang memastikan bahwa data selalu memenuhi aturan bisnis dan kriteria validitas yang didefinisikan.

Migrasi dan Integrasi Data

Migrasi data adalah proses pemindahan data dari satu sistem penyimpanan ke sistem penyimpanan lain, sementara integrasi data adalah proses penggabungan data dari beberapa sumber menjadi satu sistem informasi.

Proses migrasi dan integrasi data melibatkan perencanaan, pemetaan, dan transformasi struktur data untuk memastikan bahwa data tetap konsisten, relevan, dan mudah diakses di lingkungan yang baru. Migrasi dan integrasi data sering diperlukan dalam proyek penggantian sistem, konsolidasi sistem, atau integrasi aplikasi perusahaan.

5.7 Tantangan Masa Kini dan Masa Depan Pemodelan Data

Pemodelan data adalah proses pemetaan data dari dunia nyata ke dalam model data yang dapat digunakan untuk mengorganisir, menyimpan, dan memanipulasi data secara efektif. Terdapat berbagai pendekatan dalam pemodelan data, seperti pemodelan relasional, hierarkis, jaringan, objek, semantik, dan fisik. Setiap pendekatan memiliki kelebihan dan kelemahan dalam memodelkan data, dan dipilih berdasarkan kebutuhan dan konteks penggunaannya.

Tantangan dalam pemodelan data di masa sekarang dan masa depan terkait dengan volume, kecepatan, dan variasi data yang semakin meningkat. Dalam era Big Data, jumlah data yang dihasilkan dan disimpan semakin besar, sehingga diperlukan pemodelan data yang lebih kompleks dan efisien untuk mengakses dan memproses data tersebut. Selain itu, perkembangan teknologi seperti *Internet of Things* (IoT) dan *Artificial Intelligence* (AI) juga memperkenalkan data dengan karakteristik yang berbeda dan memerlukan pemodelan data yang lebih maju.

Selain itu, tantangan lain dalam pemodelan data adalah masalah privasi dan keamanan data. Seiring dengan meningkatnya kebutuhan akan data, juga semakin meningkat kekhawatiran akan privasi dan keamanan data. Oleh karena itu, pemodelan data harus memperhatikan aspek privasi dan keamanan data untuk memastikan data yang disimpan dan diolah aman dari ancaman dan penyalahgunaan.

Secara keseluruhan, pemodelan data tetap menjadi salah satu bidang yang sangat penting dalam pengembangan sistem informasi dan teknologi di masa sekarang dan masa depan. Pengembangan teknologi dan perkembangan Big Data menuntut pengembangan pemodelan data yang lebih canggih dan efisien untuk mengatasi tantangan dalam memproses data dengan volume yang semakin besar dan kompleksitas yang semakin tinggi.

Selain itu, pemodelan data juga harus memperhatikan aspek privasi dan keamanan data untuk memastikan data yang disimpan dan diolah aman dari ancaman dan penyalahgunaan.

Bab 6

Evaluasi Model

6.1 Pendahuluan

Evaluasi data mining digunakan untuk mengukur ketepatan atau jumlah eror yang ada pada model yang dibangun. Hal ini bertujuan untuk mengetahui seberapa optimal model yang digunakan untuk memecahkan suatu permasalahan. Selain itu bisa juga digunakan untuk menguji keandalan data set yang dimiliki. Apakah data set tersebut relevan dan handal untuk diekstrak pengetahuannya.

Evaluasi data mining ini juga dapat digunakan untuk melakukan komparasi atau perbandingan antara algoritma yang digunakan. Semisal perhitungan estimasi data konsumsi bahan bakar minyak suatu kendaraan dengan menggunakan algoritma Regresi linier, kemudian dicoba dengan model lain menggunakan algoritma neural network.

Dengan evaluasi tersebut bisa dibandingkan algoritma mana yang bisa memberikan model yang handal dan sesuai dengan kebutuhan.

6.2 Model Evaluasi

Evaluasi model data mining merupakan tahap ke empat dari proses data mining. Berikut ini evaluasi model data mining sesuai dengan peran utama:

Estimation (Estimasi)

estimasi evaluasi dalam data mining digunakan untuk mengukur tingkat eror. Biasanya pengukuran tingkat eror dilakukan dengan menggunakan metode *Root Mean Square Error* (RMSE).

Root Mean Square Error (Goel, 2011) adalah metode pengukuran dengan mengukur perbedaan nilai dari prediksi sebuah model sebagai estimasi atas nilai yang diobservasi. RMSE adalah hasil dari akar kuadrat *Mean Square Error*. Keakuratan metode estimasi kesalahan pengukuran ditandai dengan adanya nilai RMSE yang kecil. Metode estimasi yang mempunyai RMSE lebih kecil dikatakan lebih akurat daripada metode estimasi yang mempunyai RMSE lebih besar.

Nilai RMSE dapat dihitung dengan persamaan sebagai berikut:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

Di mana:

At = Nilai data aktual
Ft = Nilai hasil peramalan
N = Banyaknya data
∑ = Summation (jumlah keseluruhan nilai)

Prediction/Forecasting (Prediksi/Peramalan)

Prediksi merupakan suatu proses untuk meramalkan atau memperkirakan suatu variabel di masa yang akan datang. Dalam kasus prediksi biasanya data yang sering digunakan adalah data kuantitatif. Prediksi tidak harus menghasilkan suatu jawaban yang pasti kejadian, melainkan berusaha untuk mencari jawaban yang sedekat mungkin dengan kejadian yang akan terjadi (Putro, Tanzil Furqon dan Wijoyo, 2018).

1. Mean Squared Error (MSE)
   Mean Squared Error adalah Rata-rata kesalahan kuadrat antara nilai aktual dan nilai peramalan. Metode MSE secara umum digunakan untuk mengecek estimasi berapa nilai kesalahan pada peramalan. Nilai MSE yang rendah atau nilai yang mendekati nol, hal ini menunjukkan bahwa hasil peramalan sesuai dengan data aktual dan bisa dijadikan untuk perhitungan peramalan untuk periode mendatang.
   Metode MSE ini biasanya digunakan untuk mengevaluasi metode pengukuran dengan model regresi atau model peramalan seperti *Moving Average, Weighted Moving Average* dan Analisis *Trendline*. Cara menghitung MSE adalah melakukan pengurangan nilai data aktual dengan data peramalan dan hasilnya dikuadratkan (squared) kemudian dijumlahkan secara keseluruhan dan membaginya dengan banyaknya data yang ada.
2. Mean Absolute Percentage Error (MAPE)
   Dalam referensi lain MAPE dikenal juga dengan *Mean Absolute Percentage Deviation* (MAPD) adalah persentase kesalahan rata-rata secara mutlak (absolut). Pengertian MAPE adalah Pengukuran statistik tentang akurasi perkiraan (prediksi) pada metode peramalan. Pengukuran dengan menggunakan MAPE dapat digunakan oleh masyarakat luas karena MAPE mudah dipahami dan diterapkan dalam memprediksi akurasi peramalan.
   Metode MAPE memberikan informasi seberapa besar kesalahan peramalan dibandingkan dengan nilai sebenarnya dari series tersebut. Semakin kecil nilai presentasi kesalahan (percentage error) pada MAPE maka semakin akurat hasil peramalan tersebut.
   Berikut ini adalah kriteria nilai MAPE:
   a. Jika nilai MAPE kurang dari 10% maka kemampuan model peramalan sangat baik.
   b. Jika nilai MAPE antara 10% - 20% maka kemampuan model peramalan baik.

c. Jika nilai MAPE kisaran 20% - 50% maka kemampuan model peramalan layak (cukup baik).
d. Jika nilai MAPE kisaran lebih dari 50% maka kemampuan model peramalan buruk (tidak akurat).

Dari nilai tersebut kita bisa memahami bahwa nilai MAPE masih bisa digunakan apabila tidak melebihi 50%. Ketika nilai MAPE sudah diatas 50% maka model peramalan sudah tidak bisa digunakan lagi. Nilai MAPE dapat dihitung dengan persamaan sebagai berikut:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100\%$$

Di mana:

n = ukuran sampel
A_i = Nilai data Aktual
F_i = Nilai data peramalan

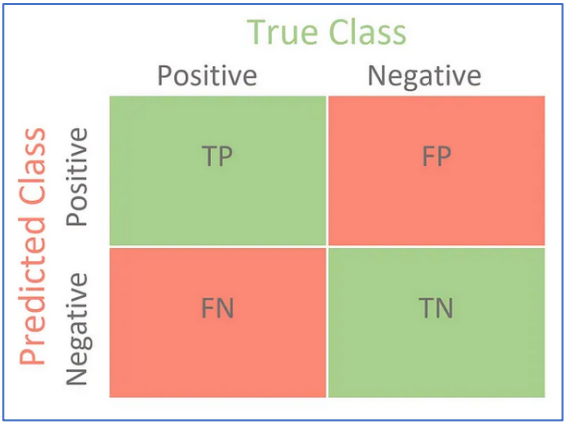
Sebaiknya MAPE tidak digunakan untuk data yang nilainya kecil. Misalnya data yang nilai aktualnya adalah 1, sedangkan hasil peramalannya adalah 2, maka persentase kesalahan absolutnya adalah 1-21=50%. Walaupun sebenarnya ramalan tersebut tidak terlalu meleset, namun persentase kesalahan absolutnya kelihatan sangat besar sehingga bisa saja nilai MAPE nantinya lebih besar dari 100% akibat banyaknya nilai penyebut yang sangat kecil.

Classification

Classification pada dasarnya adalah mengelompokkan data yang tidak diketahui label kelasnya ke dalam sejumlah kelompok tertentu sesuai ukuran kemiripannya.

1. Confusion Matrix: Accuracy
   Confusion matrix juga sering disebut eror matrix adalah matrik yang mendefinisikan hasil perbandingan antara nilai *predicted* dengan nilai aktual (Saifudin dan Wahono, 2015). *Confusion matrix* ini akan merepresentasikan kebenaran dari sebuah prediksi (Wardani, 2020). Pada dasarnya confusion matrix memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil

klasifikasi sebenarnya. Berikut merupakan gambaran *Confusion Matrix* dan terminologi di dalamnya:



Gambar 6.1: Klasifikasi Confusion Matrik (Mohajon, 2020)

Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada confusion matrix. Keempat istilah tersebut adalah *True Positive (TP), True Negative (TN), False Positive (FP)* dan *False Negative (FN)*.

Tabel 6.1: Metode Penilaian Untuk Klasifikasi Confusion Matrik

Representasi	Penjelasan
TP (True Positive)	Jumlah Nilai True diklasifikasikan dengan akurat
TN (True Negative)	Jumlah Nilai False diklasifikasikan secara akurat
FP (False Positive)	Jumlah Nilai False diklasifikasikan sebagai True
FN (False Negative)	Jumlah Nilai True diklasifikasikan sebagai False

Rumus akurasi berdasarkan jumlah total prediksi benar:

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

Rumus perhitungan kinerja klasifikasi:

$$TP = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FN = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$
$$TN = \frac{TN}{Actual\ Negatif} = \frac{TN}{TN + FP}$$
$$FP = \frac{FP}{Actual\ Negatif} = \frac{FP}{TN + FP}$$

Di mana:

TP = True Positif
TN = True Negatif
FP = False Positive
FN = False Negative

2. Kurva ROC (Receiver Operation Characteristic)
   Kurva ROC merupakan teknik untuk memvisualisasikan dan menguji kinerja pengklasifikasian. Kurva ROC mengekspresikan *confusion matrix*. Kurva ROC adalah grafik dua dimensi dengan *false positive* sebagai garis horizontal dan *true positive* untuk mengukur perbedaan performansi metode yang digunakan (Gorunescu, 2011). Model klasifikasi yang lebih baik akan menunjukkan ROC lebih besar.
3. Kurva AUC (Area Under Curve)
   Kurva AUC digunakan untuk menghitung luas daerah di bawah kurva ROC. AUC memiliki nilai antara 0,0 dan 1,0, semakin tinggi AUC maka akan semakin baik. AUC di dekat angka 1 akan semakin baik sedangkan yang mendekati angka 0 maka modelnya tidak baik. Untuk perhitungannya dapat didefinisikan dengan persamaan:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

   Menurut Gorunescu (2011) performa keakurasian AUC dapat diklasifikasikan menjadi lima kelompok, yaitu:
   - 0,90 – 1,00 = sangat baik (excellent classification)
   - 0,80 – 0,90 = baik (good classification)
   - 0,70 – 0,80 = sama (fair classification)
   - 0,60 – 0,70 = rendah (poor classification)
   - 0,50 – 0,60 = gagal (failure)

Clustering

Clustering adalah proses pembagian data ke dalam kelas atau cluster berdasarkan tingkat kesamaannya. Clustering merupakan pekerjaan yang memisahkan data atau vektor ke dalam sejumlah kelompok atau cluster menurut karakteristiknya masing-masing. Data-data yang memiliki kemiripan karakteristik akan berkumpul dalam kelompok atau cluster yang sama. Data-data yang memiliki perbedaan karakteristik, akan berkumpul dalam kelompok atau cluster yang berbeda.

Tujuan utama dari metode clustering adalah pengelompokan sejumlah data atau objek ke dalam cluster (group) sehingga dalam setiap cluster akan berisi data yang semirip mungkin. objek-objek data ini biasanya direpresentasikan sebagai sebuah titik dalam ruang multidimensi (Fajar, 2013).

Berikut ini adalah salah satu metode evaluasi internal yang dapat digunakan dalam menentukan evaluasi cluster:

1. DaviesBouldin Index (DBI)
   Davies Bouldin Index merupakan salah satu metode evaluasi internal yang mengukur evaluasi cluster pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap centroid dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar centroid dari clusternya.
   Tahapan dari perhitungan Davies Bouldin Index adalah sebagai berikut:
   a. Sum of Square Within-Cluster (SSW)
      Untuk mengetahui kohesi dalam sebuah cluster ke-i adalah dengan menghitung nilai dari SSW. Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari sebuah cluster yang diikuti. Persamaan yang digunakan untuk memperoleh nilai Sum of Square Within cluster adalah sebagai berikut:

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i)$$

Di mana:

SSW_i = minimal jarak antara titik cluster ke-i
$d(x_j,c_i)$ = jarak antara data ke-j dengan centroid cluster ke-i
m = jumlah yang berada dalam cluster ke-i

b. Sum of Square Between-Cluster (SSB)
Perhitungan SSB bertujuan untuk mengetahui separasi antar cluster. Persamaan yang digunakan untuk menghitung nilai SSB adalah sebagai berikut:

$$SSB_{i,j} = d(c_i,c_j)$$

Di mana:

SSB_{ij} = maksimal jarak di antara cluster c_i, dan c_j

dc_i,c_j = jarak antara cluster ke-*i* dengan cluster ke-*j*

c. Rasio
Bertujuan untuk mengetahui nilai perbandingan antara cluster ke-*i* dan cluster ke-*j*. Untuk menghitung nilai rasio yang dimiliki oleh masing-masing cluster, digunakan persamaan berikut:

$$R_{i,j} = \frac{SSW_1 + SSW_j}{SSB_{i,j}}$$

Di mana:

R_{ij} = rasio nilai perbandingan antara cluster ke-*i* dan cluster ke-*j*

SSW_i = minimal jarak antara titik cluster ke-*i*

SSW_j = minimal jarak antara titik cluster ke-*j*

SSW_{ij} = maksimal jarak di antara cluster c_i dan c_j

d. Davies Bouldin Indeks
Nilai rasio yang diperoleh akan digunakan untuk mencari nilai Davies Bouldin Index (DBI) dengan menggunakan persamaan berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\,(R_{i,j})$$

Di mana:

DBI = Nilai skalar evaluasi DBI

R_{ij} = Rasio nilai perbandingan antara cluster ke-*i* dan cluster ke-*j*

k = Jumlah cluster yang digunakan

Dari persamaan tersebut, *k* adalah jumlah cluster. Semakin kecil nilai Davies Bouldin Index (DBI) yang diperoleh (non-negatif >= 0), maka semakin baik cluster yang diperoleh dari pengelompokan menggunakan algoritma clustering (Bates, A. & Kalita, 2016)

Association

Analisis asosiasi dikenal juga sebagai *market basket analysis* merupakan salah satu teknik data mining untuk menemukan aturan asosiatif antara kombinasi item dengan item lainnya. Analisis ini sering dipakai untuk menganalisis isi keranjang belanja konsumen dalam suatu pasar swalayan.

Contoh penerapan dari aturan asosiatif adalah analisa pembelian produk pada sebuah toko alat tulis, pada analisa itu misalkan dapat diketahui berapa besar kemungkinan seorang pelanggan membeli pensil bersamaan dengan membeli penghapus. Penerapan aturan asosiasi dalam kasus tersebut dapat membantu pemilik toko untuk mengatur penempatan barang, mengatur persediaan atau membuat promosi pemasaran dengan menerapkan diskon untuk kombinasi barang tertentu. (Adie Wahyudi Oktavia Gama, Ketut Gede Darma Putra, 2016).

Analisis asosiasi didefinisikan sebagai suatu proses untuk menemukan semua aturan asosiasi yang memenuhi syarat minimum untuk support (minimum support) dan syarat minimum untuk *confidence* (minimum confidence) (Ulumuddin dan Juanita, 2018). Salah satu tahap analisis asosiasi yang menarik perhatian banyak peneliti untuk menghasilkan algoritma yang efisien adalah analisis pola frekuensi tinggi (frequent pattern mining).

1. Support adalah nilai penunjang atau persentase kombinasi sebuah item dalam database. Rumus support sebagai berikut:

$$Support\ (A) = \frac{jumlah\ transaksi\ mengandung\ A}{total\ transaksi} \times 100\%$$

Sementara itu, untuk mencari nilai support dari 2 item diperoleh dari rumus berikut:

$$Support\ (A, B) = P\ (A \cap B)$$

$$Support\ (A,B) = \frac{jumlah\ transaksi\ mengandung\ A\ dan\ B}{total\ transaksi} \times 100\%$$

4. Confidence adalah nilai kepastian yaitu kuatnya hubungan antar item dalam aturan asosiasi. Confidence bisa dicari setelah pola frekuensi munculnya sebuah item ditemukan. Misalkan ditemukan aturan A → B maka:

$$Confidence\ (A,B) = \frac{jumlah\ transaksi\ mengandung\ A\ dan\ B}{Jumlah\ transaksi\ mengandung\ A} \times 100\%$$

5. Lift Ratio

   Lift ratio adalah salah satu cara penghitungan yang lebih baik untuk melihat kuat tidaknya aturan asosiasi. Untuk mencari nilai confidence dapat dihitung dengan rumus (Fitriyanto, 2017):

$$Expected\ Confidence = \frac{Transaksi\ yang\ mengandung\ Support\ B}{Total\ transaksi}$$

   Lift ratio dapat dihitung dengan cara membandingkan antara confidence untuk suatu aturan dibagi dengan *expected confidence*. Berikut rumus dari lift ratio (Fitriyanto, 2017):

$$Lift\ Ratio = \frac{Confidence}{Expected\ Confidence}$$

6. Precision

   Precision merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Precision dapat juga diartikan sebagai tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

   Berikut adalah rumus untuk menghitung precision:

$$P = \frac{TP}{TP + FP}$$

   Di mana:

   P = Precision

   TP = True Positif

   FP = False Positive

7. Recall

   Recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Recall juga dapat didefinisikan

sebagai tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Berikut adalah rumus untuk menghitung recall:

$$R = \frac{TP}{TP + FN}$$

Di mana:

R = Recall

TP = True Positif

FN = False Negative

8. Specificity

Merupakan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif. Berikut adalah rumus untuk menghitung specificity:

$$S = \frac{TN}{TN + FP}$$

Di mana:

S = Specificity

9. Accuracy

Accuracy dapat diartikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual. Berikut adalah rumus untuk menghitung accuracy:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Di mana:

A = Accuracy

10. F-Measure (F1-Score)

F-Measure merupakan salah satu perhitungan evaluasi yang mengombinasikan *recall* dan *precision*. Nilai *recall* dan *precision* pada suatu keadaan dapat memiliki bobot yang berbeda. Ukuran yang menampilkan timbal balik antara *recall* dan *precision* adalah F-

Measure yang merupakan bobot *harmonic mean* dan *recall* dan *precision*. Range dari F-Measure adalah 0 sampai dengan 1.

Berikut adalah rumus untuk menghitung F-Measure:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Pemilihan acuan performansi:

Jika membandingkan nilai antara *accuracy, precision, recall* dan *specificity*, sebaiknya mana yang seharusnya dipilih? Apa yang digunakan sebagai acuan ?

1. Pilih algoritma yang memiliki accuracy tinggi jika akurasi sangat bagus kita gunakan sebagai acuan performansi algoritma jika dataset kita memiliki jumlah data false negatif dan false positif yang sangat mendekati (symmetric). Namun jika jumlahnya tidak mendekati, maka sebaiknya gunakan f1 score sebagai acuan.
2. Pilih algoritma yang memiliki recall tinggi jika kita lebih memilih false positif lebih baik terjadi daripada false negatif.
3. Pilih algoritma yang memiliki precision tinggi jika lebih menginginkan terjadinya true positif dan sangat tidak menginginkan terjadinya false positif.
4. Pilih algoritma yang memiliki specificity tinggi jika tidak menginginkan terjadinya false positif.
5. Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna.

Bab 7

Pengklasifikasian

7.1 Pendahuluan

Penambangan data atau data mining adalah proses menemukan pola dan hubungan dari data set berukuran besar yang sebelumnya tidak diketahui melalui teknis analisis data. Model statistik, algoritma matematika, dan metode pembelajaran mesin dapat dihubungkan ke dalam data mining. Data mining lebih dari sekedar mengumpulkan dan mengatur data, tetapi juga mencakup analisis dan prediksi.

Klasifikasi merupakan teknik dalam data mining yang dapat menangani rentang data yang lebih luas daripada regresi. Klasifikasi merupakan penempatan objek-objek ke salah satu dari beberapa kategori yang telah ditetapkan sebelumnya. Dalam menyelesaikan permasalahan klasifikasi, penggunaan metode atau teknik bertujuan untuk mempermudah proses klasifikasi.

Beberapa algoritma yang sering digunakan dalam kasus klasifikasi yaitu *Decision Tree, Support Vector Machine, Naive Bayes, Multilayer Perceptron*.

7.2 Teknik Data Mining

Teknik data mining merupakan suatu proses utama yang digunakan saat metode diterapkan untuk menemukan pola tertentu dari sebuah data. Terdapat beberapa teknik dan sifat analisis yang dapat digolongkan dalam data mining yaitu, sebagai berikut:

1. Clustering
   Clustering atau klasterisasi adalah metode pengelompokan data. Clustering adalah sebuah proses untuk mengelompokkan data ke dalam beberapa cluster/kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum. Clustering merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan cluster.
   Objek di dalam cluster memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan cluster yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma clustering. Oleh karena itu, clustering sangat berguna dan bisa menemukan grup atau kelompok yang tidak dikenal dalam data. Clustering dapat digunakan untuk memberikan label pada kelas data yang belum diketahui, sehingga clustering sering digolongkan sebagai metode *unsupervised learning*.
2. Association
   Tujuan dari metode ini untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang berhubung kuat satu dengan yang lainnya. Sebagai contoh *association analysis* dapat digunakan untuk menentukan produk yang datang secara bersamaan oleh banyak pelanggan, atau bisa juga disebut dengan basket analisis.
3. Regression
   Regression mirip dengan klasifikasi. Perbedaan utamanya adalah terletak pada atribut yang diprediksi berupa nilai yang kontinu.
4. Forecasting Prediksi (Forecasting)
   Berfungsi untuk melakukan kejadian yang akan datang berdasarkan data sejarah yang ada.

5. Sequence Analysis
6. Tujuan dari metode ini adalah untuk mengenali pola dari data diskrit. Sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.
7. Deviation Analysis
   Tujuan dari metode ini adalah untuk menemukan penyebab perbedaan antara data yang satu dengan data yang lain dan biasa disebut sebagai *outlier detection*. Sebagai contoh adalah apakah sudah terjadi penipuan terhadap pengguna kredit dengan melihat catatan transaksi yang tersimpan dalam basis data perusahaan kartu kredit
8. Klasifikasi
   Klasifikasi adalah suatu pengelompokan data di mana data yang digunakan tersebut mempunyai kelas label atau target. Sehingga algoritma-algoritma untuk menyelesaikan masalah klasifikasi dikategorisasikan ke dalam supervised learning. Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

Tahapan dari klasifikasi dalam data mining terdiri dari:

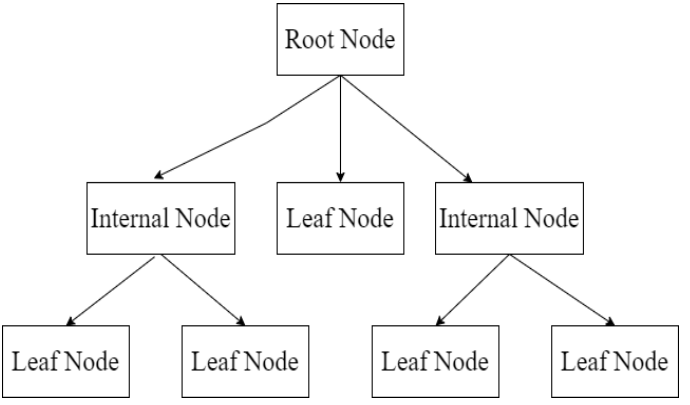
1. Pembangunan model - Pada tahapan ini dibuat sebuah model untuk menyelesaikan masalah klasifikasi *class* atau atribut dalam data. Tahap ini merupakan fase pelatihan, di mana data latih di analisa menggunakan algoritma klasifikasi, sehingga model pembelajaran direpresentasikan dalam bentuk aturan klasifikasi.
2. Penerapan model - Pada tahapan ini model yang sudah dibangun sebelumnya digunakan untuk menentukan atribut/kelas dari sebuah data baru yang atribut/kelasnya belum diketahui sebelumnya. Tahap ini digunakan untuk memperkirakan keakuratan aturan klasifikasi terhadap data uji. Jika model dapat diterima, maka aturan dapat diterapkan terhadap klasifikasi data baru

7.3 Decision Tree

Decision tree atau pohon keputusan adalah teknik klasifikasi yang sederhana yang banyak digunakan. Metode ini tidak memerlukan proses pengelolaan pengetahuan terlebih dahulu dan dapat menyelesaikan kasus-kasus yang memiliki dimensi yang besar (Ade Putranto and Wuryandari, 2015). Metode pengklasifikasian ini menggunakan contoh bentuk struktur pohon, di mana node yang menggambarkan tiap atribut, daun menggambarkan tiap kelas, dan setiap cabangnya menggambarkan nilai dari tiap kelas (Robianto, 2021).

Node akar atau root node menyatakan node yang berada paling atas dari pohon. Internal node merupakan node percabangan, pada node ini hanya terdapat satu input dan mempunyai output minimal dua output. *Leaf node* adalah node terakhir, hanya mempunyai satu masukan, dan tidak mempunyai keluaran. Pohon keputusan pada tiap *leaf node* menyatakan label tiap kelas. Pohon keputusan pada tiap cabangnya menyatakan keadaan yang harus diisi dan tiap puncak pohonnya menggambarkan nilai kelas data.

Contoh dari model pohon keputusan seperti pada gambar 7.1 berikut:



Gambar 7.1: Model Decision Tree

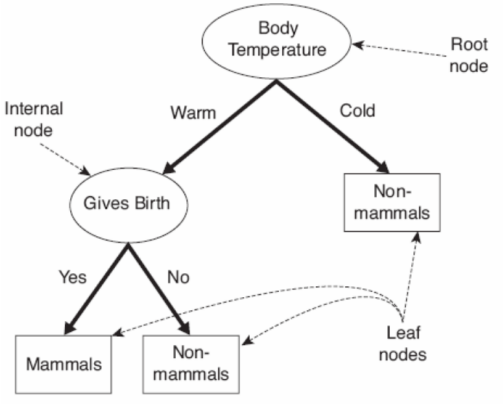
Pada umumnya decision tree melakukan strategi pencarian secara top-down untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari node akar (root) sampai node akhir (leaf) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu.

Decision tree juga berguna untuk eksplorasi data, menemukan hubungan antara sejumlah calon variabel input dengan sebuah variabel target. Decision tree eksplorasi data dan pemodelan yang salah langkah pertama yang sangat baik dalam proses pemodelan yang digunakan sebagai model akhir untuk beberapa teknik lainnya. Kelebihan lain dari metode ini adalah mampu mengeliminasi perhitungan atau data-data yang tidak diperlukan. Karena sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu.

Berikut ilustrasi dari cara kerja pohon keputusan, dengan contoh klasifikasi vertebrata dalam bentuk yang lebih sederhana (Fauzi, 2017). Terdapat dua kategori dalam klasifikasi Vertebrata yaitu mamalia dan non-mammalia. Misalnya spesies baru ditemukan oleh peneliti dan ingin diketahui apakah spesies tersebut mamalia atau non-mamalia. Salah satu pendekatan yang dapat dilakukan adalah dengan mengajukan serangkaian pertanyaan tentang karakteristik spesies.

Pertanyaan pertama yang dapat diajukan adalah apakah spesies tersebut *cold* atau *warm-blooded*. Jika spesies tersebut *cold-blooded*, maka spesies tersebut bukan mamalia. Lainnya dapat merupakan mamalia atau burung. Pertanyaan selanjutnya yang dapat diajukan adalah apakah spesies betinanya melahirkan anak, jika iya maka spesies tersebut adalah mamalia, lainnya bukan mamalia (kecuali untuk spesies mamalia tertentu, egg-laying mammal).

Contoh pohon keputusan untuk masalah klasifikasi mamalia ditunjukkan pada gambar 7.2.



Gambar 7.2: Pohon Keputusan Klasifikasi Mamalia

Beberapa model Decision Tree yang sudah dikembangkan antara lain ID3, C4.5 dan CART. Algoritma C4.5 ini merupakan pengembangan dari algoritma ID3 yang dibuat oleh J. Ross Quinlan, seorang peneliti di bidang *machine learning* dengan beragam peningkatan. Beberapa peningkatan ini di antaranya adalah penanganan atribut-atribut numerik, missing value, dan noise pada dataset, dan aturan-aturan yang dihasilkan dari model pohon yang terbentuk (Sutoyo, 2018).

Metode C4.5 dan ID3 memiliki perbedaan dalam nilai tiap atribut. Metode C4.5 menggunakan atribut yang bernilai kategorikal dan numerikal, sedangkan metode ID3 menggunakan atribut yang bernilai kategorikal. Algoritma ID3 membentuk pohon keputusan dengan metode *divide-and-conquer* data secara rekursif dari atas ke bawah.

CART merupakan singkatan dari *Classification and Regression Tree*. CART mempunyai dua Langkah penting yang harus diikuti untuk mendapatkan tree dengan performansi yang optimal. Langkah pertama adalah pemecahan objek secara berulang berdasarkan atribut tertentu. Langkah kedua, pruning (pemangkasan) dengan menggunakan data validasi.

7.4 Support Vector Machine (SVM)

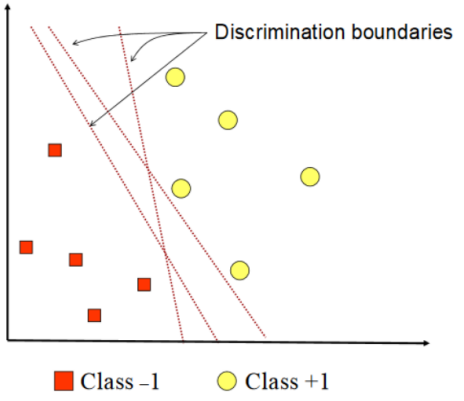
Support Vector Machine (SVM) adalah suatu metode yang handal dalam menyelesaikan masalah klasifikasi data. Permasalahan SVM dipecahkan dengan menyelesaikan persamaan Lagrangian yang merupakan bentuk dual dari SVM melalui *quadratic programming* (Fiska, 2017).

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Fungsi pemisah yang terbaik yaitu mengoptimalkan nilai margin yang merupakan *separating hyperplane* pada setiap kelas dan posisi ini dapat tercapai apabila garis pemisah tersebut berada tepat posisinya di tengah-tengah, membagi antar kelas negatif dan kelas positif (Handayanto et al., 2019).

Gambar 7.3 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah class, yaitu: +1 dan –1. *Pattern* yang tergabung pada class –1 disimbolkan dengan warna merah (kotak), sedangkan *pattern* pada class +1, disimbolkan dengan warna kuning (lingkaran). Problem klasifikasi dapat

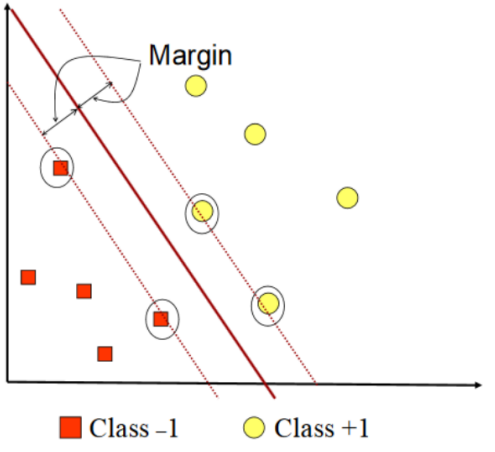
diterjemahkan dengan usaha menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut.

Berbagai alternatif garis pemisah (discrimination boundaries) ditunjukkan pada Gambar 7.3.



Gambar 7.3: Berbagai Alternatif Garis Pemisah

Garis solid pada Gambar 7.4 menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah - tengah kedua class, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*.



Gambar 7.4: Hyperplane

Langkah awal suatu algoritma SVM adalah pendefinisian persamaan suatu hyperplane pemisah yang dituliskan dengan persamaan berikut:

$$w.X + b = 0$$

Keterangan:

w = bobot vektor, w = $\{x_1, x_2, \ldots, x_n\}$

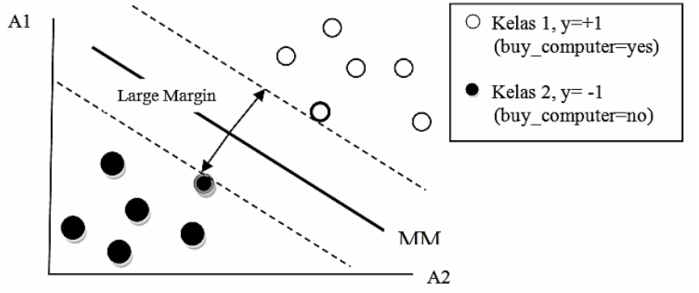
b = skalar yang disebut dengan bias. Jika berdasarkan pada atribut A_1, A_2 dengan permisalan tupel pelatihan

X = (x_1, x_2), x_1, x_2 merupakan nilai dari atribut A_1 dan A_2

jika b dianggap sebagai suatu bobot tambahan, maka persamaan suatu hyperplane pemisah dapat ditulis ulang seperti pada persamaan berikut:

$$w_0 + w_1x_1 + w_2x_2 = 0$$

Setelah persamaan dapat didefinisikan, nilai $x_1$dan x_2 dapat dimasukkan ke dalam persamaan untuk mencari bobot w_1, w_2 dan w_0 atau b. Grafik pemisahan dua kelas data dengan margin maksimum dapat dilihat pada Gambar 7.5



Gambar 7.5: Pemisahan Dua Kelas Data Dengan Margin Maksimum

Pada gambar di atas, SVM menemukan *hyperplane* pemisah maksimum, yaitu *hyperplane* yang mempunyai jarak maksimum antara tupel pelatihan terdekat. *Support vector* ditunjukkan dengan batasan tebal pada titik tupel.

Dengan demikian, setiap titik yang letaknya di atas *hyperplane* pemisah memenuhi persamaan berikut:

$$w_0 + w_1x_1 + w_2x_2 > 0$$

Sedangkan, titik yang letaknya di bawah *hyperplane* pemisah memenuhi rumus seperti pada persamaan berikut:

$$w_0 + w_1 x_1 + w_2 x_2 < 0$$

Jika dilihat dari dua kondisi di atas, maka didapatkan dua persamaan hyperplane, seperti pada persamaan yang ada di bawah ini:

$$H_1 : w_0 + w_1 x_1 + w_2 x_2 \geq 0 \text{ untuk } y_i = +1$$

$$H_2 : w_0 + w_1 x_1 + w_2 x_2 \leq 0 \text{ untuk } y_i = -1$$

Dengan demikian, setiap tuple yang berada di atas H1 memiliki kelas +1, dan setiap tuple yang berada di bawah H2 memiliki kelas -1.

7.5 Naive Bayes

Naive Bayes Classifier merupakan suatu model independen yang membahas mengenai klasifikasi sederhana berdasarkan teorema Bayes. *Naive Bayes* merupakan suatu algoritma yang dapat mengklasifikasikan suatu variabel tertentu dengan menggunakan metode probabilitas dan statistik (Kurniawan, 2018).

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum dengan persamaan berikut:

$$P(H|E) = \frac{P(H)\, P(E|H)}{P(E)}$$

Keterangan:

P(H|E): Probabilitas bebas bersyarat (conditional probability) suatu hipotesis H jika diberikan bukti (Evidence) E terjadi.

P(E|H): Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H.

P(H): Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun

P(E): Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atas peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati (Husaini, 2016).

Ada beberapa hal penting dalam aturan Bayes tersebut yaitu:

1. Sebuah probabilitas awal/prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Sebagai contoh, dalam suatu peramalan cuaca untuk memperkirakan terjadinya hujan, ada faktor yang memengaruhi terjadinya hujan, yaitu mendung. Jika diterapkan dalam *Naive Bayes*, probabilitas terjadinya hujan, jika bukti mendung sudah diamati, dinyatakan dengan:

$$P(Hujan|Mendung) = \frac{P(Hujan)\ P(Mendung|Hujan)}{P(Mendung)}$$

P(Hujan|Mendung) adalah nilai probabilitas hipotesis hujan terjadi jika bukti mendung sudah diamati. P(Mendung|Hujan) adalah nilai probabilitas bahwa mendung yang diamati akan memengaruhi terjadinya hujan. P(Hujan) adalah probabilitas awal hujan tanpa memandang bukti apapun, sementara P(Mendung) adalah probabilitas terjadinya mendung. Contoh tersebut dapat dikembangkan dengan menambahkan beberapa observasi yang lain sebagai bukti. Semakin banyak bukti yang dilibatkan, maka semakin baik hasil prediksi yang diberikan.

Namun, tentu saja bukti tersebut harus benar-benar berkaitan dan memberi pengaruh pada hipotesis. Dengan kata lain, penambahan bukti yang diamati tidak sembarangan. Bukti gempa tentu saja tidak berkaitan dengan hujan sehingga penambahan bukti gempa dalam prediksi cuaca akan memberikan hasil yang salah.

Walaupun ada bukti lain yang memengaruhi cuaca seperti suhu udara, tetap saja ada nilai probabilitas P(Suhu) yang harus dinilai secara independen dalam teorema bayes, yang sulit dilakukan karena suhu udara juga dipengaruhi oleh faktor lain seperti cuaca kemarin, mendung, polusi, dan sebagainya.

Jadi, penilaian probabilitas tersebut tidak memandang faktor lain. Inilah sebabnya disebut *Naive Bayes* (Bayes Naif). Klasifikasi dengan *Naive Bayes* bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas.

Oleh karena itu karakteristik *Naive Bayes* adalah sebagai berikut:

1. Metode *Naive Bayes* bekerja teguh (robust) terhadap data-data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda (outliner). *Naive Bayes* juga bisa menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan model dan prediksi (Husaini, 2016).
2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi *Naive Bayes* karena asumsi independensi atribut tersebut sudah tidak ada.

Algoritma Naive Bayes cocok diterapkan pada data yang berskala ordinal. Jenis data ordinal tersebut mempunyai variabel yang nilainya berupa simbol tetapi bisa diurutkan, tidak bisa diukur jaraknya dan tidak bisa dijumlahkan hasil dari semuanya (Derajad Wijaya and Dwiasnati, 2020). Kelemahan Naive Bayes adalah harus mengasumsi bahwa antar fitur tidak terkait (independent).

Jika dalam realita keterkaitan itu ada, keterkaitan tersebut tidak dapat dimodelkan oleh Naive Bayesian Classifier (Ashari Muin, 2016).

Bab 8

Regresi

8.1 Pendahuluan

Regresi adalah metode analisis data yang digunakan untuk memprediksi hubungan antara variabel terikat/kriterium (dependen) dan variabel bebas/prediktor (independen). Selain itu, metode ini dapat digunakan untuk menilai kekuatan hubungan antar variabelnya. Analisis regresi memiliki 3 jenis analisis yaitu analisis regresi linear, analisis regresi linear majemuk (berganda), dan analisis regresi non linear.

Aplikasi analisis regresi dapat dilakukan di berbagai disiplin ilmu seperti keuangan, investasi, pemasaran, dan bidang ilmu yang lain (Niko Ramadhani, 2021).

Fungsi Regresi

Regresi memiliki fungsi sebagai berikut:

1. Memprediksi masa depan
   Pada fungsi ini, regresi dapat menganalisis hal-hal yang akan terjadi di masa depan seperti meramalkan risiko dan peluang, maka dari itu regresi banyak digunakan dalam dunia bisnis.

Contoh: Analisis masa depan yang berhubungan dengan permintaan produk, banyaknya jumlah produk yang akan dibeli oleh konsumen, dan mengestimasi status kredit dari nasabah serta perkiraan angka klaim dana dalam periode tertentu pada perusahaan asuransi.

2. Memperbaiki eror

   Fungsi regresi dapat memperbaiki kekeliruan seperti pembuatan keputusan sehingga masalah tersebut dapat teratasi.

   Contoh: Ketika manajer merasa bahwa memperpanjang waktu buka toko bisa meningkatkan penjualan. Ternyata setelah menghitung regresi, keputusan tadi malah merugikan bujet perusahaan. Jadi fungsi analisis regresi dalam menghindarkan kesalahan sangat membantu si manajer.

3. Memberikan wawasan baru

   Pencarian data bisa memberikan wawasan yang baru dan segar. Para pebisnis sering mengumpulkan data-data terkait pelanggan mereka. Namun, tanpa analisis regresi yang tepat, semua data tersebut tidak berarti apapun.

   Contoh: Mencari data lewat analisis regresi dapat menunjukkan lonjakan penjualan selama hari-hari tertentu dalam seminggu dan penurunan di hari lainnya. Manajer dapat membuat penyesuaian untuk kompensasi. Mulai dari menyediakan stok yang tepat pada masing-masing hari, mencari bantuan ekstra, atau bahkan memastikan ketersediaan staf dan produk pemasaran terbaik di hari-hari tersebut.

4. Meningkatkan efisiensi operasional

   Perusahaan dapat menggunakan fungsi regresi dalam mengoptimalkan operasional bisnis.

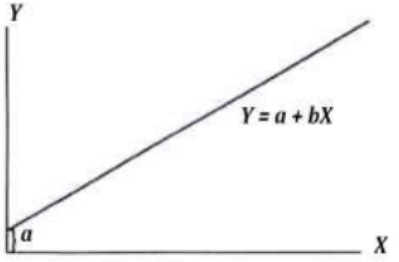
   Contoh: Seorang manajer pabrik menghitung regresi untuk mengetahui dampak dari suhu oven saat memanggang roti, seperti berapa lama waktu penyimpanannya setelah matang. Jadi, mereka tidak perlu mengandai-andai dampak tanpa data riil.

8.2 Jenis dan Rumus Regresi

Adapun jenis dan rumus regresi adalah sebagai berikut:

Regresi Linier Sederhana

Regresi linear sederhana merupakan jenis regresi yang hanya menghubungkan satu variabel dependen/terikat (Y) dengan satu variabel independen/bebas (X) di mana keduanya adalah data kuantitatif dan biasanya digambarkan dengan garis lurus.



Gambar 8.1: Garis Regresi Linier (I Made Yuliara, 2016b)

Rumus persamaan regresi linear sederhana sebagai berikut:

$$Y = a + bX + e$$

Keterangan:

Y=Variabel dependen/kriterium/ respons
X=Variabel independen (bebas)/prediktor
a=Konstanta
b=Koefisien Regresi
e=Eror (Residu)

Di mana besarnya a dan b ditentukan menggunakan persamaan sebagai berikut:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = Y - bX \text{ atau } a = \frac{(\sum Y)(\sum X^2) - (X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

Bentuk persamaan regresi tersebut sering dibaca sebagai regresi X atas Y. Koefisien arah regresi linier dinyatakan dengan huruf b. Bila harga b positif, maka variabel Y akan mengalami kenaikan atau pertambahan. Sebaliknya jika b negatif maka variabel Y akan mengalami penurunan.

Contoh:

Terdapat persamaan regresi antara pengunjung (X) dan pembeli (Y) yaitu $Y = 9 + 0{,}5X$ yang artinya karena b positif maka hubungan fungsionalnya menjadi positif.

Misal jika pengunjung bertambah 30 orang maka rata-rata pembeli (Y) akan bertambah menjadi:

Y=0+0,5*30=24 Orang

Sehingga dapat disimpulkan bahwa semakin banyak pengunjung semakin banyak pula pembelinya.

Contoh Soal 1:

Seorang Engineer ingin mempelajari hubungan antara suhu ruangan dengan jumlah cacat yang diakibatkannya, sehingga dapat memprediksi atau meramalkan jumlah cacat produksi jika suhu ruangan tersebut tidak terkendali. Engineer tersebut kemudian mengambil data selama 10 hari terhadap rata-rata suhu ruangan dan jumlah cacat produksi.

Penyelesaian:

1. Langkah 1: penentuan tujuan
   Tujuan: memprediksi jumlah cacat produksi jika suhu ruangan tidak terkendali.
2. Langkah 2: identifikasikan variabel penyebab dan akibat
   Variabel faktor penyebab (X): suhu ruangan dan variabel akibat (Y): jumlah cacat produksi.
3. Langkah 3: pengumpulan data
   Berikut ini adalah data yang dikumpulkan selama 10 hari (berbentuk tabel).

Tabel 8.1: Rata-Rata Suhu Ruangan dan Jumlah Cacat

Tanggal	Rata-Rata Suhu Ruangan	Jumlah Cacat
1	24	10
2	22	5
3	21	6
4	20	3
5	22	6
6	19	4
7	20	5
8	23	9
9	24	11
10	25	13

4. Langkah 4: hitung X^2, Y^2, XY dan total dari masing-masingnya
   Berikut ini adalah tabel yang dilakukan perhitungan X2, Y2, XY dan totalnya:

Tanggal	Rata-Rata Suhu Ruangan (X)	Jumlah Cacat (Y)	X^2	Y^2	XY
1	24	10	576	100	240
2	22	5	484	25	110
3	21	6	441	36	126
4	20	3	400	9	60
5	22	6	484	36	132
6	19	4	361	16	76
7	20	5	400	25	100
8	23	9	529	81	207
9	24	11	576	121	264
10	25	13	625	169	325
Total	220	72	4876	618	1640

5. Langkah 5: hitung *a* dan *b* berdasarkan rumus regresi linier sederhana.
   Menghitung konstanta *(a)*:

$$= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{(72)(2876) - (220)(1640)}{10(2876) - (220)^2} = 17{,}3185$$

   Menghitung koefisien regresi (b)

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{10(1640) - (220)(72)}{10(2876) - (220)^2} = 0{,}0285$$

6. Langkah 6: buat model persamaan regresi

$$Y = a + bX$$
$$Y = 17{,}3185 + 0{,}0285X$$

7. Langkah 7: lakukan prediksi terhadap variabel faktor penyebab atau variabel akibat
   a. Prediksikan jumlah cacat produksi jika suhu dalam keadaan tinggi (variabel X), contohnya: 30° C

$$Y = 17{,}3185 + 0{,}0285\ (30)$$
$$Y = 18{,}1735$$

   Jadi jika suhu ruangan mencapai 30°C, maka akan diprediksikan terdapat 18.1735 unit yang cacat produksi

   b. Jika cacat produksi (variabel Y) yang ditargetkan hanya boleh 4 unit, maka berapa suhu ruangan yang diperlukan untuk mencapai target tersebut?

$$4 = 17{,}3185 + 0{,}0285\ (x)$$
$$4 - 17{,}3185 = 0{,}0285\ (x)$$
$$-13{,}3185 = 0{,}0285\ (x)$$
$$-476{,}316 = x$$

   Jadi prediksi suhu ruangan yang paling sesuai untuk mencapai target cacat produksi adalah -476.316.

Contoh Soal 2:

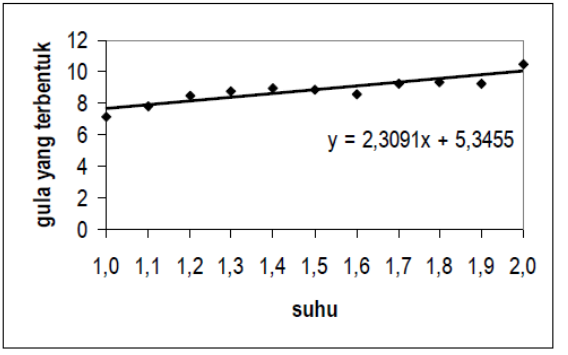
Sebuah penelitian ingin menguji apakah suhu (°C) (X) memengaruhi banyaknya gula yang terbentuk (Y):

X	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
Y	7,1	7,8	8,5	8,8	9,0	8,9	8,6	9,2	9,3	9,2	10,5

$$b = \frac{11(147{,}89) - (16{,}5)(96{,}9)}{11(25{,}85) - (16{,}5)^2} = 2{,}309$$
$$a = \left(\frac{96{,}9}{11}\right) - 2{,}309\left(\frac{16{,}5}{11}\right) = 5{,}345$$

Dengan demikian, garis regresinya adalah:

$$Y = 5{,}345 + 2{,}309X$$



Gambar 8.2: Garis Regresi Hubungan X dan Y

Regresi Linier Berganda (Majemuk)

Regresi linear berganda merupakan jenis regresi yang menunjukkan hubungan antara satu variabel dependen/terikat (Y) terhadap dua atau lebih variabel independen/bebas (X) dengan jenis data kuantitatif. Tujuan dari uji regresi linier berganda adalah untuk memprediksi nilai variabel tak bebas/respons (Y) apabila nilai-nilai variabel bebasnya/predictor (X1,X2,X3..., Xn) diketahui.

Berikut representasi matematis dari model regresi linear berganda:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + e$$

Keterangan:

Y	=Variabel dependen
X1,X2,X3	=Variabel independen (bebas)
A	=Konstanta
b1,b2,b3	=Koefisien Regresi
e	=Eror (Residu)

Di mana besarnya a dan b_1,b_2 ditentukan menggunakan persamaan sebagai berikut:

$$b_1 = \frac{[\sum {X_2}^2 \, x \sum X_1Y - \sum X_2Y \, x \sum X_1X_2]}{[\sum {X_1}^2 x \sum {X_2}^2 - (\sum X_1X_2)^2}$$

$$b_2 = \frac{[\sum {X_1}^2 \, x \sum X_2Y - \sum X_1Y \, x \sum X_1X_2]}{[\sum {X_1}^2 x \sum {X_2}^2 - (\sum X_1X_2)^2}$$

$$a = \frac{(\sum Y) - (b_1 x \sum X_1) - (b_2 x \sum X_2)}{n}$$

Regresi Non-Linear

Regresi non-linear adalah jenis regresi yang menghubungkan antara variabel Y dengan X yang tidak linear. Ada berbagai macam model dalam regresi non-linear, di antaranya:

1. Regresi Power
   Regresi Power direpresentasikan sebagai:
   $$y = ax^b$$
   Linierisasi kurva lengkung tersebut dapat dilakukan sebagai berikut:
   $$\log y = \log ax^b$$
   $$\log y = \log a + \log x^b$$
   $$\log y = \log a + b \log x$$
   dengan pemisalan: A = log a, p = log y dan q = log x maka regresi power dapat direpresentasikan sebagai:
   $$p = A + bq$$
2. Regresi Eksponensial
   Regresi Eksponensial direpresentasikan sebagai:
   $$y = ae^{bx}$$
   Untuk regresi eksponensial kita menggunakan LN (baca: len):
   $$\ln y = \ln ae^{bx}$$
   $$\ln y = \ln a + \ln e^{bx}$$
   $$\ln y = \ln a + bx$$
   dengan pemisalan: $A = \ln a, p = \ln y$ dan $q = x$ maka regresi eksponensial dapat direpresentasikan (sama seperti saat merepresentasikan regresi Power):
   $$p = A + bq$$
3. Regresi Polinomial
   Regresi Polinomial orde r direpresentasikan sebagai:
   $$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_r x^r$$
   Untuk mengingat kembali bagaimana mencari nilai a dan b, maka ditulis kembali formulanya:
   $$b = \frac{\sum x_i \sum y_i - n \sum x_i y_i}{(\sum x_i)^2 - n \sum {x_i}^2}$$

dan

$$a = \frac{1}{n}\sum y_i - \frac{1}{n}\sum x_i\, b$$

Tetapi kita rubah variabel x dan y menjadi q dan p sehingga menjadi:

$$b = \frac{\sum q_i \sum p_i - n\sum q_i p_i}{(\sum q_i)^2 - n\sum {q_i}^2}$$

dan

$$a = \frac{1}{n}\sum p_i - \frac{1}{n}\sum q_i\, b$$

Contoh soal:

Jumlah Karyawan (Orang)	Waktu (Hari)
3	40
5	36
7	33
10	28
15	22
21	20
28	19

Jika dimisalkan variabel jumlah orang menjadi x dan variabel waktu menjadi y, maka perhatikan langkah-langkah dalam tabel berikut:

x	y	$p = \log x$	$q = \log x$	pq	q^2
3	40	1,602059991	0,477121255	0,764377	0,227645
5	36	1,556302501	0,698970004	1,087809	0,488559
7	33	1,51851394	0,84509804	1,283293	0,714191
10	28	1,447158031	1	1,447158	1
15	22	1,342422681	1,176091259	1,578812	1,383191
21	20	1,301029996	1,322219295	1,720247	1,748264
28	19	1,278753601	1,447158031	1,850559	2,094266
Jumlah (sigma)		10,04624074	6,966657884	9,732254	7,656115

Setelah seluruh variabel formula sudah didapatkan, maka nilai A dan nilai b dapat dicari:

$$b = \frac{(6{,}966657884 * 10{,}04624074) - (7 * 9{,}732254)}{(6{,}966657884)^2 - (7 * 7{,}656115)} = -0{,}36828$$

$$A = (\frac{1}{7} * 10{,}04624074) - (\frac{1}{7} * 6{,}966657884 * -0{,}36828) = 1{,}801704$$

Karena A=*log a* maka=10^{A}=$10^{1{,}801704}$=63,34382

Sehingga persamaan regresi powernya $y=63{,}34382x^{-0{,}3628}$

Untuk persamaan regresi exponential perhatikan tabel berikut:

x	y	$p = \ln y$	$q = x$	pq	q^2
3	40	3,68887945	3	11,0666	9
5	36	3,58351894	5	17,9176	25
7	33	3,49650756	7	24,4756	49
10	28	3,33220451	10	33,322	100
15	22	3,09104245	15	46,3656	225
21	20	2,99573227	21	62,9104	441
28	19	2,94443898	28	82,4443	784
Jumlah (sigma)		23,1323242	89	278,502	1633

Setelah seluruh variabel formula sudah didapatkan, maka nilai A dan nilai b dapat dicari:

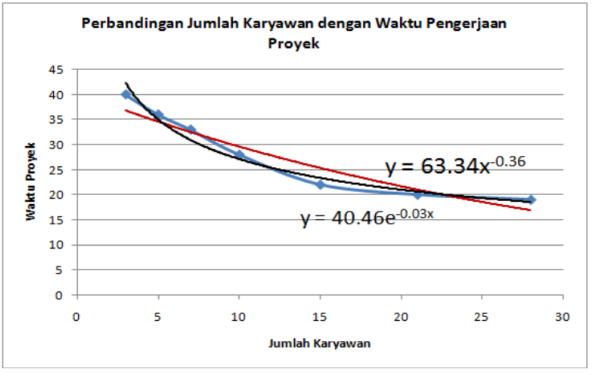
$$b = \frac{(89 * 23{,}1323242) - (7 * 278{,}502)}{(89)^2 - (7 * 1633)} = -0{,}03113$$

$$A = (\frac{1}{7} * 23{,}1323242) - (\frac{1}{7} * 89 * -0{,}03113) = 3{,}7004$$

Karena $A=ln\ a$ maka $a=e^A=e^{3{,}7004}=40{,}4634$

Sehingga persamaan regresi eksponensialnya $y=40{,}4634e^{-0{,}03113x}$

Dengan menggunakan fasilitas *trendline* dari MS Excel diperoleh persamaan regresi sebagai berikut:



Kurva *trendline* berwarna merah adalah kurva regresi linier eksponensial dan kurva *trendline* berwarna hitam adalah kurva Regresi Power.

Bab 9

Clustering

9.1 Pendahuluan

Clustering merupakan bagian dari algoritma Data Mining yang membagi data menjadi beberapa kelompok yang berguna, atau bermakna. Semakin besar perbedaan antara cluster dan semakin besar homogenitas (atau kesamaan) dalam cluster, maka semakin baik hasil pengelompokannya (Mostafa, 2020).

Clustering disebut juga dengan segmentasi (pengelompokan). Clustering ini termasuk dalam *unsupervised method*, maksudnya tidak ada pelabelan dalam atribut yang digunakan, berbeda dengan klasifikasi yang menggunakan atribut sebagai label. Sebagai ilustrasi terkait clustering ini adalah semisal kita mempunyai beberapa bangun datar, yaitu sebuah persegi panjang berwarna biru, sebuah persegi panjang berwarna merah, sebuah persegi berwarna biru, sebuah elips berwarna merah, sebuah elips berwarna biru dan sebuah lingkaran berwarna merah.

Jika bangun datar tersebut dibuat clustering bisa dijadikan beberapa kelompok sesuai dengan kemiripannya, misalnya di cluster berdasarkan warna dan kemiripan bentuk (tidak harus sama persis tapi mirip). Jika di cluster berdasarkan warna, maka akan ada dua kelompok yaitu kelompok bangun datar yang berwarna merah dan kelompok bangun datar yang berwarna biru.

Jika di cluster berdasarkan kemiripan bentuk, maka juga akan ada dua kelompok yaitu, kelompok bangun datar yang berbentuk kotak (persegi dan persegi panjang) dan kelompok bangun datar yang berbentuk bundar (elips dan lingkaran). Clustering sudah lama digunakan dalam berbagai bidang, misalnya Machine Learning, Pencarian Informasi, Data Mining, Pengenalan Pola, Biologi, Psikologi, dan lain sebagainya.

Contoh clusterisasi dalam bisnis dan penelitian di antaranya (Kusrini & Luthfi, 2009):

1. Clusterisasi untuk mendapatkan kelompok konsumen untuk menentukan target marketing dari suatu produk bagi perusahaan yang tidak memiliki dana marketing yang besar.
2. Clusterisasi dalam bidang akuntansi, misalnya untuk memisahkan perilaku finansial yang baik dan yang terindikasi tidak baik atau mencurigakan.
3. Clusterisasi dalam bidang biologi, misalnya untuk mendapatkan kemiripan perilaku gen dalam jumlah yang besar.

9.2 Konsep Dasar Clustering

Hasil clustering yang baik akan menghasilkan tingkat kesamaan yang tinggi dalam satu kelas dan tingkat kesamaan yang rendah antar kelas. Kesamaan yang dimaksud merupakan pengukuran secara numerik terhadap dua buah objek. Nilai kesamaan antara dua objek akan semakin tinggi jika kedua objek yang dibandingkan memiliki kemiripan yang tinggi. Begitu juga sebaliknya (Irwansyah, 2017). Kualitas hasil clustering sangat bergantung pada metode yang dipakai.

Dalam clustering dikenal empat tipe data. Keempat tipe data pada tersebut adalah:

1. Variabel berskala interval.
2. Variabel biner.
3. Variabel nominal, ordinal, dan rasio.
4. Variabel dengan tipe lainnya.

Metode clustering juga harus dapat mengukur kemampuannya sendiri dalam usaha untuk menemukan suatu pola tersembunyi pada data yang sedang diteliti. Terdapat berbagai metode yang dapat digunakan untuk mengukur nilai kesamaan antar objek-objek yang dibandingkan.

Salah satunya ialah dengan *Weighted Euclidean Distance*. *Euclidean distance* menghitung jarak dua buah point dengan mengetahui nilai dari masing-masing atribut pada kedua poin tersebut.

Berikut formula yang digunakan untuk menghitung jarak dengan *Euclidean distance:*

$$Distance(p,q) = (\sum_{k}^{n} \mu k \mid pk - qk|r)1/r$$

Keterangan:

N = Jumlah record data
K= Urutan field data
r= 2
μk= Bobot field yang diberikan user

Jarak adalah pendekatan yang umum dipakai untuk menentukan kesamaan atau ketidaksamaan dua vektor fitur yang dinyatakan dengan rangking. Apabila nilai rangking yang dihasilkan semakin kecil nilainya maka semakin dekat/tinggi kesamaan antara kedua vektor tersebut. Teknik pengukuran jarak dengan metode *Euclidean* menjadi salah satu metode yang paling umum digunakan.

Pengukuran jarak dengan metode *euclidean* dapat dituliskan dengan persamaan berikut:

$$j(v_1, v_2) = \sqrt{\sum_{k=1}^{N} \left(v_1(k) - v_2(k)\right)^2}$$

di mana v_1 dan v_2 adalah dua vektor yang jaraknya akan dihitung dan N menyatakan panjang vektor.

9.3 Mengapa Clustering Digunakan Dalam Data Mining?

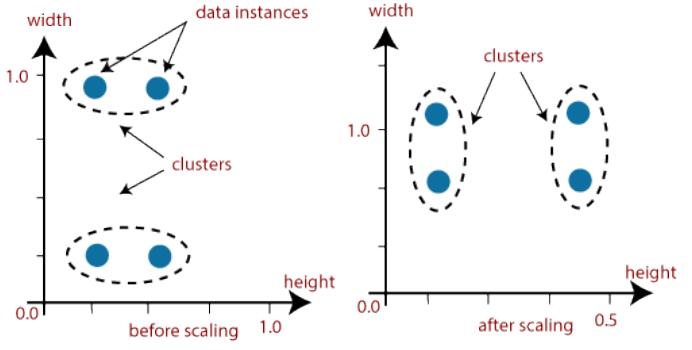
Analisis clustering telah menjadi permasalahan yang berkembang dalam data mining karena beragam aplikasinya. Munculnya berbagai alat pengelompokan data dalam beberapa tahun terakhir dan penggunaannya yang komprehensif dalam berbagai aplikasi, termasuk pemrosesan gambar, komputasi biologi, komunikasi seluler, kedokteran, dan ekonomi, hal tersebut berkontribusi pada perkembangan popularitas algoritma ini.

Masalah utama dengan algoritma clustering adalah tidak dapat di standarisasi. Algoritma tingkat lanjut dapat memberikan hasil terbaik dengan satu jenis kumpulan data, tetapi mungkin gagal atau berkinerja buruk dengan jenis kumpulan data lainnya.

Meskipun banyak upaya telah dilakukan untuk membakukan algoritma yang dapat bekerja dengan baik di semua situasi, tetapi sejauh ini belum ada pencapaian yang signifikan. Banyak alat clusterisasi telah diusulkan sejauh ini. Namun, setiap algoritma memiliki kelebihan atau kekurangan dan tidak dapat bekerja pada semua situasi nyata.

Alat clusterisasi tersebut di antaranya adalah sebagai berikut:

1. Scalability - Scalability (skalabilitas) dalam clusterisasi menyiratkan bahwa saat jumlah objek data meningkat, maka waktu yang dibutuhkan untuk melakukan clusterisasi juga harus disesuaikan dengan urutan kompleksitas algoritmanya.



Gambar 9.1: Contoh Kesalahan Scalability (Jaiswal, 2023)

Sebagai contoh, jika kita melakukan K-*means clustering*, kita tahu itu adalah O(n), di mana n adalah jumlah objek dalam data. Jika kita menaikkan jumlah objek data 10 kali lipat, maka waktu yang dibutuhkan untuk mengelompokkannya juga akan meningkat kira-kira 10 kali lipat. Artinya harus ada hubungan yang linier. Jika tidak seperti itu maka akan ada permasalahan dalam proses penerapannya.
Dari gambar 9.1 di atas menunjukkan bahwa data harus *scalable* jika tidak *scalable*, maka kita tidak bisa mendapatkan hasil yang sesuai. Angka tersebut mengilustrasikan contoh grafis yang dapat menyebabkan hasil yang salah.

2. Interpretability - Interpretability (Interpretabilitas) adalah kemampuan dari dua atau lebih sistem atau komponen untuk berbagi pakai data/ informasi. Hasil pengelompokan harus dapat ditafsirkan, dipahami, dan dapat digunakan.
3. Penemuan cluster dengan bentuk atribut - Algoritma klasterisasi harus dapat menemukan bentuk cluster yang berubah-ubah. Mereka tidak boleh dibatasi hanya pada pengukuran jarak yang cenderung menemukan gugus bola berukuran kecil.
4. Kemampuan untuk menangani berbagai jenis atribut - Algoritma harus dapat diterapkan pada data apa pun seperti data berdasarkan interval (numerik), data biner, dan data kategorikal.
5. Kemampuan untuk menangani data noise - Database yang berisi data noise, hilang atau salah. Beberapa algoritma peka terhadap data semacam itu dan dapat menghasilkan kluster yang berkualitas buruk.
6. High Dimensionality (Dimensi Tinggi) - Alat pengelompokan seharusnya tidak hanya mampu menangani ruang data berdimensi tinggi tetapi juga ruang berdimensi rendah.

9.4 Tipe Algoritma Clustering

Karena tugas clustering ini bersifat subjektif, sarana yang dapat digunakan untuk mencapai tujuan ini sangat banyak. Setiap metodologi mengikuti seperangkat aturan yang berbeda untuk menentukan 'kemiripan' di antara titik data. Faktanya, ada lebih dari 100 algoritma pengelompokan yang dikenal. Tetapi hanya sedikit dari algoritma yang digunakan secara populer, mari kita lihat secara mendetail:

Model Konektivitas

Seperti namanya, model ini didasarkan pada gagasan bahwa titik data yang lebih dekat dalam ruang data menunjukkan lebih banyak kesamaan satu sama lain daripada titik data yang terletak lebih jauh. Model ini dapat mengikuti dua pendekatan. Pada pendekatan pertama, mereka mulai dengan mengklasifikasikan semua titik data ke dalam kluster terpisah dan kemudian menggabungkannya seiring dengan berkurangnya jarak.

Dalam pendekatan kedua, semua titik data diklasifikasikan sebagai cluster tunggal dan kemudian dipartisi seiring bertambahnya jarak. Juga, pilihan fungsi jarak bersifat subjektif. Model-model ini sangat mudah diinterpretasikan tetapi tidak memiliki skalabilitas untuk menangani kumpulan data besar. Contoh model ini adalah algoritma pengelompokan hierarkis dan variannya.

Model Centroid

Ini adalah algoritma pengelompokan iteratif di mana gagasan kesamaan berasal dari kedekatan titik data ke pusat cluster. Algoritma pengelompokan K-Means adalah algoritma populer yang termasuk dalam kategori ini. Dalam model ini, nomor cluster yang diperlukan pada akhirnya harus disebutkan sebelumnya, yang membuatnya penting untuk memiliki pengetahuan sebelumnya tentang kumpulan data. Model ini dijalankan secara iteratif untuk menemukan local optima.

Model Distribusi

Model pengelompokan ini didasarkan pada gagasan tentang seberapa besar kemungkinan semua titik data dalam kluster memiliki distribusi yang sama (Misalnya: Normal, Gaussian). Model ini sering mengalami overfitting. Contoh populer dari model ini adalah algoritma Ekspektasi-maksimisasi yang menggunakan distribusi normal multivariat.

Model Kepadatan

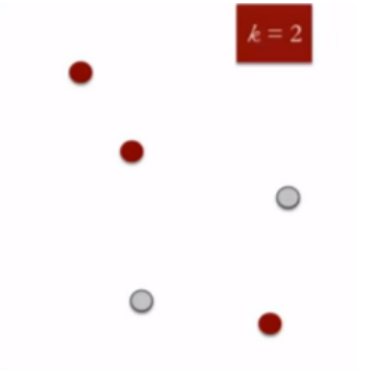
Model ini mencari ruang data untuk area dengan kepadatan titik data bervariasi dalam ruang data. Ini mengisolasi berbagai wilayah kepadatan yang berbeda dan menetapkan titik data dalam wilayah ini dalam cluster yang sama. Contoh model kerapatan yang populer adalah DBSCAN dan OPTIK.

Berikut kami berikan contoh pembahasan dua algoritma clustering yang paling populer yaitu *K Means clustering* dan *Hierarchical clustering*:

9.4.1 K-Means Clustering

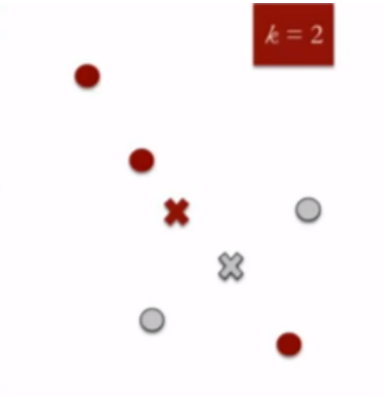
K-Mean adalah algoritma pengelompokan iteratif yang bertujuan untuk menemukan maxima lokal pada setiap iterasi. Algoritma ini bekerja dalam 5 langkah berikut:

1. Menentukan jumlah cluster K yang diinginkan, misal kluster kita tetapkan k=2 untuk 5 titik data ini dalam ruang 2-D. Secara acak kita tetapkan tiga titik di cluster 1 yang ditunjukkan menggunakan warna merah dan dua titik di cluster 2 yang ditampilkan menggunakan warna abu-abu seperti gambar 9.5.1 berikut ini:



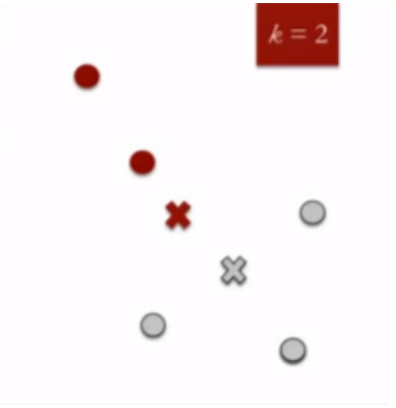
Gambar 9.2: Nomor Cluster K=2 (Kaushik, 2023)

2. Menghitung centroid cluster
   Centroid titik data di cluster merah ditampilkan menggunakan tanda silang merah dan titik data di cluster abu-abu menggunakan tanda silang abu-abu. Berikut gambarnya:



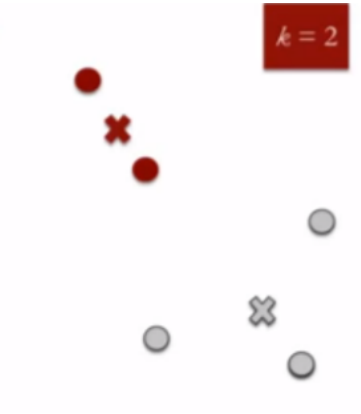
Gambar 9.3: Centroid Cluster (Kaushik, 2023)

3. Tetapkan kembali setiap titik ke pusat cluster terdekat: Perhatikan bahwa hanya titik data di bagian bawah yang ditugaskan ke cluster merah meskipun lebih dekat ke pusat cluster abu-abu. Jadi, kami menetapkan titik data tersebut ke dalam cluster abu-abu. Berikut gambarnya:



Gambar 9.4: Cluster Terdekat (Kaushik, 2023)

4. Selanjutnya menghitung ulang centroid klaster untuk kedua kluster, seperti gambar berikut:



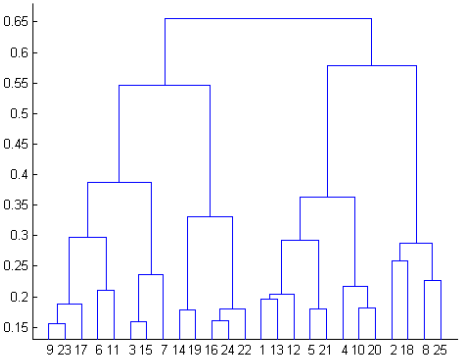
Gambar 9.5: Hasil Hitung Ulang Centroid Cluster (Kaushik, 2023)

5. Ulangi langkah 3 dan 4 hingga tidak ada peningkatan yang mungkin dilakukan hingga mencapai optimal global. Ketika tidak akan ada peralihan lebih lanjut dari titik data antara dua klaster untuk dua pengulangan berturut-turut. Ini akan menandai penghentian algoritma jika tidak disebutkan secara eksplisit.

9.4.2 Hierarchical Clustering

Hierarchical Clustering, seperti namanya, adalah algoritma yang membangun hierarki cluster. Algoritma ini dimulai dengan semua titik data yang ditetapkan ke cluster mereka sendiri. Kemudian dua klaster terdekat digabungkan menjadi klaster yang sama. Pada akhirnya, algoritma ini berhenti ketika hanya ada satu cluster yang tersisa.

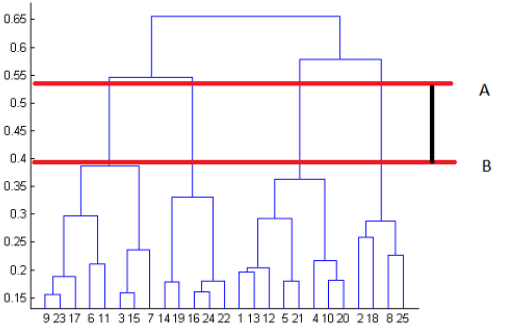
Hasil clustering hirarkis dapat ditampilkan dengan menggunakan dendrogram berikut:



Gambar 9.6: Dendrogram Hierarchical Clustering (Kaushik, 2023)

Pada gambar 9.6 di atas, di bagian bawah, dimulai dengan 25 titik data, masing-masing ditugaskan ke cluster terpisah. Dua cluster terdekat kemudian digabungkan, dan seterusnya sampai kita hanya memiliki satu cluster di atas. Ketinggian dendrogram tempat dua klaster digabungkan mewakili jarak antara dua klaster dalam ruang data.

Keputusan nomor cluster yang dapat menggambarkan kelompok yang berbeda dapat dipilih dengan mengamati dendrogram. Pilihan terbaik dari nomor cluster adalah nomor garis vertikal dalam dendrogram dipotong oleh garis horizontal yang dapat melintasi jarak maksimum secara vertikal tanpa memotong suatu cluster. Dalam contoh di atas, pilihan terbaik nomor cluster akan menjadi 4 karena garis horizontal merah pada dendogram di bawah mencakup jarak vertikal maksimum AB. Perhatikan gambar berikut!



Gambar 9.7: Dendrogram Hierarchical Clustering – Vertikal Maksimum (Kaushik, 2023)

Dua hal penting yang harus diketahui tentang pengelompokan hierarkis adalah:

1. Algoritma ini telah diimplementasikan menggunakan pendekatan *bottom up*. Dimungkinkan juga untuk mengikuti pendekatan *top-down* yang dimulai dengan semua titik data yang ditetapkan dalam kluster yang sama dan melakukan pemisahan secara rekursif hingga setiap titik data diberi kluster terpisah.
2. Keputusan penggabungan dua klaster diambil berdasarkan kedekatan klaster tersebut. Ada beberapa metrik untuk menentukan kedekatan dua klaster:
   a. Jarak Euclidean: ||a-b||2 = √(Σ(ai-bi))
   b. Kuadrat jarak Euclidean: ||a-b||22 = Σ((ai-bi)2)
   c. Jarak Manhattan: ||a-b||1 = Σ|ai-bi|
   d. Jarak maksimum:||a-b||INFINITY = max|ai-bi|
   e. Jarak Mahalanobis: √((a-b)T S-1 (-b)) {di mana, s: matriks kovarians}

Bab 10

Association Rule

10.1 Pendahuluan

Association rule merupakan sebuah teknik data mining untuk menemukan pola kemunculan item secara bersama-sama dengan frekuensi yang tinggi (Ye, 2014). Association rule menyediakan informasi berharga dalam menilai korelasi antar data secara signifikan. Kasus klasik yang menjadi dasar penerapan association rule adalah *market basket analysis*. Pada kasus tersebut dilakukan pengamatan sejumlah item barang yang dibeli oleh pelanggan. Association rule akan mengungkap pasangan item barang yang sering dibeli secara bersamaan oleh pelanggan pada sebuah transaksi tunggal (Kantardzic, 2011; Ye, 2014; Aggarwal, 2015).

Association rule dapat diekspresikan dengan X→Y. Variabel X merupakan himpunan item, sedangkan Y merupakan sebuah item tunggal (Olson and Delen, 2008). Ekspresi tersebut dapat juga diungkapkan dalam bentuk aturan seperti ”Jika X, maka Y”. Keberadaan himpunan item X akan menyebabkan adanya item Y. Variabel X disebut juga sebagai anteseden, sedangkan variabel Y disebut sebagai konsekuen. Variabel X dan Y tidak memiliki item yang sama (Ye, 2014).

Pada kasus market basket analysis dapat ditemukan korelasi antara item barang yang dibeli oleh pelanggan. Hasil association rule dapat digunakan untuk

mengetahui pasangan item barang yang sering dibeli secara bersamaan oleh pelanggan pada sebuah transaksi tunggal. Pengetahuan tersebut bermanfaat untuk melakukan pengambilan keputusan yang akan meningkatkan nilai penjualan.

Contoh keputusan yang dapat diambil seperti pengaturan posisi barang-barang yang memiliki asosiasi kuat pada rak yang berdekatan. Contoh lain yang dapat dilakukan adalah melakukan promosi penjualan barang secara kombinasi dengan harga yang lebih murah dibandingkan jika pelanggan membelinya secara terpisah (Kusrini and Luthfi, 2009).

Model yang populer digunakan pada association rule adalah menghitung frekuensi himpunan item sebagai kuantifikasi level asosiasi. Himpunan item yang diungkap adalah *large itemsets, frequent itemsets*, atau *frequent patterns*. Analisis pola frekuensi tinggi (frequent pattern mining) ini dapat diterapkan pada berbagai bidang, yaitu: data supermarket, text mining, generalisasi tipe data yang berorientasi pada ketergantungan, serta berbagai persoalan data mining lainnya (Aggarwal, 2015).

Aplikasi pada data supermarket merupakan kasus awal di mana *frequent pattern* mining diusulkan. Association rule akan memberikan *insight* yang bermanfaat untuk menemukan kumpulan item barang yang biasanya dibeli secara bersamaan oleh pelanggan pada sebuah transaksi. *Insight* tersebut bisa digunakan oleh para manajer supermarket untuk melakukan inovasi pada penjualan produk mereka sehingga menghasilkan keuntungan ekonomi yang lebih besar (Aggarwal, 2015).

Pada bidang text mining, pemanfaatan frequent pattern mining juga memberikan insight yang bermanfaat. Dokumen teks merupakan data yang tidak terstruktur. Fitur/ciri pada sebuah dokumen teks berupa kata-kata yang ada pada dokumen tersebut. Data teks biasanya direpresentasikan dengan model *bag of words*. Frequent pattern mining dapat membantu menemukan kemunculan bersama kata atau keyword pada dokumen teks. Informasi tentang kemunculan bersama tersebut sangat populer digunakan pada aplikasi text mining (Aggarwal, 2015).

Frequent pattern mining juga dapat digunakan secara general pada tipe data yang berorientasi pada ke bergantungan. Beberapa contoh data yang berorientasi pada ketergantungan adalah *data time series*, *data sekuensial*, *data spatial*, dan *data graph*. Pemanfaatan association rule pada data-data tersebut memerlukan sedikit modifikasi sehingga dapat digunakan pada analisis log

yang dihasilkan oleh aplikasi web, mendeteksi adanya *bug* pada perangkat lunak, dan lain sebagainya (Aggarwal, 2015).

10.2 Analisis Pola Frekuensi Tinggi

Pada analisis pola frekuensi tinggi, terdapat dua parameter yang sangat penting, yaitu *support* dan *confidence*. Kedua nilai ini akan digunakan sebagai garansi dalam pembentukan aturan asosiasi. Kedua nilai ini harus menyatakan bahwa jumlah kejadian pada transaksi cukup tinggi atau memenuhi nilai ambang yang ditentukan. Nilai support mewakili cakupan kejadian pada sebuah transaksi. Nilai confidence dapat mewakili nilai akurasi dari aturan asosiasi yang terbentuk (Witten, Frank and Hall, 2011).

Aturan asosiasi yang terbentuk dapat digambarkan sebagai korelasi antara anteseden dengan konsekuen, serta dilengkapi dengan nilai support dan nilai confidence. Apabila diketahui bahwa seorang pelanggan supermarket akan membeli roti tawar dan mentega secara bersamaan pada sebuah transaksi tunggal, maka aturan asosiasi dapat dituliskan sebagai berikut:

$$\{Roti\ Tawar\} \rightarrow \{Mentega\}\ (support=40\%, confidence=50\%)$$

Aturan tersebut dapat diterjemahkan sebagai berikut:

1. Jumlah transaksi yang memuat item roti tawar dan mentega sekaligus adalah 40% dari total transaksi yang tercatat.
2. Jumlah transaksi yang memuat item mentega adalah 50% dari total transaksi yang memuat item roti tawar.
3. Seorang pelanggan yang membeli roti tawar memiliki kemungkinan 50% juga membeli mentega. Hal ini cukup signifikan karena mewakili 40% dari total transaksi yang tercatat pada supermarket tersebut.

10.2.1 Nilai Support

Nilai support menyatakan proporsi kemunculan kombinasi item dari total data yang tercatat (Larose and Larose, 2005). Apabila diketahui pada kasus pembelian barang di supermarket, pelanggan yang membeli barang X juga akan membeli barang Y. Aturan asosiasi yang terbentuk adalah X→Y.

Nilai support untuk aturan tersebut adalah sebagai berikut:

$$Support(X \rightarrow \mathrm{Y}) = \frac{|X \cup Y|}{\lceil N \rceil} \quad (10.1)$$

Pada Persamaan (10.1), parameter $|X \cup Y|$ merupakan jumlah transaksi pembelian pada supermarket tersebut yang berisi item X dan Y sekaligus pada sebuah transaksi tunggal. Parameter $\lceil N \rceil$ merupakan total transaksi yang tercatat pada supermarket tersebut.

Nilai support sangat penting dalam pembentukan aturan asosiasi. Nilai support yang tinggi menandakan bahwa jumlah kejadian pada data set sangat signifikan. Sebaliknya jika nilai support kecil menandakan bahwa kejadian tersebut sangat jarang terjadi pada data set.

Pada Tabel 10.1 ditunjukkan contoh sebuah data set pembelian barang di sebuah supermarket.

Tabel 10.1: Contoh Data Set Pembelian Barang Di Sebuah Supermarket

ID Transaksi	Item Barang
T-001	Beras, Gula
T-002	Beras, Gula, Kopi
T-003	Beras, Gula, Kopi, Susu
T-004	Beras, Kopi, Minyak
T-005	Beras, Minyak
T-006	Gula, Kopi
T-007	Gula, Kopi, Minyak
T-008	Gula, Kopi, Minyak, Susu
T-009	Gula, Kopi, Susu
T-010	Kopi, Susu

Apabila diketahui sebuah aturan asosiasi {Kopi}→{Gula}, maka nilai support untuk aturan tersebut adalah sebagai berikut:

$$Support(\{Kopi\} \rightarrow \{\mathrm{Gula}\}) = \frac{6}{10} = 0{,}6 = 60\%$$

Jumlah transaksi yang berisi item kopi dan gula sekaligus adalah 6, yaitu pada transaksi: T-002, T-003, T-006, T-007, T-008, dan T-009. Sedangkan total transaksi yang tercatat pada data set adalah 10. Nilai support untuk {Kopi}→{Gula} cukup signifikan, melebihi setengah dari keseluruhan transaksi.

10.2.2 Nilai Confidence

Nilai *confidence* dapat dianggap sebagai nilai akurasi dari sebuah aturan asosiasi. Nilai confidence menggambarkan seberapa kuat asosiasi antar item. Nilai confidence untuk aturan asosiasi X→Y (yang dibahas pada bagian 10.2.1) dapat dihitung sebagai berikut:

$$Confidence(X \rightarrow \mathrm{Y}) = \frac{|X \cup Y|}{[X]} \tag{10.2}$$

Pada Persamaan (10.2), parameter X∪Y merupakan jumlah transaksi pembelian pada supermarket tersebut yang berisi item X dan Y sekaligus pada sebuah transaksi tunggal. Parameter [X] merupakan total transaksi yang memuat item X (Larose and Larose, 2005).

Berdasarkan Persamaan (10.2), dapat dihitung nilai confidence untuk aturan asosiasi {Kopi}→{Gula}. Pada Tabel 10.1 diketahui jumlah transaksi yang memuat item kopi dan sekaligus adalah 6. Diketahui pula jumlah transaksi yang memuat item kopi adalah 8.

Berdasarkan data tersebut, maka nilai confidence untuk {Kopi}→{Gula} adalah sebagai berikut:

$$Confidence(\{Kopi\} \rightarrow \{\mathrm{Gula}\}) = \frac{6}{8} = 0{,}75 = 75\%$$

Nilai confidence untuk {Kopi}→{Gula} cukup signifikan yaitu 75%. Hasil tersebut dapat diartikan bahwa 75% pelanggan yang membeli kopi juga akan membeli gula. Aturan asosiasi {Kopi}→{Gula} cukup kuat dan didukung dengan data bahwa transaksi yang memuat kopi dan gula sekaligus adalah 60% dari seluruh transaksi yang tercatat.

10.3 Algoritma Association Rule

Secara umum, algoritma *association rule* pada data mining terdiri atas dua tahap seperti yang dijelaskan oleh (Han and Kamber, 2006; Aggarwal, 2015), yaitu:

1. Menemukan semua itemset yang frequent, yaitu itemset yang memiliki frekuensi kemunculan sama dengan atau lebih besar dari nilai support minimal yang ditetapkan (min_sup).
2. Membangkitkan semua association rule yang kuat, yaitu aturan yang memenuhi persyaratan nilai support minimal dan nilai confidence minimal yang telah ditetapkan.

Ada beberapa algoritma association rule yang populer digunakan pada data mining. Algoritma Apriori dan FP-Growth merupakan dua algoritma yang sering digunakan untuk menyelesaikan kasus association rule. Kedua algoritma tersebut mampu meningkatkan efisiensi komputasi pada association rule data mining.

Apriori merupakan sebuah algoritma yang akan menemukan semua itemset yang memenuhi persyaratan nilai support minimal yang ditentukan (Wu and Kumar, 2009). Apriori akan membangkitkan kandidat aturan dengan nilai k itemset lebih kecil terlebih dahulu. Nilai k merupakan jumlah kombinasi item. Nilai support dari kandidat aturan yang dihasilkan akan dihitung terlebih dahulu untuk memastikan memenuhi persyaratan nilai support minimal yang ditentukan (Aggarwal, 2015).

Pada buku ini akan dibahas secara khusus contoh perhitungan algoritma Apriori untuk menentukan aturan asosiasi dari sebuah kasus pembelian barang di supermarket seperti pada Tabel 10.1.

Algoritma FP-Growth merupakan salah satu algoritma association rule data mining yang efisien digunakan untuk menganalisis frequent itemset dari sebuah dataset yang sangat besar (Kantardzic, 2011). Seperti halnya Apriori, algoritma FP-Growth juga melakukan perhitungan frequent itemset pada tahap awal. Kemudian akan dilanjutkan dengan pembuatan struktur pohon sesuai dengan itemset yang ada pada dataset (Witten, Frank and Hall, 2011).

Prinsip Dasar Apriori

Pada pembuatan kombinasi item dari sebuah transaksi, memungkinkan akan muncul banyak itemset. Jumlah itemset yang banyak akan membuat proses komputasi lebih lama. Hal tersebut akan mengurangi efisiensi komputasi, sehingga diperlukan algoritma tertentu untuk mengatasi hal tersebut.

Prinsip dasar yang digunakan pada Apriori adalah jika sebuah itemset Z tidak frequent, maka semua bagian dari itemset tersebut juga tidak frequent. Menambahkan item lain A ke dalam Z tidak akan membuat Z menjadi frequent. Prinsip ini akan membuat Aprori lebih efisien dalam proses perhitungan karena mempersempit ruang pencarian kandidat itemset yang frequent melalui proses seleksi kandidat (Larose and Larose, 2005).

Contoh Kasus Perhitungan Apriori

Contoh perhitungan Apriori untuk kasus transaksi pada Tabel 10.1 akan dijelaskan secara bertahap pada bagian ini. Pada contoh ini ditentukan nilai support minimal adalah 3 atau dalam hal ini 30% dari total transaksi dan nilai confidence minimal adalah 70%. Penentuan kedua nilai tersebut dapat disesuaikan tetapi akan memengaruhi hasil akhir algoritma Apriori.

Langkah pertama yang dilakukan adalah menentukan itemset untuk k=1, yaitu item tunggal dari seluruh transaksi. Seluruh item tunggal beserta nilai frequent ditunjukkan pada Tabel 10.2. Nilai frequent menunjukkan jumlah kemunculan item/itemset dari seluruh transaksi yang tercatat. Pada Tabel 10.2 seluruh item memenuhi nilai support minimal sehingga dapat diproses untuk nilai k berikutnya.

Tabel 10.2: Pembentukan Itemset Untuk k=1

Itemset	Frequent
Beras	5
Gula	7
Kopi	8
Minyak	4
Susu	4

Berdasarkan data pada Tabel 10.2, akan dilanjutkan dengan pembuatan itemset untuk k=2, yaitu kombinasi 2 item. Pada Tabel 10.3 ditunjukkan hasil pembentukan itemset untuk k=2 beserta nilai frequent. Pada hasil pembuatan itemset untuk k=2, beberapa itemset memenuhi nilai support minimal dan beberapa lainnya tidak memenuhi.

Pada Tabel 10.3 terdapat empat itemset yang tidak memenuhi nilai support minimal, yaitu: {Beras, Minyak}, {Beras, Susu}, {Gula, Minyak}, dan {Minyak, Susu}. Keempat itemset tersebut tidak akan diproses lagi untuk k=3. Hal ini sesuai dengan prinsip dasar Apriori bahwa pada itemset yang tidak frequent, maka menambahkan item pada itemset tersebut tidak akan membuatnya menjadi frequent.

Tabel 10.3: Pembentukan Itemset Untuk k=2

Itemset	Frequent
{Beras, Gula}	3
{Beras, Kopi}	3
{Beras, Minyak}	2
{Beras, Susu}	1
{Gula, Kopi}	6
{Gula, Minyak}	2
{Gula, Susu}	3
{Kopi, Minyak}	3
{Kopi, Susu}	4
{Minyak, Susu}	1

Proses dilanjutkan dengan pembuatan itemset untuk k=3, yaitu kombinasi 3 item. Pembuatan itemset pada langkah ini disesuaikan dengan data yang sesuai pada Tabel 10.3 dan hasilnya ditunjukkan pada Tabel 10.4.

Tabel 10.4: Pembentukan Itemset Untuk k=3

Itemset	Frequent
{Beras, Gula, Kopi}	2
{Gula, Kopi, Susu}	3

Terdapat beberapa kombinasi item yang tidak muncul pada Tabel 10.4. Kombinasi {Beras, Gula, Minyak} tidak muncul untuk k=3, karena subset itemset tersebut yaitu {Beras, Minyak} dan {Gula, Minyak} tidak memenuhi nilai support minimal seperti yang ditunjukkan pada Tabel 10.3. Berdasarkan prinsip dasar Apriori, menambahkan item pada itemset yang tidak frequent tidak akan membuatnya menjadi frequent. Hal ini membuat Apriori dapat melakukan komputasi lebih efisien.

Proses dilanjutkan dengan pembuatan itemset untuk k=4, yaitu kombinasi 4 item. Pada Tabel 10.1 kombinasi item pada transaksi paling tinggi adalah 4, sehingga proses akan berhenti pada k=4. Berdasarkan data pada Tabel 10.4, itemset yang memenuhi nilai support minimal adalah {Gula, Kopi, Susu}.

Kombinasi k=4 yang memungkinkan adalah {Gula, Kopi, Susu, Beras} dan {Gula, Kopi, Susu, Minyak}.

Akan tetapi itemset {Gula, Kopi, Susu, Beras} tidak frequent karena {Gula, Kopi, Beras} diketahui tidak frequent berdasarkan Tabel 10.4. Sedangkan itemset {Gula, Kopi, Susu, Minyak} juga tidak frequent karena {Gula, Minyak} dan {Minyak Susu} diketahui tidak frequent berdasarkan Tabel 10.3. Dengan demikian, tidak ada aturan yang memenuhi persyaratan untuk k=4.

Setelah menemukan semua frequent itemset, maka langkah selanjutnya adalah membangkitkan semua aturan yang memenuhi persyaratan nilai support minimal dan nilai confidence minimal yang telah ditentukan. Pada Tabel 10.5 ditunjukkan semua aturan yang dibangkitkan berdasarkan hasil frequent itemset. Aturan yang kuat diseleksi berdasarkan nilai support dan nilai confidence. Aturan yang dibangkitkan dikumpulkan dari frequent itemset mulai dari k=1 hingga k=3.

Tabel 10.5: Daftar Aturan Asosiasi Berdasarkan Frequent Itemset

Itemset	Rule	Support	Confidence
{Beras, Gula}	{Beras}→{Gula}	3/10 = 30%	3/5 = 60%
{Beras, Gula}	{Gula}→{Beras}	3/10 = 30%	3/7 = 43%
{Beras, Kopi}	{Beras}→{Kopi}	3/10 = 30%	3/5 = 60%
{Beras, Kopi}	{Kopi}→{Beras}	3/10 = 30%	3/8 = 38%
{Gula, Kopi}	{Gula}→{Kopi}	6/10 = 60%	6/7 = 86%
{Gula, Kopi}	{Kopi}→{Gula}	6/10 = 60%	6/8 = 75%
{Gula, Susu}	{Gula}→{Susu}	3/10 = 30%	3/7 = 43%
{Gula, Susu}	{Susu}→{Gula}	3/10 = 30%	3/4 = 75%
{Kopi, Minyak}	{Kopi}→{Minyak}	3/10 = 30%	3/8 = 38%
{Kopi, Minyak}	{Minyak}→{Kopi}	3/10 = 30%	3/4 = 75%
{Kopi, Susu}	{Kopi}→{Susu}	4/10 = 40%	4/8 = 50%
{Kopi, Susu}	{Susu}→{Kopi}	4/10 = 40%	4/4 = 100%
{Gula, Kopi, Susu}	{Gula, Kopi}→{Susu}	3/10 = 30%	3/6 = 50%
{Gula, Kopi, Susu}	{Susu}→{Gula, Kopi}	3/10 = 30%	3/4 = 75%
{Gula, Kopi, Susu}	{Gula, Susu}→{Kopi}	3/10 = 30%	3/3 = 100%
{Gula, Kopi, Susu}	{Kopi}→{Gula, Susu}	3/10 = 30%	3/8 = 38%
{Gula, Kopi, Susu}	{Kopi, Susu}→{Gula}	3/10 = 30%	3/4 = 75%
{Gula, Kopi, Susu}	{Gula}→{Kopi, Susu}	3/10 = 30%	3/7 = 43%

Berdasarkan hasil pembangkitan aturan pada Tabel 10.5, maka diketahui aturan-aturan yang memenuhi persyaratan nilai support minimal dan nilai confidence minimal adalah sebagai berikut:

1. {Gula}→{Kopi}: IF membeli Gula THEN membeli Kopi dengan support = 60% dan confidence = 86%.
2. {Kopi}→{Gula}: IF membeli Kopi THEN membeli Gula dengan support = 60% dan confidence = 75%.
3. {Susu}→{Gula}: IF membeli Susu THEN membeli Gula dengan support = 30% dan confidence = 75%.
4. {Minyak}→{Kopi}: IF membeli Minyak THEN membeli Kopi dengan support = 30% dan confidence = 75%.
5. {Susu}→{Kopi}: IF membeli Susu THEN membeli Kopi dengan support = 40% dan confidence = 100%.
6. {Susu}→{Gula, Kopi}: IF membeli Susu THEN membeli Gula AND Kopi dengan support = 30% dan confidence = 75%.
7. {Gula, Susu}→{Kopi}: IF membeli Gula AND Susu THEN membeli Kopi dengan support = 30% dan confidence = 100%.
8. {Kopi, Susu}→{Gula}: IF membeli Kopi AND Susu THEN membeli Gula dengan support = 30% dan confidence = 75%.

Demikianlah contoh perhitungan algoritma apriori untuk association rule. Proses yang dilakukan meliputi penentuan itemset yang frequent dan pembangkitan aturan asosiasi yang kuat. Itemset yang frequent diseleksi berdasarkan nilai support minimal yang ditentukan. Dari itemset yang frequent akan dibangkitkan aturan asosiasi yang kuat, yaitu memenuhi persyaratan nilai support minimal dan nilai confidence minimal.

Aturan asosiasi yang kuat dapat dijadikan sebagai acuan bahwa itemset tersebut memiliki nilai korelasi yang kuat. Aturan asosiasi kuat yang dihasilkan pada langkah terakhir merupakan sebuah pengetahuan yang akan digunakan pada pengambilan keputusan sesuai kasus yang diamati.

Bab 11

Time Series Analysis

11.1 Pendahuluan

Era digitalisasi sejatinya sudah berkembang dari ratusan tahun yang lalu. Setelah revolusi industri 4.0, sekarang berkembang menuju era 5.0. Era 5.0 disebut sebagai era society yang didefinisikan sebagai era yang memusatkan manusia yang menyelesaikan masalah sosial dengan mengintegrasikan kehidupan sosial dengan teknik saat revolusi industri misalnya *Artificial Intelligence* (AI) ataupun Big Data.

Oleh karena itu, secara sederhana era 5.0 dapat didefinisikan sebagai era yang menggunakan teknologi untuk menyelesaikan permasalahan dalam kehidupan sosial. Salah satu bentuk penerapan teknologi yaitu peramalan kejadian masa yang akan datang pada sebuah kasus tertentu pada data dengan metode yang ada dalam ilmu data mining.

Secara definisi, peramalan dan prediksi adalah sesuatu yang sama. Peramalan adalah proses untuk memperkirakan sebuah kejadian atau fenomena yang akan terjadi di masa depan berdasarkan data yang telah dimiliki di masa lalu dan sekarang. Peramalan dilakukan untuk memberikan pertimbangan terhadap keputusan yang akan diambil. Peramalan tidak mengharuskan adanya jawaban pasti, namun mengusahakan agar akurasi data mendekati yang semestinya (Kafil, 2019).

Peramalan atau prediksi umumnya dilakukan pada data dengan deret waktu (time series). Data deret waktu (time series) adalah data dengan periode waktu tertentu. Salah satu bentuk peramalan atau prediksi maupun analisa trend adalah saat menentukan jumlah pasien COVID-19 di masa yang akan datang. Peramalan telah diterapkan ke dalam berbagai bidang ilmu seperti bisnis, industri, pemerintahan, maupun kesehatan.

Secara umum, periode waktu peramalan dibagi menjadi tiga jenis jangka waktu yaitu pendek, menengah, dan panjang. Jangka pendek umumnya meramalkan sesuatu dalam jangka waktu harian, mingguan dan bulanan pada masa depan. Jangka menengah umumnya meramalkan sesuatu dalam jangka waktu lebih panjang yaitu satu tahun atau dua tahun. Sedangkan jangka panjang dapat dilakukan untuk beberapa tahun ke depan (Montgomery, et al., 2015). Periode waktu ini menyebabkan jenis data deret waktu (time series) yang berbeda.

Di dalam bab ini, nantinya akan dibahas mengenai (1) konsep dasar analisis data deret waktu (time series), (2) jenis data deret waktu (time series), (3) tahapan analisa data deret waktu (time series), (4) metode dalam analisis data deret waktu (time series).

11.2 Konsep Dasar Analisis Data Deret Waktu (Time Series)

Analisa data deret waktu (time series analysis) adalah sebuah proses untuk mendapatkan informasi bermakna dalam sebuah data yang memiliki waktu dan dilakukan secara statistik untuk memprediksi kejadian di masa depan dari pola kejadian masa lalu (Nielsen, 2020). Analisis deret waktu (time series) adalah salah satu bentuk tugas dalam ilmu data mining yang masuk dalam bidang peramalan. Peramalan data deret waktu (time series) sudah diterapkan dalam berbagai bidang seperti bisnis, industri, pemerintahan, maupun kesehatan (Montgomery, et al., 2015).

Data deret waktu (time series) diklasifikasikan menjadi dua yaitu kontinu dan diskrit. Data deret waktu (time series) kontinu dapat dijadikan dalam bentuk diskrit, adapun caranya adalah sebagai berikut (Kusdarwati, et al., 2022):

1. Menjumlahkan data dengan satu kurun waktu. Contohnya adalah hasil produksi susu tahunan adalah jumlah produksi susu dalam satu tahun.
2. Menghitung rata – rata data dengan satu kurun waktu tertentu. Contohnya adalah rata – rata jumlah kunjungan pasien bulanan.
3. Mengukur sebuah data dengan titik waktu tertentu. Contohnya adalah menghitung tekanan darah rutin setiap bulan atau pada waktu melakukan cek kesehatan.

Data deret waktu (time series) terdapat di berbagai bidang kehidupan, dan utamanya digunakan untuk kegiatan peramalan maupun prediksi kejadian di masa yang akan datang. Prediksi tersebut didasarkan pada pola kejadian di masa lalu.

Adapun bidang dengan data deret waktu (time series) seperti data pada bidang ilmu bisnis untuk meramalkan trend penjualan bisnis, industri, pemerintahan, maupun kesehatan. Data deret waktu (time series) digambarkan dalam satuan deret waktu dengan nilai amatan (Kusdarwati, et al., 2022). Visualisasi data deret waktu dapat dilakukan dengan berbagai macam aplikasi seperti Ms.Excel, R Studio, Orange Data Mining maupun Weka.

11.3 Jenis Data Deret Waktu (Time Series)

Data deret waktu (Time Series) dapat divisualisasikan dalam bentuk grafik menggunakan alat bantu aplikasi seperti Ms.Excel, R Studio, Orange Data Mining maupun Weka. Ketika divisualisasikan, data deret waktu (time series) memiliki pola tersendiri.

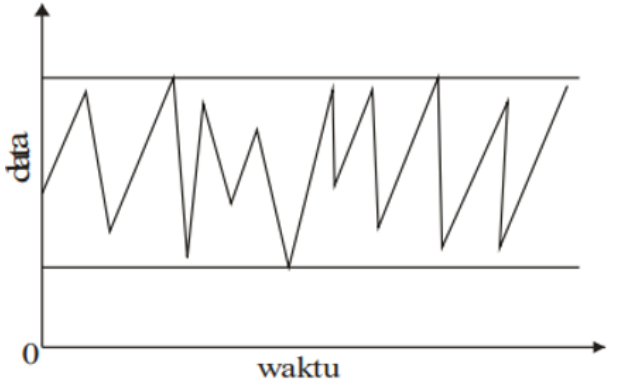
Pola ini dapat dipengaruhi berbagai macam aspek seperti kondisi ekonomi, kondisi pemerintahan, ataupun kondisi lain yang terjadi di masa yang akan datang. Pola Data Deret Waktu (Time Series) dibagi menjadi empat jenis,

antara lain yaitu pola horizontal (stationer), pola trend, pola musiman (seasonal) dan pola sillis (Hanke & Wichern, 2005).

Pola Data Horizontal (Stationer)

Pola Data Horizontal (Stasioner) adalah pola data yang terjadi ketika perubahan data yang diamati (observasi) berada pada rata – rata yang konstan. Perubahan data tidak ada yang terlalu mencolok naik ataupun mencolok turun dan cenderung konsisten. Contohnya adalah rata-rata kunjungan pasien rawat jalan di sebuah fasilitas pelayanan kesehatan yang konstan.

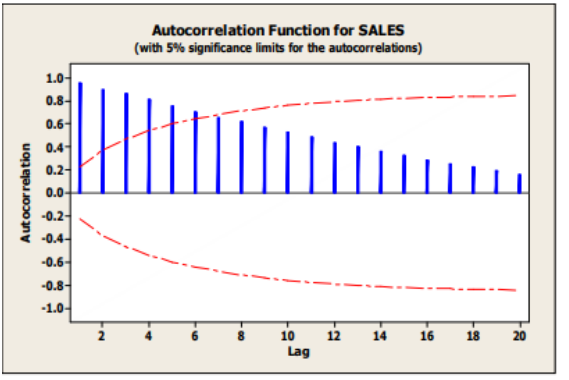
Adapun jika divisualisasikan adalah sebagai berikut:



Gambar 11.1: Pola Data Horizontal (Stationer) (Munawaroh, 2010)

Pola Data Trend

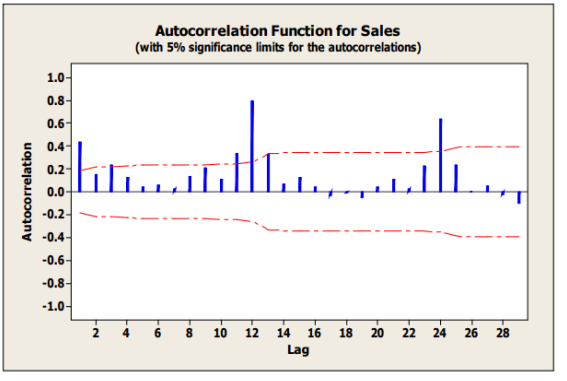
Pola Data Trend adalah pola data yang terjadi ketika perubahan data yang diamati (observasi) berada pada tingkatan dan cenderung menaik ataupun menurun pada satu periode waktu. Sehingga ketika ditarik sebuah garis regresi akan ditemukan garis lurus. Contohnya adalah data jumlah penduduk sebuah negara yang cenderung naik dan tidak turun. Adapun jika divisualisasikan adalah sebagai berikut:



Gambar 11.2: Pola Data Trend (Santoso, 2009)

Pola Data Musiman (Seasonal)

Pola Data Musiman (seasonal) adalah pola data yang terjadi ketika perubahan data yang diamati (observasi) dipengaruhi oleh kondisi musiman sehingga dapat diamati sebagai kejadian berulang. Contohnya adalah data penjualan buah durian akan meningkat ketika musim durian itu tiba. Adapun jika divisualisasikan adalah sebagai berikut:

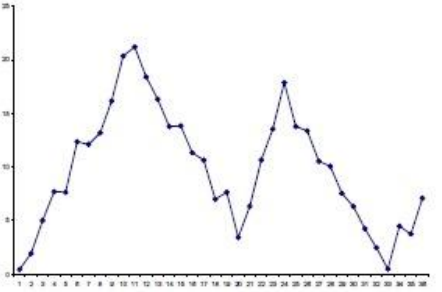


Gambar 11.3: Pola Data Musiman (Hanke & Wichern, 2005).

Pola Data Siklis

Pola Data Siklis adalah pola data yang terjadi ketika perubahan data yang diamati (observasi) berubah dengan sangat cepat (fluktuatif) dan bergelombang. Pola ini erat kaitannya dengan pola data trend. Contohnya adalah data penjualan barang yang dipengaruhi oleh kondisi ekonomi.

Adapun jika divisualisasikan adalah sebagai berikut:



Gambar 11.3: Pola Data Siklis (Montgomery, et al., 2015).

11.4 Tahapan Analisis Data Deret Waktu (Time Series)

Analisa Data Deret Waktu (time series) merupakan salah satu kegiatan yang dapat diselesaikan dalam bidang ilmu Data Mining yaitu peramalan. Oleh karena itu, tahapan dalam analisis data deret waktu dilakukan berdasarkan tahapan pengolahan data dalam bidang ilmu data mining. Kasus tersebut kemudian diselesaikan menggunakan metode perhitungan data deret waktu seperti *regresi linear*, *exponential smoothing*, maupun *moving average*.

Tahapan pengolahan data dalam bidang ilmu data mining untuk peramalan tentunya terdiri atas serangkaian proses yang saling terkoneksi satu sama lain untuk menghasilkan sebuah tujuan (output). Setiap tahapan harus dilalui dan tidak boleh terlewat.

Tahapan tersebut antara lain ialah:

1. Identifikasi masalah
   Tahapan identifikasi masalah merupakan tahapan untuk menganalisis dan memahami kejadian yang akan diramalkan. Tahapan identifikasi masalah harus mengakomodasi keinginan pengguna. Peramalan apakah ditujukan dalam periode harian, mingguan, ataupun bulanan. Selain itu, dalam tahapan ini harus dipertimbangkan pendekatan yang dibutuhkan untuk pemodelan kasus.

2. Pengumpulan data
   Tahapan pengumpulan data merupakan tahapan untuk mengumpulkan data yang dibutuhkan. Tahapan ini juga harus memperhatikan atribut yang relevan serta rentang waktu yang diinginkan. Umumnya dalam tahapan ini akan menemukan data yang outlier, missing value, atau tidak sesuai dengan bentuk keseluruhan data. Maka data tersebut wajib masuk ke preprocessing data terlebih dahulu agar mendapatkan data yang berkualitas.
3. Analisa data
   Tahapan analisa data adalah tahapan visualisasi data untuk memperoleh model peramalan yang cocok dengan data. Tahapan ini adalah tahapan terpenting dalam peramalan. Setiap jenis data deret waktu (time series) dapat diselesaikan dengan metode yang berbeda. Data dengan jenis trend model peramalannya akan berbeda dengan jenis data musiman.
4. Seleksi model
   Tahapan seleksi model adalah tahapan untuk memilih model peramalan yang cocok dengan data. Model peramalan akan memengaruhi tingkat akurasi. Ketika sudah menemukan model peramalan yang akan digunakan, maka kita harus mempertimbangkan parameter yang akan digunakan dalam model tersebut.
5. Validasi model
   Tahapan validasi model adalah tahapan evaluasi model peramalan yang dipilih dengan cara mengukur keakuratan model tersebut. Tahapan validasi model ini ditujukan untuk melihat kesesuaian model yang digunakan dengan studi kasus. Kesesuaian ini tentunya mengukur tingkat kesalahan model.
6. Pembangunan model peramalan
   Tahapan pembangunan model peramalan adalah tahapan yang ditujukan untuk memastikan bahwa model peramalan berhasil untuk digunakan user, serta kinerjanya akan dipantau secara terus menerus berkelanjutan.

7. Evaluasi akurasi peramalan
   Tahapan Evaluasi Akurasi Peramalan merupakan tahapan yang ditujukan untuk memantau tingkat kesalahan peramalan model. Metode evaluasi akurasi peramalan antara lain yaitu *Mean Absolute Deviation* (MAD), *Mean Absolute Percentage Error* (MAPE), dan masih banyak lainnya.

Sub bab selanjutnya adalah bagian yang membahas tentang metode yang digunakan untuk memodelkan data deret waktu. Adapun metode yang dibahas yaitu (1) Moving Average; (2) Exponential Smoothing.

11.5 Metode Dalam Analisis Data Deret Waktu (Time Series)

Seperti yang telah dijelaskan pada sub bab sebelumnya, metode dalam peramalan adalah hal yang paling berpengaruh dan berdampak pada tingkat akurasi. Dalam analisis data deret waktu (time series) terdapat beberapa metode yang sering digunakan yaitu moving average, exponential smoothing dan regressi linear.

Bab ini akan memaparkan konsep dasar, rumus, dan cara perhitungan dari masing – masing metode.

11.5.1 Moving Average (Metode Rata – Rata Bergerak)

Moving Average adalah sebuah metode yang cara perhitungannya berdasarkan rata – rata bergerak dari suatu data dengan deret waktu suatu data yang bergelombang. Rata – rata bergerak dihitung dengan cara merata – rata sebuah data secara berturut – turut, baik dalam periode harian, bulanan, maupun tahunan.

Dengan demikian, maka akan ditemukan rata – rata bergerak secara teratur atas dasar jumlah tahun tertentu. Moving Average tidak menentukan jangka waktu yang harus dicari rata – ratanya. Jika yang dicari adalah 4 tahun rata – rata bergerak, peramalannya menggunakan rata – rata dari 4 tahun sebelumnya (Arifin, 2007).

Moving Average menurut Pangestu (2013), adalah sebuah metode yang kurang cocok untuk digunakan pada data deret waktu dengan jenis trend dan musiman (seasonal). Moving Average terbagi menjadi *Single Moving Average* dan *Double Moving Average*. Semakin panjang periode waktu rata – rata bergerak, maka hasil peramalan akan semakin baik. Hal ini sejalan dengan sifat dari rata – rata, di mana mempertimbangkan semua nilai pengamatan dan sangat dipengaruhi oleh nilai ekstrem besar atau kecil (Purwanto, 1994).

Adapun rumus untuk menghitung Moving Average adalah sebagai berikut:

$$P_t = \frac{D_t + D_{t-1} + \cdots + D_{t-n+1}}{n}$$

P_t adalah peramalan ke t, D_t adalah data aktual pada periode t, D_{t1} adalah data aktual pada periode t-1, n banyaknya periode dalam rata – rata bergerak. Sebagai contoh pada Tabel 11.1 terdapat data penjualan barang dari bulan Januari hingga Juni. Peramalan di masa yang akan datang dilakukan dengan metode Moving Average.

Hitung Moving Average 3 bulan pada kasus tersebut.

Tabel 11.1: Data Penjualan Barang (Sumber Data Pribadi)

No	Bulan	Jumlah Penjualan
1	Januari	22
2	Februari	23
3	Maret	21
4	April	19
5	Mei	24
6	Juni	26

Maka ditemukan:

Moving Average (April) $: \frac{22+23+21}{3} = 22$

Moving Average (Mei) $: \frac{23+21+19}{3} = 21$

Moving Average (Juni) $: \frac{21+19+24}{3} = 21{,}3$

Hasil perhitungan diatas, merupakan hasil peramalan dengan metode Moving Average. Hasil peramalan memiliki hasil yang cukup baik karena mendekati dengan data aktual.

11.5.2 Single Exponential Smoothing

Metode peramalan yang kedua yang sering digunakan dalam analisis data deret waktu (time series) adalah Metode Single Exponential Smoothing. Metode ini menggunakan konstanta pemulusan (α) agar mendapatkan hasil peramalan yang lebih mulus dan berdampak pada tingkat kesalahan yang lebih kecil.

Metode ini cocok digunakan pada data deret waktu (time series) dengan jenis horizontal (stasioner) dan bukan pada data trend. Selain itu, metode ini umumnya memberikan hasil peramalan lebih tinggi daripada data pengamatan.

Adapun rumus metode Single Exponential Smoothing adalah sebagai berikut:

$$F_{t+1} = \alpha\, X_{t-1} + (1-\alpha) F_{t-1}$$

Keterangan:

- F_{t+1} = ramalan untuk periode ke t+1
- X_t = nilai riil periode ke t
- α = konstanta pemulusan ($0 < \alpha < 1$)
- F_{t-1} = ramalan untuk periode ke t-1

Metode Single Exponential Smoothing mengacu pada konstanta pemulusan (α) untuk mendapatkan tingkat akurasi yang baik pada sebuah kasus. Nilai konstanta tersebut diubah – ubah (trial dan error) agar mendapatkan konstanta terendah. Nilai konstanta berkisar antara 0.1 sampai dengan 0.9. Sebagai contoh pada Tabel 11.2 terdapat data penjualan barang selama 3 minggu. Ramalkan penjualan barang minggu depan menggunakan metode single exponential smoothing !

Tabel 11.2: Data Penjualan Barang (Sumber Data Pribadi)

No	Minggu Ke-	Jumlah Penjualan
1	1	105
2	2	113
3	3	111
4		

Maka ditemukan:

Asumsikan peramalan pada minggu ke-1 adalah sama dengan data actual dan nilai konstanta pemulusan adalah 0.5, sehingga:

Peramalan minggu ke-2 : 105 + (1-0.5) * (105-105) = 105

Peramalan minggu ke-3 : 113 + (1-0.5) * (105-113) = 105

Peramalan minggu ke-4 : 111 + (1-0.5) * (109-0) = 110

Hasil perhitungan diatas, merupakan hasil peramalan dengan metode Single Exponential Smoothing. Hasil peramalan memiliki hasil yang cukup baik, karena mendekati dengan data aktual. Untuk mengetahui apakah hasil peramalan cukup akurat dengan kedua metode yang telah dijelaskan sebelumnya, maka pengguna dapat mengukur akurasi dengan *Mean Absolute Deviation* (MAD), *Mean Absolute Percentage Error* (MAPE), maupun *Mean Square Error* (MSE) ataupun dengan ukuran akurasi lainnya.

Bab 12

Text Mining

12.1 Pendahuluan

Text mining adalah sebuah bidang interdisipliner yang menggabungkan beberapa lintas ilmu pengetahuan. Seperti IF (Information Retrieval), data mining, machine learning, statistika, dan computational linguistic. Text mining merupakan sebuah proses ekstraksi pengetahuan dari data dalam bentuk teks (Jo, 2015).

Data teks berada dalam variasi format dokumen, seperti berkas doc/docx, berkas PDF, dan juga teks dari HTML. Tantangan utama dalam text mining terletak pada analisis dan memodelkan bahasa yang tidak terstruktur dari sekumpulan data teks (Zong, Xia and Zhang, 2021). Hal ini muncul karena data yang disimpan di dalam basis data berbentuk semi terstruktur dan perlu teknik data mining yang spesial untuk menggali informasi penting dari sekumpulan data tersebut yang dinamakan dengan text mining.

Data teks disebut dengan data yang tidak/semi terstruktur karena data tersebut masih perlu dilakukan pra proses untuk menghasilkan data terstruktur. Jika data dapat dituliskan dalam format tabel, maka data tersebut dapat dikatakan terstruktur dengan baik (Villepastour, 2019).

Pada proses data mining, data perlu ditransformasikan menjadi data matang yang dapat direpresentasikan dalam bentuk tabular. Hal inilah yang menjadi

salah satu proses di dalam text mining, yaitu membuat data yang tidak/semi terstruktur diubah ke dalam bentuk yang terstruktur.

Relasi Text Mining Dengan Data Mining

Bidang kajian text mining adalah sub-bidang dari bidang data mining. Relasi dari text mining dengan data mining adalah variasi pembelajaran dalam menghasilkan informasi dari teks dengan aspek yang berbeda.

Di antaranya adalah sebagai berikut:

1. Categorization of documents - Memberikan kategori terhadap dokumen. Contohnya mengategorikan artikel pada koran, memberikan label pada surat elektronik sebagai spam atau tidak.
2. Clustering - Pengelompokan dokumen berdasarkan kemiripannya. Contohnya untuk mengidentifikasi dokumen.
3. Summarization - Mencari bagian terpenting dari beberapa dokumen.
4. Information retrieval - Mendapatkan dokumen yang cocok dengan kueri yang merepresentasikan informasi dari koleksi dokumen yang besar.
5. Extracting the meaning of documents of their part - Mengidentifikasi topik tersembunyi, analisa sentimen, pendapat dan emosi.
6. Information extraction - Ekstraksi informasi yang terstruktur seperti entitas, event.
7. Association mining - Mencari asosiasi di antara konsep dan aturan dalam teks.
8. Trend analysis - Mencari tahu bagaimana konsep yang dikandung di dalam dokumen seiring berjalannya waktu.
9. Machine translation - Mengubah teks yang ditulis dalam sebuah bahasa ke bahasa lainnya.

Proses Pada Text Mining

Untuk mendapatkan informasi yang berguna dari teks pada sekumpulan dokumen dengan menggunakan beberapa tahapan. Tahapan tersebut adalah text mining process yang meliput beberapa hal berikut:

1. Mendefinisikan masalah
   Membuat dan memperjelas lingkup permasalahan (problem domain) sehingga pertanyaan dapat dijawab dan didefinisikan dengan jelas.
2. Mengumpulkan data
   Sumber data harus ditentukan dengan jelas sesuai dengan kebutuhan pada permasalahan. Data dapat bersumber dari basis data, api, dan sumber luar lainnya.
3. Mendefinisikan fitur
   Fitur yang bagus adalah fitur yang merepresentasikan objek dari permasalahan yang telah dipilih. Fitur harus mengkarakteristikkan teks dan sesuai terhadap permasalahan yang telah didefinisikan sebelumnya.
4. Menganalisis data
   Tahap ini adalah mencari pola dari data berdasarkan tipe permasalahan yang akan diselesaikan. Contohnya dengan klasifikasi, klastering atau regresi.
5. Menginterpretasikan hasil
   Hasil di sini diperoleh dari analisa data. Pada tahap ini harus lebih berhati-hati karena mencakup langkah-langkah verifikasi dan validasi untuk meningkatkan keandalan dari model text mining yang dibuat.

Tugas Utama Text Mining

Text mining adalah domain ilmu lintas teknologi. Pada praktiknya, text mining selalu mengombinasikan beberapa teknologi yang saling berhubungan untuk dapat menyelesaikan permasalahannya. Sebagai contoh sebuah mesin penerjemah. Sistem membutuhkan kamus penerjemah, memiliki penyaring kata supaya mendapatkan kata dasar dari sebuah kalimat, dan sistem juga memiliki generator untuk menerjemahkannya kalimat yang dimasukkan. Proses dari tiap teknologi tersebut memiliki tugasnya masing-masing, sehingga text mining melakukan kolaborasi beberapa teknologi pada penerapannya.

Berikut adalah beberapa tugas yang dapat dilakukan oleh text mining:

1. Text classification
   Klasifikasi teks adalah penerapan dari pengenalan pola. Berdasarkan pola yang terkandung dalam data, data akan diklasifikasikan ke

dalam beberapa kelas yang sudah ditentukan sebelumnya. Contohnya sebuah artikel ditulis dan sistem secara otomatis mengklasifikasikan artikel tersebut sebagai blog, berita, novel, atau Entertainment.

2. Text clustering
   Tujuan dari klasterisasi teks adalah mengelompokkan data sekumpulan teks menjadi beberapa kategori.
3. Topic model
   Secara umum setiap artikel memiliki topik dan beberapa sub-topik. Topik tersebut dapat mengekspresikan sekelompok dokumen dengan korelasi yang kuat terhadap topik yang sama.
4. Text sentiment analysis dan opinion mining
   Sentimen teks merujuk pada informasi subjektif yang dikandung pada teks berdasarkan sudut pandang penulisnya. Tujuan utama dari sentimen teks analisis adalah untuk menggali pendapat (opinion mining) dari penulisnya. Contohnya adalah cuitan pada Twitter tentang sebuah produk. Dengan sentimen teks analisis, kita dapat mengetahui seberapa banyak orang yang setuju terhadap produk tersebut dan akan tetap menggunakannya atau sebaliknya.
5. Topic detection dan tracking
   Deteksi topik biasanya merujuk pada kuantitas dari laporan sebuah berita atau komentar. Topik dengan banyaknya orang peduli dan memperhatikannya dan seberapa banyak orang yang mencarinya, dinamakan dengan hot topics. Biasanya hal ini berkaitan dengan viral atau trendingnya sebuah konten di dunia maya.
6. Information extraction
   Ekstraksi informasi merujuk pada ekstraksi informasi faktual seperti entitas, atribut, relasi antar entitas dari sekumpulan dokumen. Ekstraksi informasi bertugas untuk rekognisi entitas atau ekstraksi hubungan antar entitas. Contohnya pada bidang biomedis. Terdapat gejala yang mirip/serupa antara satu penyakit dengan penyakit lainnya. Dengan ekstraksi informasi dapat mengetahui spesifik penyakit berdasarkan simtom yang ditunjukkan dan penanganan yang paling tepat.

7. Automatic text summarization
   Otomasi ringkasan teks ini merujuk pada teknologi yang dapat membuat ringkasan kalimat secara otomatis. Contohnya adalah membuat narasi untuk sebuah konten dengan diberikan beberapa kata kunci tertentu yang mendeskripsikan konten tersebut.

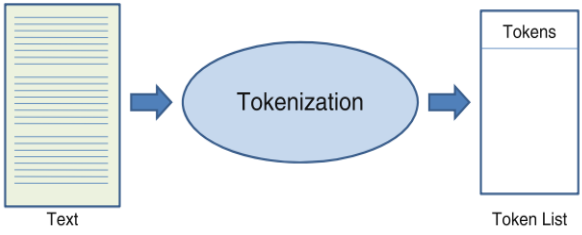
12.2 Representasi Teks

Pada umumnya text mining membutuhkan input yang direstrukturisasi, contohnya dengan melakukan parsing, menghilangkan kata yang tidak penting atau mengekstrak kata dasar dari kata yang memiliki imbuhan (Han, Kamber and Pei, 2012).

Berikut adalah beberapa cara untuk merepresentasikan teks ke di dalam text mining:

Tokenization

Tokenisasi adalah sebuah proses untuk menyegmentasi kata dari sekumpulan teks menjadi token (kata tunggal) yang dipisahkan dengan spasi atau tanda baca (Chen, 2020). Gambar 12.1 diatas menunjukkan bahwa sebuah dokumen berisi sekumpulan teks kemudian dilakukan proses tokenisasi sehingga menghasilkan daftar token.



Gambar 12.1: Proses Tokenisasi Pada Dokumen (Jo, 2019)

Sebagai contoh, lihat Tabel 12.1 berikut, di sana terdapat satu buah teks yang berisi dua buah kalimat.

Tabel 12.1: Contoh Teks Yang Akan Dilakukan Tokenisasi

Teks	Buku yang saya beli sungguh menarik dan menambah wawasan. Buku mudah untuk didapatkan.

Teks pada Tabel 12.1 kemudian dilakukan tokenisasi sehingga menghasilkan tabel token seperti pada Tabel 12.2 berikut. Tiap-tiap kata hasil tokenisasi pada tabel dapat disebut juga dengan term.

Tabel 12.2: Hasil Tokenisasi

Tokenisasi
Buku
yang
saya
beli
sungguh
…
didapatkan

Stemming

Stemming adalah mengubah kata hasil tokenisasi menjadi kata dasar. Pada literatur lain, stemming juga disebut dengan lemmatization dan sering digunakan sebagai praproses teks (Dalianis, 2018). Lihat Tabel 12.3 sebagai contoh proses dari stemming.

Kata pada kolom tokenisasi adalah hasil tokenisasi dan masih memiliki kata dengan imbuhan awalan dan akhiran. Setelah dilakukan stemming, hasilnya terdapat pada kolom stemming. Pada kolom tersebut berisi kata dasar atau kata akar.

Tabel 12.3: Hasil Stemming

Tokenisasi	Stemming
Memberi	beri
makan	makan
hewan	hewan
bertanduk	tanduk

Filtering

Filtering adalah tahap untuk menyaring terms. Pada kenyataannya di dalam suatu dokumen terdapat kata yang tidak memiliki kontribusi pada kalimat tersebut. Hal ini dapat memengaruhi performa dari proses mining yang akan dilakukan. Pada filtering ini, kata tersebut akan dihilangkan atau dihapuskan.

Kata yang akan dihilangkan disebut dengan stop words. Contohnya adalah kata tunjuk atau kata imbuhan. Seperti ini, itu, dia, dan lain-lain. Tabel 12.4 menunjukkan penerapan dari filtering. Dapat dilihat bahwa kata *woi*, ke dihilangkan pada tahap ini.

Tabel 12.4: Contoh Hasil Filtering

text	woi Kemarin saya pergi ke rumah kamu @xyz
filtering	kemarin saya pergi rumah kamu

Normalization

Proses normalisasi di sini adalah mengubah semua terms sehingga menghasilkan bentuk yang seragam. Salah satu caranya case folding. Ada dua opsi dalam melakukan transformasi bentuknya, dengan mengubah terms menjadi huruf besar (kapital) atau huruf kecil semuanya. Tabel 12.5 di bawah adalah contoh penerapan dari normalisasi.

Tabel 12.5: Contoh Hasil Normalisasi

teks	Produk ini bagus dan Penjual ramah
case folding	produk ini bagus dan penjual ramah

Bag-of-Words

BOW (bag-of-words) adalah sebuah cara untuk mengelompokkan kata atau beberapa kata ke dalam satu grup tertentu. Masing-masing dokumen pada dataset direpresentasikan sebagai bag.

Sebagai contoh lihat pada Tabel 12.6 di bawah:

Tabel 12.6: Contoh Sekumpulan Data Teks

kalimat 1	produk ini bagus
kalimat 2	sangat sangat bagus
kalimat 3	tidak bagus
kalimat 4	produk ini biasa

Tabel diatas memiliki 4 buah kalimat (dokumen) yang berisi seperti yang ditunjukkan pada tabel. Pada BOW, kata (term) akan dipisahkan dan menjadi kamus tersendiri. Kemudian masing-masing dokumen akan dilakukan perhitungan banyaknya kemunculan kata sehingga dibentuklah tabel matrix seperti berikut.

Tabel 12.7: Hasil Matrix Fitur Dari BOW

dokumen	produk	ini	bagus	sangat	tidak	biasa
kalimat 1	1	1	1	0	0	0
kalimat 2	0	0	1	2	0	0
kalimat 3	0	0	1	0	1	0
kalimat 4	1	1	0	0	0	1

Berdasarkan Tabel 12.7 diatas, angka-angka di dalamnya adalah atribut/fitur yang akan digunakan pada tahap mining berikutnya. Angka 1 berarti kemunculan kata pada dokumen tersebut terjadi sekali. Begitu juga dengan angka 2 yang menunjukkan kata tersebut muncul sebanyak dua kali. Contohnya pada kalimat 2 pada kata "sangat". Jika angka menunjukkan nol berarti kata tersebut tidak muncul sama sekali.

BOW memiliki batasan dan kekurangan. Salah satunya adalah fitur bisa menjadi sangat banyak. Bahkan pada dokumen yang kecil sekalipun bisa jadi memiliki banyak kata unik yang dikandungnya. Sehingga dimensi dari data set bisa membesar. Tingginya jumlah fitur berpengaruh ke dalam kompleksitas komputasi dari komputer sehingga membutuhkan waktu yang cenderung lebih lama saat memprosesnya (Žižka, Dařena and Svoboda, 2019).

Term Weighting

Term weighting adalah salah satu cara dalam memberikan bobot pada kata. Hal ini berguna untuk mengonversi data ke dalam bentuk vektor numerik di mana tiap vektornya merepresentasikan teks data (Tripathy, Agrawal and Rath, 2015). Salah satu teknik yang sering digunakan adalah TF-IDF (term frequency-inverse document frequency). TF-IDF merefleksikan kata penting di dalam corpus (kumpulan kata) atau dokumen (Tripathy, Agrawal and Rath, 2016).

TF-IDF bekerja dengan mencari frekuensi kemunculan kata dari sekumpulan dokumen. TF (term frequency) yaitu frekuensi banyaknya kemunculan kata dari tiap dokumen. IDF (inverse document frequency) digunakan untuk menghitung bobot kata yang berbeda dari dokumen. Kata yang langka / jarang muncul akan memiliki nilai IDF yang tinggi. Hasil akhir dari pembobotan dengan TF-IDF adalah fitur dalam bentuk matrik vector.

Cara menghitung TI-IDF ditunjukkan pada persamaan berikut:

$$tfidf(t_j, d_i) = tf(t_j, d_i) \times (\log \frac{N}{df(t_j)}) + 1$$

di mana (t_j, d_i) adalah frekuensi kemunculan kata t_j dalam spesifik dokumen, $df(t_j)$ adalah banyaknya dokumen di mana t_j itu muncul dan N adalah nilai keseluruhan banyaknya dokumen. Berdasarkan persamaan diatas, rumus IDF $= (\log \frac{N}{df(t_j)}) + 1$. Penambahan 1 supaya nilai IDF tidak bernilai 0. Apabila tanpa penambahan 1, maka hasil perkalian TF dengan IDF akan menghasilkan nilai 0.

Bab 13

Data Mining Dalam Big Data

13.1 Pendahuluan

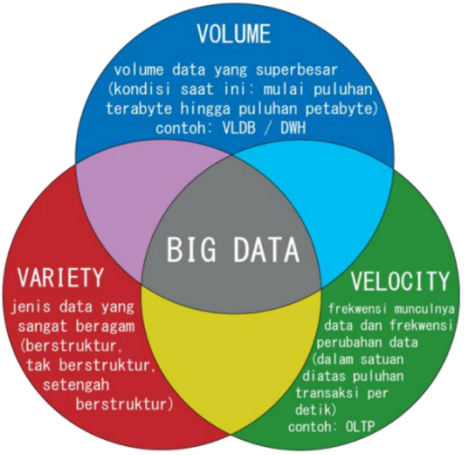
Fenomena big data menarik banyak perhatian dan jadi bahan perbincangan, khususnya persoalan data yang tidak hanya mencakup bagaimana data ditemukan dan ditempatkan, melainkan bagaimana data bisa dimanfaatkan oleh banyak orang.

Pentingnya big data terbukti dalam fenomena di mana pendapatan di berbagai sektor industri yaitu melalui mengekstrak informasi yang paling cocok dan juga bermanfaat dalam menciptakan peluang maupun strategi bisnis yang dikelola. Informasi diperoleh dari berbagai sumber seperti melalui sosial media, data transaksi, hingga data pelayanan. Semakin kompleks informasi yang ingin diperoleh dan di analisis, semakin besar data yang akan diekstrak (Mantik and Awaludin, 2023).

Big data identik dengan istilah kumpulan data yang sangat besar dan kompleks yang tidak dapat diproses menggunakan alat pengelola database konvensional atau aplikasi pemrosesan data lainnya.

Big data memiliki 3 karakteristik yang membedakannya dengan data lainnya yaitu:

1. Volume, mengacu pada jumlah data yang perlu dikelola dalam skala yang sangat besar.
2. Variety, mengacu pada sifat-sifat sumber data yang sangat berbeda, baik data terstruktur maupun data tidak terstruktur.
3. Velocity, menunjukkan kecepatan di mana pemrosesan data harus mengatasi peningkatan jumlah data yang cepat.



Gambar 13.1: Karakteristik Big Data (Sawitri, 2019)

Perbedaan big data dengan konsep data masif dan data yang sangat besar dapat dilihat berdasarkan 3 tipe definisi yaitu:

1. Definisi atribut yang menjelaskan dimensi dari data.
2. Definisi arsitektur yang menyatakan bahwa big data adalah tempat di mana kemampuan untuk melakukan analisis yang efektif menggunakan pendekatan tradisional dibatasi oleh volume data, kecepatan akuisisi, atau representasi data.
3. Definisi komparatif, big data didefinisikan sebagai data yang memiliki ukuran di luar kemampuan perangkat lunak basis data biasa dalam menangkap, menganalisis, menyimpan dan mengelola.

Proses menggali dan mengolah data menjadi informasi yang berharga untuk mengambil keputusan disebut data mining. Data mining menggambarkan sebuah pengumpulan beberapa teknik dengan tujuan untuk menemukan pola-pola yang tidak diketahui pada data yang telah dikumpulkan (Sawitri, 2019).

Data mining merupakan proses semi automatic yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang tersimpan di dalam database.

Data mining adalah bagian dari proses KDD (Knowledge Discovery in Databases) yang secara umum juga dikenal sebagai pangkalan data, terdiri dari beberapa tahapan seperti pemilihan data, pra pengolahan, transformasi, data mining, dan evaluasi hasil.

13.2 Fungsi dan Tujuan Data Mining

Data Mining mempunyai 5 fungsi antara lain:

Classification

Classification, adalah pengelompokan yang sistematis mengenai objek atau gagasan ke dalam kelas atau golongan tertentu berdasarkan kesamaan karakteristik. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkannya ke dalam kelas yang sebelumnya sudah didefinisikan.

Menurut Han dalam bukunya, klasifikasi secara umum terdiri atas dua tahapan. Tahapan learning (proses belajar) merupakan model yang dibuat dalam menggambarkan himpunan kelas atau konsep yang sudah ditentukan sebelumnya. Model klasifikasi dibuat berdasarkan analisa record pada database yang digambarkan dalam bentuk atribut. Record yang diasumsikan akan masuk ke dalam suatu kelas yang telah ditentukan sebelumnya, yang disebut atribut kelas (Frederika et al., 2022).

Contoh algoritma yang bisa digunakan untuk melakukan klasifikasi adalah algoritma naive bayes, K-Nearest Neighbor, Support vector machine, algoritma random forest, Decision Tree, C4.5, ID3, dan lain sebagainya.

Clustering

Clustering, adalah metode untuk mengelompokkan data yang banyak digunakan sebagai salah satu metode data mining. Clustering juga dapat diartikan sebagai proses partisi satu set objek data ke dalam himpunan bagian yang disebut cluster.

Oleh karena itu, metode clustering ini sangat berguna untuk menemukan kelompok yang tidak dikenal dalam data (Prastiwi, Pricilia and Raswir, 2022). Contoh algoritma yang bisa digunakan untuk melakukan clustering adalah algoritma K-Means, algoritma *Hierarchical Clustering* (AHC), K-Medoids, X-Means, dan lain sebagainya.

Association

Association, merupakan proses untuk menemukan atribut yang muncul dalam suatu waktu. Tujuan asosiasi adalah mencari hubungan antar suatu kombinasi item dalam suatu set data yang telah ditentukan. Asosiasi atau yang bisa juga disebut dengan association rules digunakan untuk menemukan pola-pola yang terjadi pada kumpulan data.

Asosiasi memiliki istilah *antecedent* dan *consequent*. Antecedent mewakili bagian "jika" dan consequent mewakili bagian "maka" asosiasi berbentuk C→D, di mana C dan D adalah dua itemset terpisah (disjoint) yang masing-masing disebut sebagai LHS (Left-Hand Side) dan RHS (Right-Hand Side). Antecedent juga bisa disebut dengan LHS (Left Hand Side) sedangkan consequent bisa disebut dengan RHS (Right Hand Side), dengan interpretasi di mana setiap pembelian item pada LHS memungkinkan adanya pembelian pada RHS (Frederika et al., 2022).

Contoh algoritma yang bisa digunakan dalam association adalah algoritma Apriori, algoritma Pincer Search, Frequent Pattern-Growth, algoritma Hash-Based, dan lain sebagainya.

Sequencing

Sequencing, atau fungsi pengurutan data mining digunakan untuk mengidentifikasikan perubahan pola yang telah terjadi dalam jangka waktu tertentu (Tarigan, 2023). Contoh algoritma yang biasa digunakan dalam sequencing adalah algoritma *Generalized Sequential Pattern* (GSP), algoritma *Sequential Pattern Using Discovery Equivalent Classes* (SPADE), dan lain sebagainya.

Forecasting

Forecasting, merupakan salah satu metode data mining untuk memprediksi suatu peristiwa di masa yang akan datang, sering diterapkan dalam bidang bisnis untuk melakukan proses pengambilan keputusan. Pada bidang pemasaran, forecasting bisa melihat dan mengetahui trend penjualan produk sehingga bisa memprediksi teknik pemasaran yang tepat yang akan digunakan di masa yang akan datang.

Mengutip jurnal Inovtek Polbeng, forecasting dibagi menjadi forecasting jangka pendek, jangka panjang, dan jangka menengah. Forecasting jangka pendek memprediksi menggunakan periode waktu (bulanan, mingguan bahkan harian) untuk masa depan. Forecasting jangka menengah menggunakan waktu satu hingga dua tahun.

Kebanyakan forecasting menggunakan metode time series yang menggunakan data historis berdasarkan kecenderungan datanya dan memprediksikan data tersebut untuk masa depan (Wijaya and Gantini, 2019). Contoh algoritma yang bisa digunakan adalah algoritma K-Nearest Neighbor, apriori, Iterative Dichotomiser Tree (Id3), C4.5, Support Vector Machine (SVM), Simple Moving Average, dan lain sebagainya.

Karakteristik Data Mining antara lain (Wahyudi, Azizah and Saputro, 2022):

1. Data sering kali terkubur dalam database yang sangat besar, yang terkadang berisi data selama bertahun-tahun. Tidak sedikit kasus, data dibersihkan lalu disatukan ke dalam data warehouse.
2. Environment data mining pada umumnya adalah arsitektur client-server atau arsitektur sistem informasi berbasis web.
3. Tool baru yang canggih, termasuk berbagai tool visualisasi, membantu mengangkat informasi yang terkubur dalam file-file korporat atau record-record arsip. Untuk mendapatkannya akan melibatkan memoles dan melakukan sinkronisasi data untuk mendapatkan hasil yang tepat. Data miners yang mutakhir juga melakukan pemeriksaan kegunaan data (misalnya, teks yang tidak terstruktur yang disimpan dalam tempat-tempat seperti database Lotus Notes, atau file-file teks di internet).

4. Dalam menemukan pola sering kali menemukan hasil yang tidak diharapkan dan meminta end user untuk berpikir secara kreatif dalam menjalankan proses, termasuk interpretasi terhadap temuan.
5. Banyak tool data mining siap dikombinasikan dengan berbagai spreadsheet dan tool development software lainnya. Sehingga data bisa dianalisis dan diterapkan dengan mudah, dan cepat.
6. Karena jumlah data yang sangat besar dan usaha pencarian yang masif, kadang-kadang perlu menggunakan pemrosesan paralel untuk data mining.

Tujuan Data Mining antara lain:

1. Explanatory, artinya memberikan penjelasan beberapa kondisi penelitian, contoh: mengapa penjualan mobil di Indonesia meningkat.
2. Exploratory, artinya melakukan analisis data untuk hubungan baru yang tidak diharapkan, contoh: pola apa yang cocok untuk kasus penggelapan kartu kredit.
3. Confirmatory, Untuk mempertegas hipotesis, contoh: dua kali pendapatan karyawan lebih suka dipakai untuk membeli peralatan rumah tangga, dibandingkan dengan satu kali pendapatan karyawan.

Implementasi Data Mining

Data Mining dapat diterapkan dalam bidang:

1. Keuangan, contoh: Penerapan Crisp-Dm Menggunakan Mlr K-Fold Pada Data Saham PT. Telkom Indonesia (Persero) Tbk (Tlkm) (Studi Kasus: Bursa Efek Indonesia Tahun 2015-2022)(Pambudi, Abidin and Permata, 2023).
2. Telekomunikasi, contoh: Implementasi RapidMiner menggunakan Metode K-Means dalam Penentuan Kluster Gangguan Jaringan WIFI Provider PT. XYZ di Daerah Karawang (Widianto et al., 2022).
3. Asuransi, contoh: Penerapan Operasional-CRM Untuk Pelayanan Asuransi pada PT. Asuransi FPG Indonesia kantor Cabang Lampung (Ardiyansyah and Santoso, 2022).

4. Analisa Perusahaan dan Manajemen Risiko, contoh: Analisis Risiko Pinjaman dengan Metode Support Vector Machine, Artificial Neural Network dan Naïve Bayes (Pernama, Dwi Purnomo and Satya Wacana, 2023).
5. Olahraga, contoh: Penerapan Aturan Asosiasi untuk Rule Mining pada Piala Dunia FIFA 2022 (Tahalea and Permadi, 2023).
6. Dan lain sebagainya.

13.4 Metodologi Data Mining

Komponen Data Mining

Komponen perencanaan data mining antara lain:

1. Analisa masalah (analyzing the problem)
   Data sumber atau data asal harus bisa ditaksir untuk mengetahui apakah data tersebut memenuhi kriteria data mining atau tidak. Kualitas data adalah faktor utama untuk memutuskan apakah data tersebut cocok dan tersedia sebagai tambahan. Hasil yang diharapkan dari dampak data mining harus dimengerti dan dipastikan bahwa data yang diperlukan membawa informasi yang bisa diekstrak.
2. Mengekstrak dan membersihkan data (extracting dan cleansing the data)
   Data pertama kali diekstrak dari data aslinya, seperti dari OLTP basis data, text file, Microsoft Access Database, dan bahkan dari spreadsheet, lalu data tersebut diletakan dalam data warehouse yang mempunyai struktur yang sesuai dengan data model secara khas. *Data Transformation Service* (DTS) digunakan untuk mengekstrak dan membersihkan data yang tidak kompatibel dan tidak konsisten dengan format yang sesuai.

3. Validitas data (validating the data)
   Data yang telah diekstrak dan dibersihkan, dilatih untuk menelusuri model yang telah kita ciptakan dan memastikan bahwa semua data yang ada adalah data sekarang dan tetap.
4. Membuat dan melatih model (creating and training the model)
   Struktur sudah dibangun saat algoritma diterapkan pada model. Hal tersebut sangat penting untuk melihat data yang telah dibangun dan memastikan bahwa data tersebut menyerupai fakta di dalam data sumber.
5. Query data dari model data mining (querying the model data)
   Ketika model yang telah cocok diciptakan dan dibangun, data yang sudah dibuat tersedia untuk mendukung keputusan, biasanya melibatkan penulisan front end query aplikasi dengan program aplikasi/suatu program basis data.
6. Evaluasi validitas dari mining model (maintaining the validity of the data mining model). Setelah model data mining terkumpul, lewat beberapa waktu, karakteristik data awal seperti granularities dan validitas mungkin berubah.

Proses Data Mining

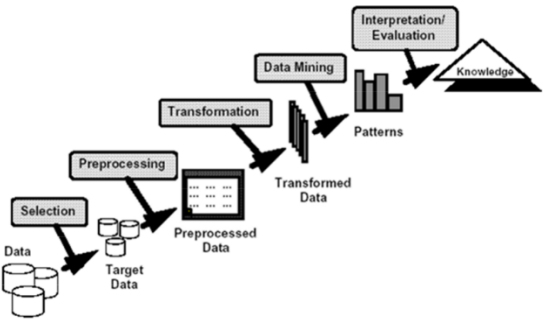
Tahapan data mining yang memproses data mentah menjadi pengetahuan atau informasi antara lain:

1. Data Cleansing.
2. Data Integration.
3. Data Selection.
4. Data Transformation.
5. Data Mining.
6. Pattern Evaluation.
7. Knowledge Representation.

Terdapat 5 tahapan proses data mining yang sama dengan proses yang dilakukan pada knowledge discovery antara lain:

1. Seleksi data - Bertujuan untuk mengekstrak data dari gudang data yang besar menjadi data yang relevan dengan analisis data mining.

2. Data preprocessing - Merupakan tahapan pembersihan data dan persiapan tugas untuk memastikan hasil dengan benar, seperti missing value data dan memastikan bahwa tidak ada nilai data palsu.
3. Transformasi data - Merupakan tahapan dalam mengubah data dalam bentuk atau format yang sesuai untuk kebutuhan data mining. Proses normalisasi biasanya diperlukan dalam tahap ini.
4. Data mining - Bertujuan untuk menganalisis data sesuai algoritma yang digunakan sehingga menemukan hasil berupa pola atau aturan yang bermakna serta menghasilkan model prediksi. Tahapan tersebut merupakan elemen inti dari siklus KDD.
5. Interpretasi dan evaluasi - Bertujuan untuk memilih model yang berguna dalam membuat keputusan bisnis masa depan, karena algoritma data mining biasanya menghasilkan jumlah yang tidak terbatas dari hasil aturan pada proses tersebut yang mungkin tidak berguna atau bermakna.



Gambar 13.2: Proses Data Mining (Widaningsih and Yusuf, 2022)

Teknik Data Mining

Sebelum mengetahui teknik apa saja yang bisa digunakan dalam data mining, terdapat empat operasi yang bisa dihubungkan dengan data mining antara lain sebagai berikut:

1. Predictive modelling, ada dua teknik yang dapat dilakukan antara lain:
   a. Classification.

b. Value Prediction.

2. Database Segmentation, untuk memartisi database menjadi beberapa segmen, cluster atau record yang sama di mana record tersebut diharapkan homogen.
3. Link Analysis, untuk membuat relasi antar record yang individual atau sekumpulan record dalam database. Aplikasi pada link analysis meliputi direct marketing, product affinity analysis, dan stock price movement.

Terdapat dua tipe teknik dalam data mining antara lain:

1. Teknik Klasik (Classical Technique), terdiri atas Statistic, Nearest Neighbour, dan Clustering.
2. Teknik Next Generation, terdiri atas Decision Tree dan lain sebagainya.

Bab 14

Aplikasi Data Mining

14.1 Pendahuluan

Aplikasi data mining adalah proses penggalian data dari kumpulan data yang besar dan kompleks, dengan tujuan menemukan pola, tren, atau hubungan yang bermanfaat untuk diambil kesimpulan atau membuat prediksi. Aplikasi Data Mining sangat penting dalam berbagai bidang seperti bisnis, keuangan, pemerintahan, kesehatan, pendidikan, dan lain-lain. Contoh penerapan Data Mining dalam bisnis adalah untuk memprediksi perilaku konsumen, mengoptimalkan strategi pemasaran, dan meningkatkan keuntungan perusahaan.

Teknik Data Mining yang sering digunakan antara lain Clustering, Association Rules, Decision Trees, dan Neural Networks (Mujiasih, 2011). Teknik-teknik tersebut dapat membantu dalam menggali informasi yang bermanfaat dari data, seperti pola pembelian konsumen, preferensi produk, dan kecenderungan pasar.

Dalam era digital dan informasi saat ini, Data Mining semakin penting dan diperlukan untuk mengolah data yang semakin banyak dan kompleks. Oleh karena itu, pemahaman tentang Data Mining dan aplikasinya sangat penting bagi organisasi dan individu yang ingin mengambil keuntungan dari data yang mereka miliki.

Beberapa implementasi aplikasi Data Mining dalam kehidupan sehari-hari

Pemasaran

Data mining dapat membantu perusahaan mengidentifikasi pelanggan potensial dan memberikan penawaran yang sesuai dengan minat mereka. Data mining juga dapat digunakan untuk mengidentifikasi pola pembelian dan perilaku konsumen yang dapat membantu perusahaan dalam merancang kampanye pemasaran yang lebih efektif.

Salah satu contoh aplikasi implementasi data mining dalam bidang pemasaran adalah untuk melakukan analisis perilaku konsumen dan segmentasi pasar (Nurdiawan and Salim, 2018). Data mining memungkinkan perusahaan untuk mengidentifikasi pola dan tren dalam data pelanggan yang tersedia, sehingga dapat memahami preferensi dan kebutuhan pelanggan dengan lebih baik.

Dengan memahami perilaku pelanggan, perusahaan dapat mengembangkan strategi pemasaran yang lebih efektif untuk meningkatkan penjualan dan keuntungan.

Contohnya adalah sebagai berikut:

1. Analisis klaster (cluster analysis)
   Analisis klaster digunakan untuk membagi pelanggan menjadi kelompok berdasarkan preferensi dan kebiasaan pembelian. Dengan menggunakan data yang telah dikumpulkan, perusahaan dapat membangun profil pelanggan dan mengelompokkannya ke dalam kelompok yang sama. Kemudian, perusahaan dapat mengembangkan strategi pemasaran yang lebih efektif untuk masing-masing kelompok.
2. Analisis asosiasi (association analysis)
   Analisis asosiasi digunakan untuk menemukan hubungan antara produk dan layanan yang dijual oleh perusahaan. Dengan menggunakan data transaksi, perusahaan dapat mengetahui produk apa yang sering dibeli bersamaan atau produk apa yang paling sering dibeli setelah membeli produk tertentu. Kemudian, perusahaan dapat membuat strategi bundling atau penawaran khusus untuk meningkatkan penjualan.

3. Analisis regresi (regression analysis)
   Analisis regresi digunakan untuk menentukan hubungan antara dua variabel atau lebih. Dalam pemasaran, analisis regresi digunakan untuk memprediksi penjualan berdasarkan faktor-faktor seperti harga, promosi, dan faktor lingkungan lainnya. Dengan memahami faktor-faktor yang memengaruhi penjualan, perusahaan dapat mengoptimalkan strategi pemasaran mereka dan meningkatkan penjualan.
4. Analisis sentimen (sentiment analysis)
   Analisis sentimen digunakan untuk menentukan pandangan konsumen terhadap produk atau layanan tertentu. Dalam pemasaran, analisis sentimen digunakan untuk menilai umpan balik pelanggan dari media sosial, survei, atau forum online. Dengan memahami sentimen konsumen, perusahaan dapat membuat perubahan yang diperlukan untuk meningkatkan kepuasan pelanggan dan menjaga reputasi merek.

Keuangan

Data mining dapat membantu perusahaan dalam melakukan analisis risiko kredit dan mendeteksi penipuan. Data mining juga dapat digunakan untuk melakukan analisis tren pasar dan mengidentifikasi pola investasi yang berpotensi menguntungkan.

Salah satu contoh aplikasi implementasi data mining dalam bidang keuangan adalah untuk menganalisis risiko kredit (Nurzahputra and Muslim, 2017). Data mining memungkinkan bank atau lembaga keuangan untuk mengumpulkan dan menganalisis data historis pelanggan dan kredit untuk mengidentifikasi pola dan tren yang dapat membantu dalam memprediksi risiko kredit di masa depan.

Dengan memahami risiko kredit secara lebih baik, lembaga keuangan dapat mengambil keputusan yang lebih cerdas dan meminimalkan risiko kredit yang tinggi.

Berikut ini adalah beberapa contoh aplikasi data mining dalam bidang keuangan:

1. Analisis risiko kredit (Credit risk analysis)
   Data mining digunakan untuk menganalisis risiko kredit dengan mempelajari data historis pelanggan dan kredit, termasuk data keuangan seperti riwayat pembayaran dan saldo kredit. Dengan mempelajari pola dalam data, perusahaan keuangan dapat mengidentifikasi pelanggan yang memiliki risiko kredit yang tinggi dan mengambil tindakan yang tepat untuk meminimalkan risiko kredit tersebut.
2. Analisis fraud (Fraud analysis)
   Data mining juga digunakan untuk menganalisis fraud atau penipuan dalam transaksi keuangan. Data mining memungkinkan lembaga keuangan untuk mempelajari pola dan trend dalam data transaksi, termasuk jenis transaksi yang sering menjadi sasaran penipuan, dan mengambil tindakan untuk meminimalkan risiko penipuan.
3. Analisis pasar (Market analysis)
   Data mining digunakan untuk menganalisis pasar dan memprediksi tren pasar di masa depan. Data mining memungkinkan perusahaan keuangan untuk mempelajari pola dan trend dalam data historis pasar dan mengidentifikasi faktor yang memengaruhi perubahan pasar. Dengan memahami tren pasar dan faktor-faktor yang memengaruhinya, perusahaan keuangan dapat mengambil tindakan yang tepat untuk mengoptimalkan investasi mereka.
4. Analisis portfolio (Portfolio Analysis)
   Data mining digunakan untuk menganalisis portofolio investasi dan memprediksi kinerja portofolio di masa depan. Dengan mempelajari pola dalam data historis portofolio, perusahaan keuangan dapat mengidentifikasi investasi yang paling efektif dan mengambil tindakan yang tepat untuk mengoptimalkan portofolio mereka.

Ilmu Kesehatan

Data mining dapat membantu dalam pengembangan obat baru, peningkatan kualitas perawatan pasien, dan deteksi penyakit. Data mining juga dapat membantu dalam identifikasi pola epidemiologi dan pemantauan kesehatan masyarakat (Nabila et al., 2021).

Salah satu contoh aplikasi implementasi data mining dalam bidang kesehatan adalah untuk menganalisis data kesehatan pasien untuk membantu dalam diagnosis, pengobatan, dan pencegahan penyakit. Data mining dapat digunakan untuk mengidentifikasi pola dan tren dalam data medis yang besar dan kompleks, yang memungkinkan para ahli kesehatan untuk mengambil keputusan yang lebih baik dan meningkatkan perawatan pasien.

Berikut ini adalah beberapa contoh aplikasi data mining dalam bidang kesehatan:

1. Analisis data genomik (Genomic data analysis)
   Data mining dapat digunakan untuk menganalisis data genomik, yaitu data yang dihasilkan dari pengujian DNA pasien. Dengan mempelajari data genomik, ahli kesehatan dapat mengidentifikasi faktor risiko genetik yang terkait dengan penyakit tertentu, dan mengambil tindakan yang tepat untuk mencegah atau mengobati penyakit tersebut.
2. Analisis data medis (Medical data analysis)
   Data mining dapat digunakan untuk menganalisis data medis pasien, seperti catatan medis elektronik dan hasil tes medis. Dengan mempelajari pola dalam data medis, ahli kesehatan dapat mengidentifikasi penyakit atau kondisi yang sering terjadi, memprediksi perkembangan penyakit, dan mengambil tindakan yang tepat untuk mencegah atau mengobati penyakit.
3. Analisis penggunaan obat (Drug usage analysis)
   Data mining dapat digunakan untuk menganalisis penggunaan obat oleh pasien, termasuk dosis obat dan efek samping yang mungkin terjadi. Dengan mempelajari data penggunaan obat, ahli kesehatan dapat mengidentifikasi pola dan tren dalam penggunaan obat,

memprediksi efek samping yang mungkin terjadi, dan mengambil tindakan yang tepat untuk meminimalkan efek samping tersebut.

4. Analisis pencegahan penyakit (Disease prevention analysis)
   Data mining dapat digunakan untuk menganalisis data kesehatan populasi, seperti vaksinasi dan riwayat kesehatan keluarga, untuk membantu mencegah penyebaran penyakit. Dengan mempelajari data pencegahan penyakit, ahli kesehatan dapat mengidentifikasi faktor risiko dan memprediksi potensi penyebaran penyakit, dan mengambil tindakan yang tepat untuk mencegah penyebaran penyakit tersebut.

Pendidikan

Data mining juga dapat diterapkan dalam bidang pendidikan untuk mengumpulkan dan menganalisis data pendidikan. Dengan menerapkan teknik data mining pada data yang dihasilkan oleh sistem pendidikan, maka dapat membantu pengambilan keputusan yang lebih baik oleh para pengambil kebijakan dalam meningkatkan kualitas pendidikan.

Data mining dapat membantu dalam pengembangan strategi pembelajaran yang lebih efektif dan pemantauan kinerja siswa (Diponegoro, Kusumawardani and Hidayah, 2021). Data mining juga dapat digunakan untuk mengidentifikasi masalah di dalam sistem pendidikan dan memberikan solusi yang tepat.

Berikut ini adalah beberapa contoh aplikasi implementasi data mining dalam bidang pendidikan:

1. Prediksi kinerja akademik siswa (Academic performance prediction)
   Data mining dapat digunakan untuk memprediksi kinerja akademik siswa. Dengan menganalisis data seperti riwayat pelajaran, nilai ujian, dan faktor-faktor lain yang memengaruhi hasil belajar siswa, maka dapat memprediksi tingkat keberhasilan siswa di masa depan. Hal ini dapat membantu guru dan staf pendidikan untuk menyediakan bimbingan dan dukungan yang tepat kepada siswa yang membutuhkan.
2. Pemetaan perkembangan siswa (Student progress mapping)
   Data mining juga dapat digunakan untuk membuat pemetaan perkembangan siswa dalam rangka meningkatkan kualitas

pendidikan. Dengan mengumpulkan dan menganalisis data seperti nilai ujian, keterlambatan tugas, dan absensi, maka dapat membuat profil siswa yang lebih rinci. Hal ini dapat membantu guru dan staf pendidikan untuk mengenali masalah yang dialami oleh siswa dan memperbaiki proses belajar mengajar secara umum.

3. Analisis preferensi pelajaran (Course preference analysis)
   Data mining dapat digunakan untuk menganalisis preferensi pelajaran siswa. Dengan menganalisis data seperti kuesioner dan data pendaftaran siswa, maka dapat memahami lebih dalam mengenai preferensi siswa terhadap mata pelajaran tertentu. Hal ini dapat membantu guru dan staf pendidikan untuk menyusun program belajar yang lebih efektif dan meningkatkan motivasi siswa.
4. Evaluasi kualitas pengajaran (Teaching quality evaluation)
   Data mining dapat digunakan untuk menganalisis evaluasi kualitas pengajaran oleh siswa. Dengan menganalisis data seperti kuesioner evaluasi pengajaran, maka dapat memahami lebih dalam mengenai kekuatan dan kelemahan guru dalam proses belajar mengajar. Hal ini dapat membantu guru dan staf pendidikan untuk memperbaiki kualitas pengajaran secara keseluruhan.

Transportasi

Data mining dapat diterapkan dalam berbagai bidang, termasuk di bidang transportasi. Di bidang ini, data mining dapat membantu pengambil keputusan dalam meningkatkan efisiensi transportasi dan mengoptimalkan operasi transportasi.

Aplikasi data mining di transportasi dapat digunakan untuk analisis permintaan transportasi, peramalan lalu lintas, dan pengoptimalan rute. Teknik regresi dapat digunakan untuk memprediksi permintaan transportasi, sementara teknik clustering dapat digunakan untuk mengelompokkan titik-titik pemberhentian transportasi berdasarkan pola penggunaan.

Berikut ini adalah beberapa contoh aplikasi implementasi data mining dalam bidang transportasi:

1. Prediksi keterlambatan transportasi (Transportation delay prediction)
   Data mining dapat digunakan untuk memprediksi keterlambatan transportasi. Dengan menganalisis data seperti data cuaca, data lalu lintas, dan data operasi transportasi, maka dapat memprediksi kemungkinan keterlambatan transportasi. Hal ini dapat membantu pengelola transportasi untuk membuat perencanaan dan strategi operasi yang lebih baik.
2. Optimasi rute transportasi (Transportation route optimization)
   Data mining dapat digunakan untuk mengoptimalkan rute transportasi. Dengan menganalisis data seperti data lalu lintas, data jarak, dan data waktu tempuh, maka dapat menemukan rute yang paling efisien. Hal ini dapat membantu pengelola transportasi untuk mengurangi waktu dan biaya operasi.
3. Pemantauan kondisi kendaraan (Vehicle condition monitoring)
   Data mining dapat digunakan untuk memantau kondisi kendaraan. Dengan menganalisis data seperti data perawatan, data suhu mesin, dan data penggunaan bahan bakar, maka dapat memprediksi kerusakan dan melakukan perawatan secara tepat waktu. Hal ini dapat membantu pengelola transportasi untuk mengurangi biaya perawatan dan meningkatkan efisiensi operasi.
4. Prediksi permintaan transportasi (Transportation demand prediction)
   Data mining dapat digunakan untuk memprediksi permintaan transportasi. Dengan menganalisis data seperti data demografi dan data sejarah penggunaan transportasi, maka dapat memprediksi permintaan transportasi di masa depan. Hal ini dapat membantu pengelola transportasi untuk membuat perencanaan dan strategi operasi yang lebih baik.

Sosial Media

Contoh aplikasi data mining di media sosial adalah penggunaan teknik clustering untuk mengelompokkan pengguna berdasarkan minat atau perilaku mereka. Dengan menganalisis pola perilaku pengguna media sosial,

perusahaan dapat membuat kampanye iklan yang lebih tepat sasaran dan meningkatkan interaksi dengan pengguna.

Dalam aplikasi data mining, beberapa teknik yang umum digunakan adalah clustering, klasifikasi, regresi, asosiasi, dan deteksi anomali. Data mining dapat membantu perusahaan dan organisasi dalam membuat keputusan yang lebih baik dengan menganalisis data dan mengidentifikasi pola dan tren yang tersembunyi. Namun, perlu diingat bahwa aplikasi data mining juga harus memperhatikan etika dan privasi data, karena penggunaan data yang tidak etis dapat berdampak negatif pada individu atau masyarakat.

14.2 Tahapan Pembuatan Aplikasi Data Mining

Membuat aplikasi data mining melibatkan beberapa tahapan, antara lain:

1. Pemahaman bisnis
   Tahap ini dilakukan untuk memahami kebutuhan bisnis dan apa yang ingin dicapai melalui aplikasi data mining. Misalnya, perusahaan ingin meningkatkan penjualan atau mengurangi biaya produksi dengan memanfaatkan data yang dimilikinya.
2. Pengumpulan data
   Tahap ini dilakukan untuk mengumpulkan data yang diperlukan untuk aplikasi data mining. Data dapat berasal dari berbagai sumber, seperti database perusahaan, data publik, atau data dari media sosial.
3. Preprocessing data
   Data yang telah dikumpulkan harus diproses dan disiapkan sebelum dapat digunakan untuk aplikasi data mining. Tahap ini meliputi pembersihan data dari duplikasi, kesalahan, atau data yang tidak relevan, serta transformasi data ke format yang sesuai.
4. Pemilihan algoritma
   Algoritma yang tepat harus dipilih untuk memproses data dan menghasilkan hasil yang diinginkan. Beberapa algoritma data mining

yang umum digunakan adalah clustering, klasifikasi, regresi, asosiasi, dan deteksi anomali.

5. Pemodelan data
   Setelah algoritma dipilih, data harus dimodelkan untuk melihat pola dan tren yang tersembunyi. Pemodelan dapat dilakukan dengan menggunakan perangkat lunak khusus seperti R, Python, atau Weka.
6. Evaluasi model
   Model yang telah dibuat harus dievaluasi untuk memastikan kualitas dan akurasi hasilnya. Evaluasi dilakukan dengan membandingkan hasil model dengan data aktual dan mengidentifikasi kesalahan atau ketidakakuratan dalam model.
7. Implementasi
   Setelah model dievaluasi, aplikasi data mining harus diimplementasikan ke dalam lingkungan bisnis atau organisasi. Implementasi dapat dilakukan dengan cara membangun aplikasi berbasis web, atau dengan menyediakan antarmuka untuk akses ke hasil data mining.
8. Monitoring
   Proses data mining harus terus dipantau dan dievaluasi untuk memastikan konsistensi dan keakuratan hasilnya. Perlu juga memastikan bahwa data yang digunakan tetap relevan dan up-to-date.

Membuat aplikasi data mining membutuhkan pengetahuan teknis dan keahlian dalam pengolahan data, analisis data, dan pemrograman. Oleh karena itu, jika Anda tidak memiliki pengetahuan teknis yang cukup, Anda mungkin perlu mempertimbangkan untuk menghubungi ahli atau perusahaan yang menyediakan layanan data mining untuk membantu Anda dalam membuat aplikasi data mining yang efektif.

14.3 Aplikasi Data Mining Yang Sudah Diimplementasikan

Berikut adalah beberapa contoh aplikasi data mining yang sudah diimplementasikan dan bisa diakses oleh orang banyak:

1. Netflix
   Netflix menggunakan teknik data mining untuk merekomendasikan film dan acara TV yang relevan dengan preferensi pengguna. Netflix menganalisis data dari perilaku menonton dan peringkat yang diberikan oleh pengguna untuk menemukan pola dan tren yang dapat digunakan untuk merekomendasikan konten baru.
2. Amazon
   Amazon menggunakan teknik data mining untuk merekomendasikan produk yang relevan dengan preferensi pengguna. Amazon menganalisis data dari riwayat pembelian dan perilaku pengguna untuk menemukan pola dan tren yang dapat digunakan untuk merekomendasikan produk baru.
3. Spotify
   Spotify menggunakan teknik data mining untuk merekomendasikan lagu dan artis yang sesuai dengan preferensi pengguna. Spotify menganalisis data dari riwayat pemutaran dan perilaku pengguna untuk menemukan pola dan tren yang dapat digunakan untuk merekomendasikan konten baru.
4. Google Maps
   Google Maps menggunakan teknik data mining untuk memperkirakan waktu perjalanan dan memberikan rute tercepat ke tujuan. Google Maps menganalisis data dari lalu lintas jalan dan informasi cuaca untuk memprediksi waktu perjalanan dan mengoptimalkan rute.
5. Facebook
   Facebook menggunakan teknik data mining untuk menampilkan iklan yang relevan dengan minat dan perilaku pengguna. Facebook

menganalisis data dari perilaku pengguna dan preferensi untuk menampilkan iklan yang relevan dan meningkatkan interaksi dengan pengguna.

6. Kaggle
   Kaggle adalah platform kompetisi data mining yang memungkinkan pengguna untuk mengembangkan model prediksi dan algoritma untuk menyelesaikan masalah bisnis atau sosial. Kaggle memberikan akses terhadap data dan tantangan terbuka yang dapat diakses oleh orang banyak untuk mengembangkan keterampilan dan pengalaman di bidang data mining.

Dalam buku ini akan dibahas sedikit lebih dalam terkait implementasi Data Mining yang ada pada aplikasi Netflix

Netflix

Netflix adalah sebuah platform streaming video online yang memungkinkan pengguna untuk menonton berbagai jenis konten hiburan melalui internet. Konten yang disediakan oleh Netflix meliputi film, serial TV, dan program acara dari berbagai genre, termasuk drama, komedi, horor, dokumenter, dan banyak lagi.

Pengguna dapat mengakses Netflix melalui berbagai perangkat, seperti komputer, smartphone, tablet, smart TV, dan perangkat streaming media seperti Roku dan Amazon Fire TV. Dengan cara ini, pengguna dapat menonton konten Netflix di mana saja dan kapan saja selama terhubung ke internet.

Netflix adalah sebuah layanan berlangganan, di mana pengguna membayar biaya bulanan untuk mengakses konten yang tersedia di platform. Selain itu, Netflix juga menyediakan opsi untuk mendownload konten dan menontonnya secara offline, sehingga pengguna dapat menonton film dan program acara favorit mereka tanpa koneksi internet.

Untuk memudahkan pengguna dalam menemukan konten yang tepat untuk ditonton, Netflix menggunakan berbagai teknologi seperti data mining dan machine learning. Dengan menggunakan data pengguna seperti riwayat tontonan dan rating, Netflix dapat merekomendasikan film dan program acara yang sesuai dengan preferensi masing-masing pengguna. Selain itu, Netflix juga membuat konten original mereka sendiri, seperti serial TV Stranger

Things dan The Crown, yang menjadi sangat populer di kalangan pengguna platform ini.

Netflix menggunakan teknologi data mining untuk merekomendasikan film dan acara televisi kepada pelanggan berdasarkan riwayat tontonan mereka dan penilaian yang diberikan.

Berikut adalah beberapa teknik data mining yang digunakan oleh Netflix:

Collaborative Filtering

Netflix menggunakan teknik *Collaborative Filtering* untuk merekomendasikan film dan acara televisi berdasarkan preferensi tontonan dan penilaian pengguna yang serupa (Zhou et al., 2008). Teknik ini menganalisis data pengguna untuk menemukan kesamaan antara preferensi tontonan mereka dan merekomendasikan konten yang relevan.

Collaborative Filtering (CF) adalah teknik data mining yang digunakan oleh Netflix untuk merekomendasikan konten yang relevan kepada pelanggan. CF menganalisis riwayat tontonan dan penilaian pengguna untuk menemukan kesamaan antara preferensi tontonan mereka dan merekomendasikan konten yang serupa. Teknik ini digunakan oleh Netflix untuk meningkatkan akurasi dan personalisasi rekomendasi konten bagi pengguna.

CF pada aplikasi Netflix dapat dibagi menjadi dua jenis:

1. User-based Collaborative Filtering
   User-based CF membandingkan preferensi tontonan dan penilaian antara pengguna untuk menemukan kesamaan. Teknik ini menganalisis data pengguna untuk menemukan pengguna lain yang memiliki preferensi tontonan dan penilaian yang serupa. Ketika seorang pengguna menonton atau memberikan penilaian positif pada suatu konten, Netflix akan merekomendasikan konten yang disukai oleh pengguna lain yang memiliki preferensi tontonan dan penilaian yang mirip. Contoh: Jika pengguna A menonton dan memberikan penilaian positif pada serial TV drama Korea, Netflix akan merekomendasikan serial TV drama Korea yang disukai oleh pengguna lain yang memiliki preferensi tontonan dan penilaian yang mirip dengan A.

2. Item-based Collaborative Filtering
   Item-based CF membandingkan kesamaan antara konten (film dan acara televisi) berdasarkan preferensi tontonan dan penilaian pengguna. Teknik ini menganalisis konten yang disukai oleh pengguna dan merekomendasikan konten yang mirip dengan konten tersebut. Contoh: Jika pengguna A menonton dan memberikan penilaian positif pada film aksi Hollywood, Netflix akan merekomendasikan film aksi Hollywood lainnya yang memiliki karakteristik dan genre yang sama dengan film yang disukai oleh A.

Untuk meningkatkan akurasi dan personalisasi rekomendasi konten, Netflix menggunakan teknik *Hybrid Collaborative Filtering*. Teknik ini menggabungkan dua jenis CF, User-based CF dan Item-based CF, untuk memberikan rekomendasi konten yang lebih relevan dan sesuai dengan preferensi dan perilaku tontonan pengguna.

Dengan menerapkan teknologi Collaborative Filtering, Netflix dapat meningkatkan kepuasan pelanggan dan memperkuat posisinya di pasar layanan streaming. Teknik ini membantu Netflix merekomendasikan konten yang sesuai dengan preferensi dan perilaku tontonan pengguna, sehingga meningkatkan pengalaman tontonan dan kepuasan pelanggan.

Content-based Filtering

Netflix menggunakan teknik *content-based filtering* untuk merekomendasikan film dan acara televisi berdasarkan analisis konten atau metadata seperti genre, sutradara, pemain, dan sinopsis (Havolli, Maraj and Fetahu, 2022). Teknik ini mengidentifikasi kemiripan antara konten yang disukai oleh pengguna dan merekomendasikan konten yang serupa.

Content-based Filtering adalah salah satu teknik dalam sistem rekomendasi yang digunakan oleh Netflix untuk memberikan rekomendasi film dan acara TV kepada penggunanya. Teknik ini bekerja dengan menganalisis dan mengevaluasi konten film dan acara TV, kemudian mencocokkan preferensi pengguna dengan konten yang relevan.

Content-based Filtering memerlukan metadata yang berkualitas dari konten film dan acara TV, seperti judul, genre, sinopsis, sutradara, aktor, dan elemen lain yang berkaitan dengan konten tersebut. Netflix menggunakan metadata ini

untuk membuat profil konten yang kaya, detail, dan terperinci, sehingga memudahkan sistem untuk memahami preferensi pengguna.

Setelah profil konten dibuat, sistem akan membandingkan preferensi pengguna dengan metadata konten, kemudian memberikan rekomendasi yang sesuai dengan preferensi pengguna. Sebagai contoh, jika pengguna menyukai film aksi dengan aktor Bruce Willis, maka sistem akan merekomendasikan film aksi lain dengan Bruce Willis sebagai pemeran utama.

Content-based Filtering memiliki beberapa kelebihan, seperti dapat memberikan rekomendasi yang relevan dan personalisasi yang tinggi, serta tidak memerlukan data dari pengguna lain. Namun, kelemahannya adalah terbatasnya variasi dan variasi konten yang ditawarkan kepada pengguna. Sebagai contoh, jika pengguna hanya menyukai genre tertentu, maka rekomendasi yang diberikan juga hanya terbatas pada genre tersebut.

Oleh karena itu, Netflix menggunakan teknik rekomendasi lain, seperti Collaborative Filtering, untuk melengkapi sistem rekomendasi yang mereka miliki. Dengan menggunakan beberapa teknik rekomendasi, Netflix dapat memberikan rekomendasi yang lebih akurat dan personalisasi yang tinggi kepada penggunanya.

Berikut adalah langkah-langkah untuk melakukan Collaborative Filtering pada Netflix:

1. Mengumpulkan data pengguna dan preferensi program-program Netflix. Data ini dapat berupa penilaian pengguna atau riwayat tontonan.
2. Membuat model Collaborative Filtering
   Ada dua jenis Collaborative Filtering yang umum digunakan: User-Based Collaborative Filtering dan Item-Based Collaborative Filtering. Pada User-Based Collaborative Filtering, model akan merekomendasikan program-program berdasarkan preferensi pengguna lain yang memiliki preferensi yang mirip. Pada Item-Based Collaborative Filtering, model akan merekomendasikan program-program berdasarkan kesamaan program dengan program yang telah disukai oleh pengguna.

3. Melatih model Collaborative Filtering menggunakan data yang dikumpulkan. Model ini akan belajar untuk merekomendasikan program-program Netflix berdasarkan preferensi pengguna lain.
4. Menggunakan model Collaborative Filtering untuk merekomendasikan program-program Netflix untuk pengguna. Model akan merekomendasikan program-program yang dianggap cocok untuk pengguna berdasarkan preferensi pengguna lain.
5. Menguji dan memperbarui model Collaborative Filtering secara teratur untuk memastikan bahwa merekomendasikan program-program yang relevan untuk pengguna.

Dalam praktiknya, Collaborative Filtering dapat membantu Netflix dalam meningkatkan pengalaman pengguna di platform mereka. Netflix dapat merekomendasikan program-program yang dapat disukai pengguna berdasarkan preferensi pengguna lain, dan dengan demikian, meningkatkan kepuasan pengguna dan meningkatkan loyalitas pelanggan. Selain itu, Collaborative Filtering juga dapat membantu Netflix dalam mengambil keputusan tentang program-program baru yang akan diproduksi atau diberikan kepada pengguna.

Sentiment Analysis

Sentiment Analysis merupakan ilmu yang berguna untuk menganalisis pendapat seseorang, sentiment seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis (Saputra, Adji and Permanasari, 2015). Sentiment Analysis sudah banyak diimplementasikan ke dalam berbagai lini kehidupan, di antaranya terkait manajemen reputasi calon presiden (Saputra, Nurbagja and Turiyan, 2022), Review Film (Putri, Anisa and Saputra, 2022), dan pandangan masyarakat dengan keberadaan Self Driving Car (Saputra, 2019), Pendidikan (Saputra, 2016).

Netflix menggunakan teknik Sentiment Analysis untuk menganalisis penilaian yang diberikan oleh pengguna dan mengevaluasi sentimen positif atau negatif terhadap film dan acara televisi tertentu (Bagkar, Borude and Aga, 2021). Teknik ini membantu Netflix memahami preferensi pengguna dan merekomendasikan konten yang lebih relevan. Teknik analisis teks yang digunakan oleh Netflix untuk mengukur sentimen atau emosi yang diungkapkan oleh pengguna melalui review dan komentar. Sentiment Analysis digunakan untuk memahami apa yang dikatakan pengguna tentang film dan

acara TV tertentu, dan juga untuk mengukur respons pengguna terhadap konten baru yang ditawarkan oleh Netflix.

Teknik ini bekerja dengan menggunakan algoritma pemrosesan bahasa alami (Natural Language Processing atau NLP) untuk menganalisis kata-kata yang digunakan oleh pengguna dalam review dan komentar. Algoritma ini akan mencari pola kata-kata positif, negatif, atau netral dalam teks, kemudian memberikan skor sentimen yang sesuai dengan kata-kata tersebut.

Sentiment Analysis memerlukan data teks dari review dan komentar pengguna untuk dilakukan analisis. Netflix mengumpulkan data ini dari berbagai sumber, seperti situs web, aplikasi, dan media sosial. Data ini kemudian diproses dan dianalisis menggunakan algoritma Sentiment Analysis.

Hasil dari analisis sentimen kemudian digunakan oleh Netflix untuk memperbaiki sistem rekomendasi mereka. Misalnya, jika banyak pengguna memberikan sentimen negatif pada suatu konten, maka Netflix akan mengevaluasi konten tersebut dan memutuskan apakah akan tetap menampilkan konten tersebut di platform mereka atau tidak.

Sentiment Analysis juga dapat membantu Netflix untuk mengidentifikasi tren dan preferensi pengguna secara real-time, dan memperbaiki sistem rekomendasi mereka berdasarkan data yang dianalisis. Hal ini memungkinkan Netflix untuk memberikan pengalaman yang lebih personalisasi dan memuaskan kepada pengguna mereka.

Untuk melakukan Sentiment Analysis pada Netflix, langkah-langkah yang dapat dilakukan adalah sebagai berikut:

1. Mengumpulkan data teks yang relevan dengan Netflix, seperti ulasan pengguna atau tweet yang berkaitan dengan Netflix.
2. Membuat dataset yang terdiri dari teks dan label sentimen. Label sentimen dapat berupa positif, negatif, atau netral, dan dapat diberikan berdasarkan analisis manusia atau menggunakan algoritma machine learning.
3. Menggunakan algoritma machine learning untuk melatih model Sentiment Analysis. Ada berbagai jenis algoritma yang dapat digunakan untuk ini, seperti *Naive Bayes*, *Support Vector Machine* (SVM), atau *Recurrent Neural Networks* (RNN).

4. Menggunakan model yang dilatih untuk melakukan Sentiment Analysis pada data teks baru. Model akan mengeluarkan label sentimen untuk setiap teks yang diberikan.
5. Menganalisis hasil dari Sentiment Analysis untuk mendapatkan wawasan tentang bagaimana pengguna merespons program-program yang tersedia di Netflix.

Dalam praktiknya, Sentiment Analysis pada Netflix dapat membantu Netflix dalam mengambil keputusan tentang program-program yang akan mereka produksi atau berikan kepada pengguna. Netflix dapat menggunakan informasi ini untuk mengetahui apa yang disukai oleh pengguna dan apa yang harus mereka lakukan untuk meningkatkan pengalaman pengguna di platform mereka.

Predictive Analytics

Netflix menggunakan teknik *Predictive Analytics* untuk memprediksi perilaku pengguna, seperti apakah mereka akan menonton film atau acara televisi tertentu, dan merekomendasikan konten yang sesuai (Tanuwijaya, Alamsyah and Ariyanti, 2021). Teknik ini menganalisis data pengguna untuk memprediksi preferensi dan perilaku tontonan pengguna di masa depan.

Predictive Analytics adalah salah satu teknik data mining yang digunakan oleh Netflix untuk memprediksi perilaku pengguna dan memperbaiki sistem rekomendasi mereka. Teknik ini melibatkan analisis data historis dan pola perilaku pengguna untuk memprediksi perilaku di masa depan.

Proses Predictive Analytics di Netflix dimulai dengan mengumpulkan data historis tentang preferensi, perilaku menonton, dan interaksi pengguna dengan platform. Data ini kemudian diolah menggunakan teknik Machine Learning dan model prediktif untuk memprediksi perilaku pengguna di masa depan.

Model prediktif ini dapat digunakan untuk memprediksi hal-hal seperti apa konten yang akan diminati oleh pengguna tertentu, kapan pengguna akan menonton, dan seberapa lama pengguna akan menonton. Hasil dari analisis prediksi ini kemudian digunakan untuk meningkatkan sistem rekomendasi Netflix dan memberikan pengalaman yang lebih personalisasi dan memuaskan kepada pengguna.

Salah satu contoh penggunaan Predictive Analytics di Netflix adalah fitur "Recommended for You". Fitur ini menggunakan algoritma prediktif untuk

memprediksi jenis konten yang mungkin diminati oleh pengguna berdasarkan preferensi dan perilaku menonton mereka yang telah tercatat di masa lalu.

Dalam konteks Predictive Analytics, Netflix juga menggunakan teknik Big Data Analytics untuk mengelola dan menganalisis volume data yang besar dan kompleks dari berbagai sumber, seperti log interaksi pengguna, informasi konten, dan data demografi.

Predictive Analytics membantu Netflix untuk memahami perilaku dan preferensi pengguna mereka secara lebih mendalam dan memberikan pengalaman yang lebih personalisasi dan relevan. Teknik ini memungkinkan Netflix untuk memperbaiki sistem rekomendasi mereka secara terus-menerus, meningkatkan loyalitas pelanggan, dan mengoptimalkan keuntungan bisnis mereka.

Dengan menerapkan teknologi data mining, Netflix dapat memberikan pengalaman tontonan yang lebih personal dan relevan bagi pelanggannya. Netflix dapat merekomendasikan konten yang sesuai dengan preferensi dan perilaku tontonan pengguna, sehingga meningkatkan kepuasan pelanggan dan memperkuat posisi Netflix di pasar layanan streaming.

Daftar Pustaka

A. S. Sajedi and S. H. Ghazi, "Application of data mining techniques in stock market forecasting: A systematic literature review," Journal of Financial Data Science, vol. 2, no. 1, pp. 94-113, 2020.

Ade Putranto, R. and Wuryandari, T. (2015) 'PERBANDINGAN ANALISIS KLASIFIKASI ANTARA DECISION TREE DAN SUPPORT VECTOR MACHINE MULTICLASS UNTUK PENENTUAN JURUSAN PADA SISWA SMA', 4, pp. 1007–1016. Available at: http://ejournal-s1.undip.ac.id/index.php/gaussian.

Adie Wahyudi Oktavia Gama, Ketut Gede Darma Putra, I. P. A. B. (2016) "Implementasi Algoritma Apriori Untuk menemukan Frequent Itemset Dalam Keranjang Belanja," Teknologi Elektro, 15(2), hal. 27–31.

Aggarwal, C.C. (2015) Data Mining: The Textbook. Berlin: Springer International Publishing.

Anonim. (2015). Komputasi Numeris: Analisis Regresi Sederhana Non-Linear. Jati.Itda.Ac.Id.

Ardiyansyah, R. and Santoso, A.B. (2022) 'PENERAPAN OPERASIONAL-CRM UNTUK PELAYANAN ASURANSI PADA PT ASURANSI FPG INDONESIA KANTOR CABANG LAMPUNG', Teknologiterkini.org, 2(4), pp. 1–13.

Arifin, J., (2007). Aplikasi Excel untuk Perencanaan Bisnis. Jakarta: Elex Media Komputindo.

Ashari Muin, A. (2016) 'Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)', Jurnal Ilmiah Ilmu Komputer, 2(1). Available at: http://ejournal.fikom-unasman.ac.id.

Bagkar, P., Borude, A. and Aga, Z. (2021) Sentiment Analysis on Netflix, IRE Journals |.

Bates, A. & Kalita, J. (2016) "Counting clusters in twitter posts. Proceedings of the 2nd," in International Conference on Information Technology for Competitive Strategies.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American.

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). The Unified Modeling Language User Guide. Addison-Wesley.

Caruana, R., Lawrence, S., & Giles, L. (2015). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Advances in neural information processing systems (pp. 402-408).

Chang, P. C., Wang, Y. W. dan Liu, C. H. (2007) "The development of a weighted evolving fuzzy neural network for PCB sales forecasting," Expert Systems with Applications, 32(1), hal. 86–96. doi: 10.1016/j.eswa.2005.11.021.

Charu C.Anggarwal (2015) Data Mining, Springer Cham Heidelberg New York Dordrecht London. Springer Cham Heidelberg New York Dordrecht London. doi: 10.1007/978-3-319-14142-8.

Chen, L.-P. (2020) Text mining in practice with R, Journal of Statistical Computation and Simulation. doi: 10.1080/00949655.2019.1630887.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13(6), 377-387.

Connolly, T., & Begg, C. (2021). Database Systems: A Practical Approach to Design, Implementation, and Management. Pearson.

Dalianis, H. (2018) Clinical text mining: Secondary use of electronic patient records, Clinical Text Mining: Secondary Use of Electronic Patient Records. doi: 10.1007/978-3-319-78503-5.

Date, C. J. (2004). An Introduction to Database Systems. Pearson.

Deepak Kumar Sharma, Mayukh Chatterjee, Gurmehak Kaur, S. V. (2022) "Deep Learning for Medical Applications with Unique Data," in Sinha, V. B. and G. R. (ed.). Academic Press, hal. 31–51. doi: https://doi.org/10.1016/C2020-0-00679-1.

Derajad Wijaya, H. and Dwiasnati, S. (2020) 'Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat', JURNAL

INFORMATIKA, 7(1). Available at: http://ejournal.bsi.ac.id/ejurnal/index.php/ji.

Diponegoro, M.H., Kusumawardani, S.S. and Hidayah, I. (2021) 'Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada Prediksi Kinerja Murid', Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 10(2), pp. 131–138. Available at: https://doi.org/10.22146/JNTETI.V10I2.1417.

Elmasri, R., & Navathe, S. B. (2016). Fundamentals of Database Systems. Pearson.

Fajar, H. A. (2013) DATA MINING. Vol. 2. Yogyakarta: ANDI.

Fauzi, R. (2017) KLASIFIKASI SISWA YANG AKAN MENGIKUTI LOMBA OLIMPIADE SAINS NASIONAL (OSN) MENGGUNAKAN ALGORITMA C4.5. Universitas Muhammadiyah Gresik.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (2014). From data mining to knowledge discovery in databases. In Advances in knowledge discovery and data mining (pp. 1-35). Springer US.

Fiska, R.R. (2017) Penerapan Teknik Data Mining dengan Metode Support Vector Machine (SVM) untuk Memprediksi Siswa yang Berpeluang Drop Out (Studi Kasus di SMKN 1 Sutera), SATIN-Sains dan Teknologi Informasi. Available at: http://jurnal.stmik-amik-riau.ac.id.

Fitriyanto, E. T. (2017) Penentuan aturan asosiasi pada transaksi penjualan obat menggunakan algoritma apriori (studi kasus pada RSUD Dr. Soetrasno Rembang). Sanata Dharma University.

Frederika, A.A. et al. (2022) 'Classification Based Association (CBA) Menggunakan R', JITTER-Jurnal Ilmiah Teknologi dan Komputer, 3(1).

Goel, A. (2011) "ANN-Based Approach for Predicting Rating Curve of an Indian River.," International Scholarly Research Network ISRN Civil Engineering, 2011, hal. 4.

Gorunescu, F. (2011) Data Mining Concepts, Models And Techniques. 12 ed. Berlin: Springer - Verlag Berlin Heidelberg.

Grossman, D. and Frieder, O. (2016) Text mining in practice with R. New Jersey: John Wiley & Sons.

Gunawan Sudarsono, B. et al. (2021) 'ANALISIS DATA MINING DATA NETFLIX MENGGUNAKAN APLIKASI RAPID MINER', JBASE - Journal of Business and Audit Information Systems, 4(1), pp. 13–21. Available at: https://doi.org/10.30813/JBASE.V4I1.2729.

Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann.

Han, J. and Kamber, M. (2019) Data mining: concepts and techniques. London: Elsevier Inc.

Han, J., & Kamber, M. (2018). Data mining: concepts and techniques. Morgan Kaufmann Publishers.

Han, J., Kamber, M. and Pei, J. (2012) Data Mining: Concepts and Techniques, San Francisco, CA, itd: Morgan Kaufmann. doi: 10.1016/B978-0-12-381479-1.00001-0.

Handayanto, A. et al. (2019) Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi (Analysis and Application of Algorithm Support Vector Machine (SVM) in Data Mining to Support Promotional Strategies).

Hanke, J. & Wichern, (2005). Business Forecasting. Prentice Hall: New York.

Havolli, A., Maraj, A. and Fetahu, L. (2022) 'Building a content-based recommendation engine model using Adamic Adar Measure; A Netflix case study', 2022 11th Mediterranean Conference on Embedded Computing, MECO 2022 [Preprint]. Available at: https://doi.org/10.1109/MECO55406.2022.9797139.

Hernandez, M. J. (2003). Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design. Addison-Wesley.

Hoberman, S. (2009). Data Modeling for the Business: A Handbook for Aligning the Business with IT using High-Level Data Models. Technics Publications.

Hoffer, J. A., George, J. F., & Valacich, J. S. (2021). Essentials of Systems Analysis and Design. Pearson.

Husaini, F. (2016) ALGORITMA KLASIFIKASI NAÏVE BAYES UNTUK MENILAI KELAYAKAN KREDIT (Studi Kasus : Bank Mandiri Kredit Mikro). Universitas Muhammadiyah Jember.

I Made Yuliara. (2016). Modul Regresi Linier Berganda.

Irwansyah, E. (3. Maret 2017). School of Computer Science. Von Bina Nusantara: https://socs.binus.ac.id/2017/03/09/clustering/ abgerufen

Jaiswal, S. (9. March 2023). Clustering in Data Mining. Von Java T Point: https://www.javatpoint.com/data-mining-cluster-analysis abgerufen

Jo, T. (2015) Text Mining. Edited by J. Kacprzyk. doi: https://doi.org/10.1007/978-3-319-91815-0.

Jo, T. (2019) Text Categorization: Approaches, Studies in Big Data. doi: 10.1007/978-3-319-91815-0_6.

Johnson, M. E., & Singh, A. (2018). Exploring the benefits of data mining in healthcare: A systematic literature review. Journal of Medical Systems, 42(12), 249. https://doi.org/10.1007/s10916-018-1105-2

Joyce Jackson (2002) 'Data Mining: A Conceptual Overview', Communications of the Association for Information Systems, 8, pp. 267–296.

Kafil, M., (2019). Penerapan Metode K-Nearest Neighbors untuk Prediksi Penjualan Berbasis Web pada Boutiq Dealove Bondowoso. Jurnal Mahasiswa Teknik Informatika (JATI), pp. 59-66.

Kantardzic, M. (2011) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Inc.

Kaushik, S. (7. February 2023). Clustering. Von Analytics Vidya: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ abgerufen

Kavitha, S., & Sivanandam, S. N. (2019). An extensive survey on data exploration techniques in data mining. Journal of Ambient Intelligence and Humanized Computing, 10(3), 1013-1031. doi: 10.1007/s12652-018-0842-5

Kononenko, I. and Kukar, M. (2007) Machine learning and data mining: introduction to principles and algorithms. New York: Horwood Publishing.

Kurniawan, C. (2019) 'A Survey on Big Data Analytics Model', ITEJ, 4(1), pp. 1–13.

Kurniawan, Y.I. (2018) 'Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining', Jurnal Teknologi Informasi dan Ilmu Komputer, 5(4), p. 455. Available at: https://doi.org/10.25126/jtiik.201854803.

Kusdarwati, H., Effendi, U. & Handoyo, S., (2022). Analisis Deret Waktu Univariat Linier. Malang: UB Press.

Kusrini and Luthfi, E. T. (2009) Algoritma Data Mining. Yogyakarta: Penerbit Andi Yogyakarta.

Larose, D. T. and Larose, C. D. (2005) Discovering Knowledge in Data: An Introduction to Data Mining, Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc. doi: 10.1002/9781118874059.

Larose, D.T. (2006) Data Mining Methods and Models. New Jersey: John Wiley & Sons.

Mahalakshmi, G., Sridevi, S. and Rajaram, S. (2016) 'A survey on forecasting of time series data', 2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016. doi: 10.1109/ICCTIDE.2016.7725358.

Makridakis, W. & M. (1999) Metode dan Aplikasi Peramalan (terjemahan). Jakarta: Binarupa Aksara.

Mantik, H. and Awaludin, M. (2023) 'REVOLUSI INDUSTRI 4.0: BIG DATA, IMPLEMENTASI PADA BERBAGAI SEKTOR INDUSTRI (BAGIAN 2)', JSI (JURNAL SISTEM INFORMASI), 10(1), pp. 107–121.

Mohajon, J. (2020) "'Confusion Matrix for Your Multi-Class Machine Learning Model,'" Towards Data Science.

Montgomery, D. C., Jennings, C. L. & Kulahci, M., (2015). Introduction to TIme Series Analysis and Forecasting. United States of America: John Willey and Sons.

Mostafa, S. M. (2020). Clustering Algorithms: Taxonomy, Comparison, and Empirical Analysis in 2D. Journal on Artificial Intelligence, 189-215.

Mujiasih, S. (2011) 'PEMANFATAN DATA MINING UNTUK PRAKIRAAN CUACA', Jurnal Meteorologi dan Geofisika, 12(2). Available at: https://doi.org/10.31172/JMG.V12I2.100.

Munawaroh, A. N., (2010). Peramalan Jumlah Penumpang pada PT. Angkasa Pura 1 (Persero) Kantor Cabang Bandar Udara Internasional Adisucipto Yogyakarta dengan Metode Winters Exponential Smoothing dan Seasonal ARIMA. Skripsi.

Muslim, M.A. et al. (2019) Data Mining Algoritma C4.5.

Nabila, Z. et al. (2021) 'ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS', Jurnal Teknologi dan Sistem Informasi, 2(2), pp. 100–108. Available at: https://doi.org/10.33365/JTSI.V2I2.868.

Navisa, S., Hakim, L. and Nabilah, A. (2021) 'Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM', Jurnal Sistem Cerdas, 4(2), pp. 114–125. Available at: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3591.

Nawrin, S., Rahman, M.R. & Akhter, S. (2017) "xploreing k-means with internal validity indexes for data clustering in traffic management system.," International Journal of Advanced Computer Science and Applications, 8(3), hal. 264–272.

Nielsen, A., (2020). Practical Time Series Analysis : Prediction with Statistics & Machine Learning. United States of America: O'Reilly Media Inc.

Niko Ramadhani. (2021, August 30). Regresi Adalah dalam Statistik: Fungsi dan Rumusnya. Akseleran.Co.Id.

Nurdiawan, O. (Odi) and Salim, N. (Noval) (2018) 'Penerapan Data Mining pada Penjualan Barang Menggunakan Metode Metode Naive Bayes Classifier untuk Optimasi Strategi Pemasaran', Jurnal Teknologi Informasi dan Komunikasi, 13(1), pp. 84–95. Available at: https://www.neliti.com/publications/320659/ (Accessed: 25 March 2023).

Nurzahputra, A. and Muslim, M.A. (2017) 'PENINGKATAN AKURASI PADA ALGORITMA C4.5 MENGGUNAKAN ADABOOST UNTUK MEMINIMALKAN RESIKO KREDIT', Prosiding SNATIF, 0(0), pp. 243–247. Available at:

https://jurnal.umk.ac.id/index.php/SNA/article/view/1494 (Accessed: 25 March 2023).

Olson, D. L. and Delen, D. (2008) Advanced Data Mining Techniques, Advanced Data Mining Techniques. Springer. doi: 10.1007/978-3-540-76917-0.

Pambudi, A., Abidin, Z. and Permata, P. (2023) 'PENERAPAN CRISP-DM MENGGUNAKAN MLR K-FOLD PADA DATA SAHAM PT. TELKOM INDONESIA (PERSERO) TBK (TLKM) (STUDI KASUS: BURSA EFEK INDONESIA TAHUN 2015-2022)', JDMSI, 4(1), pp. 1–14.

Pangestu, S., (2013). Forecasting Konsep dan Aplikasi. Yogyakarta: BPPE.

Peng, B., Zhang, X., & Huang, J. Z. (2021) Deep adversarial learning for imbalanced classification with partially missing data. Information Sciences, 560, pp. 343-357.

Pernama, B., Dwi Purnomo, H. and Satya Wacana, K. (2023) 'Analisis Risiko Pinjaman dengan Metode Support Vector Machine, Artificial Neural Network dan Naïve Bayes', Jurnal Teknologi Informasi dan Komunikasi), 7(1), pp. 92–99. Available at: https://doi.org/10.35870/jti.

Prasetyo, E. (2014) Data Mining Mengubah Data Menjadi Informasi.

Prastiwi, H., Pricilia, J. and Raswir, E. (2022) 'Implementasi Data Mining Untuk Menentuksn Persediaan Stok Barang Di Mini Market Menggunakan Metode K-Means Clustering', Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM) , 1(2), pp. 141–148.

Purwanto, H., (1994). Pengantar Statistik Keperawatan. Jakarta: EGC.

Putri, S.B., Anisa, Y.N. and Saputra, N. (2022) 'ANALISIS SENTIMEN FILM KULIAH KERJA NYATA (KKN) DI DESA PENARI MENGGUNAKAN METODE NAIVE BAYES', JuSiTik : Jurnal Sistem dan Teknologi Informasi Komunikasi, 5(2), pp. 22–26. Available at: https://doi.org/10.32524/JUSITIK.V5I2.704.

Putro, B., Tanzil Furqon, M. dan Wijoyo, S. H. (2018) "Prediksi Jumlah Kebutuhan Pemakaian Air Menggunakan Metode Exponential Smoothing (Studi Kasus : PDAM Kota Malang)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(11), hal. 4679–4686. Tersedia pada: http://j-ptiik.ub.ac.id.

Rafi Muttaqin, M. et al. (2022) 'Penerapan K-Means Clustering dan Cross-Industry Standard Process For Data Mining (CRISP-DM) untuk Mengelompokan Penjualan Kue', Journal.Unpak.Ac.Id, 19(1), pp. 38–53. Available at: http://journal.unpak.ac.id/index.php/komputasi/article/view/3976.

Ramadhani, A., (2022). Sistem Prediksi Penjualan dengan Metode Single Exponential Smoothing dan Trend Parabolik. Tangerang Selatan: Pascal Books.

Robianto, Sitorus, S.H. and Ristian, U. (2021) 'PENERAPAN METODE DECISION TREE UNTUK MENGKLASIFIKASIKAN MUTU BUAH JERUK BERDASARKANFITUR WARNA DAN UKURAN', Jurnal Komputer dan Aplikasi , 09, pp. 76–86.

Rukmana, M. and Ramdani, F. (2018) 'Implementasi Algoritme Dijkstra pada Webgis untuk Pencarian Lokasi SPBU di Kota Malang', Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(6), pp. 2141–2149. Available at: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1520.

Saifudin, A. dan Wahono, R. S. (2015) "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," Journal of Software Engineering, 1(1).

Santoso, S., (2009). Business Forecasting : Metode Peramalan Bisnis Masa Kini dengan Minitab dan SPSS. Jakarta: PT. Elex Media Komputindo.

Saputra, N. (2016) 'Analisis Sentimen Mahasiswa Terhadap Universitas', in Seminar Nasional Dinamika Informatika. Available at: https://nanopdf.com/download/seminar-nasional-universitas-pgri-yogyakarta-2016-isbn-978-3_pdf (Accessed: 26 October 2020).

Saputra, N. (2019) 'Analisis Sentimen Self-driving car dengan Sentiment Confident Terbaik', in Seminar Nasional Dinamika Informatika. Yogyakarta: Universitas PGRI Yogyakarta, pp. 40–44. Available at: http://prosiding.senadi.upy.ac.id/index.php/senadi/article/view/101.

Saputra, N., Adji, T.B. and Permanasari, A.E. (2015) 'ANALISIS SENTIMEN DATA PRESIDEN JOKOWI DENGAN PREPROCESSING NORMALISASI DAN STEMMING MENGGUNAKAN METODE NAIVE BAYES DAN SVM', Jurnal Dinamika Informatika, 5(1).

Available at: http://ojs.upy.ac.id/ojs/index.php/dinf/article/view/113 (Accessed: 26 October 2020).

Saputra, N., Nurbagja, K. and Turiyan, T. (2022) 'Sentiment Analysis of Presidential Candidates Anies Baswedan and Ganjar Pranowo Using Naïve Bayes Method', JURNAL SISFOTEK GLOBAL, 12(2), pp. 114–119. Available at: https://doi.org/10.38101/SISFOTEK.V12I2.552.

Sawitri, D. (2019) 'REVOLUSI INDUSTRI 4.0 : BIG DATA MENJAWAB TANTANGAN REVOLUSI INDUSTRI 4.0', JURNAL ILMIAH MAKSITEK, 4(3), pp. 1–9.

Shelly, G. B., Rosenblatt, H. J., & Vermaat, M. E. (2021). Systems Analysis and Design. Cengage.

Simsion, G., & Witt, G. (2005). Data Modeling Essentials. Elsevier.

Singgalen, Y. A. (2022) 'Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM', Building of Informatics, Technology and Science (BITS), 4(3), pp. 1634–1646. doi: 10.47065/bits.v4i3.2766.

Sitompul, B. J. (2018) Peningkatan hasil evaluasi clustering davies-bouldin index dengan penentuan titik pusat cluster awal Algoritma K-Means. UNIVERSITAS SUMATERA UTARA.

Suad A. Alasadi and Wesam S. Bhaya (2017) 'Review of Data Preprocessing Techniques in Data Mining', Journal of Engineering and Applied Sciences, 12(16), pp. 4102–4107.

Sutoyo, I. (2018) 'IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI DATA PESERTA DIDIK', 14(2). Available at: www.bsi.ac.id.

Tahalea, S.P. and Permadi, V.A. (2023) 'Penerapan Aturan Asosiasi Untuk Rule Mining pada Piala Dunia FIFA 2022', Progresif: Jurnal Ilmiah Komputer, 19(1), pp. 165–172.

Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Addison-Wesley Professional.

Tanuwijaya, S., Alamsyah, A. and Ariyanti, M. (2021) 'Mobile Customer Behaviour Predictive Analysis for Targeting Netflix Potential Customer', 2021 9th International Conference on Information and Communication

Technology, ICoICT 2021, pp. 348–352. Available at: https://doi.org/10.1109/ICOICT52021.2021.9527487.

Tarigan, V. (2023) 'Pembuatan Aplikasi Data Mining Untuk Memperediksi Masa Studi Mahasiswa Menggunakan Algoritma Naive Bayes', INFORMATIKA, 11(1), pp. 54–62.

Tripathy, A., Agrawal, A. and Rath, S. K. (2015) 'Classification of Sentimental Reviews Using Machine Learning Techniques', International Conference on Recent Trends in Computing 2015, 57, pp. 821–829. doi: 10.1016/j.procs.2015.07.523.

Tripathy, A., Agrawal, A. and Rath, S. K. (2016) 'Classification of sentiment reviews using n-gram machine learning approach', Expert Systems With Applications, 57, pp. 117–126. doi: 10.1016/j.eswa.2016.03.028.

Ulumuddin, A. dan Juanita, S. (2018) "Implementasi Data Mining Dengan Metode Association Rule Pada Aplikasi Business Analytic Data Penjualan," Skanika, 1(3), hal. 1212–1218.

Villepastour, A. (2019) The Cuban Lexicon Lucumi and African language Yoruba: Musical and historical connections, Handbook of the Changing World Language Map. doi: 10.1007/978-3-030-02438-3_183.

Wahyudi, A.K., Azizah, N. and Saputro, H. (2022) 'DATA MINING KLASIFIKASI KEPRIBADIAN SISWA SMP NEGERI 5 JEPARA MENGGUNAKAN METODE DECISION TREE ALGORITMA C4.5', Journal of Information System and Computer, 2(2), pp. 8–13. Available at: https://journal.unisnu.ac.id/JISTER/.

Wardani, N. W. (2020) Penerapan Data Mining dalam Analytic CRM, Yayasan Kita Menulis. Tersedia pada: https://www.researchgate.net/publication/351776255.

Widaningsih, S. and Yusuf, S. (2022) 'Penerapan Data Mining Untuk Memprediksi Siswa Berprestasi Dengan Menggunakan Algoritma K Nearest Neighbor', Jurnal Teknik Informatika dan Sistem Informasi, 9(3), pp. 2598–2611. Available at: http://jurnal.mdp.ac.id.

Widianto, J. et al. (2022) 'Penerapan RapidMiner dengan Metode K-Means dalam Penentuan Kluster Ganguan Jaringan WIFI Provider PT.XYZ di Daerah Karawang', Jurnal Informatika dan Rekayasa Perangkat Lunak , 4(1), pp. 31–35.

Wijaya, A.D. and Gantini, T. (2019) 'Analisis Forecasting dengan Implementasi Dashboard Business Intelligence Untuk Data Penjualan Pada PT. "X"', Jurnal Strategi, 1(2), pp. 457–470.

Witten, I. H., & Frank, E. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers.

Witten, I. H., Frank, E. and Hall, M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques, Encyclopedia of Ecology, Five-Volume Set. Elsevier Inc. doi: 10.1016/B978-008045405-4.00153-1.

Witten, I.H., Frank, E. and Hall, M.A. (2016) Data mining: practical machine learning tools and techniques. United States: Morgan Kaufmann.

Wu, X. and Kumar, V. (2009) The Top Ten Algorithms in Data Mining, Journal of Physics A: Mathematical and Theoretical. Taylor and Francis Group.

Wulandari, A. et al. (2022) 'Pengembangan Coverage 5G Wilayah Depok Memanfaatkan Analisis Big Data Multi-Parameter', in Prosiding Seminar Nasional Teknik Elektro dan Informatika (SNTEI) 2022-Teknik Telekomunikasi, pp. 116–122.

Xu, Z., Zhang, Y., & Ji, J. (2018). An intelligent framework for data preprocessing in big data analytics. Information Sciences, 423, 1-20. doi: 10.1016/j.ins.2017.08.028

Ye, N. (2014) Data Mining: Theories, Algorithms, and Examples, Data Mining: Theories, Algorithms, and Examples. Taylor and Francis Group.

Yu, J. et al. (2018) 'Modelling domain relationships for transfer learning on retrieval-based question answering systems in E-commerce', WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining, 2018-Febua, pp. 682–690. doi: 10.1145/3159652.3159685.

Zafarani, R., Abbasi, M.A. and Liu, H. (2014) Social media mining: an introduction. Cambridge: Cambridge University Press.

Zai, C. and Komputer, T. (2022) 'IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA', Portaldata.org, 2(3), pp. 1–12.

Zheng, H. and Lytras, M. (2018) 'Web data mining and the development of knowledge-based decision support systems: a case study in the telecom industry.', Journal of Decision Systems, 27(1), pp. 38–47.

Zhou, Y. et al. (2008) 'Large-scale parallel collaborative filtering for the netflix prize', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5034 LNCS, pp. 337–348. Available at: https://doi.org/10.1007/978-3-540-68880-8_32/COVER.

Žižka, J., Dařena, F. and Svoboda, A. (2019) Text Mining with Machine Learning. CRC Press.

Zong, C., Xia, R. and Zhang, J. (2021) Text Classification, Text Data Mining. doi: 10.1007/978-981-16-0100-2_5.

Biodata Penulis

Muttaqin, S.T., M.Cs

Lahir dan besar di Aceh. Pendidikan TK hingga SMA diselesaikan di Kabupaten Aceh Utara dan Kabupaten Bireun Provinsi Aceh. Menyelesaikan Pendidikan D3 Instrumentasi & Komputasi di Universitas Syiah Kuala, S1 Teknik Informatika Sekolah Tinggi Teknik Bina Cendikian Banda Aceh, dan S2 Ilmu Komputer Universitas Gadjah Mada. Mengajar mata kuliah Sistem Operasi Komputer, Kecerdasan Buatan, Sistem Informasi Geografis, Pemodelan Sistem Informasi, Teknik Digital, Pemograman C++, Sistem Basis Data, E-Commerce. Tidak terasa menulis atau menghasilkan karya diawali tahun 2019 sampai Sekarang tahun 2023 baru melahirkan 27 buku: E-Commerce: Implementasi, strategi & Inovasinya (2019), Biometrika Teknologi Identifikasi (2020), Panduan Belajar Manajemen Referensi dengan Mendeley (2020), MOOC: Platform Pembelajaran Daring di Abad 21 (2020), Sistem Pendukung Keputusan: Metode & Implementasi (2020), Sistem Informasi Manajemen (2020), Pembelajaran Daring untuk Pendidikan: Teori dan Penerapan (2020), Etika Profesi: Membangun Profesionalisme Diri (2020), Tren Teknologi Masa Depan (2020),Pengenalan Teknologi Informasi (2020), Keamanan Data dan Informasi (2020), Pengantar Forensik Teknologi Informasi (2021), Statistika Bidang Teknologi Informasi (2021), Sistem Informasi (2021), Hukum dan Cybercrime (2021), Internetworking dan TCP/IP (2021), Teknologi Jaringan Nirkabel (2022), Perancangan Basis Data (2022). BIG DATA: Informasi Dalam Dunia Digital (2022), Dasar-Dasar Teknologi Internet of Things (IoT) (2022), Teknologi Jaringan Komputer (2022), Teknologi Cloud Computing (2022), Google Workspace for Education Platform Pendidikan Digital: Konsep dan Praktik (2022), Konsep Dasar Kecerdasan Buatan (2023), Internet of Things (IoT): Teori dan Implementasi (2023), Digital Learning (2023) Semuanya diterbitkan oleh Penerbit Kita Menulis. Email penulis muttaqin.ugm@gmail.com, Hp/WA. +6285260409204.

Wahyu Wijaya Widiyanto

Lahir di Kudus, pada 18 September 1986. Ia tercatat sebagai lulusan Universitas Amikom Yogyakarta dengan predikat Cumloude pada tahun 2019. Pria yang kerap disapa Wijaya ini adalah anak dari pasangan Suharto (ayah) dan Siti Aminah (ibu). Wahyu Wijaya Widiyanto bukanlah orang baru di dunia Pendidikan. Ia merupakan lulusan D3 Manajemen Informatika Politeknik Indonusa Surakarta pada tahun 2011, S1 Sistem Informasi STMIK Duta Bangsa Surakarta (2017) dan prestasi cumloude selalu disandangnya setiap jenjang pendidikan. Saat ini Wahyu Wijaya Widiyanto aktif sebagai dosen tetap di Politeknik Indonusa Surakarta pada Program Studi Sarjana Terapan Manajemen Informasi Kesehatan. Penulis buku selain bentuk implementasi Tridharma sebagai dosen, Wahyu Wijaya Widiyanto juga aktif dalam melakukan penelitian, hal ini terlihat dari hasil publikasi yang telah dibuatnya, dan dapat dilihat pada akun google scholar di ID NHEh0ZgAAAAJ, ID Sinta 6689713, dan ID Scopus 57215302744.

Muhammad Munsarif

Meraih gelar Magister Ilmu Komputer dari Universitas Dian Nuswantoro (UDINUS) pada tahun 2002. Saat ini menjadi dosen Teknik Informatika di Universitas Muhammadiyah Semarang (UNIMUS). Minat penelitiannya meliputi computer vision, datascience, dan technopreneuership.

Email:m.munsarif@unimus.ac.id

Green Ferry Mandias

Lahir di Kawangkoan 4 Februari 1981. Menjadi anggota Aptikom pada tahun 2014 sampai saat ini. Menyelesaikan sarjana komputer di Universitas Klabat pada tahun 2003 dan Master of Computer Science di Universitas Gadjah Mada 2011. Aktif sebagai dosen di Fakutas Ilmu Komputer Universitas Klabat dan menjadi Ketua Program Studi Informatika Universitas Klabat pada tahun 2018 – saat ini. Bidang kajian yang diminati ialah kajian data mining, database dan datawarehouse.

Selain aktif sebagai dosen, aktif pula dalam penelitian dan pengembangan di bidang informatika dan sistem informasi. Telah mempublikasikan banyak makalah ilmiah di jurnal nasional dan konferensi international. Sering diundang untuk menjadi pembicara dalam seminar dibeberapa daerah dan mempunyai komitmen untuk memberikan pendidikan yang berkualitas dan menginspirasi mahasiswa menjadi ahli dibidangnya.

Stenly Richard Pungus, S.Kom, MT, MM.

Dosen Program Studi Sistem Informasi Universitas Klabat, Airmadidi, Manado. Saya alumni S2 Rekayasa Perangkat Lunak Institut Teknologi Bandung dan memiliki gelar S2 pada bidang manajemen dari Universitas Klabat. Saat ini sedang mengambil Program Doktoral di Universiti Kebangsaan Malaysia dalam area Data Modelling

Agung Widarman

Lahir di Purwakarta, pada tanggal 6 Mei 1982. Saat ini tercatat aktif sebagai Dosen di Sekolah Tinggi Teknologi Wastukancana Purwakarta sejak tahun 2006. Pendidikan terakhir S2 di Universitas Pasundan Bandung, tahun 2016. Memiliki ketertarikan dalam keilmu Manajemen, Teknik Industri, Teknologi Informasi dan Komunikasi. Sejak bergabung dengan Komunitas Kita Menulis sudah 9 kali ikut serta dalam menulis Buku Referensi untuk buku berjudul Manajemen Strategi Kontemporer, Manajemen Operasi, Dasar Komunikasi Organisasi, Konsep Dasar Sistem Informasi Dalam Dunia Usaha, Pengantar Teknologi Informasi dan Komputer, Bisnis Inovasi dan Kreatif, Strategi Digital Marketing untuk Bisnis Digital, Ekonomi Teknik, Mengenal Bisnis Kuliner, Pengantar Internet dan sepertinya akan mencoba untuk terus menulis.

Wiranti Kusuma Hapsari

Lahir di Banyumas, pada 16 Juli 1994. Tercatat sebagai lulusan S1 Teknik Informatika, Universitas Muhammadiyah Purwokerto dan S2 Ilmu Komputer, Universitas Gadjah Mada. Saat ini berprofesi sebagai dosen Tetap Fakultas Teknologi, Universitas Pertiwi.

Siska Aprilia Hardiyanti

Lahir di Banyuwangi, pada tanggal 1 April 1992. Ia tercatat sebagai lulusan strata 1 pendidikan matematika di Universitas Jember tahun 2014 dan magister Matematika di Institut Teknologi Sepuluh Nopember Surabaya tahun 2016. Wanita yang kerap disapa Siska ini diangkat menjadi dosen di Politeknik Negeri Banyuwangi tahun 2016 dan ditempatkan di Jurusan Teknik Sipil pada program studi Teknik Sipil.

Aslam Fatkhudin, S.Kom. M.Kom. kelahiran Pekalongan, 16 Mei 1982 ini merupakan lulusan Program Studi Sarjana Teknik Informatika Universitas Abadi Karya Indonesia (UNAKI) Semarang tahun 2003, melanjutkan pendidikan Pasca Sarjananya di Program Studi Magister Sistem Informasi Universitas Diponegoro (MSI UNDIP) lulus tahun 2014. Saat ini sebagai dosen tetap pada Program Studi Sarjana Informatika yang juga menjabat sebagai Wakil Rektor III Bidang Kemahasiswaan Universitas Muhammadiyah Pekajangan Pekalongan. Sebagai Dosen, beberapa jurnal ilmiah sudah beliau hasilkan, termasuk menjadi narasumber dalam berbagai seminar atau workshop seputar Teknologi Informasi maupun Sistem Informasi. Beberapa buku yang sudah pernah ditulisnya adalah buku dengan judul "Membangun Web E-Commerce dengan Quick Cart" Penerbit Komunitas Gemulun Indonesia tahun 2022 dan buku dengan judul "Kemanan Siber Tantangan di Era Revolusi Industri 4.0" Penerbit Yayasan Kita Menulis tahun 2022.

Pasnur

Lahir di Parepare pada tanggal 15 April 1980. Penulis menyelesaikan pendidikan Strata I (S1) di Jurusan Elektro Fakultas Teknik Universitas Hasanuddin Makassar pada tahun 2004. Penulis menyelesaikan pendidikan Magister (S2) di Jurusan Teknik Informatika Fakultas Teknologi Informasi Institut Teknologi Sepuluh Nopember Surabaya pada tahun 2015. Penulis aktif mengajar di Universitas Teknologi Akba Makassar (Unitama) sejak tahun 2009 sebagai Dosen Tetap pada program studi Teknik Informatika. Fokus penelitian yang dilakukan adalah bidang Natural Language Processing, Information Retrieval, Computer Vision, Image Processing, dan Data Mining. Penulis juga aktif menjadi reviewer pada beberapa Jurnal Nasional Terakreditasi (SINTA 2 – SINTA 4). Selain itu penulis merupakan Kepala UPT ICT Unitama dengan salah satu tugas adalah melakukan pengembangan sistem informasi manajemen terintegrasi.

Eva Firdayanti Bisono

Lahir di Sidoarjo, pada 11 Juni 1992. Ia tercatat sebagai lulusan Universitas Brawijaya Malang dan Institut Teknologi Sepuluh Nopember Surabaya.. Wanita yang kerap disapa Eva ini adalah anak dari bungsu dari dua bersaudara. Eva Firdayanti Bisono bukanlah orang baru di dunia Data Mining. Ia kerap melakukan penelitian seputar penyelesaian kasus Data Mining.

Mochammad Anshori

Lahir di Banyuwangi, Jawa Timur, bulan Maret 1994. Mengawali menulis pada tahun 2022 ini. Anak terakhir dari dua bersaudara. Menginjak pendidikan dasar di SD Muhammadiyah (2001), lanjut ke SMPN 1 Glagah (2007) dan MAN Banyuwangi (2010). Kemudian memasuki pendidikan tinggi S1 Informatika di Universitas Muhammadiyah Malang (2012) dan S2 Ilmu Komputer di Universitas Brawijaya (2017). Pernah bekerja secara freelance dan memulai karir mandiri dengan mengerjakan beberapa proyek dengan bahasa pemrograman Java berbasis desktop. Kemudian menjadi asisten trainer pada program Digitalent Scholarship yang diselenggarakan oleh Kominfo saat masih kuliah S2 sebanyak 3 batch berturut-turut sejak tahun 2019. Terakhir kali menjadi trainer dengan tema Orange for Data Mining yang diselenggarakan oleh IAIS. Pernah mengikuti kegiatan ilmiah prosiding internasional sebanyak 3 kali. Saat ini penulis sedang aktif bekerja sebagai dosen di kampus Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW Malang, tepatnya pada program studi S1 Informatika

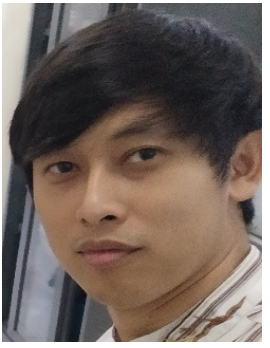
Suryani, S.Kom., M.T.

Lahir di Maros, pada 4 januari 1987. Ia tercatat sebagai lulusan Universitas Dipa Makassar dan Universitas Hasanuddin Makassar. Wanita yang kerap disapa Surya ini adalah anak pertama dari pasangan Zainuddin Dg. Ajang (ayah) dan Nurjannah Dg. Ngai (ibu). Suryani berprofesi sebagai salah satu dosen Universitas Dipa Makassar mulai 1 Maret 2016 hingga saat ini. Bermula mengecap pendidikan di perguruan tinggi program Diploma Tiga (D3) Jurusan Manajemen Informatika di Universitas Dipa Makassar ex STMIK Dipanegara pada tahun 2008, kemudian mendapatkan beasiswa wisudawan terbaik utama melanjutkan study program Strata Satu (S1) Jurusan Teknik Informatika di kampus yang sama, dan selesai dengan predikat Terbaik Utama atau Cum

Laude dengan IPK Tertinggi yaitu 4.00. Dengan mendapatkan Beasiswa Unggulan (BU) Calon Dosen dari LLDIKTI, Ia menyelesaikan study program Pasca Sarjana atau S2 Program Study Teknik Elektro Jurusan Teknik Informatika di Universitas Hasanuddin makassar.

Email Penulis: suryani187@undipa.ac.id

Nurirwan Saputra



Lahir di Serang, pada 20 Mei 1988. Ia tercatat sebagai lulusan Prodi Teknik Informatika, Universitas Islam Indonesia dan Magister Teknologi Informasi, Universitas Gadjah Mada. Saat ini sudah bekerja di Prodi Informatika Universitas PGRI Yogyakarta dari tahun 2015-sekarang. Penelitian yang menjadi fokusnya adalah terkait Analisis Sentimen.

Pengenalan Data Mining

Buku ini, berjudul "Pengenalan Data Mining", dirancang sebagai panduan praktis bagi siapa saja yang ingin mempelajari konsep-konsep dasar data mining dan menerapkannya dalam praktik. Buku ini mencakup berbagai topik penting dalam data mining, mulai dari pendahuluan, pengumpulan data, preprocessing data, eksplorasi data, hingga teknik-teknik modeling, evaluasi model, dan aplikasi data mining dalam berbagai bidang.

Setiap bab dalam buku ini dirancang untuk memberikan pemahaman yang mendalam mengenai konsep-konsep dasar, teknik-teknik, dan aplikasi data mining yang relevan. Selain itu, setiap bab juga dilengkapi dengan contoh kasus dan latihan-latihan praktis, sehingga pembaca dapat memperoleh pengalaman langsung dalam menerapkan teknik-teknik data mining dalam praktik.

Secara lengkap buku ini membahas :

YAYASAN KITA MENULIS
press@kitamenulis.id
www.kitamenulis.id

ISBN 978-623-342-791-3