BAB II

LANDASAN TEORI

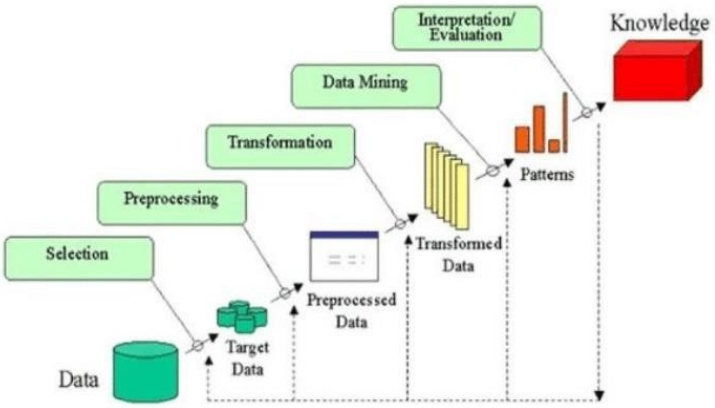
2.1. Data Mining

Data mining adalah proses ekstraksi pola, pengetahuan, dan informasi yang bermanfaat dari dataset yang besar dan kompleks [1]. Tujuan utama dari data mining adalah untuk mengungkap pola tersembunyi, hubungan, dan tren yang tidak selalu terlihat secara langsung [2]. Dalam analisis data ini, berbagai metode dan teknik seperti clustering, klasifikasi, asosiasi, dan regresi digunakan untuk merinci struktur data dan membuat prediksi yang dapat mendukung pengambilan keputusan [3]. Data mining memiliki penerapan luas di berbagai bidang, termasuk bisnis, ilmu pengetahuan, kesehatan, keamanan, dan lainnya, memberikan pemahaman yang lebih dalam dan bernilai tambah terhadap informasi yang terkandung dalam dataset yang besar dan kompleks.

Dalam proses data mining, langkah-langkah awal melibatkan pemahaman masalah, pengumpulan data, dan eksplorasi data untuk mempersiapkan dataset sebelum analisis lebih lanjut. Setelah itu, preprocessing data diterapkan untuk membersihkan data, menangani missing values, dan melakukan normalisasi agar data siap digunakan oleh algoritma analisis seperti Naive Bayes, K-Means, atau metode lainnya. Hasil analisis kemudian dievaluasi dan divalidasi untuk memastikan keakuratan dan keandalan model yang dikembangkan. Selain itu, dokumentasi yang cermat dari seluruh proses data mining menjadi kunci untuk memberikan transparansi, memfasilitasi pengulangan analisis, dan memperkuat interpretasi hasil agar dapat memberikan wawasan yang berarti kepada pemangku

kepentingan. Dengan memanfaatkan data mining secara efektif, organisasi dapat membuat keputusan yang lebih informasional, memprediksi tren masa depan, dan meningkatkan pemahaman mereka terhadap lingkungan bisnis atau domain lainnya.

2.2. Knowledge Discovery in Database (KDD)



Gambar 2.1 Knowledge Discovery in Database

Knowledge Discovery in Databases (KDD) adalah proses menyeluruh yang mencakup beberapa langkah untuk mengidentifikasi, memahami, dan memanfaatkan pengetahuan yang bermanfaat dari data. Langkah pertama dalam KDD adalah pemahaman masalah, di mana tujuan dan persyaratan bisnis dijelaskan dengan jelas. Setelah itu, dilakukan pemilihan dan pengumpulan data yang relevan dari berbagai sumber. Proses eksplorasi data kemudian digunakan untuk mengidentifikasi pola, tren, dan anomali dalam data tersebut. Langkah berikutnya adalah preprocessing data, di mana data dibersihkan, diubah, dan disiapkan untuk analisis. KDD melibatkan penggunaan algoritma dan teknik data mining, seperti clustering, klasifikasi, asosiasi, dan regresi, untuk mengekstrak

pengetahuan yang tersembunyi dalam dataset. Evaluasi model dan validasi kemudian dilakukan untuk memastikan keandalan hasil yang diperoleh.

Dokumentasi dari setiap tahap proses KDD sangat penting, tidak hanya untuk memberikan transparansi terhadap analisis yang dilakukan tetapi juga sebagai referensi untuk pemahaman ulang, replikasi, atau pengembangan lebih lanjut. Dengan menerapkan KDD secara efektif, organisasi dapat mengoptimalkan pengambilan keputusan, meramalkan tren pasar, dan meningkatkan pemahaman mereka terhadap dinamika dalam data mereka. Keseluruhan, KDD memberikan kerangka kerja yang komprehensif untuk mengubah data menjadi pengetahuan berharga dan actionable.

2.3. Metode Naïve Bayes

Metode Naive Bayes adalah salah satu algoritma klasifikasi yang sangat populer dalam bidang data mining dan machine learning [4]. Algoritma ini berdasarkan pada teorema probabilitas Bayes dan mengasumsikan bahwa setiap fitur dalam dataset adalah independen satu sama lain, meskipun dalam kenyataannya, asumsi ini seringkali tidak sepenuhnya terpenuhi [5]. Meskipun sederhana, Naive Bayes dapat memberikan kinerja yang baik terutama pada dataset dengan dimensi tinggi [6]. Proses pelatihan model Naive Bayes melibatkan perhitungan probabilitas kelas target dan probabilitas setiap fitur dalam kelas tersebut. Selama fase prediksi, model ini menggunakan probabilitas yang dihitung untuk menentukan kelas yang paling mungkin untuk suatu instance data baru.

Kelebihan dari Naive Bayes adalah kecepatan pelatihan dan prediksi yang tinggi, serta ketangguhan terhadap keberadaan fitur yang tidak relevan.

Namun, Naive Bayes juga memiliki keterbatasan, terutama dalam mengatasi ketergantungan antar-fitur yang sebenarnya. Meskipun demikian, algoritma ini sering digunakan dalam aplikasi seperti klasifikasi teks, deteksi spam email, dan analisis sentimen, di mana kecepatan dan ketepatan prediksi sangat dihargai. Kesederhanaan Naive Bayes membuatnya menjadi pilihan yang baik untuk tugas klasifikasi pada dataset besar dengan fitur yang berkisar dan kompleksitas yang moderat.

2.3.1. Uji Performa

Uji performa menggunakan confusion matrix pada metode Naive Bayes adalah langkah kritis dalam mengevaluasi keakuratan model klasifikasi. Confusion matrix memetakan hasil prediksi model terhadap kelas sebenarnya dalam empat kategori: true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). Dari sini, berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score dapat dihitung, memberikan gambaran yang holistik tentang kinerja model Naive Bayes. Pemahaman yang mendalam terhadap confusion matrix memungkinkan para praktisi untuk mengidentifikasi kelemahan model, memperbaiki parameter yang tepat, dan meningkatkan kemampuan prediksi secara keseluruhan.

		Kelas Prediksi	
Kelas Atribut	Kelas Benar	Belar True Positive (TP)	Salah False Positive (FP)
	Salah	False Negative (FN)	True Negative (TN)

$$Acuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\%$$

$$Presisi = \frac{TP}{100\% TP + FP} \times$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

2.4. Metode K-Means

Metode K-Means adalah salah satu algoritma clustering yang digunakan dalam analisis data [7]. Tujuannya adalah mengelompokkan data ke dalam k kelompok atau cluster sedemikian rupa sehingga data dalam satu kelompok memiliki kemiripan yang tinggi dan berbeda dengan kelompok lainnya. Algoritma ini bekerja dengan cara menginisialisasi pusat kluster secara acak, kemudian mengelompokkan data ke kluster terdekat dan memperbarui pusat kluster berdasarkan rata-rata data dalam setiap kluster. K-Means meminimalkan inersia atau jumlah kuadrat jarak antara setiap data dengan pusat kluster yang terkait. Proses ini dilakukan secara iteratif hingga konvergensi, di mana tidak ada perubahan yang signifikan dalam alokasi data ke kluster atau perubahan dalam pusat kluster. Kelebihan K-Means meliputi kemudahan implementasi, kecepatan

konvergensi, dan keterbatasannya dalam penanganan data dengan bentuk dan ukuran yang kompleks. Namun, K-Means juga memiliki beberapa keterbatasan, seperti sensitivitas terhadap inisialisasi pusat kluster awal dan ketidakmampuannya menangani kelompok yang memiliki bentuk atau ukuran yang tidak bulat. Selain itu, penentuan jumlah kluster k juga merupakan tantangan yang perlu diperhatikan secara hati-hati. Meskipun demikian, K-Means tetap menjadi algoritma clustering yang umum digunakan dalam berbagai bidang, termasuk analisis data, ilmu komputer, dan pengolahan citra.

Dalam metode K-Means, penentuan kluster untuk suatu data dilakukan dengan menghitung jarak Euclidean antara data tersebut dan pusat kluster yang terkait. Jarak Euclidean diukur sebagai jarak linier atau sejati antara dua titik dalam ruang berdimensi n. Rumus jarak Euclidean umumnya dinyatakan sebagai akar kuadrat dari jumlah kuadrat perbedaan antara setiap dimensi. Proses ini memungkinkan penentuan kluster dengan menemukan pusat kluster yang memiliki jarak minimum dengan setiap data, sehingga data tersebut akan terelokasi ke kluster dengan pusat terdekat berdasarkan jarak Euclidean tersebut. Meskipun sederhana, penggunaan jarak Euclidean dalam K-Means memberikan pemahaman yang jelas tentang kedekatan data dengan pusat kluster, yang menjadi dasar utama dalam pembentukan kelompok yang homogen.

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

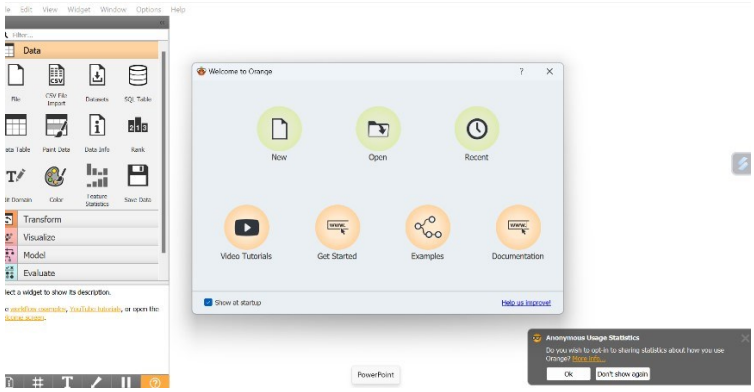
2.5. Alat Bantu Pemrograman/Tools Pendukung

2.5.1. Orange

Orange adalah sebuah aplikasi open-source yang populer untuk analisis data dan data mining, dirancang dengan antarmuka pengguna yang intuitif serta berbasis visual. Aplikasi ini memungkinkan pengguna, baik yang berpengalaman maupun pemula, untuk melakukan analisis data secara cepat dan efisien tanpa perlu menulis kode. Dengan menyediakan berbagai widget yang dapat dihubungkan seperti blok bangunan, Orange memudahkan pengguna dalam mengembangkan alur kerja

analisis data, mulai dari pembersihan data, visualisasi, hingga penerapan algoritma pembelajaran mesin. Hal ini membuat Orange menjadi alat yang sangat berguna dalam proses eksplorasi data serta pengambilan keputusan berbasis data.

Orange mendukung berbagai teknik data mining seperti klasifikasi, klasterisasi, regresi, dan asosiasi, sehingga cocok untuk berbagai jenis proyek analisis data. Pengguna dapat dengan mudah mengimpor data dari berbagai sumber, termasuk file CSV, Excel, SQL, dan bahkan sumber online. Setelah data diimpor, Orange menawarkan berbagai opsi untuk menganalisis dan memvisualisasikan data, seperti diagram alir, pohon keputusan, peta panas, dan banyak lagi. Kombinasi antara kesederhanaan penggunaan dan kemampuan analisis yang kuat menjadikan Orange sebagai pilihan ideal bagi para peneliti, akademisi, dan profesional yang ingin melakukan analisis data tanpa harus terlibat dalam kompleksitas pemrograman.



Gambar 2.2. Tampilan Awal Aplikasi Orange

2.6. Metodologi Penelitian

2.6.1. Penelitian Terdahulu

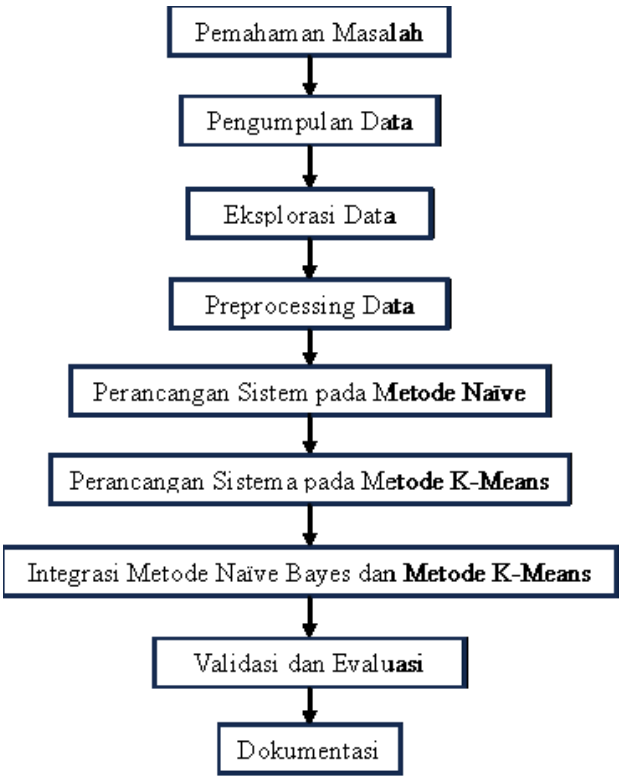
Tabel 2.1 Penelitian Terdahulu

Referensi Penelitian	1
Judul	Application of the K-Means Clustering Agorithm to Group Train Passengers in Labuhanbatu
Nama Penulis	Indri Cahaya Indah1)*, Mila Nirmala Sari2), Muhammad Halmi Dar3)
Tahun	2023
Hasil	Transportasi memiliki peran krusial dalam memindahkan benda dan individu dari satu lokasi ke lokasi lain. Dalam konteks transportasi darat di Indonesia, khususnya kereta api, kami berencana melakukan klasterisasi data penumpang menggunakan metode K-Means. Fokus klasterisasi ini adalah mengidentifikasi pola dan preferensi penumpang kereta api dalam tiga kelompok menggunakan aplikasi oranye, dengan

	harapan mendapatkan wawasan yang lebih dalam mengenai karakteristik pengguna, baik dari kelas bisnis maupun eksekutif [7].
Referensi Penelitian	2
Judul	Data Mining Sales of Skin Care Products Using the K-Means Method
Nama Penulis	Dasril Aldo1)
Tahun	2023
Hasil	Data mining, sebagai kemajuan metode dalam komputerisasi, menjadi solusi dalam penelitian mengenai penjualan produk kecantikan dan skincare yang saat ini masih dilakukan secara manual. Penelitian ini memfokuskan pada penggunaan metode KMeans untuk mengelompokkan data penjualan menjadi dua kategori: produk yang paling laris dan produk yang tidak terjual. Dengan tujuan untuk menyelaraskan stok produk dengan permintaan pasar, penelitian ini menggunakan sampel 30 data, di mana hasil perhitungan manual dan menggunakan aplikasi RapidMiner menunjukkan kemiripan 100%. Dengan demikian, dapat disimpulkan bahwa algoritma K-Means efektif sebagai solusi untuk mengoptimalkan manajemen stok dan penjualan produk kecantikan dan skincare [8].
Referensi Penelitian	3

Judul	Clustering Analysis of Tweets About COVID19 Using the K-Means Algorithm
Nama Penulis	Andi1)* , Carles Juliandy2) , David3)
Tahun	2023
Hasil	Penelitian ini membahas pengelompokan tweet COVID-19 tahun 2020-2022 menggunakan metode K-Means untuk mengatasi ketidakteraturan dan pencampuran tweet. Dengan dataset dari Kaggle dan Bright Data (4.103 data), hasil analisis Elbow menunjukkan k = 5 sebagai jumlah cluster optimal. Hasilnya menunjukkan cluster terbesar adalah cluster 4 (1.185 tweet), diikuti oleh cluster 1 (1.047 tweet), cluster 2 (757 tweet), cluster 3 (744 tweet), dan cluster terkecil adalah cluster 5 (370 tweet). Pendekatan ini memberikan kontribusi pada organisasi dan keterbacaan tweet COVID-19 di Twitter, mempermudah akses informasi bagi pengguna [9].

2.7. Kerangka Penelitian



Gambar 2.3. Kerangka Kerja Penelitian

Pada gambar diatas merupakan kerangka kerja penelitian ini, adapun untuk penjelasan dari setiap tahapannya yaitu sebagai berikut:

1. Pemahaman Masalah

Dalam pemahaman masalah pada data mining, metode Naive Bayes digunakan untuk klasifikasi data dengan memanfaatkan probabilitas kondisional. Naive Bayes efektif dalam mengidentifikasi pola dan hubungan antar variabel, terutama pada dataset dengan variabel-variabel kategorikal. Di sisi lain, metode KMeans digunakan untuk analisis clustering dengan mengelompokkan data ke dalam cluster-cluster yang memiliki kesamaan. K-Means cocok untuk mengidentifikasi struktur tersembunyi dalam data, memungkinkan

pengelompokan berdasarkan kesamaan karakteristik. Integrasi kedua metode ini dapat memberikan pemahaman holistik terhadap pola dan struktur yang ada dalam dataset, menghasilkan informasi yang lebih komprehensif.

2. Pengumpulan Data

Dalam pengumpulan data pada data mining, metode Naive Bayes membutuhkan dataset yang mencakup variabel target dan atribut-atribut yang relevan untuk melatih model probabilitas klasifikasi. Sebaliknya, metode K-Means memerlukan data numerik yang mencerminkan pola atau kesamaan dalam struktur data. Pengumpulan data yang baik untuk kedua metode ini harus memperhatikan keberagaman dan representativitas dataset, serta memastikan keberlanjutan dan integritas variabel yang digunakan dalam analisis untuk menghasilkan hasil yang akurat dan berarti.

3. Eksplorasi Data

Dalam eksplorasi data pada data mining, metode Naive Bayes memanfaatkan visualisasi dan analisis statistik deskriptif untuk memahami distribusi dan tren dalam dataset, memungkinkan identifikasi variabel yang relevan untuk klasifikasi. Sementara itu, metode K-Means menggunakan visualisasi data untuk memahami pola dan hubungan antar data, fokus pada analisis clustering untuk mengidentifikasi kelompok data yang memiliki kesamaan. Kedua metode ini bergantung pada langkah eksplorasi data untuk memberikan wawasan awal sebelum penerapan

model analisis lebih lanjut.

4. Preprocessing Data

Dalam preprocessing data pada data mining, baik metode Naive Bayes maupun K-Means memerlukan tahapan normalisasi atau standarisasi data guna memastikan konsistensi dan akurasi analisis. Selain itu, keduanya melibatkan encoding variabel kategorikal dan pemisahan data menjadi subset training dan testing untuk mempersiapkan data secara optimal sebelum pelatihan model. Preprocessing data yang cermat menjadi kunci dalam memastikan bahwa data yang digunakan sesuai dengan persyaratan masing-masing metode, sehingga meningkatkan kualitas analisis yang dihasilkan.

5. Perancangan Sistem pada Metode Naive Bayes

Perancangan sistem pada metode Naive Bayes dalam data mining melibatkan langkah-langkah untuk membangun model probabilitas klasifikasi berdasarkan data training. Sebaliknya, dalam integrasi dengan metode K-Means, perlu mempertimbangkan pengaturan jumlah cluster yang sesuai agar output dari kedua metode dapat diintegrasikan secara efektif. Desain sistem ini penting untuk memastikan kesesuaian antara model klasifikasi Naive Bayes dan analisis clustering K-Means, sehingga dapat menghasilkan solusi yang holistik dan informatif terhadap data yang diamati.

6. Perancangan Sistem pada Metode K-Means

Dalam perancangan sistem pada metode K-Means dalam data mining, fokus utama adalah menentukan jumlah cluster yang optimal untuk mengelompokkan data berdasarkan kesamaan karakteristiknya. Saat mengintegrasikan dengan metode Naive Bayes, perlu dipertimbangkan cara menggabungkan hasil clustering

dengan model probabilitas klasifikasi, sehingga desain sistem mencakup langkahlangkah integrasi yang efisien untuk memperoleh pemahaman yang holistik terhadap struktur dan pola dalam dataset.

7. Integrasi Metode Naive Bayes dan Metode K-Means

Integrasi metode Naive Bayes dan K-Means pada data mining menghasilkan pendekatan analisis yang holistik. Dengan menyatukan kemampuan klasifikasi probabilitas Naive Bayes dan kemampuan pengelompokan K-Means, sistem dapat memberikan wawasan yang lebih dalam terhadap pola, hubungan, dan struktur data, menciptakan solusi analisis yang lebih komprehensif dan informatif.

8. Validasi dan Evaluasi

Dalam validasi dan evaluasi pada data mining dengan metode Naive Bayes dan K-Means, perlu dilakukan penilaian terhadap akurasi model klasifikasi Naive Bayes serta efektivitas pengelompokan K-Means. Proses ini membantu memastikan bahwa hasil analisis dapat diandalkan dan sesuai dengan tujuan awal, memungkinkan identifikasi area perbaikan atau pengembangan lebih lanjut guna meningkatkan kualitas dan kebermanfaatan solusi yang diberikan oleh kedua metode tersebut.

9. Dokumentasi

Dokumentasi pada data mining dengan metode Naive Bayes dan K-Means memiliki peran penting dalam merekam secara rinci langkah-langkah analisis, parameter yang digunakan, serta hasil dan temuan signifikan.