BAB II

LANDASAN TEORI

2.1. *Data Mining*

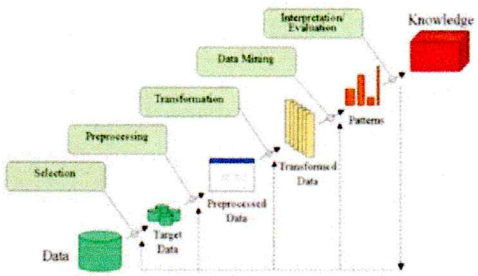
Data mining adalah proses menganalisis dan mengekstraksi informasi berharga dari kumpulan data yang besar [1]. Proses ini melibatkan penggunaan berbagai teknik statistik, matematika, dan algoritma komputer untuk menemukan pola, tren, dan hubungan dalam data yang tidak langsung terlihat [2]. Data mining digunakan dalam berbagai bidang seperti pemasaran, keuangan, kesehatan, dan pendidikan untuk membantu pengambilan keputusan yang lebih baik [3]. Dengan kemampuan untuk mengolah data dalam jumlah besar dan kompleks, data mining memungkinkan organisasi untuk mengidentifikasi peluang baru, mengoptimalkan operasi, dan meningkatkan strategi bisnis [4].

Metode yang sering digunakan dalam data mining termasuk klasifikasi, klastering, asosiatif, dan regresi. Salah satu algoritma yang populer adalah Naive Bayes, yang didasarkan pada teorema Bayes dan digunakan untuk klasifikasi data berdasarkan probabilitas. Metode ini sangat efektif dalam menangani berbagai jenis data dan sering digunakan dalam aplikasi seperti filter spam, diagnosis medis, dan analisis perilaku konsumen. Dalam konteks pendidikan, data mining dapat digunakan untuk menganalisis data siswa, mengidentifikasi faktor-faktor yang mempengaruhi prestasi akademis, dan mengembangkan strategi pembelajaran yang lebih efektif. Dengan memanfaatkan data mining, institusi pendidikan dapat lebih memahami kebutuhan dan minat siswa, serta meningkatkan kualitas pengajaran dan hasil belajar.

2.2. *Knowledge Discovery in Database (KDD)*

Knowledge Discovery in Databases (KDD) adalah proses penggalian pengetahuan atau informasi yang bermanfaat dari data yang besar dan kompleks. KDD melibatkan langkah-langkah seperti seleksi, pre-processing, transformasi, data mining, interpretasi, dan evaluasi. Tujuan utama dari KDD adalah untuk menemukan pola, tren, atau pengetahuan yang dapat digunakan untuk pengambilan keputusan yang lebih baik. Proses ini sering kali melibatkan kombinasi dari berbagai teknik seperti data mining, statistik, kecerdasan buatan, dan analisis pola.

Salah satu tahap penting dalam KDD adalah data mining, yang merupakan proses pencarian pola atau informasi yang signifikan dari data. Data mining menggunakan berbagai algoritma dan teknik untuk mengidentifikasi pola-pola yang tidak terduga dalam data, yang kemudian dapat digunakan untuk membuat prediksi atau mengambil keputusan. Contoh penerapan KDD dan data mining termasuk dalam bidang pemasaran, keuangan, kesehatan, dan ilmu pengetahuan. Dengan kemampuannya untuk mengungkap pengetahuan yang tersembunyi dalam data, KDD memberikan nilai tambah yang signifikan bagi organisasi dalam menghadapi tantangan pengolahan dan analisis data yang semakin kompleks.



Gambar 2. 1. Knowledge Discovery in Database

Sumber Gambar: *https://www.researchgate.net*

Interpretation/Evaluation : Interpretation dalam KDD (Knowledge Discovery in Databases) merupakan proses penting untuk menguraikan dan memahami pola atau penemuan yang ditemukan melalui analisis data, membantu mengambil wawasan yang berarti dari data yang telah diproses.

Data Mining : Data mining pada Konferensi Penemuan Pengetahuan dan Data (KDD) adalah proses mengidentifikasi pola yang bermanfaat dan informasi yang berguna dari data besar untuk mendukung pengambilan keputusan. Metode ini melibatkan penggunaan teknik statistik, machine learning, dan kecerdasan buatan untuk menganalisis dan memahami pola yang tersembunyi dalam data.

Transformation : Transformasi pada KDD (Knowledge Discovery in Databases) adalah proses mengubah data mentah menjadi bentuk yang lebih berguna dan mudah dipahami untuk analisis lebih lanjut. Transformasi melibatkan penggabungan, normalisasi, penghapusan noise, dan langkah-langkah lain untuk mempersiapkan data untuk tahapan selanjutnya dalam proses penemuan pengetahuan.

Preprocessing	:	Preprocessing pada KDD (Knowledge Discovery in Databases) adalah tahapan penting yang meliputi pembersihan, transformasi, dan penggabungan data untuk mempersiapkan data mentah menjadi sesuatu yang lebih mudah dianalisis oleh algoritma pembelajaran mesin. Tahap ini membantu memastikan keakuratan dan kualitas data sebelum dilakukan proses analisis lebih lanjut.
Selection	:	Selection pada KDD (Knowledge Discovery in Databases) adalah proses penting yang bertujuan untuk memilih subset data yang relevan dan signifikan untuk analisis lebih lanjut. Proses ini melibatkan pemilihan atribut atau instance yang sesuai dengan kriteria yang telah ditentukan.

KDD melibatkan langkah-langkah yang kompleks dan berurutan untuk menghasilkan pengetahuan yang berarti. Setelah proses data mining selesai, langkah berikutnya adalah interpretasi dan evaluasi hasil untuk memastikan kebenaran dan kegunaan penemuan tersebut. Interpretasi melibatkan pemahaman yang mendalam tentang konteks bisnis atau ilmiah di mana pengetahuan tersebut akan digunakan, serta menghubungkan hasil dengan tujuan yang ingin dicapai. Evaluasi dilakukan untuk mengukur sejauh mana pengetahuan yang ditemukan dapat diandalkan dan berguna dalam konteks aplikasi yang dimaksud.

Selain itu, KDD juga mencakup langkah-langkah untuk menyajikan informasi yang ditemukan secara efektif kepada pengguna akhir. Ini melibatkan representasi visual dari pola-pola yang ditemukan, seperti grafik, diagram, atau laporan yang mudah dimengerti. Presentasi yang efektif dari pengetahuan yang ditemukan sangat penting untuk memastikan bahwa informasi tersebut dapat digunakan dengan baik oleh para pengambil keputusan. Dengan demikian, KDD bukan hanya tentang menggali pengetahuan dari data, tetapi juga tentang menyajikan pengetahuan tersebut secara efektif sehingga dapat memberikan dampak yang signifikan dalam pengambilan keputusan.

2.3. Metode Algoritma Naïve Bayes

Algoritma Naive Bayes adalah salah satu metode klasifikasi yang paling sederhana dan efektif dalam data mining. Metode ini didasarkan pada teorema Bayes yang menggunakan probabilitas untuk klasifikasi data. Meskipun sederhana, algoritma ini sering kali memberikan hasil yang baik dalam berbagai aplikasi, terutama dalam klasifikasi teks dan analisis sentimen, filter spam, serta diagnosis medis. Salah satu keunggulan utama dari algoritma Naive Bayes adalah kemampuannya untuk mengatasi masalah dimensi tinggi, di mana jumlah fitur atau variabel independen yang digunakan untuk klasifikasi sangat besar. Algoritma ini juga cepat dalam melakukan pelatihan model dan prediksi, membuatnya cocok untuk digunakan dalam aplikasi real-time. Namun, kelemahan utamanya terletak pada asumsi "naif" atau sederhana yang dibuat, yaitu menganggap bahwa setiap fitur independen satu sama lain, meskipun dalam kenyataannya ada ketergantungan antara fitur-fitur tersebut.

$$\mathbf{P(A \mid B):} \frac{P(B \mid A)\, P(A)}{P(B)}$$

Information:

A	:	hipotesis data A (kelas tertentu)
B	:	data dengan kelas yang tidak diketahui
P (A | B)	:	Probabilitas hipotesis berdasarkan kondisi B
P (A)	:	Kemungkinan hipotesis A
P (B | A)	:	Probabilitas B ketika kondisi A
P (B)	:	Probabilitas

Ada beberapa variasi dari algoritma Naive Bayes, termasuk Naive Bayes Gaussian, Naive Bayes Multinomial, dan Naive Bayes Bernoulli, yang masing-masing memiliki karakteristik dan asumsi yang berbeda. Naive Bayes Gaussian cocok untuk data yang memiliki distribusi Gaussian (normal), sementara Naive Bayes Multinomial cocok untuk data yang berupa angka yang mewakili jumlah kemunculan fitur dalam dokumen teks. Naive Bayes Bernoulli, di sisi lain, cocok untuk data biner yang hanya memiliki dua nilai, seperti data yang digunakan dalam filter spam. Dalam pemilihan jenis Naive Bayes yang sesuai, penting untuk mempertimbangkan distribusi data dan karakteristik dari masalah klasifikasi yang dihadapi. Selain itu, algoritma Naive Bayes juga dapat digunakan untuk menangani masalah klasifikasi yang melibatkan data yang tidak seimbang (imbalanced data). Data yang tidak seimbang terjadi ketika jumlah sampel dalam setiap kelas tidak seimbang, misalnya, dalam kasus deteksi fraud di mana jumlah transaksi yang sah jauh lebih banyak daripada transaksi yang curang. Algoritma Naive Bayes dapat memberikan hasil yang memuaskan dalam kasus ini karena menganggap bahwa distribusi kelas dalam data tidak mempengaruhi probabilitas fitur-fitur yang ada. Namun, tetap perlu dilakukan penyesuaian atau teknik khusus seperti pengaturan

ulang bobot kelas atau penggunaan metode resampling untuk memperbaiki masalah ketidakseimbangan data dalam konteks algoritma Naive Bayes. Dengan kelebihan dan kemudahan implementasinya, algoritma Naive Bayes tetap menjadi salah satu pilihan yang populer dalam berbagai aplikasi klasifikasi.

2.3.1. Uji Performa

Confusion Matrix adalah salah satu metode yang digunakan untuk mengukur performa dari sebuah model klasifikasi, termasuk model yang dibangun dengan menggunakan algoritma Naive Bayes. Confusion Matrix menggambarkan hasil klasifikasi berdasarkan empat kemungkinan hasil: true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). True positive adalah jumlah kasus positif yang terklasifikasi dengan benar oleh model, sedangkan true negative adalah jumlah kasus negatif yang terklasifikasi dengan benar. False positive adalah jumlah kasus negatif yang salah terklasifikasi sebagai positif, sementara false negative adalah jumlah kasus positif yang salah terklasifikasi sebagai negatif. Dari Confusion Matrix, berbagai metrik evaluasi performa dapat dihitung, seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran yang lebih komprehensif tentang seberapa baik model melakukan klasifikasi.

Tabel 2. 1. Kelas Atribut Confusion Matrix

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

Dimana tabel 1 berisi:

- TP (True Positive), yaitu jumlah data positif yang memiliki nilai benar.

- TN (True Negative), yaitu jumlah data negatif yang memiliki nilai benar.
- FN (False Negative), yaitu jumlah data negatif tetapi yang memiliki nilai salah.
- FP (False Positive), yaitu jumlah data yang positif tetapi yang memiliki nilai salah.

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

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

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

Selain Confusion Matrix, terdapat juga metode evaluasi performa lainnya yang dapat digunakan untuk menguji model klasifikasi, seperti Receiver Operating Characteristic (ROC) Curve dan Area Under the Curve (AUC). ROC Curve adalah kurva yang mengilustrasikan hubungan antara true positive rate (TPR) dan false positive rate (FPR) untuk berbagai nilai threshold. AUC mengukur area di bawah kurva ROC, yang memberikan gambaran tentang seberapa baik model dapat membedakan antara kelas positif dan negatif. Metode-metode ini dapat memberikan informasi yang berharga tentang performa model klasifikasi, sehingga memungkinkan untuk melakukan penyesuaian atau perbaikan model jika diperlukan.

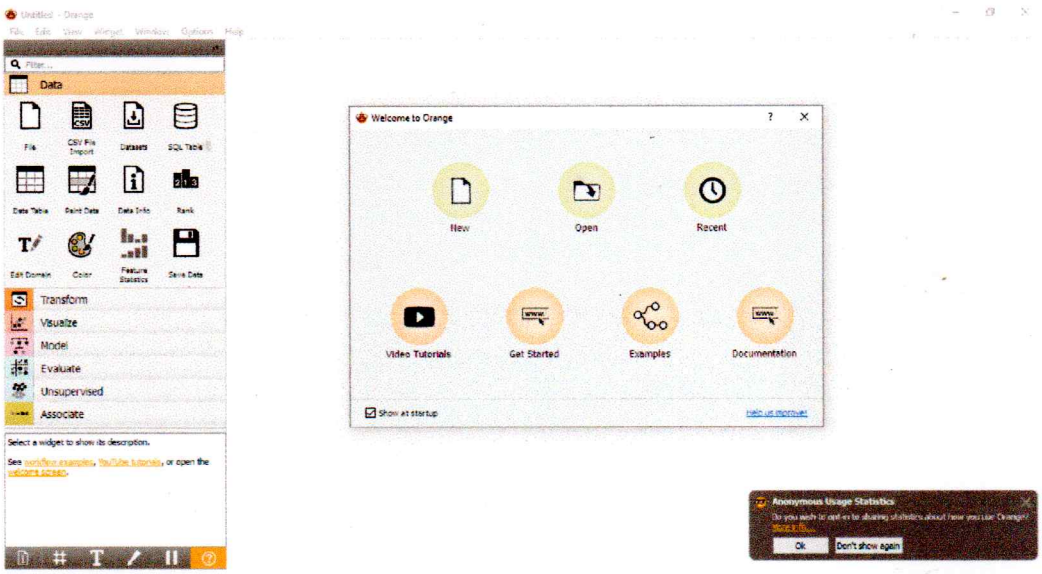
2.4. Alat Bantu Pemrograman/Tools Pendukung

2.4.1. *Aplikasi Orange*

Orange adalah salah satu perangkat lunak yang digunakan untuk analisis data dan data mining dengan antarmuka pengguna berbasis visual. Orange menyediakan berbagai alat dan fungsionalitas yang memungkinkan pengguna untuk melakukan berbagai tugas analisis data, termasuk pemrosesan data, pemodelan, visualisasi, dan

evaluasi model. Antarmuka pengguna berbasis visual memungkinkan pengguna untuk membangun alur kerja analisis data dengan cara yang intuitif, tanpa perlu menulis kode pemrograman.

Salah satu keunggulan Orange adalah keberagaman fungsionalitasnya yang mencakup berbagai teknik analisis data, mulai dari analisis statistik dasar hingga pembelajaran mesin yang lebih kompleks. Orange juga dilengkapi dengan berbagai algoritma pembelajaran mesin yang populer, seperti decision tree, regresi, clustering, dan lain-lain. Selain itu, Orange juga menyediakan berbagai pilihan visualisasi data yang menarik dan informatif, seperti scatter plot, bar chart, dan heatmap, yang memudahkan pengguna untuk memahami pola-pola dalam data mereka.



Gambar 2. 2. Tampilan Utama Aplikasi Orange

2.5. Metodologi Penelitian

Metodologi penelitian ini didasarkan pada penggunaan metode algoritma Naive Bayes dalam menganalisis minat siswa terhadap jurusan TKJ di SMKS Al-

Washliyah 2 Merbau. Langkah pertama dalam metodologi ini adalah pengumpulan data, yang meliputi data historis pendaftaran siswa, survei siswa, wawancara dengan guru dan orang tua, serta data sekolah yang relevan. Data yang terkumpul kemudian akan disiapkan untuk analisis dengan melakukan pre-processing, seperti membersihkan data dari nilai yang hilang atau tidak valid, dan mengubah data menjadi format yang sesuai untuk analisis menggunakan algoritma Naive Bayes. Setelah data siap, langkah selanjutnya adalah melakukan analisis menggunakan algoritma Naive Bayes. Algoritma ini akan digunakan untuk mengidentifikasi pola-pola signifikan dari data yang dapat mempengaruhi minat siswa terhadap jurusan TKJ. Hasil analisis ini akan memberikan pemahaman yang lebih baik tentang faktor-faktor yang memengaruhi minat siswa, seperti latar belakang pendidikan, minat pribadi, dan persepsi terhadap prospek karier di bidang teknologi informasi.

Setelah hasil analisis didapatkan, langkah berikutnya adalah interpretasi dan evaluasi hasil. Hasil analisis akan dievaluasi untuk memastikan keakuratannya dan relevansinya dalam konteks aplikasi. Interpretasi dilakukan untuk memahami implikasi hasil analisis dalam pengembangan strategi pemasaran dan rekrutmen untuk menarik lebih banyak siswa ke jurusan TKJ. Hasil akhir dari penelitian ini akan disajikan dalam bentuk laporan yang mencakup temuan-temuan utama, rekomendasi untuk pengembangan kurikulum dan fasilitas pendidikan, serta saran untuk penelitian lebih lanjut dalam bidang ini. Dengan menggunakan metodologi ini, diharapkan penelitian ini dapat memberikan kontribusi yang berharga dalam pengembangan pendidikan kejuruan di SMKS Al-Washliyah 2 Merbau. Selain itu, dalam metodologi ini juga akan dilakukan uji performa terhadap model klasifikasi

yang dibangun menggunakan algoritma Naive Bayes. Uji performa ini dilakukan dengan menggunakan Confusion Matrix serta metrik evaluasi lainnya seperti akurasi, presisi, recall, dan F1-score. Hasil uji performa ini akan memberikan gambaran yang lebih jelas tentang seberapa baik model klasifikasi dapat memprediksi minat siswa terhadap jurusan TKJ. Dengan demikian, uji performa ini menjadi langkah penting untuk memvalidasi keefektifan model yang dibangun dan memastikan bahwa model tersebut dapat memberikan hasil yang dapat diandalkan dalam membantu pengambilan keputusan di SMKS Al-Washliyah 2 Merbau.

2.5.1. Penelitian Terdahulu

Peneliti terdahulu dalam konteks penelitian ini mungkin telah melakukan studi tentang faktor-faktor yang mempengaruhi minat siswa terhadap jurusan teknik komputer dan jaringan (TKJ) di sekolah menengah kejuruan. Mereka mungkin juga telah mengidentifikasi pola-pola yang signifikan dalam data siswa yang dapat memengaruhi pilihan mereka terhadap jurusan TKJ. Penelitian terdahulu ini mungkin juga telah menggunakan metode-metode analisis data yang relevan, termasuk mungkin menggunakan algoritma Naive Bayes, yang akan mendukung validitas dan relevansi metodologi yang digunakan dalam penelitian ini. Dengan memahami temuan dan metodologi peneliti terdahulu, penelitian ini dapat memperluas pemahaman tentang faktor-faktor yang mempengaruhi minat siswa terhadap jurusan TKJ dan memberikan kontribusi yang berharga dalam pengembangan strategi rekrutmen dan pengajaran di SMKS Al-Washliyah 2 Merbau.

Tabel 2. 2. Penelitian Terdahulu

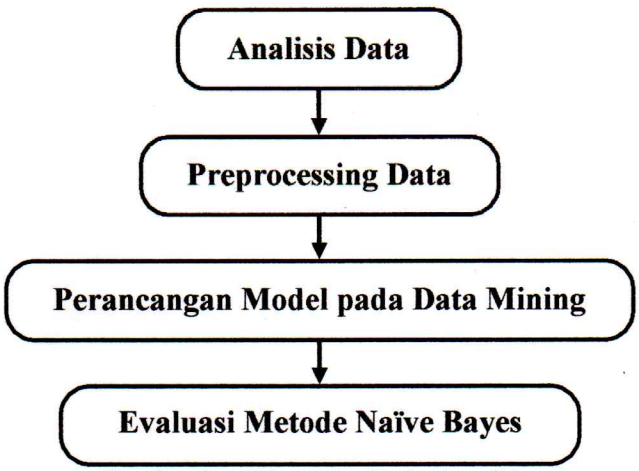
Referensi Penelitian	Penelitian 1
Judul	Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes
Nama	Alam1)*, Divi Adiffia Freza Alana. 2) , Christina Juliane3
Tahun	2023
Hasil	Diabetes mellitus adalah masalah kesehatan global yang mendesak dengan dampak besar di seluruh dunia, ditandai oleh tingginya kadar glukosa dalam darah karena gangguan produksi atau penggunaan insulin. Penelitian ini bertujuan untuk mendeteksi diabetes secara dini dengan akurasi tinggi agar pengobatan bisa segera dilakukan, mengurangi risiko kematian, dan membandingkan dua algoritma dengan tingkat akurasi terbaik. Algoritma yang digunakan adalah C4.5 Decision Tree dan Naïve Bayes. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes dapat digunakan untuk model deteksi dini diabetes dengan rata-rata akurasi 90,745%. Determinasi atribut, jumlah dimensi atribut, dan jumlah sampel sangat mempengaruhi kinerja model yang dibangun [5].
Referensi Penelitian	Penelitian 2

Judul	Performance of Various Naïve Bayes Using GridSearch Approach In Phishing Email Dataset
Nama	Rizki Rahman1)* , Ferian Fauzi Abdulloh2)
Tahun	2023
Hasil	Penelitian ini membandingkan kinerja empat varian Naive Bayes dalam mengklasifikasikan email phishing. Metode melibatkan pra-pemrosesan data, di mana email dikumpulkan, dibersihkan, dan dikonversi menjadi fitur numerik. GridSearch digunakan untuk mencari parameter terbaik. Hasil menunjukkan Bernoulli memiliki akurasi terbaik 97,34%, dan setelah tuning parameter, Bernoulli mencapai 97,38%. Studi ini memberikan wawasan tentang efektivitas Naive Bayes dalam deteksi phishing, serta panduan untuk memilih varian terbaik [6].
Referensi Penelitian	Penelitian 3
Judul	Implementation of the Naïve Bayes Method to Determine Student Interest in Gaming Laptops
Nama	Rico Fadly Nasution1)*, Muhammad Halmi Dar2) , Fitri Aini Nasution3)
Tahun	2023
Hasil	Dalam penelitian ini, penulis membahas penggunaan metode Naïve Bayes untuk mengklasifikasikan minat

	mahasiswa terhadap laptop gaming. Sebanyak 100 data mahasiswa digunakan dalam penelitian ini. Hasil klasifikasi menunjukkan bahwa 55 mahasiswa (55% representasi) tertarik dengan laptop gaming, sedangkan 45 mahasiswa (45% representasi) tidak tertarik. Meskipun laptop gaming memiliki desain dan spesifikasi yang bagus, hasil menunjukkan bahwa tidak semua mahasiswa tertarik dengan laptop gaming [7].
Referensi Penelitian	Penelitian 4
Judul	Digital Forensic Investigates Sexual Harassment on Telegram using Naïve Bayes
Nama	Meyti Eka Apriyani1)*, Rahmad Alfian Maskuri2), Muhammad Hasyim Ratsanjani3), Agung Nugroho Pramudhita4), Rawansyah5)
Tahun	2023
Hasil	Penelitian ini mengkaji penggunaan metode Naïve Bayes dalam menganalisis percakapan di Telegram yang berisi pelecehan seksual. Metode ini menggunakan metodologi National Institute of Justice (NIJ) untuk melakukan analisis forensik dan mengklasifikasikan percakapan. Dalam pengujian, algoritma Naïve Bayes mampu mengklasifikasikan

	percakapan dengan akurasi 91.3%, Presisi 100%, dan Recall 90% [8].

2.6. Kerangka Penelitian



Gambar 2. 3. Kerangka Kerja Penelitian

1. Analisis Data

Analisis data adalah proses penguraian, penyelidikan, transformasi, dan pemodelan data untuk mendapatkan informasi yang berguna, mendukung pengambilan keputusan, dan mengidentifikasi pola atau hubungan yang tersembunyi dalam data. Tujuan utamanya adalah untuk mendapatkan pemahaman yang lebih baik tentang data yang ada, mengidentifikasi tren atau pola yang dapat digunakan untuk membuat keputusan yang lebih baik di masa depan, serta menguji hipotesis atau mengonfirmasi asumsi. Analisis data dapat melibatkan berbagai teknik, mulai dari statistik deskriptif sederhana hingga analisis inferensial yang kompleks, dan sering menggunakan perangkat lunak dan alat khusus untuk membantu dalam prosesnya.

2. *Preprocessing Data*

Preprocessing data adalah langkah awal dalam analisis data yang bertujuan untuk membersihkan, mentransformasi, dan mempersiapkan data mentah agar siap untuk analisis lebih lanjut. Langkah-langkah dalam preprocessing data dapat mencakup penghapusan data yang tidak lengkap atau tidak relevan, penanganan nilai yang hilang, normalisasi data untuk menghilangkan perbedaan skala, dan konversi data ke format yang lebih sesuai untuk analisis, seperti mengubah data kategorikal menjadi bentuk numerik. Preprocessing data yang baik dapat meningkatkan kualitas analisis dan memastikan bahwa hasil yang dihasilkan dari analisis tersebut dapat diandalkan.

3. *Perancangan Model pada Data Mining*

Perancangan model dalam data mining melibatkan langkah-langkah yang sistematis untuk mengembangkan dan mengevaluasi model yang dapat mengungkap pola atau informasi berharga dari data. Proses ini mencakup pemilihan teknik atau algoritma yang sesuai dengan tujuan analisis, pemrosesan data yang tepat seperti pemilihan atribut atau fitur, pemisahan data menjadi set pelatihan dan pengujian, serta penyesuaian parameter untuk meningkatkan kinerja model. Evaluasi model dilakukan dengan menguji model pada data yang belum pernah dilihat sebelumnya untuk mengukur keakuratannya dan memastikan keberlanjutannya dalam mengidentifikasi pola yang relevan dalam data.

4. *Evaluasi Metode Naive Bayes*

Evaluasi metode Naive Bayes dilakukan dengan menggunakan berbagai metrik kinerja seperti akurasi, presisi, recall, dan F1-score. Metode ini juga dapat

dievaluasi melalui kurva ROC (Receiver Operating Characteristic) dan area di bawah kurva (AUC) untuk klasifikasi biner. Selain itu, dalam konteks klasifikasi multikelas, evaluasi dilakukan dengan menggunakan matriks konfusi untuk memahami seberapa baik model dapat mengklasifikasikan setiap kelas. Evaluasi juga mencakup analisis terhadap asumsi dasar Naive Bayes, yaitu asumsi independensi fitur, untuk memastikan bahwa asumsi tersebut memang berlaku pada data yang digunakan.