BAB II
LANDASAN TEORI

2.1 Data Mining

Data mining adalah suatu proses pengumpulan dan pengolahan data menggunakan teknik atau metode tertentu yang bertujuan mengekstraksi atau mengambil suatu informasi dari data tersebut dengan memperhatikan pola dan hubungan antar data dalam jumlah besar (Sinaga, et al., 2021). Menurut Jassim dan Abdulwahid (2021), data mining adalah kumpulan metode yang diterapkan pada basis data yang besar dan kompleks untuk menemukan pola tersembunyi dalam data tersebut. Menurut Charles Zai (2022), data mining merupakan sebuah proses perhitungan semi otomatik yang digunakan untuk mengekstraksi dan menggali informasi dalam suatu data besar dengan menggunakan teknik statistik, kecerdasan buatan, dan machine learning.

Data mining dalam pengembangannya, menurut Ginantara dkk (2021) memiliki siklus yang dapat dibagi menjadi enam tahapan yaitu sebagai berikut.

1. Business understanding adalah tahap untuk memahami tujuan serta menganalisis kebutuhan informasi yang diperlukan.
2. Data understanding merupakan tahap proses pengumpulan data yang bertujuan untuk mengidentifikasi data dan menemukan pola tersembunyi yang ada pada data.
3. Data preparation adalah tahap membangun dan mempersiapkan dataset. Hal tersebut meliputi pemilihan atribut dan pembersihan data yang tidak diperlukan.
4. Modeling adalah tahap pemilihan metode yang tepat untuk digunakan pada proses penggalian informasi.
5. Evaluation merupakan proses evaluasi untuk mengetahui keefektifan model yang digunakan.

6. Deployment adalah penyajian hasil yang didapatkan dari proses pengambilan ke dalam bentuk yang sederhana dan mudah dipahami oleh user/pengguna.(Ginantara et al., 2021)

Data yang diperoleh dari proses pengambilan data berpotensi belum memadai untuk diolah secara langsung yang dapat disebabkan oleh dataset tersebut belum lengkap maupun terdapat noisy data sehingga diperlukan preprocessing data. Ada beberapa tahap penting dalam preprocessing data yang disebutkan Randi Farmana Putra dkk (2023) dalam buku yaitu sebagai berikut :

1. *Data Cleaning* (Pembersihan Data)

   Proses ini difokuskan kepada perbaikan kesalahan dengan mengisi atau melengkapi nilai yang hilang (missing value) atau menghapus hal-hal yang tidak bisa diperbaiki dan dilengkapi. Jika permasalahan yang ditemui adalah data yang tidak konsisten, maka strategi data cleaning yang dapat dilakukan adalah standarisasi pemformatan, yaitu melakukan standarisasi variable dan memastikan konsistensi format data, ejaan, dan singkatan.

2. *Data Integration* (Integrasi Data)

   Pada tahap ini dilakukan penggabungan atau penyatuan data yang memiliki tipe atau sumber yang berbeda menjadi sekelompok dataset. Proses integrasi data dapat dilakukan setelah data cleaning. Namun, dalam perspektif dan pertimbangan lain, tahap ini disarankan untuk dilakukan sebelum tahap data cleaning.

3. *Data Transformation* (Transformasi Data)

   Pada proses ini dilakukan pengubahan struktur data, tipe data, format berkas, dan skema metadata. Hal ini dilakukan untuk menyesuaikan kebutuhan analisis atau pemodelan maupun pembelajaran mesin.

4. *Data Reduction* (Reduksi Data)

   Pada tahap ini akan dilakukan pengurangan volume atau variable atau kompleksitas data dengan tetap mempertahankan informasi penting. Data akan lebih terfokus kepada pokok permasalahan sehingga lebih mudah dalam pengambilan informasi.

2.2 Klasifikasi

Klasifikasi adalah salah satu model pembelajaran dalam data mining dimana setiap data diberi label kelas. Kumpulan data tersebut akan dibagi menjadi data latih data uji dengan tujuan memprediksi kelas baru yang akan muncul (Salmi & Rustam, 2019). Dalam data mining terdapat beberapa yaitu ekstraksi pola, *clustering*, dan klasifikasi. Klasifikasi dapat dilakukan dengan *clustering*. Perbedaan dari kedua Teknik terdebut adalah klasifikasi membutuhkan data yang sudah diberi label kelas (Purba, 2012). Menurut Putri dan Wijayanto (2022), klasifikasi merupakan proses untuk menemukan fungsi yang dapat menjelaskan dan membedakan kelas data. Klasifikasi adalah proses evaluasi data yang bertujuan untuk mengkategorikan data ke dalam salah satu dari beberapa kelas yang telah ditentukan. Tujuan dari metode klasifikasi adalah memaham sebuah dataset sehingga kita dapat membuat aturan atau model yang mampu mengklasifikasikan data baru yang belum pernah kita lihat sebelumnya.

Klasifikasi juga dapat dijelaskan sebagai proses untuk mengidentifikasi suatu objek data sebagai anggota dari salah satu kategori yang telah didefinisikan sebelumnya (Rahayu, et al., 2024. Klasifikasi termasuk dalam supervised learning, yang membutuhkan sampel data beserta hasilnya untuk mencari hubungan antar data dengan tujuan meningkatkan akurasi keberhasilan yang akan diperoleh (Srirahayu & Pribadie, 2023). Ada beberapa algoritma dalam klasifikasi yang sering

digunakan yaitu Artificial Neural Network, SVM, Native Bayes, dan Decision Tree (Prasetyo, 2012).

2.3 Naïve Bayes

Naïve Bayes merupakan suatu metode pengklasifikasian dalam data mining dengan menggunakan probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dataset yang diberikan (Alkhairi et al., 2021). Dalam algoritma naïve bayes, nilai atribut dari data diasumsikan bersifat saling bebas atau independen secara kondisional dan disertai dengan nilai output (Putri & Wijayanto, 2022). Data yang dihasilkan dari perhitungan naïve bayes adalah data yang terklasifikasi dengan metode statistik dan probabilitas untuk dapat memprediksikan masa datang berdasarkan dengan data masa lalu (Irmayani, 2021). Algoritma naïve bayes sering digunakan karena memiliki metode yang sederhana namun memiliki hasil dengan tingkat akurasi yang tinggi (A'yiniyah, et al., 2022). Keuntungan menggunakan metode naïve bayes adalah bahwa metode ini tidak membutuhkan jumlah data training yang sangat besar untuk proses training sistem.

Naïve bayes sendiri adalah teknik yang berdasar dari penerapan teorema bayes dengan asumsi independensi yang kuat (Salmi & Rustam, 2019). Teorema bayes adalah metode yang digunakan untuk menghitung ketidakpastian data dari bagian data tertentu dengan membandingkan dua kumpulan data (Zulham, et al., 2023). Teorema Bayes dirumuskan dengan formula sebagai berikut :

$$P(X \mid H) = \frac{P(X \mid H) \, . \, P(H)}{P(X)} \qquad (2.1)$$

Dimana :

X : Data dengan class yang belum diketahuiH : Hipotesis data X

P(H | X) : Probabilitas hipotesis H berdasarkan kondisi X

P(X | H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(H) : Probabilitas hipotesis H

P(X) : Probabilitas X

Nilai X (posterior density) adalah data testing yang belum diketahui kelasnya. Nilai H adalah hipotesis data X yang merupakan suatu kelas yang lebih spesifik. Nilai P(X | H) adalah probabilitas hipotesis X berdasarkan kondisi H yang disebut likelihood. Nilai P(H) disebut juga dengan prior probability adalah probabilitas hipotesis H. Nilai P(X) disebut juga dengan predict prior probability adalah probabilitas X.

Klasifikasi dalam prosesnya memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis. Oleh karena itu, untuk menjelaskan teorema naïve bayes diperlukan penyesuaian teorema bayes seperti persamaan 2.2 berikut:

$$P(C \mid F_1 \ldots F_n = \frac{P(C)P(F_1 \ldots F_n \mid C)}{P(F_1 \ldots F_n)} \quad (2.2)$$

Dimana variabel C adalah kelas dan variable $F_1 \ldots F_n$ adalah karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka, rumus diatas menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (posterior) adalah peluang munculnya kelas C (sebelum masuk sampel disebut prior) dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (likelihood), lalu dibagi dengan peluang kemunculan karakteristik sampel secara global (evidence). Selanjutnya, penjabaran $P(C \mid F_1 \ldots F_n)$ dapat disederhanakan menjadi persamaan 3 dan 4 sebagai berikut :

$$P(C \mid F_1 \ldots F_n = P(C)\, P(F_1 \mid C)\, P(F_2 \mid C) \ldots \quad (2.3)$$

$$P(C \mid F_1 \ldots F_n = P(C) \prod_{i=1}^{n} P(F_1 \mid C) \quad (2.4)$$

Persamaan diatas merupakan model teorema naïve bayes yang selanjutnya digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinu digunakan rumus Densitas Gauss sebagai berikut:

$$P(X_i = x_i \mid Y = y_i) = \frac{1}{2\pi\sigma_{ij}{}^2} e\left(-\frac{(x_i-\mu_i)^2}{2\pi\sigma_{ij}{}^2\sigma\sigma}\right) \quad (2.5)$$

Keterangan:

P : Peluang

Xi : atribut ke i

xi : nilai atribut ke i

Y : kelas yang dicari

yi : subkelas Y yang dicari

π : mean, rata-rata dari seluruh atribut

σ : standar deviasi, menyatakan variasi dari seluruh atribut.

2.4 Review Artikel

Tabel 2. 1 Review Artikel

No.	Landasan Literatur	Metode	Masalah	Hasil Penelitian
1.	"Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes". Agus Fajar Riany dan Gusmelia Testiana (2023).	Algoritma Naïve Bayes	Masalah penyakit stroke perlu perhatian yang lebih serius karena jumlah kasus yang meningkat dan mempunyai angka kematian yang tinggi.	klasifikasi pasien penderita stroke mengggunakan algoritma naïve bayes berhasil dan mendapat hasil akurasi yang tinggi
2.	"Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di	Algoritma Naïve Bayes	Meningkatnya kasus penyebaran penyakit COVID-19 membuat Indonesia	Data uji dari provinsi Aceh dikatakan bahwa Tingkat penyebaran COVID-19 dengan kasus

	Indonesia". Alvina Felicia Watratan dkk. (2020)		dalam status waspada terhadap ancaman virus corona.	pasien yang terkena positif lebih rendah.
3.	"Prediksi Stunting Pada Balita di Rumah Sakit Kota Semarang Menggunakan Naïve Bayes". Widya Cholid Wahyudin dkk. (2023)	Algoritma Naïve Bayes	Berdasarkan hasil Riset Kesehatan Dasar (Riskesdas) mencatat bahwa kasus stunting semakin meningkat dan dapat memberikan dampak jangka pendek dan panjang.	Dari 300 data klasifikasi, didapatkan hasil 252 data diprediksi mengalami stunting.
4.	"Klasifikasi Penyakit Hati dengan Menggunakan Metode Naïve Bayes". Wahyugi Fadri (2023)	Algoritma Naïve Bayes	Penyakit liver dapat disebabkan oleh gaya hidup yang tidak sehat. Jika tidak segera ditangani, penyakit ini	Akurasi tinggi dari penggunaan algoritma naïve bayes mempermudah diagnosis penyakit hati sehingga pasien dapat

			dapat mencapai stadium tinggi yang semakin berbahaya.	ditangani dengan segera.
5.	"Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik". Qurotul A'yuniyah dkk. (2022)	Algoritma Naïve Bayes	Dengan tingginya kasus penyakit ginjal kronik, dibutuhkan klasifikasi untuk memprediksi PGK terhadap gejala yang menuju kepada indicator terkena PGK.	Klasifikasi penyakit ginjal kronik berhasil dilakukan dengan Tingkat akurasi tinggi sehingga mudah untuk mendiagnosis PGK.