BAB II

LANDASAN TEORI

2.1 DATA

Data adalah catatan atas kumpulan fakta[1].Data juga merupakan elemen fundamental dalam analisis penjualan, khususnya dalam penerapan metode Naïve Bayes. Menurut [2], pengolahan data yang efektif dan efisien diperlukan untuk memprediksi hasil penjualan dan menentukan strategi pemasaran yang tepat.

Dalam penelitian ini, data yang digunakan berasal dari laporan penjualan Toko Bolen One Cake Rantau Prapat. Data tersebut mencakup informasi seperti tanggal penjualan, jenis produk, jumlah produk yang terjual, harga per unit, dan total pendapatan harian. Informasi ini esensial untuk memahami tren penjualan dan perilaku konsumen.

Sebelum analisis dilakukan, data akan melalui tahap pra-pemrosesan untuk memastikan kualitas dan konsistensinya. Langkah-langkah ini meliputi penanganan data yang hilang, normalisasi data, dan transformasi data ke dalam format yang sesuai untuk analisis lebih lanjut. Sebagaimana disampaikan oleh [3], pra-pemrosesan data adalah langkah krusial untuk meningkatkan akurasi dalam klasifikasi data penjualan.

Setelah tahap pra-pemrosesan, data akan dianalisis menggunakan metode Naïve Bayes. Metode ini dipilih karena kesederhanaannya dan kemampuannya dalam memberikan akurasi yang tinggi dalam klasifikasi data. Sebagaimana dinyatakan oleh [4], Naïve Bayes efektif dalam mengklasifikasikan data penjualan untuk menentukan strategi pemasaran yang optimal.

Selain itu, data penjualan akan dieksplorasi untuk mengidentifikasi pola musiman atau tren tertentu. Misalnya, apakah terdapat peningkatan penjualan pada periode tertentu atau produk tertentu yang lebih diminati oleh konsumen. Analisis semacam ini penting untuk memahami dinamika penjualan dan perilaku konsumen.

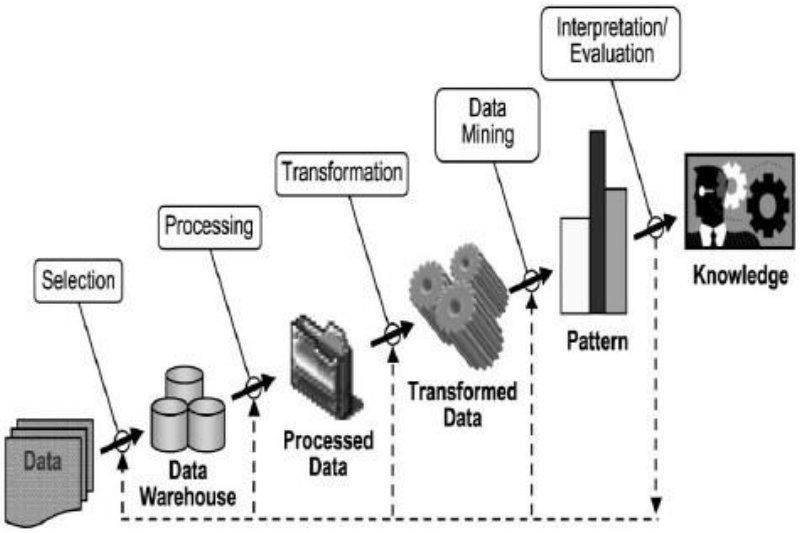
Selanjutnya, data akan dibagi menjadi data pelatihan dan data pengujian untuk membangun dan menguji model Naïve Bayes. Pembagian ini penting untuk memastikan bahwa model yang dibangun memiliki kemampuan generalisasi yang baik dan dapat memprediksi data baru dengan akurat.

Dengan data yang telah dipersiapkan dan dianalisis dengan baik, diharapkan penelitian ini dapat memberikan wawasan yang berharga bagi Toko Bolen One Cake Rantau Prapat dalam memahami tren penjualan dan perilaku konsumen, serta membantu dalam pengambilan keputusan strategis yang lebih efektif.

2.2 Knowledge Discovery in Database (KDD)

Knowledge Discovery in Databases (KDD) adalah proses yang digunakan untuk mengidentifikasi pola atau pengetahuan yang berguna dari kumpulan data yang besar. KDD merupakan sebutan dari Data Mining[5]. Proses KDD ini melibatkan beberapa tahapan, termasuk pembersihan data, integrasi data, seleksi data, transformasi data, data mining, evaluasi pola, dan presentasi pengetahuan. Menurut [6], KDD digunakan untuk secara otomatis mengeksplorasi dan menganalisis sumber data yang besar.yang bertujuan untuk menemukan pola

dalam sejumlah data besar dengan tujuan untuk melakukan klasifikasi, estimasi, prediksi, asosiasi, dan klaster. Penambangan data termasuk sebagai Knowledge Discovery in Database (KDD)[7].



Gambar 2.1 Proses *Knowledge Discovery in Database (KDD)*

Gambar 2.1 menunjukkan langkah-langkah dalam proses *KDD*, yang dimulai dari pengumpulan data mentah hingga menghasilkan pengetahuan yang bermanfaat. Berikut adalah tahapan yang digambarkan:

1. Selection (Seleksi Data) : Data yang relevan dipilih dari berbagai sumber untuk dianalisis.
2. Data Warehouse (Gudang Data) :Data yang telah dipilih disimpan dalam gudang data untuk pengelolaan yang lebih terstruktur.

3. Processing (Pemrosesan Data) : Data diproses untuk menghilangkan anomali dan menyiapkannya untuk analisis lebih lanjut.
4. Processed Data (Data yang Telah Diproses) : Data yang telah melalui tahap pemrosesan siap untuk ditransformasikan.
5. Transformation (Transformasi Data**)** : Data diubah atau dikonversi agar sesuai untuk proses *Data Mining*.
6. Data Mining : Teknik dan algoritma diterapkan untuk menemukan pola tersembunyi dalam data.
7. Pattern (Pola yang Ditemukan) : Hasil dari *Data Mining* berupa pola atau tren tertentu yang dapat digunakan dalam pengambilan keputusan.
8. Interpretation/Evaluation (Interpretasi dan Evaluasi) : Pola yang ditemukan dianalisis dan dievaluasi untuk memastikan keakuratan dan relevansinya.
9. Knowledge (Pengetahuan) : Hasil akhir dari proses ini berupa wawasan yang dapat diterapkan dalam berbagai bidang, seperti bisnis, kesehatan, dan penelitian.

Data mining adalah inti dari proses KDD, di mana teknik-teknik seperti klasifikasi, klastering, dan asosiasi digunakan untuk menemukan pola dalam data. Salah satu metode yang sering digunakan dalam data mining adalah algoritma Naïve Bayes, yang efektif untuk tugas klasifikasi. Setelah data mining, tahap evaluasi pola (pattern evaluation) dilakukan untuk mengidentifikasi pola yang menarik dan relevan. Terakhir, presentasi pengetahuan (knowledge presentation) menyajikan pengetahuan yang ditemukan kepada pengguna dengan cara yang

mudah dipahami. Seperti yang dijelaskan oleh Utami (2020), data mining berfungsi untuk memfasilitasi pekerjaan dengan data yang banyak dan efektif digunakan tidak hanya di lingkungan bisnis tetapi juga di berbagai bidang lainnya.

Dalam konteks analisis data penjualan pada Toko Bolen One Cake Rantau Prapat, penerapan proses KDD dapat membantu dalam mengidentifikasi pola pembelian pelanggan, tren penjualan produk, dan informasi penting lainnya yang dapat mendukung pengambilan keputusan bisnis. Dengan mengikuti tahapan KDD secara sistematis, toko dapat menggali pengetahuan berharga dari data penjualan yang ada.

Sebagai contoh, pada tahap data mining, algoritma Naïve Bayes dapat digunakan untuk mengklasifikasikan produk berdasarkan tingkat penjualan atau preferensi pelanggan. Hasil dari klasifikasi ini kemudian dievaluasi untuk memastikan akurasi dan relevansinya, sebelum akhirnya disajikan dalam bentuk laporan atau visualisasi yang informatif bagi manajemen toko.

Dengan demikian, penerapan KDD tidak hanya membantu dalam memahami data yang ada, tetapi juga memberikan wawasan strategis yang dapat digunakan untuk meningkatkan kinerja penjualan dan kepuasan pelanggan di Toko Bolen One Cake Rantau Prapat.

2.3 Data Mining

Data mining adalah proses analisis data untuk menemukan pola, tren, dan informasi berharga dari kumpulan data besar dan kompleks. Proses ini menggunakan berbagai algoritma dan teknik statistik untuk mengidentifikasi

hubungan yang sebelumnya tidak terlihat, sehingga dapat mendukung pengambilan keputusan strategis. Dalam konteks bisnis, penerapan data mining telah terbukti meningkatkan efisiensi operasional dan memberikan wawasan yang mendalam [8].menurut Yuli Mardi Data mining merupakan suatu cara untuk menemukan pola atau informasi yang menarik dalam kumpulan data terpilih dengan menerapkan teknik atau pendekatan tertentu. Berbagai teknik, pendekatan, atau algoritma yang digunakan dalam data mining sangat beragam.

Pemilihan pendekatan atau algoritma yang sesuai sangat dipengaruhi oleh tujuan dan keseluruhan proses Knowledge Discovery in Database (KDD).. Sedangkan Data mining menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT adalah analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut[9]. Proses data mining mengkombinasikan tahapan KDD dengan Naïve Bayes, mulai dari seleksi data, preprocessing, hingga evaluasi[10].

Proses data mining terdiri dari beberapa tahapan utama:

1. Persiapan Data

   Tahapan ini melibatkan pengumpulan, pembersihan, dan transformasi data untuk memastikan kualitas data yang optimal [11]. Proses ini mencakup penghapusan data duplikat, penanganan nilai hilang, dan normalisasi data.

2. Eksplorasi Data

   Pada tahap ini, data dianalisis untuk memahami karakteristiknya, seperti distribusi, korelasi antar-atribut, dan deteksi anomali [12].

3. Pemodelan

   Teknik seperti clustering, classification, dan regression digunakan untuk menemukan pola dalam data. Metode seperti K-Means dan algoritma Apriori sering digunakan untuk menganalisis data pelanggan dan pola pembelian [13].

4. Evaluasi

   Model yang telah dibuat dievaluasi untuk memastikan keakuratan dan relevansi hasilnya. Teknik validasi silang dan pengujian data digunakan untuk mengevaluasi kinerja model [14].

5. Implementasi

   Model yang telah divalidasi diterapkan pada data baru untuk menghasilkan prediksi atau rekomendasi. Proses ini memungkinkan perusahaan untuk mengambil keputusan berbasis data secara lebih cepat dan akurat [15].

tabel 2. 1 Metode Populer dalam Data Mining

Metode	**Deskripsi**	**Contoh Algoritma**	**Referensi**
Clustering	Mengelompokkan data ke dalam segmen berdasarkan kesamaan karakteristik.	K-Means, DBSCAN	[13]
Association	Menemukan hubungan antar item dalam dataset.	Apriori, FP-Growth	[13]

Classification	Memprediksi label untuk data baru berdasarkan data historis.	Decision Tree, SVM	[16]

2.4 Algoritma Naïve Bayes

Algoritma Naïve Bayes adalah metode klasifikasi probabilistik yang didasarkan pada Teorema Bayes, dengan asumsi independensi antar fitur yang diberikan kelas..Pendekatan ini memungkinkan algoritma untuk menghitung probabilitas suatu kelas berdasarkan pola kemunculan fitur dalam dataset[17]. Naïve Bayes adalah sebuah teknik dalam pembelajaran mesin yang berlandaskan pada perhitungan kemungkinan.

Metode ini menggunakan prinsip-prinsip probabilitas dan statistik yang diperkenalkan oleh ilmuwan asal Inggris, Thomas Bayes, dengan tujuan untuk meramalkan kemungkinan di masa depan berdasarkan pengalaman yang telah ada sebelumnya.[18].Keunggulan utama Naïve Bayes adalah kemampuannya dalam menangani dataset dengan dimensi tinggi dan efisiensi komputasi yang tinggi. Oleh karena itu, algoritma ini sering digunakan dalam berbagai aplikasi, seperti analisis sentimen, prediksi pasar, dan diagnosis medis [19]. Model ini juga menghasilkan prediksi kategori Laris dan Kurang Laris yang cukupakurat[20].

2.4.1 Confusion Matrix

Confusion matrix merupakan tabel yang memudahkan melihat hasil kerja dari suatu algoritma klasifikasi, dengan cara membandingkan prediksi yang dihasilkan model terhadap kategori yang sebenarnya.dipenelitian ini diketahui bahwa Confusion Matrix sangat berperan penting dalam penelitian ini,yang mana Confusion Matrix ini dapat memperlihatkan produk yang terlaris dan yang tidak larisnya,serta dapat mempermudah peneliti mengklasifikasikannya.adapun struktur confusion metrix yaitu :

1. True Positive (TP): Data positif yang berhasil diprediksi positif.
2. False Positive (FP): Data negatif yang salah diprediksi positif.
3. False Negative (FN): Data positif yang salah diprediksi negatif.
4. True Negative (TN): Data negatif yang berhasil diprediksi negatif.

Adapun rumus confusion matrix ada 4 yaitu Accurasy,Precision,Recall dan F1- score untuk rumusnya ada di bawah ini.

adapun rumus yang dapat digunakan untuk mencari sebuah *accuracy* diantaranya seperti dibawah ini:

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

Adapun rumus yang digunakan untuk mencari sebuah *Precision* diantaranya seperti dibawah ini:

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}}$$

Adapun rumus yang digunakan untuk mencari sebuah *Recall* diantaranya seperti dibawah ini :

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

Adapun rumus yang digunakan untuk mencari sebuah *F1-Score* diantaranya seperti dibawah ini:

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.4.2 Prinsip Dasar Algoritma

Algoritma Naïve Bayes menghitung probabilitas gabungan fitur sebagai produk dari probabilitas individualnya berdasarkan asumsi independensi.dibawah ini merupakan rumus pada perhitungan naïve bayes .

$$P(C\backslash X) = \frac{P(X\backslash C).P(C)}{P(X)}$$

di mana P(C|X)P(C|X)P(C|X) adalah probabilitas kelas C yang diberikan fitur X, P(X|C) adalah probabilitas fitur X dalam kelas C, dan P(C) adalah probabilitas prior kelas C ([21]).

Tahapan Implementasi

Proses implementasi Naïve Bayes meliputi:

1. Pemmrosesan data

   Dataset dipersiapkan melalui pembersihan, tokenisasi (untuk teks), atau normalisasi (untuk data numerik).

2. Perhitungan Probabilitas Prior

   Probabilitas masing-masing kelas dihitung berdasarkan frekuensi kemunculannya.

3. Perhitungan Probabilitas Kondisional

   Setiap fitur dihitung probabilitasnya untuk setiap kelas.

4. Prediksi

   Probabilitas setiap kelas untuk data baru dihitung, dan kelas dengan probabilitas tertinggi dipilih sebagai hasil prediksi.

Keunggulan dan Kelemahan

Keunggulan:

1. Efisiensi komputasi untuk dataset besar.
2. Kinerja baik pada data berskala besar meskipun asumsi independensi tidak sepenuhnya terpenuhi [22]

Kelemahan:

1. Sensitif terhadap ketidakseimbangan data.
2. Kurang optimal jika fitur saling bergantung [23].

Studi Kasus

Algoritma Naïve Bayes telah diterapkan dalam berbagai studi. Contohnya adalah analisis sentimen ulasan aplikasi e-government di Google Play Store, di mana algoritma ini mampu mengklasifikasikan sentimen dengan akurasi yang

memadai [24]. Studi lainnya memanfaatkan Naïve Bayes untuk memprediksi penjualan produk di sektor ritel, menunjukkan efisiensi algoritma ini dalam menangani data berskala besar [25].

2.5 Orange

Orange adalah perangkat lunak open-source untuk analisis data dan machine learning yang dirancang untuk kemudahan penggunaan melalui antarmuka visual berbasis drag-and-drop. Orange sangat populer di kalangan akademisi, peneliti, dan mahasiswa karena memungkinkan pemrosesan data tanpa harus menulis kode pemrograman. Orange mendukung berbagai metode seperti klasifikasi, regresi, clustering, dan visualisasi data interaktif.

Menurut dokumentasi resmi Orange, aplikasi ini memiliki modul-modul visual yang disebut *widget*, yang bisa dihubungkan untuk membentuk alur kerja analisis data. Hal ini memudahkan pengguna dalam merancang eksperimen analisis data, termasuk dalam penggunaan algoritma Naive Bayes.

2.5.1 Fitur Utama Orange

Orange memiliki sejumlah fitur utama yang menjadikannya alternatif yang sangat baik untuk perangkat lunak data mining lainnya:

1. Workflow Visual (Alur Kerja Visual) : Analisis data dibangun menggunakan blok-blok modul (widget) yang dapat dihubungkan satu sama lain.

2. Pra-pemrosesan Data: Orange menyediakan widget untuk normalisasi, seleksi atribut, dan imputasi data hilang.
3. Algoritma Machine Learning : Mendukung berbagai algoritma seperti Naive Bayes, Decision Tree, Random Forest, dan SVM.
4. Visualisasi Interaktif: Menyediakan visualisasi data berupa scatter plot, box plot, histogram, dan heatmap.
5. Integrasi Python : Bagi pengguna tingkat lanjut, Orange menyediakan integrasi Python untuk membuat widget kustom.

2.5.2 Keunggulan Orange

1. Antarmuka Intuitif: Tidak memerlukan keterampilan pemrograman, cocok untuk pemula maupun peneliti.
2. Open Source: Gratis dan dapat dikembangkan lebih lanjut sesuai kebutuhan pengguna..
3. Ringan dan Portabel: Tidak memerlukan instalasi sistem database atau server terpisah.

2.5.3 Studi Kasus Penggunaan Orange

Orange telah digunakan dalam berbagai bidang seperti pendidikan, riset medis, dan analisis bisnis. Contoh penggunaannya dalam penelitian adalah untuk:

1. Menganalisis data penjualan untuk memprediksi produk terlaris.
2. Melatih dan menguji model klasifikasi menggunakan Naive Bayes dengan dataset UMKM.

3. Visualisasi tren penjualan berdasarkan waktu, produk, atau kategori.

Dengan fitur yang lengkap dan kemudahan penggunaan, Orange merupakan solusi efektif untuk penelitian ini dalam mengklasifikasikan data penjualan pada Toko Bolen One Cake menggunakan algoritma Naive Bayes.