BAB II

KAJIAN PUSTAKA

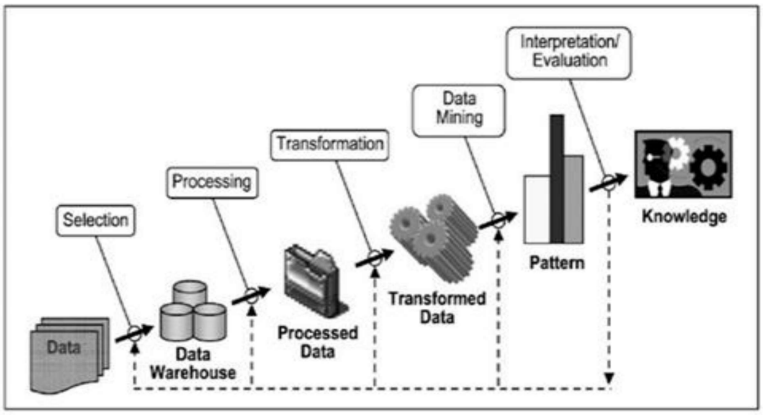
2.1 Konsep Teoritis

Penelitian ini berlandaskan teori-teori yang akan dimulai *dari Knowladge Discovery in Database* (KDD).

2.1.1 *Knowladge Discovery in Database* (KDD)

Knowladge Discovery in Database (KDD) adalah proses untuk menentukan informasi yang bermanfaat seperti berbentuk suatu data yang bersifat baru dan bisa digunakan (Gukguk & Sitohang, 2021). Data informasi yang berukuran besar berguna untuk pola-pola yang ada dalam data dan potensinya yang bermanfaat. Data *mining* merupakan salah satu langkah dari serangkaian proses iterative KDD.

Menurut (Ikhwan et al., 2015) langkah-langkah proses *Knowladge Discovery in Database* (KDD) yang dilakukan diawali pada gambar 2.2 sebagai berikut:



Gambar 2.1 *Knowladge Discovery in Database (KDD)*

1. Data *selection* (seleksi data)

   Sebelum tahap information *mining* KDD dimulai, data perlu diseleksi dari sekumpulan data operasional. Data yang dipilih akan diproses dan data disimpan dalam file.

2. *Pre-processing* / pembersihan

   Sebelum menjalankan proses data *mining*, harus melalui proses pembersihan data yang menjadi perhatian KDD. Proses pembersihan secara khusus meliputi deduplikasi, pengecekan data yang tidak konsisten, dan perbaikan kesalahan data.

3. Transformasi

   Proses transformasi untuk data yang dipilih. Hal ini membuat data cocok untuk proses data *mining*. Proses pengkodean KDD adalah proses kreatif yang bergantung pada jenis dan pola informasi yang di cari di database.

4. Data *mining*

   Proses menemukan pola atau informasi yang menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Pilihan metode dan algoritme yang benar sangat berpengaruh pada tujuan proses KDD secara semuanya.

5. interpretasi /evaluasi

   Pola informasi yang dihasilkan oleh proses data *mining* harus ditampilkan dalam format yang mudah dipahami oleh para pemangku kepentingan. Tahap ini merupakan bagian dari proses KDD dan disebut interpretasi.

2.2 Data *Mining*

2.2.1 Pengenalan Data *Mining*

Data *mining* atau penambangan data adalah proses mencari informasi penting dari data besar. Proses penambangan data selalu menggunakan teknik statistik dan matematika untuk memanfaatkan manfaat teknik kecerdasan buatan (AI) dan penggunaan data dengan hubungan yang tidak terduga (Sikumbang, 2018). Secara sederhana dapat dipahami bahwa data *mining* adalah rangkaian proses untuk mengekstraksi pola-pola menarik dari sekumpulan besar data berupa pengetahuan yang tidak diketahui buatan dan munculnya data mining karena hingga bertahun-tahun pengumpulan akumulasi data seperti data penjualan, data pembelian, data pelanggan, data transaksi, dll (Sinaga & Handoko, 2021). Tujuan dari data *mining* untuk dapat mempelajari lebih lanjut tentang perilaku data yang diamati, atau yang biasa disebut dengan deskripsi tentang apa yang terjadi, untuk dapat memprediksi situasi masa depan.

2.2.2 Proses Data *Mining*

Proses data *mining* menurut (Budiantara & Budihartanti, 2020) yang dilakukan diawali dari:

1. Description

   Identifikasi pola berulang dalam data dan pola tersebut ke dalam aturan dengan standar yang mudah dipahami. Aturan yang dihasilkan harus mudah dipahami agar dapat meningkatkan tingkat pengetahuan secara efektif dalam sistem.

2. *Prediction*

   Memprediksi penurunan pelanggan jangka pendek dan mengkategorikannya berdasarkan perilaku masa depan yang diharapkan.

3. *Estimate*

   Membangun model dengan catatan lengkap yang memberikan nilai variabel target sebagai nilai prediksi. Selain itu, pada ulasan selanjutnya, estimasi nilai variabel target didasarkan pada nilai variabel prediktor. Misalnya, kami akan memperkirakan tekanan darah sistolik pada pasien rawat inap berdasarkan usia pasien, jenis kelamin, berat badan, dan kadar natrium darah. Hubungan tekanan darah sistolik dengan nilai prediktor selama pembelajaran menghasilkan model perkiraan.

4. *Classification*

   Proses menemukan model fungsional dan menggambarkan data sebagai kelas melibatkan pemeriksaan karakteristik objek dan menetapkan objek ke salah satu kelas yang telah ditentukan.

5. *Clustering*

   Mengelompokkan data daripada berdasarkan kelas objek tertentu. Tujuannya yaitu menghasilkan pengelompokan objek dalam kelompok yang mirip satu sama lain. Semakin besar kemiripan objek dalam suatu *cluster*, maka semakin besar pula perbedaan antar setiap *cluster*, dan semakin baik kualitas analisis *cluster*.

6. *Association*

Menemukan satu kemunculan atribut dalam dunia bisnis sering disebut sebagai market basket analysis. Tugas asosiasi mencoba mengungkap aturan yang mengukur hubungan antara dua atau lebih atribut.

2.3 Metode Data *Mining*

Algoritma *Naïve Bayes* adalah cara langsung memprediksikan pola data menggunakan ramalan penjualan bulan depannya. Oleh karena itu, Naive Bayes adalah algoritma yang baik untuk meramalkan penjualan (Wijaya & Dwiasnati, 2020). Peneliti menggunakan metode algoritma *Naive Bayes* merupakan metode statistik yang berdasarkan pengenalan pola, menggunakan *Probability* dan biaya yang dihasilkan dari keputusan ini untuk menyelidiki proses klasifikasi disebut *Teorema Bayes* (Fithri, 2017).

Pengklasifikasi *Probability* dengan mudah menghitung satu set probabilitas dan kombinasi beberapa level atau hasil dari kumpulan data yang diperoleh. Kedekatan dan kecepatan Nave Bayesian dalam database data besar (Saleh, 2015). Berikut rumus 2.1 *Probability Bayes*:

$$p(x|y = \frac{p(x \cap y)}{p(y)}$$ **Rumus 2.1** *Probability Bayes*

Probability X pada Y artinya *Probability* interaksi X dan Y dari probabilitas Y. Atau disebut dengan P(X|Y) artinya persentase banya nya X didalam Y. Berikut rumus 2.2 *Teorema Bayes*:

$$\text{P (H|X)} = \frac{p(X|H)p(H)}{p(X)}$$ **Rumus 2.2** *Teorema Bayes*

Penjelasan rumus 2.2 X artinya ciri-ciri, H artinya hipotesis, P(H|X) artinya *Probability* bahwa hipotesis H benar untuk ciri X atau disebut dengan P(H|X) adalah *Probability posterior* H dengan ketentuan X, P(X|H) artinya *Probability* bahwa ciri X benar untuk hipotesis H atau *Probability posterior* X dengan ketentuan H, P(H) adalah *Probability prior* hipotesis H, dan P(X) artinya *Probability prior* bukti X. Berikut Rumus 2.3 *Probability Bayes*:

$$P(Y) = \frac{P(X \cap Y)}{P(Y)}$$ **Rumus 2.3** *Probability Bayes*

Probability X didalam Y adalah *Probability* interseksi X dan Y dari *Probability* Y, atau disebut dengan P(X|Y) artinya tingkatan banyaknya X didalam Y. Berikut rumus 2.4 *Teorema bayes*:

$$P(X) = \frac{p(H)\,p(H)}{p(X)}$$ **Rumus 2.4** *Teorema Bayes*

Penjelasan diatas X artinya ciri-ciri, H artinya hipotesis, P(H|X) artinya *Probability* bahwa hipotesis H benar untuk ciri x atau disebut dengan P(H|X) artinya *Probability posterior* H dengan ketentuan X, P(X|C) artinya *Probability*.

2.4 Software Pendukung

Dalam penelitian data mining dapat menggunakan berbagai macam software agar membantu menyokong data mining dalam proses prediksi penjualan bahan material, maka peneliti akan menggunakan software pendukung yaitu *Waikato Knowledge Analysis Environment (WEKA)* merupakan aplikasi penelitian milik

New Zealand yang telah diolehkan di *University Waikato Trust. WEKA* dapat memecahkan masalah data mining dunia nyata, terutama masalah klasifikasi dengan metode pembelajaran mesin. *WEKA* adalah software untuk memahami konsep data mining, yang menyediakan berbagai metode data mining, dimulai dengan mengeksplorasi analisis data, pembelajaran statistik, pembelajaran mesin, dan database. Tidak seperti kebanyakan perangkat lunak penambangan data, *WEKA* didasarkan pada perangkat lunak sumber terbuka, setiap orang dapat mengakses kode sumber dan menambahkan algoritmenya sendiri selama dia setuju dan menyetujui lisensi distribusi perangkat lunak (Purnamasari et al., 2013).

2.5 Penelitian Terdahulu

Penelitian ini tidak akan terwujud tanpa adanya jurnal ataupun sumber pendukung, dan beberapa topik yang berkaitan dengan judul penelitian yang di gunakan sebagai bahan referensi dalam penulisan skripsi ini adalah sebagai berikut:

1. (Nurajijah, 2019) **Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah.** e-ISSN: 2338-0403 dengan menyatakan bahwa keputusan untuk mengotorisasi pembiayaan koperasi Syariah datang dengan risiko gagal bayar yang tinggi ketika utang kredit pelanggan jatuh tempo atau dikenal sebagai kredit macet. Mempertahankan dan meminimalkan risiko ini memerlukan metode yang akurat untuk menentukan persetujuan pendanaan. Tujuan dari penelitian ini

adalah untuk mengklasifikasikan data histori pinjaman nasabah koperasi syariah menggunakan algoritma Naive Bayes, pohon keputusan dan SVM untuk memprediksi keandalan calon nasabah selanjutnya. Hasil penelitian menunjukkan keakuratan algoritma Naive Bayes.

2. (Wijaya & Dwiasnati, 2020) **Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat.** ISSN: 2355-6579 menyatakan bahwa tujuan penelitian ini untuk Menganalisis masalah penentuan produk vitamin mana yang dapat atau tidak dapat dijual berdasarkan kategori dapat menjadi panduan bagi apotek untuk menentukan berapa banyak persediaan yang harus mereka miliki di gudang farmasi. Informasi yang diharapkan dari penelitian ini adalah dengan menggunakan algoritma klasifikasi data mining yaitu algoritma Naive Bayes, untuk mendapatkan nilai akurasi data penjualan obat khususnya untuk jenis vitamin yang sering dipilih oleh pelanggan yang membutuhkan obat.
3. (Mustafa et al., 2017) **Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier**. ISSN: 2460-4259 menyatakan bahwa penelitian ini mengevaluasi kinerja akademik mahasiswa STMIK Dipanegara Makassar selama dua tahun pertama menggunakan teknik data mining algoritma *Naive Bayes Classifier* (NBC), sehingga diklasifikasikan kelulusannya dan memberikan rekomendasi kelulusan mata kuliah tepat waktu. Berdasarkan sejarah kelas yang telah diambil siswa, nilai optimal adalah yang paling tepat.

4. (Annur, 2018) **KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES**. e-ISSN: 2548-7779 menyatakan bahwa masalah utama dalam pengentasan kemiskinan saat ini terkait dengan distribusi pertumbuhan ekonomi yang tidak merata. Dalam penelitian ini, kami menggunakan teknik data mining untuk mengklasifikasikan data penduduk miskin yang diperoleh di wilayah Tibawa. Metode yang digunakan adalah metode classifier Naive Bayes yang merupakan salah satu teknik klasifikasi data mining. Hasil pengujian tabel klasifikasi sore menggunakan metode split-validation akurat 73% pada dataset yang diteliti menggunakan metode klasifikasi Naive Bayes, atau termasuk dalam kategori Baik. Nilai presisi adalah 92% dan tingkat recall adalah 86%.
5. (Nofriansyah et al., 2016) **Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di CV. Sumber Utama Telekomunikasi)**. ISSN: 1978-6603 menyatakan bahwa persaingan terjadi dalam dunia bisnis, dan pelaku harus terus mempertimbangkan strategi dan kemajuan untuk menjamin kelangsungan bisnis yang mereka operasikan. Hal ini menyebabkan persaingan di antara penyedia kartu untuk kartu internet. Penyedia kartu internet berlomba-lomba menarik pelanggan melalui berbagai strategi pemasaran. Untuk bertahan hidup tanpa mengurangi daya saing, metode klasifikasi dapat menemukan model yang membedakan konsep atau kelas data. Tujuannya adalah untuk dapat

menyimpulkan kelas objek dengan label yang tidak diketahui sehingga algoritma Naive Bayes dapat memprediksi prospek masa depan berdasarkan pengalaman sebelumnya. Hasil penelitian ini membantu perusahaan memprediksi atau memprediksi perilaku kartu internet baru. Kemampuan untuk membuat keputusan dan meningkatkan strategi pemasaran.

6. (Putro et al., 2020) **Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan**. e-ISSN: 2620-7532 menyatakan bahwa lokasi usaha memegang peranan penting dalam penjualan. Perusahaan ini berbasis di kota, sehingga memudahkan vendor untuk mendistribusikan acara kepada orang-orang. Kegiatan distribusi erat kaitannya dengan kegiatan penjualan. Jika ada penawaran penjualan untuk pelanggan potensial dan kelompok pelanggan non-pilihan. Salah satu metode yang dapat digunakan untuk klasifikasi adalah data mining. Penambangan data kategorikal yang paling umum digunakan adalah metode Naive Bayes. Atribut yang digunakan dalam proses klasifikasi pelanggan adalah jumlah pembelian, jangka waktu, dan lokasi. Sistem klasifikasi menghasilkan 23 jawaban valid dan 2 jawaban salah. Menurut hasil yang diperoleh dengan metode confusion matrix, tingkat akurasi adalah 92%, tingkat akurasi 100%, dan tingkat recall 91%.
7. (Kurniawan, 2018) **PERBANDINGAN ALGORITMA NAIVE BAYES DAN C.45 DALAM KLASIFIKASI DATA MINING**. e-ISSN: 2528-6579 menyatakan bahwa metode Naive Bayes dan C.45 diterapkan pada 4 studi kasus, yaitu kasus penerimaan “Kartu Indonesia Sehat”, identifikasi pengajuan kartu kredit di bank, penentuan usia saat lahir, dan penentuan

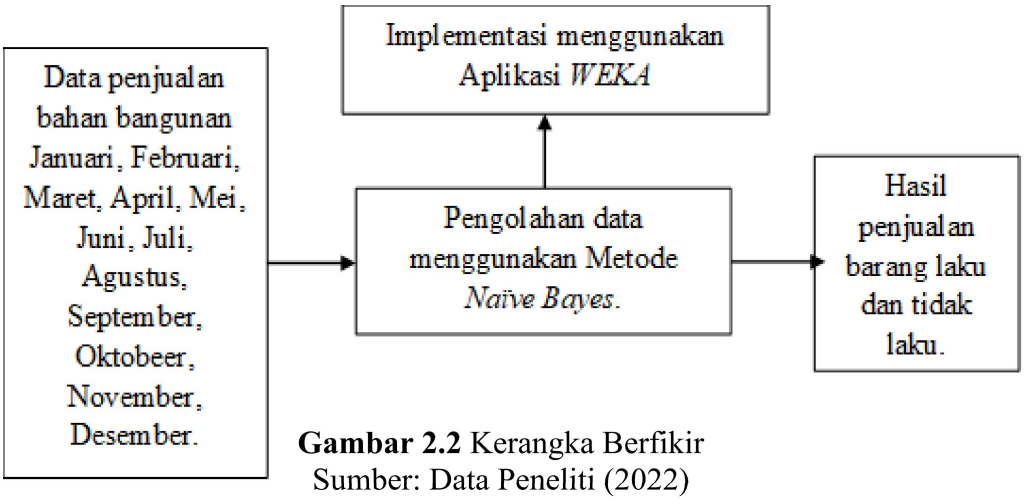
kelayakan. Temukan algoritma terbaik dalam koperasi, dalam setiap kasus. Precision, recall, dan precision kemudian dibandingkan untuk setiap data training dan test. Berdasarkan hasil implementasi, telah dibangun sebuah aplikasi untuk menerapkan algoritma Naive Bayes dan C.45 pada 4 kasus tersebut. Aplikasi diuji dengan kotak hitam dan algoritme dan hasilnya berhasil dan dapat menjalankan kedua algoritme dengan benar. Berdasarkan hasil pengujian, semakin banyak data pelatihan yang digunakan, semakin tinggi nilai presisi dan recall.

8. (K.Vembandasamy et al., 2015) **Heart Diseases Detection Using Naive Bayes Algorithm**. ISSN 2348 – 7968 menyatakan bisnis kesehatan telah menjadi bidang penting dari berbagai obat-obatan dalam industri kesehatan mengandung banyak data dan informasi tersembunyi untuk itu informasi tersembunyi ini akan digunakan untuk membuat keputusan yang efektif dengan menerapkan teknik data mining. Data mining dapat memecahkan banyak masalah kesehatan. Algoritma Naïve Bayes mewakili metode penambangan data untuk mendiagnosis penyakit jantung, dan mengusulkan sistem prediksi penyakit jantung (HDPS) berdasarkan metode penambangan data dengan menganalisis beberapa parameter untuk memprediksi penyakit jantung.
9. (Peling et al., 2017) **Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm**. e-ISSN: 2579-597X menyatakan kualitas perguruan tinggi di Indonesia, khususnya kualitas program sarjana, diukur dari akreditasi yang dilakukan mahasiswa yang

tidak lulus tepat waktu masih menjadi perbincangan hangat tentang kegagalan akademik. Ketepatan waktu mahasiswa sarjana dapat digunakan untuk menentukan pola kelulusan mahasiswa sarjana melalui teknik data mining yang dapat digunakan sebagai dasar untuk memprediksi kelulusan mahasiswa tahun depan. Studi ini menunjukkan bahwa *Naive Bayes* dapat mengklasifikasikan data uji dengan benar dengan tingkat kesalahan rata-rata masing-masing 86,16% dan 13,84%.

2.6 Kerangka Berfikir

Setelah memaparkan teori-teori di atas, penelitian ini menggunakan kerangka pemikiran yang dirancangkan oleh peneliti seperti gambar 2.2 berikut:



Gambar 2.2 Kerangka Berfikir
Sumber: Data Peneliti (2022)

Data penjualan material, diambil dari hasil data arsip PT Tanjung Uncang yang digunakan sebagai masukan dalam penelitian, lalu proses pengolahan datanya memakai metode *Naive Bayes* dengan mengimplementasikan ke aplikasi *WEKA*. Keluaran hasilnya berupa barang laku dan tidak laku.