BAB II

TINJAUAN PUSTAKA

2.1 Knowledge Discovery in Database

Knowledge Discovery In Database (KDD) merupakan metode untuk memperoleh pengetahuan dari *database* yang ada. Dalam *database* terdapat tabel - tabel yang saling berhubungan / berelasi. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan (*knowledge base*) untuk keperluan pengambilan keputusan (Yuli Mardi, 2019).

Istilah *Knowledge Discovery in Database* (KDD) dan data mining seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD secara garis besar dapat dijelaskan sebagai berikut:

1. *Data Selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery in Database* (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.

2. *Pre-processing / Cleaning*

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus *Knowledge Discovery in Database*

(KDD). Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *Knowledge Discovery in Database* (KDD), seperti data atau informasi eksternal lainnya yang diperlukan.

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam *Knowledge Discovery in Database* (KDD) merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan.

5. *I nterpretation / Evaluation*

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses *Knowledge Diccovery in Database* (KDD) yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta sebelumnya.

2.2 Data Mining

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 (Yuli Mardi, 2019).

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan terlebih dulu, data mining memiliki akar yang panjang dari bidang ilmu yang berbeda seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik, *database*, dan juga *information retrieval* (Yuli Mardi, 2019).

2.2.1 Pengelompokan *Data Maining*

Berdasarkan kegunaannya, aktivitas-aktivitas *data mining* dapat dikelompokkan sebagai berikut (Wanto et al., 2020):

1. Klastering (*Clustering)*

   Digunakan untuk mengemlompokkan atau mengidentifikasi data yang memiliki karakteristik tertentu. Contoh algoritma: *K-Means, K-Medoids,* dan lainnya.

2. Klasifikasi (*Classification)*

   Digunakan untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelsa dari suatu object yang labelnya tidak diketahui. Contoh algoritma: *C4.5*, *K-Nearest Neighbor, Naive Bayes,* dan lainnya.

3. Asosiasi (*Association*)

   Digunakan untuk mengatasi masalah bisnis yang khas, yakni dengan menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang sering kali dibeli bersamaan oleh *customer*. Contoh algoritma: *Apriori, Frequent Pattern Growth (FP-Growth)*, dan lainnya.

4. Estimasi (*Estimation*)

   Digunakan untuk memperkirakan atau menilai sesuatu hal yang belum pernah ada sebelumnya yang disajikan dalam bentuk hasil kuantiatif (angka). Contoh algortima: *Regresi Linieat, Confidence Interval Estimations*, dan lainnya.

5. Prediksi (*Prediction*)

   Digunakan untuk memperkirakan atau meramalkan suatu kejadian yang belum pernah terjadi. Contoh algoritma: *Decission Tree, KNN*, dan lainnya.

2.3 Klasifikasi

Klasifikasi adalah proses menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau data kelas, dengan tujuan untuk dapat mengubah kelas dari suatu obyek yang labelnya tidak diketahui. Dalam mencapai tujuan tersebut, prosesklasifikasi membentuk suatu model yang mampu membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu (Siswandi & Fitriana, 2019).

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu : pertama, pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan kedua, penggunaan model tersebut untuk melakukan pengenalan / klasifikasi / prediksi pada suatu

objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang mudah disimpan (Ariyanti & Iswardani, 2020).

2.4 Naive Bayes

Naive Bayes adalah algoritma klasifikasi untuk menghitung probabilitas dengan menghitung frekuensi dan kombinasi nilai dalam suatu data (Romadhon & Kurniawan, 2021). Metode ini dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes* (Pratama et al., 2022).

Naïve Bayes merupakan salah satu algoritma klasifikasi yang sederhana namun memiliki akurasi yang tinggi. *Naïve Bayes* memiliki kelemahannya sangat sensitif dalam. Metode klasifikasi berbasis fitur yang dikembangkan dalam penelitian tersebut menghasilkan akurasi yang baik. *Naïve Bayes* memiliki beberapa keunggulan yaitu sederhana, cepat dan memiliki akurasi yang tinggi (Putri et al., 2020). Adapaun persamaan dari *Teorema Bayes* ditunjukkan pada rumus berikut ini (Saputra & Herdiansyah, 2022):

$$\mathrm{P(H|X)} = \frac{P(X|H).P(H)}{P(X)}$$ **Rumus 2. 1** *Teorma Bayes*

Keterangan:

X: data dengan class yang belum diketahui.

H: hipotesis data menggunakan suatu class spesifik.

P(H|X): probabilitas hipotesis H berdasar kondisi X (parteriori probabilitas).

P(X|H): probabilitas X bedasarkan kondisi pada hipotesis H.

P(H): probabilitas hipotesis H (prior probabilitas).

P(X): probabilitas X.

2.5 *Software* Pendukung

Gambar 2. 1 Logo *Software* Tanagra

Tanagra adalah rangkaian perangkat lunak pembelajaran mesin gratis untuk tujuan penelitian dan akademik yang dikembangkan oleh Ricco Rakotomalala di Universitas Lumière Lyon 2, Prancis. Tanagra mendukung beberapa tugas penambangan data standar seperti itu sebagai: Visualisasi, Statistik deskriptif, pemilihan *Instance*, pemilihan fitur, konstruksi fitur, regresi, faktor analisis, pengelompokan, klasifikasi dan pembelajaran aturan asosiasi. Tanagra membuat kompromi yang baik antara pendekatan statistik (misalnya uji statistik parametrik dan nonparametrik), metode analisis multivariat (misalnya *factor analysis, correspondence analysis, cluster analysis, regression*) dan teknik pembelajaran mesin (misalnya. *neural network, support vector machine, decision trees, random forest*) (Naik & Samant, 2016).

Tanagra dengan logo seperti pada Gambar 2.1 merupakan perangkat lunak *open source*, sehingga setiap individu dapat dengan mudah mengakses kode sumbernya dan menambahkan algoritma mereka sendiri, asalkan mereka setuju dan mematuhi persyaratan lisensi distribusi perangkat lunak tersebut (Putra et al., 2019). Tampilan aplikasi Tanagra dapat dilihat pada Gambar 2.2.

Salah satu contoh penggunaan aplikasi Tanagra adalah pada penelitian milik (Simanjuntak et al., 2021) dengan judul "*Data Mining* Rekomendasi Pemakaian *Skincare*". Aplikasi Tanagra pada penelitian ini digunakan untuk pengimplementasian *data mining* dengan metode *Naive Bayes*. Penelitian yang

dilakukan bertujuan untuk membuat rekomendasi penggunaan produk perawatan kulit di Kota Batam.

Gambar 2. 2 Tampilan Aplikasi Tanagra

2.6 Penelitian Terdahulu

1. **Nama Pengarang:** (Rahmatullah et al., 2019)**, judul: Data Mining Untuk Menentukan Produk Terlaris Menggunakan Metode *Naive Bayes*.**

   Tahun: Vol.7 No.2 (2019) 57-64, ISSN: 2623 1247

   PT. Cipta Niaga Semesta adalah salah satu perusahaan bagian dari Mayora group yang bergerak di bidang distribusi produk makanan dan minuman ringan yang ingin meningkatkan jumlah penjualan produk- produk yang mereka miliki. Untuk membantu perusahaan ini semakin maju diperlukan sebuah sistem yang akan membantu kemajuan perusahaan dalam memaksimalkan penjualan produk mereka, maka peneliti mencoba untuk melakukan suatu penelitian terhadap data produk PT. Cipta Niaga Semesta Sub Branch dengan menggunakan sebuah metode *Naive Bayes Classifier* Pada penelitian ini penulis menggunakan metode pengumpulan data metode

wawancara (*interview*), metode Observasi, metode Dokumentasi dan studi Literalur sedangkan metode pengembangan sistem menggunakan Metode *Waterfall*. Diimplementasi dengan menggunakan bahasa pemrograman menggunakan *Boland Delphi 7* dan *database Microsoft Acces 2010*. Aplikasi produk terlaris yang dibangun menggunakan *metode Naive Bayes Classifier* ini meliputi data produk, *cluster Naive Bayes* serta pelaporan. *Data Mining* Untuk Menentukan Produk Terlaris Menggunakan Metode *Naive Bayes* Pada PT. Cipta Niaga Semesta Sub Branch Kotabumi ini bertujuan untuk Sistem pengklasifikasian pada produk terlaris dan membantu Kepala *Area Operational Supervisor* (AOS) PT. Cipta Niaga Semesta Sub Branch Kotabumi dalam pemilihan produk terlaris. Sistem ini menghasilkan penentuan produk terlaris dengan 8 atribut yang diambil dari data penjualan selama 2 tahun terakhir.

2. **Nama Pengarang:** (Budiyanto & Dwiasnati, 2018)**, judul:** ***The Prediction of Best-Selling Product Using Naïve Bayes Algorithm*** **(*A Case Study at* PT Putradabo Perkasa).**

   Tahun:Vol.5 No.6 (2018) 68-74, ISSN: 2394-2231

   Data Mining is a technique for processing and extracting large data into information which can form new data. The technique is applied by using Knowledge Discovery Process in Database (KDD). The objective of the research is determining which product is the best-selling in 2018, so that the increase of customer's demand can be anticipated in the following year. In this research, the authors employ classification method in producing the best-selling product information by using algorithms naïve Bayes. There are some

variables involved, such as type of goods, brand of goods, quality of goods, price of goods, and Target. Rapid minner Studio 9.0 is a tool for assessing data which calculated to produce a model. The analysis results are expected to be used by the company for preparation supply of the best-selling products. The findings reveal that the level of data accuracy is is 78,33% and the best-selling product based on sales is the IP Camera product with type Infinity I-993V.

3. **Nama Pengarang:** (Abdullah et al., 2022)**, judul: Penerapan Data Mining untuk Memprediksi Jumlah Produk Terlaris Menggunakan Algoritma Naive Bayes Studi Kasus (Toko Prapti).**

   Tahun: Vol.13 No.1 (2022) 20-27, ISSN: 2477-3786

   Toko Prapti adalah perusahaan kecil milik pribadi yang menjual barang kebutuhan pokok,. Selama ini, toko prapti menghasilkan data penjualan setiap hari akan tetapi hasil yang diperoleh menunjukan toko prapti belum memaksimalkan data tersebut sehingga menjadi penumpukan data. Maka dari itu, peneliti melakukan suatu penelitian terhadap data penjualan produk dengan memanfaatkan dan menerapkan data mining dengan menggunakan algoritma *naïve bayes classifier* untuk mengetahui minat pembelian barang di toko prapti.Penulis melakukan penelitian ini dengan menggunakan metode wawancara, *observasi*, dan studi pustaka tentang metode pengumpulan data. Dalam penelitian ini, penulis menggunakan metode pengembangan sistem *waterfall*.. Penulis mengimplementasikan penelitian ini menggunakan bahasa pemrograman *web* yaitu *PHP* dengan menggunakan *framework codeIgniter* dengan basis data *MySQl*. Sistem yang dibangun dengan algoritma *naïve*

bayes ini meliputi data penjualan produk, perhitungan *naïve* dari masing-masing atribut serta pelaporan. Sistem ini menghasilkan 4 atribut yang sangat mempengaruhi hasil klasifikasi.Atribut yang digunakan dalam penelitian ini yaitu atribut adalah triwulan 1, triwulan 2, triwulan 3 dan triwulan 4. Hasil prediksi yang diperoleh dengan menggunakan metode algoritma *naïve bayes* menghasilkan informasi yang dapat digunakan oleh toko untuk mengidentifikasi produk terlaris yang dibeli konsumen sehingga dapat membantu toko prapti untuk menemukan dan menentukan target pasar dengan lebih akurat. Sumber data yang diambil dari 1 tahun sebelumnya dengan keakuratan sistem menggunakan *confusion matrix* menghasilkan *accuracy* 83,3%, *precision* 84,2% dan *recall* 88,9%.

4. **Nama Pengarang:** (Pransiska et al., 2019)**, judul: Penerapan Data Mining Prediksi Penjualan Barang Elektronik Terlaris Menggunakan Algoritma Naïve Bayes (Study Kasus : Planet Cash And Credit Cabang Muara Enim).Tahun: Vol.1 No.6 (2019) 2157-2169, ISSN: 2685-2683**

   Dalam menghadapi persaingan pasar untuk menghasilkan peningkatan pendapatan toko, pihak terkait harus menentukan strategi pemasaran produk yang dijual. Para pelaku bisnis juga harus membutuhkan sebuah analisis untuk memprediksi barang yang paling banyak terjual. Analisis yang dibutuhkan harus bias membantu dan memudahkan para karyawan took untuk mengetahui barang yang mudah habis terjual dengan memprediksi data penjualan pada tahun-tahun sebelumnya dengan menerapkan *data mining* menggunakan algoritma *naive bayes* menggunakan data penjualan tahun 2014 sampai dengan tahun 2016 di PT Solusi Prima Artha planet *cash and*

credit cabang Muara Enim.

5. **Nama Pengarang:** (Pratama et al., 2022)**, judul: Sistem Klasifikasi Penjualan Produk Alat Listrik Terlaris Untuk Optimasi Pengadaan Stok Menggunakan Naïve Bayes.**

   Tahun: Vol.6 No.4 (2022) 2135-2139, ISSN: 2614-5278

   Optimasi merupakan suatu proses penyelesaian suatu masalah sehingga dapat memberikan kondisi terbaik yang mampu memberikan nilai maksimum atau minimum. Dalam sebuah bisnis, optimasi pengadaan stok menjadi hal yang penting termasuk dalam hal penjualan produk. Jika stok suatu produk kosong maka potensi penjualan menurun. Oleh karena itu perlu suatu metode untuk mengoptimalkan stok sehingga dapat memenuhi permintaan konsumen dan akhirnya dapat meningkatkan penjualan. Data mining dapat diterapkan dalam sistem penjualan dengan membuat model klasifikasi penjualan produk terlaris. Dalam penelitian ini dilakukan klasifikasi penjualan produk terlaris pada toko elektronik menggunakan *Naïve Bayes*. Data yang digunakan dalam penelitian ini adalah data penjualan produk elektronik selama 3 bulan. Pada tahap awal dilakukan *preprocessing* yaitu dengan label *encoding*. Pengujian model dilakukan menggunakan *percentage split* dan *cross validation* dengan beberapa kali percobaan. Melalui penggunaan *percentage split* diperoleh akurasi terbaik sebesar 93,3% dengan perbandingan 30% data uji dan 70% data latih. Akurasi terbaik dengan menggunakan *cross validation* diperoleh sebesar 84% untuk 7-*fold*. Sistem klasifikasi yang telah dibuat mampu melakukan klasifikasi produk terlaris setiap triwulan dalam setahun. Melalui penggunaan sistem klasifikasi produk terlaris tersebut maka pihak toko dapat

mengetahui stok produk yang terlaris sehingga stok tidak kosong. Dengan demikian pengadaan stok toko dapat lebih optimal dan penjualan menjadi lebih meningkat.

6. **Nama Pengarang:** (Wijaya & Dwiasnati, 2020)**, judul: Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat.**

   Tahun: Vol.7 No.1 (2020) 2135-2139, ISSN: 2355-6579

   Jenis obat yang makin lama makin bervariatif, dari obat yang berharga murah sampai harga yang kalau dilihat sangat kurang masuk akal namun fungsinya sangat bagus. Meningkatnya peredaran jenis obat terutama vitamin, hal ini mendorong penulis untuk melakukan penelitian untuk menentukan produk vitamin mana yang LAKU atau TIDAK LAKU yang bisa di gunakan sebagai pedoman sebuah apotek menentukan jumlah stok barang yang harus ada pada gudang apotek tersebut. Penelitian ini bertujuan untuk mendapatkan nilai accuracy untuk data penjualan obat terutama jenis-jenis vitamin dengan menggunakan algoritma klasifikasi data mining yaitu algoritma Naïve Bayes. Penelitian ini menggunakan tools Rapidminner versi 8 sebagai media untuk menguji data yang akan diolah untuk mendapatkan hasil accuracy dan ROC.

7. **Nama Pengarang:** (Soepriyanto, 2021)**, judul:** ***Comparative Analysis of K-NN and Naïve Bayes Methods to Predict Stock Prices*****.**

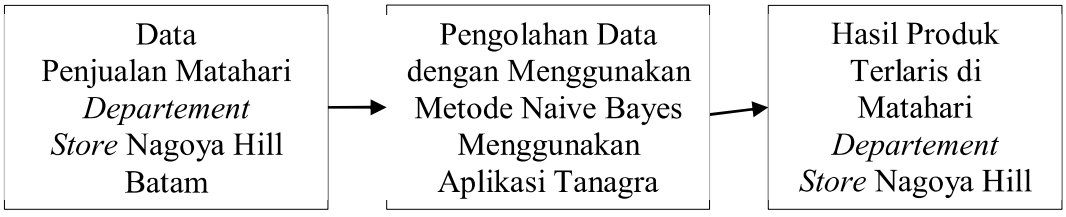
   Tahun: Vol.2 No.2 (2021) 49-53, ISSN: 2745-9659

   Buying and selling shares is a transaction that is widely carried out at this time, especially buying and selling stocks online which are widely available in the market, to make buying and selling shares require ability or knowledge so that the buying and selling of shares are profitable, to be able to help

economic players predict prices. Profit shares or not purchased in the future, this research will conduct stock price predictions using classification methods, namely K-Nearest Neighbor and Naïve Bayes, to predict the stock price data used for one month in minute levels totalling 39065 data, based on prediction results. The highest results obtained were using Naïve Bayes with an accuracy value of 69.38 then the K-Nearest Neighbor method with a K = 5 value of 67.25%, based on these results it can be concluded that the use of the K-Nearest Neighbor and Naïve Bayes methods for prediction share price not yet owned I high accuracy, so it can be combined with other methods or by using other variable predictors.

2.7 Kerangka Penelitian

Untuk mempermudah penelitian ini, peneliti telah membuat suatu kerangka penelitian seperti berikut ini:



Gambar 2. 3 Kerangka Penelitian

Pada Gambar 2.3 dapat dilihat bahwa data yang akan digunakan sebagai data *input* adalah data penjualan pada Matahari *Departement Store* Nagoya Hill Batam. Kemudian data akan diproses menggunakan metode *Naive Bayes* yang pengimplementasiannya menggunakan aplikasi Tanagra., maka akan mendapatkan hasil prediksi penjualan produk terlaris pada Matahari *Departement Store* Nagoya Hill Batam dan juga hasil akurasi untuk produk terlaris menggunakan *Naive Bayes*.