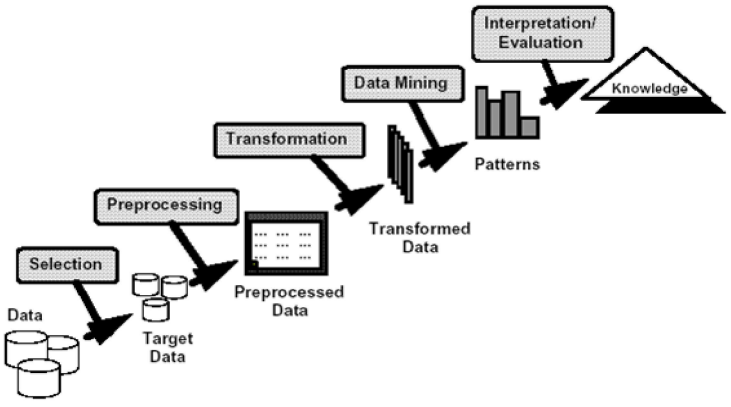
BAB II

TINJAUAN PUSTAKA

2.1 Knawladge Discovery in Database (KDD)

Data mining adalah sebuah tahapan dari proses rangkayan Knawladge Discovery in Database (KDD) yang berhubungan dengan penemuan ilmiah dan teknik integrasi, interprestasi serta visualisasi pola dari jumlah data. serangkaian dari proses tersebut mempunyai beberapa tahapan. adapun tahapan tersebut yaitu:

1. Pembersihan data (untuk membersihkan data yang tidak sesuai)
2. Integrasi data (beberapa sumber data yang digabungkan)
3. Transformasi data (data dapat dirubah bentuk yang sesuai untuk dimining)
4. Aplikasi teknik Data Mining, pola data yang ada diproses ekstraksi
5. Evaluasi pola yang ditemukan (pola proses interpretasi menjadi pengetahuan dapat digunakan untuk mengambil keputusan)
6. Presentasi pengetahuan (dengan teknik visualisasi).



Gambar 2.1 Tahapan *Knowladge Discovery in Database*

Sumber: (Peneliti 2022

Berikut ini adalah penjelasan dari tahapan *knowladge discovery in database* (KDD).

1. *Data selection*

Sekumpulan data yang dipilih untuk melakukan operasi sebelum pada tahap penggalian informasi dimulai dalam KDD, data dari hasil seleksi akan digunakan sebagai proses untuk data mining.

2. *Processing/Cleaning*

Data pada umumnya, baik dari database suatu perusahaan tidak sempurna seperti hilang, data tidak valid, selain itu juga terdapat atribut data yang tidak berguna atau relevan dan sebaiknya dihapus.

3. *Transformation*

Hasil dari transformasi pemilihan data ditentukan oleh kualitas dari data mining karena karakteristik data mining yang memperlukan teknik-teknik tertentu yang tergantug pada tahap ini.

4. *Data mining*

Data mining merupakan proses pencarian pola atau informasi yang menarik dalam suatu data terpilih dengan menggunakan metode atau teknik tertentu.

5. *Interpretation*

Pada tahapan ini mencakup pemeriksaan pola apakah informasi yang didapat bertentangan dengan hipotesis atau fakta yang ada sebelumnya.

2.2 *Data Mining*

Data mining merupakan suatu data yang diproses melalui ekstraksi (dari yang sebelumnya belum diketahui, serta tidak bermanfaat) sehingga menjadi ilmu pengetahuan atau informasi dari pola data yang besar jumlahnya (Witten, Ian H. Frank, 2011).

Data mining merupakan prosses dengan menggunakan mesin learning, teknik statistik, dan kecerdasan buatan serta matematika untuk diekstraksi dan di

identifikasi menjadi informasi yang bermanfaat dan pengetahuan yang terhubung dengan berbagai database yang besar. (Turban, dkk, 2005).

Data Mining sendiri sering disebut juga sebagai (KDD) Knowledge Discovery in Database merupakan kegiatan yang meliputi pengumpulan data historis yang digunakan untuk menemukan keteraturan pada hubungan pola dalam data set dengan berukuran yang besar (Handoko, 2016).

2.2.1 Fungsi *Data Mining*

Berikut ini adalah beberapa fungsi umum yang sering diterapkan pada data minng (Hasket, 2000).

1. *Assosiation* merupakan suatu proses untuk menentukan aturan hubungan antar kombinasi item pada waktu yang bersamaan.
2. *Sequence* yaitu proses untuk menentukan aturan hubungan antar kombinasi item pada waktu yang sama dan diterapkan pada beberapa periode.
3. *Clustering* yaitu proses untuk mengelompokan sejumllah objek atau data ke dalam suatu kelompok data hingga pada setiap kelompok berisikan data yang sama.
4. *Classification* yaitu proses untuk menemuankan fungsi/model yang membedakan dan menjelaskan konsep/kelas data, yang bertujuan untuk memprediksi kelas dari sebuah objek yang labelnya belum diketahui.
5. *Regression* yaitu proses yang memetakkan nilai data prediksi.
6. *Forecsting* yaitu proses estimasi suatu nilai prediksi yang berdasarkan pola pada sekumpulan data.

7. Solution yaitu proses untuk menemukan akar suatu masalah dan menyelesaikan persoalan bisnis tersebut atau dijadikan sebagai informasi untuk mengambil keputusan.

2.2.2 Kategori *Data Mining*

Data mining terbagi menjadi dua kategori (han dan kamber 2006), adapun kategori tersebut adalah sebagai berikut:

1. Prediktif

Prediktif bertujuan untuk memperkirakan nilai dari suatu atribut yang berdasarkan pada suatu nilai atribut lain.

2 Deskriptif

Deskriptif bertujuan menurunkan pola (cluster, trend dan anomali serta korelasi) untuk mempersingkat dalam hubungan data.

2.3 Metode *Data Mining*

2.3.1 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) adalah salah satu metode data mining yang masuk dalam algoritma klasifikasi. menurut (Harrington, 2012), algoritma K-NN memiliki beberapa kelebihan, di antaranya akurasi tinggi, insentif terhadap outler dan dugaan terhadap data tidak ada. akan tetapi algoritma ini juga mempunyai kekurangan di antaranya yaitu menentukan optimal nilai k, komputasi yang mahal dan membutuhkan banyak memori. di bawah ini adalah rumus Euclidean distance.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

Rumus 3.4.1. *Euclidean distance.*

di mana:

1) d (x, y) adalah antara jarak data x dan data y.

2) xi yaitu data testing ke-i

3) yi yaitu data training ke-i

Berikut ini adalah langkah-langkah algoritma K-NN:

1) menentukan parameter nilai k (dimana nilai k akan dipilih dngan manual).

2) untuk menghitung antara jarak data testing dan data training

3) mengurutkan data training dengan berdasarkan nilai jarak yang terkecil.

4) untuk menetapkan kelas, di mana data testing akan dipilih berdasarkan nilai k dengan jumlah terbanyak.

2.3.2 Algoritma C4.5

Dapat melakukan pengembangan diantaranya mengatasi *continue* data, *pruning* serta mengatasi missing value. Algoritma ini adalah salah satu teknik dari pohon keputusan atau *decesion tre* yang dapat digunakan untuk menghasilkan beberapa aturan. sebuah *decision tree* atau pohon keputusan memiliki tujuan untuk meningkatkan nilai prediksi dan keakuratan data yang dilakukan. algoritma C45 ini termasuk ke dalam algoritma klasifikasi.

2.3.3 Algoritma K-Means

Algoritma ini adalah sala satu algoritma yang mengelompokkan data non hirarki yang membagi data dalam bentuk 2 atau lebih kelompok. metode atau cara ini membagi data ke dalam kelompok agar data yang berkarakteristik sama dapat dimasukkan ke dalam satu kelompok. Data dengan karakteristik yang berbeda akan dikelompokkan dengan kelompok data yang lain. pengelompokkan data ini bertujuan untuk memaksimalkan variasi antar kelompok dan juga meminimalkan fungsi objek yang diatur ke dalam suatu kelompok.

2.3.4 Algoritma Naïve Bayes

Algoritma *bayesian classification* adalah statistik pengkalsifikasian untuk memprediksi nilai probabilitas keanggotaan pada suatu kelas. *Bayesian classification* adalah dasar dari teorema bayes yang juga mempunyai kemampuan untuk pengklasifikasian sama dengan neural network dan pohon keputusan.

Menurut (Sardiarinto, 2013), metode Data Mining memliki tujuan yang sesuai dan dikelompokkan menjadi dua kelas utama yaitu.

1. Supervised Model

Model supervised merupakan model yang memiliki tujuan untuk memperkirakan kejadian dan memprediksi nilai atribut angka yang terjadi secara terus menerus. moel ini terbagi atas tiga yaitu:

a) klasifikasi

Tujuan prosses ini yaitu untuk mengelompokkan hubungan antara variabel target dan variabel kriteria, klasifikasi adalah kelompok kelas yang telah diketahui sebelumnya. sebagai contoh yaitu algoritma Naive bayes, C4.5 dan ID3.

b) Prediksi

Prediksi pada umumnya hampir mirip halnya dengan klasifikasi yang membedakannya adalah hasil nilai dari prediksi keluaran akan dipergunakan pada masa mendatang. sebagai contoh algoritma Support Vector Machine dan Neural Network serta Linear Regression.

c) Estimasi

Pengelompokkan ini hampir sama halnya dengan klasifikasi. estimasi merupakan prediksi atau perkiraan. Perbedaannya terdapat pada bentuk pengelompokkan. estimasi pengelompokkan ini lebih cenderung kearah angka dari pada kearah kategori. sebagai contoh algoritma Support Vector Machine dan Linear Regression serta. Neural Network.

2. Unsupervised Model

Unsupervised merupakan pemodelan dengan atribut tidak di pandu oleh target tertentu tetapi pemodelan ini adalah model yang terarah. dalam hal ini tidak terdapat bidang keluaran tetapi hanya terdapat masukkan. yang termasuk metode model ini, yaitu:

a. Model Asosiasi

Asosiasi merupakan perserikatan, gabungan dan himpunan serta kelompok. variabel kemunculan atau yang sering muncul pada waktu bersamaan merupakan

proses pengelompokkan asosiasi. nilai confidence dapat diukur dengan besarnya kemunculan pada atribut dengan bersamaan. sebagai contoh yaitu algoritma apriori dan Fp-Growth.

b. Model Cluster

Pengelompokkan data pada pengklasteran mempunyai persamaan nilai. perolehan pengamatan adalah hasil dari bentuk data yang dihasilkan dari pengklasteran seperti perekaman objek dan data yang mempinyai kemiripan. Contoh algoritma K-Medoids, K-Means dan Self Organization Map (SOM) serta Fuzzy C-means

c. Model Deskripsi

Model ini bertujuan untuk mengubah dan mencari pola yang sering muncul dan pola tersebut akan di masukkan ke dalam aturan baru yang dipergunakan untuk mempermudah aktivitas.

2.4 Software Pendukung

2.4.1 Rapidminer

Gambar 2.3 Logo *Rapidminer*
Sumber: RapidMiner.com

. RapidMiner menyediakan metode mulai dari association, classification, clustering, dan lain-lain. Rapid Miner merupakan software atau data dalam ilmu pengetahuan dan dikembangkan oleh sebuah perusahaan yang sama dengan nama aplikasi tersebut dan menyediakan lingkungan yang terpadu untuk mesin

pembelajaran (machine learning), pembelajaran yang mendalam (deep learning), dan analisis prediktif (predictive analytics). serta penambangan teks (text mining), aplikasi ini dipergunakan sebagai penelitian, pelatihan, pendidikan, komersial, bisnis dan juga untuk membuat prototype serta pengembangan aplikasi dan mendukung semua langkah proses pembelajaran mesin termasuk untuk persiapan validasi data, pengomptimalan dan visualisasi hasil.

2.4.2 Microsoft Excel

Gambar 2.3.2 *Icon Microsoft Excel*

Sumber: softicons.com

Microsoft excel dapat juga disebut dengan Excel adalah lembar kerja yang mengolah data atau angka dengan otomatis. Microsoft Excel di distribusikan dan dibuat oleh Microsoft corporation dan dapat dijalankan melalui Mac OS dan Microsoft windows. Aplikasi ini merupakan bagian dari Microsoft Office System. Microsoft Excel dapat digunakan untuk pengolahan data, manajemen data, penggunaan rumus tertentu, perhitungan matematika dasar dan membuat grafik.

2.5 Penelitian Terdahulu

Untuk melakukan penelitian dalam memprediksi hasil penjualan terlaris menggunakan Data Mining maka penulis menjadikan beberapa jurnal sebagai bahan referensi. adapun jurnal referensi tersebut sebagai berikut:

1. (Hendri Risman, Didik Nugroho & Yustina Retno WU, 2015) Penerapan Metode K-Nearest Neighbor Pada Aplikasi Penentu Penerima Beasiswa Mahasiswa Di STMIK Sinar Nusantara Surakarta, ISSN : 2338-4018. Tujuan dari penelitian ini yaitu membuat aplikasi dalam membantu dan mempermudah tim membuat keputusan alternatif dalam menentukancalon penerima beasiswa dengan Metode klasifikasi K-Nearest Neighbor Jurnal TIKomSiN. Hasil pengijian yang dilakukan terhadap 22 data sampel sebagai acuan dan diperoleh nilai kekuratan sebesar 90,90%.
2. (Rusda Wajhillah, Irsyad Hafizh Ubaidallah & Saeful Bahri, 2019) Analisis Kelayakan Kredit Berbasis Algoritma K-Nearst Neighboar (Studi Kasus: Koperasi AKU). Infotekjar : Jurnal Nasional Informatika dan Teknologi Jaringan VOL.4NO.1. Tujuan dari penelitian ini adalah untuk menganalisis resiko terjadinya kredit bermasalah dengan menggunakan metode algoritma klasifikasikan K-Nearest Neighbor. Hasil yang didapatkan akurasi tertinggi sebesar 79,45% pada nilai K=1, dengan rata-rata akurasi 73,696%, dan nilai AUC tertinggi didapat pada K=9 dengan nilai sebesar 0,811.
3. (Diana Theresa Worung, Sherwin R.U.A. Sompie & Agustinus Jacobus, 2020). Implementasi K-Means dan K-NN pada Pengklasifikasian Citra Bunga. Jurnal Teknik Informatika p-ISSN : 2301-8364, vol. 15 no. 3. Tujuan dari penelitian ini adalah untuk mengimplementasikan pengelompokan citra pada pengklasifikasian jenis bunga dengan menggunakan gabungan metode K-Means dan metode K-NN dan mendapatkan hasil akurasi tertinggi 85%. Pengujian recall dan precision mendapatkan hasil paling tinggi 88% dan 85% .

4. (Handoko & Lesmana, 2018). Data Mining Pada Jumlah Penumpang Menggunakan Metode Clustering, Seminar Nasional Ilmu Sosial dan Teknologi Vo. 1, 97-102. penelitian bertujuan untuk pengelompokkan jumlah penumpang di bandara hang nadim batam untuk menggali data dan mendapatka pengetahuan yang baru dari data yang ada. dengan menggunakan teknik clustering dan hasil dari pada penelitian ini adalah membantu pihak bandar udara untuk memberikan informasii dan mengantisipasi jadwal padatnya lalu lintas penumpang bandar udara hang nadim pada bulan tertentu.
5. *(I Ketut Agung Enriko, Muhammad Suryanegara & Dadang Gunawan, 2016) Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters, Journal of Telecommunication, Electronic and Computer Engineering Vol. 8No. 12. This research aims to predict heart disease, to simplify parameters with remote patient monitoring. by using the K-Nearest Neighbor method. The results of the study indicate that the accuracy of the 8 parameters using the K-Nearest Neighbor algorithm is quite good, compared to 13 parameters with the K-Nearest Neighbor.*
6. *(E. Fadaei-Kermani, G. A. Barani & M. Ghaeini-Hessaroeyeh, 2017) . Drought Monitoring and Prediction using K-Nearest Neighbor Algorithm. Journal of AIand Data Mining Vol 5, No 2. This research aims to predicts the drought that occurs based on the standard rainfall index (SPI) using the k-nearest neighbor method. The results obtained indicate that the corresponding values for the correlation coefficient (r = 0.874), mean absolute error (MAE = 0.106), root mean square error (RMSE = 0.119) and residual mass coefficient (CRM = 0.0011).*

7. *(Engin Esme & Mustafa Servet Kiran. 2018). Prediction of Football Match Outcomes Based on Bookmaker Odds by Using K-Nearest Neighbor Algorithm. International Journal of Machine Learning and Computing, Vol. 8, No. 1. predict the results of soccer matches by measuring the similarities between competitions based on betting odds using the K-Nearest Neighbor method. This study gets results with an accuracy rate of 96%.*

2.6 Kerangka Pemikiran

Untuk mempermudah penelitian ini maka peneliti membuat *flowchart* sebagai berikut:

Gambar 2.4 Kerangka Pemikiran

Sumber: (Peneliti 2022)

Penjelasan dari gambar di atas adalah pada data input penjualan yang digunakan oleh PT Daya Anugrah Mandiri untuk data masukkan telah dilakukan tahapan *cleasing* data, kemudian data tersebut diprosses ke dalam algoritma *K-Nearest Neighbor* dengan menggunakan *software rapidminer.* di dalam proses algoritma *K-Nearest Neighbor* terdapat beberapa langkah yakni menentukan nilai parameter *k*, menghitung jarak data training dan data testing, mengurutkan data training berdasarkan jarak terkecil dan yang terakhir yaitu menetapkan data testing dari nilai kelas *k* terbanyak. setelah melakukan proses tersebut maka akan mendapatkan hasil prediksi penjualan untuk dijadikan sebagai bahan acuan pengambilan keputusan terhadap produk terlaris.