BAB II
LANDASAN TEORI

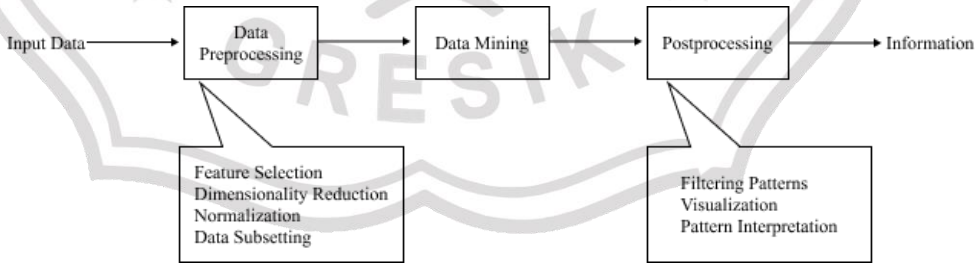
2.1 Data

Menurut (Wahyudi) data adalah dapat berupa angka-angka, huruf-huruf, gambar-gambar atau simbol-simbol apapun yang dimasukan (*input*) ke komputer dan dikeluarkan (*output*) dari komputer, karena komputer itu benda mati yang tidak memiliki kemampuan apapun termasuk kemampuan untuk mengenali mana huruf, angka, data dan informasi.

Data tentunya memiliki beberapa fungsi, salah satunya yaitu sebagai alat pengambilan keputusan, karena dengan bantuan data maka seseorang atau pengambil keputusan di suatu organisasi dapat mengetahui langkah apa yang harus dibuat agar keputusan yang diambil tidak merugikan pihak-pihak yang terlibat.

2.2 Data Mining

Secara umum data mining didefinisikan sebagai ekstraksi dari suatu informasi dari data yang disimpan dalam jumlah besar yang sebelumnya belum diketahui potensial kegunaannya menjadi suatu informasi yang berguna. Data mining memiliki beberapa sebutan atau nama lain yaitu: *Knowledge discovery in databases* (KDD), *knowledge extraction* (ekstraksi pengetahuan), analisa data/pola, *bussines intelligence* (kecerdasan bisnis).



Gambar 2.1 Proses dalam *Knowledge discovery in databases*

Data mining merupakan suatu proses ekstraksi dari suatu informasi yang disimpan dalam suatu basis data, atau data *warehouse*. Dengan demikian arsitektur dalam data mining memiliki komponen sebagai berikut:

a. Basis data

Komponen basis data atau data *warehouse* berfungsi sebagai pengambilan data yang relevan berdasarkan kebutuhan untuk selanjutnya diolah menjadi suatu informasi atau *insight* baru.

b. Basis pengetahuan

Komponen basis pengetahuan digunakan untuk mengevaluasi pola-pola yang dihasilkan dari proses pengolahan data. Basis pengetahuan ini digunakan untuk mengorganisasikan suatu atribut atau nilai kedalam level abstraksi yang berbeda.

c. Data mining *engine*

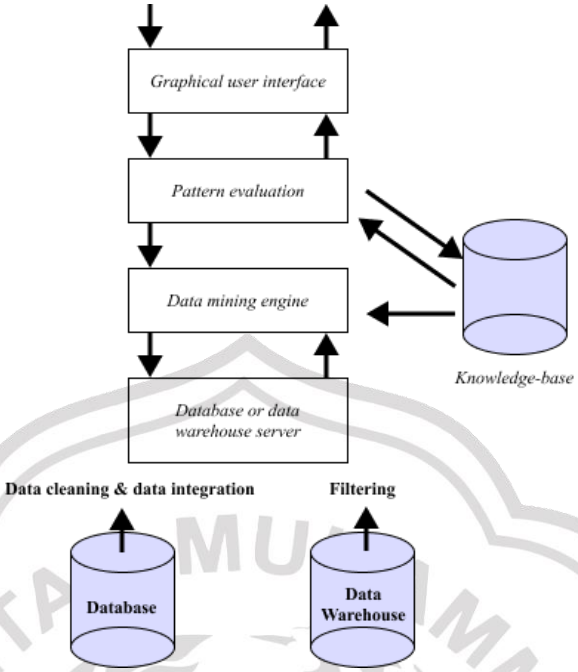
Komponen data mining *engine* merupakan komponen penting dalam arsitektur data mining. Komponen ini digunakan untuk mentransformasikan data menjadi bentuk yang sesuai untuk di mining.

d. Modul evaluasi pola

Modul evaluasi pola menggunakan threshold untuk memfilter pola yang didapat. Komponen ini berinteraksi dengan modul data mining dalam pencarian pola-pola menarik.

e. Antarmuka pengguna grafis

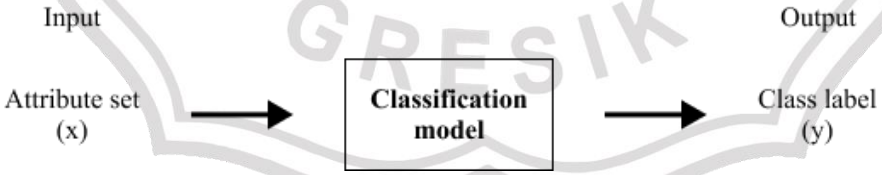
Di dalam modul ini pengguna berinteraksi dengan dengan sistem untuk menentukan *query* atau *task* pada data mining. Modul antarmuka pengguna grafis menyediakan beragam informasi untuk melakukan pencarian dan eksplorasi berdasarkan hasil data mining. Mencari struktur data dan basis data dapat dilakukan oleh komponen ini yang berfungsi melakukan evaluasi dan visualisasi pola dalam berbagai bentuk.



Gambar 2.2 Arsitektur Sistem Data Mining

2.3 Klasifikasi Data Mining

Menurut Tan *et all* (2004), klasifikasi adalah sebuah proses untuk menemukan model yang menjelaskan konsep atau kelas data dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui. Di dalam klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari beberapa atribut (kontinyu atau kategoris), salah satu atribut menunjukan kelas *record.*



Gambar 2.3 Konsep klasifikasi

Model klasifikasi memiliki suatu model pembelajaran yang dapat mengambil suatu masukan, kemudian memikirkan masukan tersebut dan memberikan jawaban sebagai keluaran dari pemikiran tersebut (Prasetyo, 2014).

2.4 Penjualan

Penjualan merupakan sebuah proses sosial dimana kelompok maupun individu mendapatkan atau membeli apa yang dibutuhkan, menciptakan, dan melakukan pertukaran produk yang memiliki nilai dengan kelompok maupun individu lain (Philip Kotler, 2008). Konsep penjualan merupakan cara untuk mempengaruhi konsumen untuk membeli produk yang ditawarkan. Dalam praktiknya penjualan dilakukan dengan cara tunai dan penjualan yang dilakukan secara angsuran.

2.5 Algoritma *K-Nearest Neighbor*

Algoritma K-Nearest Neighbor adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Untuk melakukan klasifikasi data baru, metode *K-Nearest Neighbor* menggunakan ukuran jarak yang sesuai. Jarak tetangga terdekat K dihitung dari tetangga terdekat untuk diprediksi sebagai label dari *instance* baru.

Memilih jumlah K tetangga terdekat sangat berpengaruh terhadap hasil dari akurasi *K-Nearest Neighbor*. Jika nilai dari K terlalu besar maka akan menimbulkan bias pada model dan jika nilai dari K terlalu kecil maka akan menimbulkan sensitif terhadap noise.

Pada algoritma *K-Nearest Neighbor* terdapat 5 cara untuk mencari tetangga terdekat (Prasetyo, 2014), yaitu :

1. Jarak *Euclidean*
2. Jarak *Manhattan*
3. Jarak *Cosine*
4. Jarak *Correlation*
5. Jarak *Hamming*

Dalam penelitian kali ini, penulis menggunakan perhitungan jarak *Euclidean*, rumus perhitungan jarak *euclidean* :

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2.1)$$

Berikut ini adalah langkah-langkah dalam menentukan klasifikasi menggunakan metode *K-Nearest Neighbor* :

1. Tentukan parameter K.
2. Hitung jarak antara data baru dengan seluruh data latih menggunakan rumus jarak *euclidean*.
3. Urutkan seluruh jarak berdasarkan jarak minimum dan tentukan tetangga terdekat sesuai dengan nilai K.
4. Sesuaikan klasifikasi dari kategori K dengan tetangga terdekat yang telah ditetapkan.
5. Gunakan kelas dengan jumlah terbanyak sebagai dasar menentukan kelas dari data baru.

Contoh penerapan metode *K-Nearest Neighbor* :

Tabel 2.1 Contoh Data Latih

Data	X	Y	Kelas
1	7	6	1
2	6	6	1
3	5	6	1
4	1	3	2
5	3	2	2
6	2	1	2
7	3	4	?

Pertanyaan :

Data uji adalah data (3,4), fitur X = 3, fitur Y = 4. Akan dilakukan prediksi menggunakan jarak *euclidean*. Masuk dalam kelas mana seharusnya?

Penyelesaian :

1. Tentukan dulu parameter nilai K, penulis menggunakan jumlah tetangga terdekat K = 3.
2. Hitung jarak antara data baru dengan seluruh data latih menggunakan rumus jarak *euclidean* (3,4).

Tabel 2.2 Perhitungan Menggunakan Rumus Jarak *Euclidean*.

Data	X	Y	Jarak Euclidean (3,4)
1	7	6	$\sqrt{(7-3)^2+(6-4)^2}=\sqrt{(4)^2+(2)^2}=4.47$
2	6	6	$\sqrt{(6-3)^2+(6-4)^2}=\sqrt{(3)^2+(2)^2}=3.6$
3	5	6	$\sqrt{(5-3)^2+(6-4)^2}=\sqrt{(2)^2+(2)^2}=2.82$

4	1	3	$\sqrt{(1-3)^2+(3-4)^2} = \sqrt{(-2)^2+(-1)^2} = 2.23$
5	3	2	$\sqrt{(3-3)^2+(2-4)^2} = \sqrt{(0)^2+(-2)^2} = 1.41$
6	2	1	$\sqrt{(2-3)^2+(1-4)^2} = \sqrt{(-1)^2+(-3)^2} = 3.16$

3. Urutkan seluruh jarak berdasarkan jarak minimum dan tentukan tetangga terdekat sesuai dengan nilai K = 3

   Tabel 2.3 Pengurutan Jarak Terdekat Dengan Data Training

Data	X	Y	Jarak Euclidean (3,4)	Urutan Jarak	Apakah Termasuk 3-NN?
1	7	6	4.47	6	Tidak (K>3)
2	6	6	3.6	5	Tidak (K>3)
3	5	6	2.82	3	Ya (K<3)
4	1	3	2.23	2	Ya (K<3)
5	3	2	1.41	1	Ya (K<3)
6	2	1	3.16	4	Tidak (K>3)

4. Sesuaikan klasifikasi dari kategori K dengan tetangga terdekat yang telah ditetapkan. Perhatikan data ke-3, 4, dan 5 pada tabel sebelumnya.

   Tabel 2.4 Penentuan Kelas Yang Termasuk K=3

Data	X	Y	Jarak Euclidean (3,4)	Urutan Jarak	Apakah Termasuk 3-NN?	Kelas Ya untuk KNN
1	7	6	4.47	6	Tidak (K>3)	-
2	6	6	3.6	5	Tidak (K>3)	-
3	5	6	2.82	3	Ya (K<3)	1
4	1	3	2.23	2	Ya (K<3)	2
5	3	2	1.41	1	Ya (K<3)	2
6	2	1	3.16	4	Tidak (K>3)	-

Kelas Ya untuk KNN pada baris 6 mencakup baris 3, 4, dan 5. Selanjutnya berikan kelas berdasarkan tabel awal, baris 3 memiliki kelas 1 sedangkan baris 4,5 memiliki kelas 2.

5. Gunakan kelas dengan jumlah terbanyak sebagai dasar menentukan kelas dari data baru.

   Tabel 2.5 Hasil Dari Klasifikasi

Data	X	Y	Kelas
1	7	6	1
2	6	6	1
3	5	6	1
4	1	3	2
5	3	2	2
6	2	1	2
7	3	4	**2**

   Data yang kita miliki pada baris 3, 4, dan 5 mempunyai 1 kelas 1 dan 2 kelas 2. Dari jumlah keseluruhan kelas (2 > 1) dapat disimpulkan bahwa data X = 3 dan Y = 4 termasuk dalam kelas 2.

2.6 RapidMiner

RapidMiner adalah sebuah alat bantu berbentuk *software* untuk menganalisis data mining, text mining dan analisis prediksi. *Software* bersifat *open source*. RapidMiner memiliki Graphic User Interface untuk merancang sebuah pipeline analysis. GUI ini akan menghasilkan file XML (Extensible Markup Language) yang mendefinisikan proses analisis dan menjalankannya secara otomatis.

2.7 Tinjauan Pustaka

Sebagai upaya penguatan topik penilitan, penulis melakukan analisis dari hasil riset penelitian sebelumnya yang berkaitan dengan topik penelitian. Berikut ini beberapa hasil dari penelitian sebelumnya:

a. Aisha Alfani W.P. R., Fahrur Rozi, Farid Sukmana (2021) dengan judul penelitian "Prediksi Penjualan Produk Unilever Menggunakan Metode *K-Nearest Neighbor*". Pada penelitian ini mendapatkan kesimpulan

didapatkan hasil prediksi penjualan produk Unilever, nilai akurasi tertinggi sebesar 86,6%. Sedangkan nilai akurasi terendah sebesar 40%.

b. Sebastinus Reczy S (2020) dengan judul penelitian "Penerapan Algoritma *K-Nearest Neighbor* Untuk Prediksi Harga Cabai Rawit Di Yogyakarta" mendapatkan kesimpulan, bahwa dengan mengimpelementasikan metode *K-Nearest Neighbor* mendapatkan hasil akurasi sebesar 72%.
c. Rino Indra Muhammad, Esron Rikardo Nainggolan, Jordy Lasmana Putra, Sidik, Susafa'ati dan Ummu Radiyah (2021) dengan judul penelitian "Implementasi Metode *K-Nearest Neighbor* Untuk Prediksi Penjualan Kemasan Skincare Pada PT.Universal Jaya Perkasa" mendapatkan kesimpulan, bahwa dengan menerapkan metode *K-Nearest Neighbor* didapatkan hasil prediksi penjualan produk kemasan skincare dengan akurasi sebesar 80%. Adapun akurasi ini masih bisa ditingkatkan dengan melakukan pengujian menggunakan algoritma klasifikasi lainnya seperti *neural network, random forest* dan lain-lain.
d. Saifur Rochman Cholil, Titis Handayani, Rastri Prathivi, Tria Ardianita (2021) dengan judul penelitian "Implementasi Algoritma Klasifikasi *K-Nearest Neighbor* Untuk Klasifikasi Seleksi Penerima Beasiswa" mendapatkan kesimpulan, evaluasi algoritma *K-Nearest Neighbor* menggunakan metode *confusion matrix* hasil dari perhitungan rata-rata akurasi dari metode ini sebesar 90,5%. Evaluasi perbandingan data training dan testing pada metode *K-Nearest Neighbor* dengan *cross validation sampling* memperlihatkan hasil perhitungan rata-rata *precision* sebesar 89%.
e. Dedi Handoko, Heru Satria Tambunan, Jaya Tata Hardinata (2021) dengan judul penelitian "Analisis Penjualan Produk Paket Kuota Internet Dengan Metode *K-Nearest Neighbor*". Pada penelitian ini didapatkan hasil penjualan kartu paket kuota internet dengan 4 jenis paket internet, data yang digunakan adalah data tahun 2017 – 2019 berdasarkan nilai akurasi terhadap prediksi penjualan produk untuk tahun 2020 yaitu sebesar 71,43% dan telah diuji menggunakan *tools RapidMiner*.

f. Ike Yolanda, Hasanul Fahmi (2021) dengan judul penelitian "Penerapan Data Mining Untuk Prediksi Penjualan Produk Roti Terlaris Pada PT. Nippon Indosari Corpindo Tbk Menggunakan Metode *K-Nearest Neighbor*" mendapatkan kesimpulan penerapan metode KNN digunakan sebagai aplikasi yang dapat memprediksi penjualan produk roti terlaris dengan menggunakan kriteria dan bobot agar dapat digunakan dengan metode tersebut. Berdasarkan hasil perhitungan jarak diatas, maka akan didapatkan suatu hasil keputusan yaitu Naik = 3 dan Menurun = 2. Dapat dilihat pada tabel mayoritas klasifikasi yang memiliki jumlah paling banyak adalah ">1,5" dimana variable ">1,5" merupakan kategori "Naik", dan variable "<1,5" merupakan kategori "Menurun" sehingga penjualan produk roti memiliki nilai sesuai dengan data uji yang telah dihitung diprediksi masuk kedalam kategori "Naik".
g. Yulia Rizky Amalia (2018) dengan judul penelitian "Penerapan Data Mining Untuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode *K-Nearest Neighbor*" mendapatkan kesimpulan pada penelitian ini dilakukan pemodelan menggunakan algoritma *K-Nearest Neighbor* didapatkan hasil prediksi penjualan produk elektronik terlaris sebanyak 6 jenis dari 22 produk dengan nilai akurasi terhadap hasil klasifikasi sebesar 92%.
h. Abdul Ghani Muttaqin, Karina Auliasari, Febriana Santi Wahyuni (2020) dengan judul penelitian "Penerapan Metode *K-Nearest Neighbor* Untuk Prediksi Penjualan Berbasis Web Pada PT. Wika Industry Energy" memperoleh kesimpulan hasil pengujian metode *K-Nearest Neighbor* menggunakan 8 data latih diperoleh hasil nilai akurasi 95% dan nilai error sebesar 5%
i. Sahara Abdy, Erika Roberta Br Gultom, Sri Ramadhany, Afifudin (2022) dengan judul penelitian "Prediksi Penjualan Sparepart Mobil Terlaris Menggunakan Metode *K-Nearest Neighbor*" mendapatkan hasil prediksi penjualan sparepart pada PT. Gaya Makmur Mulia dilakukan dengan

menghitung menggunakan euclidean distance dan aplikasi *Rapidminer* mendapat hasil prediksi dengan akurasi 80%.

j. Mohammad Kafil (2020) dengan judul penelitian "Penerapan Metode *K-Nearest Neighbor* Untuk Prediksi Penjualan Berbasis Web Pada Boutiq Dealove Bondowoso" memperoleh kesimpulan hasil pengujian keakuratan 83,3% dengan menggunakan 12 data training dan 12 data testing.