BAB 2
LANDASAN TEORI

2.1 Data

Data adalah hasil representasi fakta dunia nyata berasal dari pengamatan atau yang dikumpulkan. Data kemudian digunakan untuk analisis dan diolah menjadi informasi [7]. Informasi merupakan hasil dari data yang telah direkam, diorganisir dan dihubungkan sehingga menjadi sebuah arti atau makna sehingga menghasilkan sebuah kesimpulan yang tepat dari data.

Data memiliki berbagai macam bentuk, seperti angka, simbol, gambar dan teks. Data secara umum dibagi menjadi dua tipe, yaitu:

1. Data Kuantitatif, adalah jenis data yang mengukur besaran yang dapat digunakan untuk melakukan perhitungan matematika seperti: data numerik, diskret, dan kontinu.
2. Data Kualitatif, adalah data yang tidak bisa dinyatakan dalam bentuk angka dan tidak bisa dilakukan perhitungan matematika. Data kualitatif biasanya digunakan untuk menggambarkan peristiwa atau suatu hal yang lebih abstrak seperti: ordinal, nominal, dan binary.

2.2 Data Mining

Data Mining adalah proses menemukan pola menarik dan pengetahuan dari kumpulan data yang berjumlah besar [8]. Data mining merupakan proses penggalian informasi dan menemukan pola yang berguna dan tren dari data yang berjumlah besar. Data mining juga populer sebagai *Knowledge discovery from data* (KDD), *knowledge discover, information harvesting, data/pattern analysis.* Data mining menggunakan analisis matematika dan statistika untuk mendapatkan pengetahuan baru dan menemukan pola dari data.

Tujuan data mining adalah untuk menggali pengetahuan yang sebelumnya tidak diketahui dari sebuah data. Ketika pola tersebut sudah diketahui, maka dapat digunakan untuk menyelesaikan masalah. Data mining dapat digunakan

sebagai alat pendukung keputusan, meramalkan tren masa depan untuk membuat langkah yang efektif untuk menyelesaikan masalah [9].

Data mining memiliki dua jenis, yaitu *predictive analysis* dan *descriptive analysis. Predictive analysis* merupakan proses dimana suatu model yang terdiri dari beberapa variabel yang kemudian digunakan untuk memprediksi keluaran, contohnya prediksi penjualan, segmentasi pelanggan dan prediksi tren. *Descriptive analysis* merupakan proses yang menghasilkan korelasi, pola dan infromasi yang bermakna untuk pelaporan dan ringkasan dari data, contohnya *history* pembelian pelanggan, performa penjualan dan performa promosi.

2.3 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) dikembangkan oleh Daimler Chrysler, SPSS, dan NCR. CRISP-DM menyederhanakan proses penambangan data ke dalam pemecahan masalah umum untuk binsis maupun penelitian. CRISP-DM memiliki enam tahap siklus hidup [10], yaitu:

1. *Business Understanding*

   Tahap ini dilakukan untuk memahami objektif dari bisnis, memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian mendefenisikan masalah yang akan diselesaikan menggunakan data mining.
2. *Data Understanding*

   Tahap ini merupakan tahap mengumpulkan data, memahami data yang telah dikumpulkan dan memahami kualitas data.
3. *Data Preparation*

   Tahap ini merupakan tahap pemilihan data, pembersihan data dan trasnformasi data.
4. *Modeling*

   Tahap ini merupakan tahap untuk pemilihan algoritma, kemudian implementasi algoritma data mining.
5. *Evaluation*

   Tahap ini merupakan tahap untuk melakukan evaluasi dari model yang telah dibangun, pengecekan performa dari hasil algoritma yang dipilih.

6. Deployment

Tahap ini merupakan tahap pembuatan laporan atau artikel ilmiah, serta implementasi ke sebuah aplikasi yang bisa diakses oleh pengguna.

2.4 *Outlier*

Outlier merupakan data yang menyimpang secara ekstrim dari sekumpulan data lainnya. Mengidentifikasi *outlier* penting dilakukan karena beberapa metode statistik bisa terpengaruh oleh data *outlier* yang akan mempengaruhi hasil akhir. *Outlier* dapat diidentifikasi menggunakan metode statistik dan bantuan visualisasi. Metode statistik yang umum dilakukan adalah *Interquartile rang*e (IQR) dan visualisasi yang dapat digunakan adalah *boxplot. Boxplot* merupakan cara yang banyak digunakan untuk memvisualisasikan distribusi data [8].

1. *Interquartile range* (IQR)

Interquartile range (IQR) dapat dihitung dengan cara mengurangi *quartile* ke-3 dengan *quartile* ke-1.

$$IQR = Q3 - Q1 \qquad (1)$$

Untuk menghitung *quartile* dapat menggunakan persamaan berikut:

$$Qi = \frac{i\,(n+1)}{4} \qquad (2)$$

Keterangan:

i = kuartil yang ingin diketahui (1,2,3)

n = jumlah data

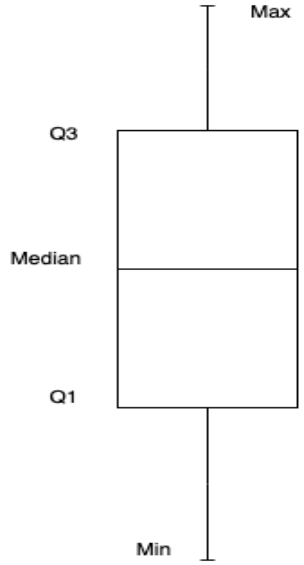
Outlier teridentifikasi ketika nilai data kurang dari *lower bound* dan melebihi *upper bound*, berikut merupakan persamaan *lower bound* dan *upper bound:*

$$lower\ bound = Q1 - 1{,}5(IQR) \qquad (3)$$

$$upper\ bound = Q3 + 1{,}5(IQR) \qquad (4)$$

2. ***Boxplot***

Boxplot merupakan grafik yang terdiri dari garis, kotak, dan satu atau lebih titik yang digunakan untuk visualisasi *five-number summary* [11] atau rangkuman lima angka yang berisi: Minimum, Q1, Median, Q3, dan Maksimum.



Gambar 2. 1 Boxplot

2.5 Penghalusan *Outlier*

Outlier yang telah teridentifikasi dapat dilakukan penghapusan atau jika ingin mempertahankan nilai *outlier* maka harus dilakukan penghalusan. Metode yang digunakan untuk penghalusan outlier yaitu metode *binning.*

Metode *binning* memiliki 4 jenis metode [10] yaitu:

1. *Equal width binning*
2. *Equal frequency binning*
3. *Binning by clustering*
4. *Binning by predictive value*

Penghalusan menggunakan metode *binning* memiliki tiga cara penghalusan yaitu:

1. Penghalusan dengan nilai rata-rata bin
2. Penghalusan dengan nilai tengah bin (median)
3. Penghalusan dengan nilai maksimum dan minimum bin

Metode *binning* memiliki tahapan sebagai berikut:

1. Mengurutkan data secara ascending
2. Membagi data menjadi beberapa bin
3. Penghalusan menggunakan *smoothing by bin means*

2.6 Normalisasi Data

Normalisasi data merupakan proses mengubah data numerik agar memiliki rentang nilai yang sama seperti (-1, 1), (0, 1), (0.0, 1.0). Tujuan dari normalisasi data yaitu untuk mencegah variabel yang memiliki rentang yang besar melebihi atribut yang memiliki rentang kecil [8]. Contoh metode normalisasi yaitu *min-max normalization, z-score normalization,* dan *normalization by scaling.* Untuk beberapa algoritma data mining, perbedaan rentang akan menyebabkan variabel yang memiliki nilai rentang lebih besar memiliki pengaruh yang lebih besar terhadap hasil algoritma.

2.7 Min-max Normalization

Min-max normalization merupakan normalisasi data dengan melakukan transformasi linier pada data. Min-max normalization dapat dilakukan dengan menggunakan persamaan berikut:

$$x = \frac{x - \min(x)}{rentang} \quad (5)$$

Keterangan:

x = nilai yang ingin dinormalisasi

min(x) = nilai minimum dari variabel

rentang = rentang nilai yang didapat dari nilai maksimum - minimum

2.8 Klasifikasi

Klasifikasi merupakan salah satu metode data mining yang digunakan untuk mendapatkan model, pola dan label dari data yang belum diketahui berdasarkan data training [9]. Data training merupakan kumpulan data yang telah diketahui

model dan label dari objek data. Model, pola dan label yang diberikan dapat berupa label ya atau tidak, label aman atau beresiko dan label lainnya. Klasifikasi dapat diterapkan pada kehidupan sehari-hari. Contoh penerapan klasifikasi yaitu klasifikasi transaksi bank, transaksi *online*, dan penyakit.

2.9 Regresi Linier

Regresi linier merupakan suatu metode dalam statistika yang digunakan untuk menentukan hubungan antara satu variabel dependen (respon) dengan satu atau lebih variabel independen (prediktor) [12]. Regresi juga merupakan bagian dari klasifikasi. Tujuan dari regresi adalah untuk membuat model matematis yang dapat memprediksi nilai variabel dependen berdasarkan nilai variabel independen. Regresi menggunakan data yang bertipe numerik atau nilai kontinu. Selain itu regresi linier juga bisa digunakan untuk mencari hubungan antara variabel dependen dengan variabel independen, sehingga dapat diketahui faktor-faktor apa saja yang mempengaruhi variabel dependen. Regresi secara umum terbagi menjadi dua jenis:

1. Regresi linier sederhana

Regresi linier sederhana digunakan untuk menentukan hubungan antara dua variabel. Regresi linier sederhana ini merupakan metode regresi yang paling sederhana. Jika hasilnya berbentuk garis melengkung, maka model tersebut dikategorikan sebagai non-linear. Namun jika hasilnya berbentuk garis lurus, maka model tersebut dikategorikan sebagai linier. Berikut ini merupakan persamaan dari regresi linier sederhana:

$$y =. bx + a \qquad (6)$$

y = variabel respon

b = konstanta

x = variabel prediktor

a = koefisien

2. Regresi linier berganda

Regresi linier berganda digunakan untuk mencari korelasi antara beberapa variabel bebas atau prediktor. Metode ini merupakan metode paling populer yang

digunakan untuk prediksi pada data mining. Persamaan regresi linier berganda dapat ditulis sebagai berikut:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k \quad (7)$$

Keterangan:

y = variabel respon

b = konstanta

b_1, b_2,b_3 = koefisien regresi

x_1, x_2, x_3 = variabel bebas

Regresi linier atau prediksi dapat diuji dengan beberapa metode seperti *mean square error* (MSE), *root mean square error* (RMSE), dan *mean absolute percentage error* (MAPE) [13]. MSE merupakan rata-rata dari selisih kuadrat antara nilai yang diprediksi dan nilai aktual. RMSE merupakan akar kuadrat dari MSE. MAPE merupakan rata-rata dari nilai absolut dari perbedaan antara nilai prediksi dan nilai aktual.

Berikut merupakan persamaan dari 3 metriks pengujian regresi:

$$MSE = \frac{1}{n} \Sigma(Yt - Y't)^2 \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma(Yt - Y't)^2} \quad (9)$$

$$MAPE = \frac{1}{n} \Sigma_{t=1}^{n} \frac{|Yt - Y't|}{Yt} \quad (10)$$

2.10 Analisis Korelasi

Analisis korelasi adalah suatu metode untuk mengukur hubungan atau asosiasi antara dua variabel atau lebih [14]. Berdasarkan jumlah variabel, jenis korelasi dapat dibagi menjadi dua kategori utama yaitu:

1. Korelasi Bivariat

Jenis korelasi yang mengukur hubungan antar dua variabel. Korelasi bivariat memberikan pengetahuan mengenai keterhubungannya antara satu variabel dengan satu variabel lainnya.

2. Korelasi Multivariat

Jenis korelasi yang mengukur hubungan antar tiga variabel atau lebih. Korelasi multivariat dapat menggambarkan hubungan yang tidak terlihar dari korelasi bivariat.

Teknik yang umum digunakan untuk analisis korelasi antara lain: *Pearson, Spearman, Kendall Tau.* Berikut merupakan penjelasan lebih lanjut dari teknik analisis korelasi:

1. Pearson

Korelasi *pearson*, yang juga dikenal sebagai korelasi produk-momen, merupakan metode yang paling umum digunakan untuk mengukur korelasi linier antara dua variabel kontinu. Koefisien korelasi *pearson*, yang disimbolkan dengan 'r', memiliki rentang nilai antara -1 dan 1. Nilai positif menandakan adanya hubungan positif antara kedua variabel (yaitu, ketika satu variabel meningkat, variabel lainnya juga cenderung meningkat), sedangkan nilai negatif menunjukkan adanya hubungan negatif (ketika satu variabel meningkat, variabel lainnya cenderung menurun). Sebaliknya, nilai 0 menunjukkan bahwa tidak ada korelasi yang terlihat antara kedua variabel tersebut. Adapun persamaan *pearson* sebagai berikut:

$$r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n \Sigma x^2 - (\Sigma x)^2\}\{n \Sigma y^2 - (\Sigma y)^2\}}} \qquad (11)$$

Keterangan:

r = nilai koefisien relasi

x = variabel independent

y = variabel dependent

n = jumlah data

Korelasi *pearson* bertujuan untuk mencari tingkat keeratan hubungan antar variabel [15]. Indikator koefisien relasi *pearson* dapat dilihat pada **Tabel 2.1** berikut ini:

Tabel 2. 1 Indikator Koefisien Relasi

Interval Koefisien	Tingkat Hubungan
0,00 – 0,199	Sangat Lemah
0,20 – 0,399	Lemah
0,40 – 0,599	Sedang

0,60 – 0,799	Kuat
0,80 – 1,00	Sangat Kuat

2. Spearman

Korelasi *spearman*, juga dikenal sebagai korelasi peringkat, digunakan untuk menilai hubungan monoton antara dua variabel. Metode ini tidak memerlukan asumsi distribusi normal data dan termasuk dalam kategori nonparametric [14]. Koefisien korelasi *spearman*, yang dapat disimbolkan dengan 'ρ' atau 'rs', memiliki rentang nilai antara -1 dan 1, sama seperti korelasi *pearson*. Persamaan *spearman* untuk data yang berukuran < 30 adalah sebagai berikut:

$$\sigma = 1 - \frac{6 \Sigma bi^2}{n(n^2-1)} \qquad (12)$$

Keterangan:

σ = Koefisien korelasi spearman

bi = Ranking data variabel x – y

n = jumlah data

3. Kendall Tau

Korelasi *kendall*, atau yang juga dikenal sebagai korelasi τ (tau *kendall*), digunakan untuk mengevaluasi hubungan ordinal antara dua variabel. Sama seperti korelasi *spearman*, korelasi *kendall* juga merupakan metode nonparametrik yang tidak memerlukan asumsi distribusi normal data. Koefisien korelasi *kendall*, yang dinyatakan sebagai 'τ', memiliki rentang nilai antara -1 dan 1, mirip dengan korelasi *pearson* dan *spearman*.