BAB II

TINJAUAN PUSTAKA

II.1. Data Mining

Bab ini akan membahas tentang pengertian, tujuan, tahap-tahap dan metode dari *data mining* sebagai berikut :

II.1.1 Pengertian Data Mining

Data Mining adalah sebuah teknik yang digunakan untuk menggali informasi berharga atau pola yang tersembunyi dari data besar. Tujuan utama dari *data mining* adalah untuk mengidentifikasi hubungan, pola, dan tren yang ada dalam data yang dapat membantu dalam pengambilan keputusan atau prediksi (Takdirillah, 2020). *Data mining* diartikan sebagai menambang data atau upaya untuk menggali informasi yang berharga dan berguna pada database yang sangat besar (Rusdianto et al., 2020)

Karakteristik *data mining* sebagai berikut :

a. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
b. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
c. *Data mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Dalam *data mining,* pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita bisa mengetahui pola *universal* data-data yang ada.

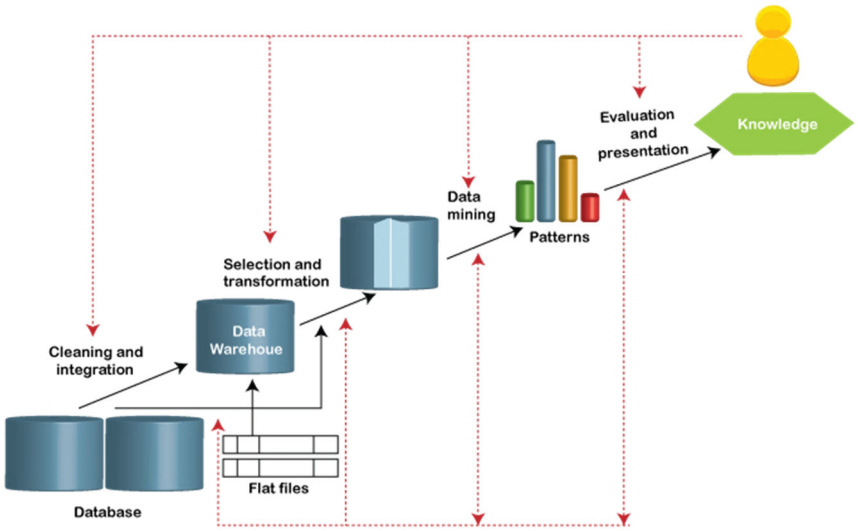
II.1.2 Tujuan Data Mining

Data Mining tentu saja memiliki beberapa tujuan. Berikut adalah beberapa tujuan dilakukannya *Data Mining* :

a. *Explanatory*, yaitu untuk menjelaskan beberapa kegiatan observasi atau suatu kondisi.
b. *Confirmatory*, yaitu untuk mengkonfirmasikan suatu hipotesis yang telah ada.
c. *Exploratory*, yaitu untuk menganalisis data baru suatu relasi yang janggal.

II.1.3 Tahap-Tahap Data Mining

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap yang diilustrasikan di gambar II.1. Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar II.1. Tahapan *Data Mining*

Tahapan *data mining* ada 7 yaitu :

1. Pembersihan Data (*Data Cleaning*)

   Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau data tidak relevan. Pembersihan data juga akan mempengaruhi performasi dari teknik *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi Data (*Data Integration*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya.

3. Seleksi Data (*Data Selection*)

   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.

4. Transformasi data (*Data Transformation*)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*.

5. Proses *Mining*

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi Pola *(Pattern Evaluation)*

   Untuk mengidentifikasi pola-pola menarik kedalam *knowledge based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

II.1.4 Metode *Data Mining*

Dengan definisi *data mining* yang luas, ada banyak jenis metode analisis yang dapat digolongkan dalam *data mining* (Sinaga et al., 2022). Pengelompokkan *data mining* berdasarkan tugas yang dapat dilakukan yaitu:

1. Deskripsi (*Description*)

   Deskripsi adalah menggambarkan pola dan kecenderungan yang terdapat dalam data yang memungkinkan memberikan penjelasan dari suatu pola atau kecenderungan tersebut.

2. Estimasi (*Estimation*)

   Estimasi hampir sama dengan klasifikasi, akan tetapi variabel target estimasi lebih ke arah numerik daripada ke arah kategori.

3. Prediksi (*Prediction*)

   Prediksi hampir sama dengan klasifikasi dan estimasi, akan tetapi dalam prediksi nilai dari hasil akan terwujud di masa yang akan datang.

4. Klasifikasi (*Clasification*)

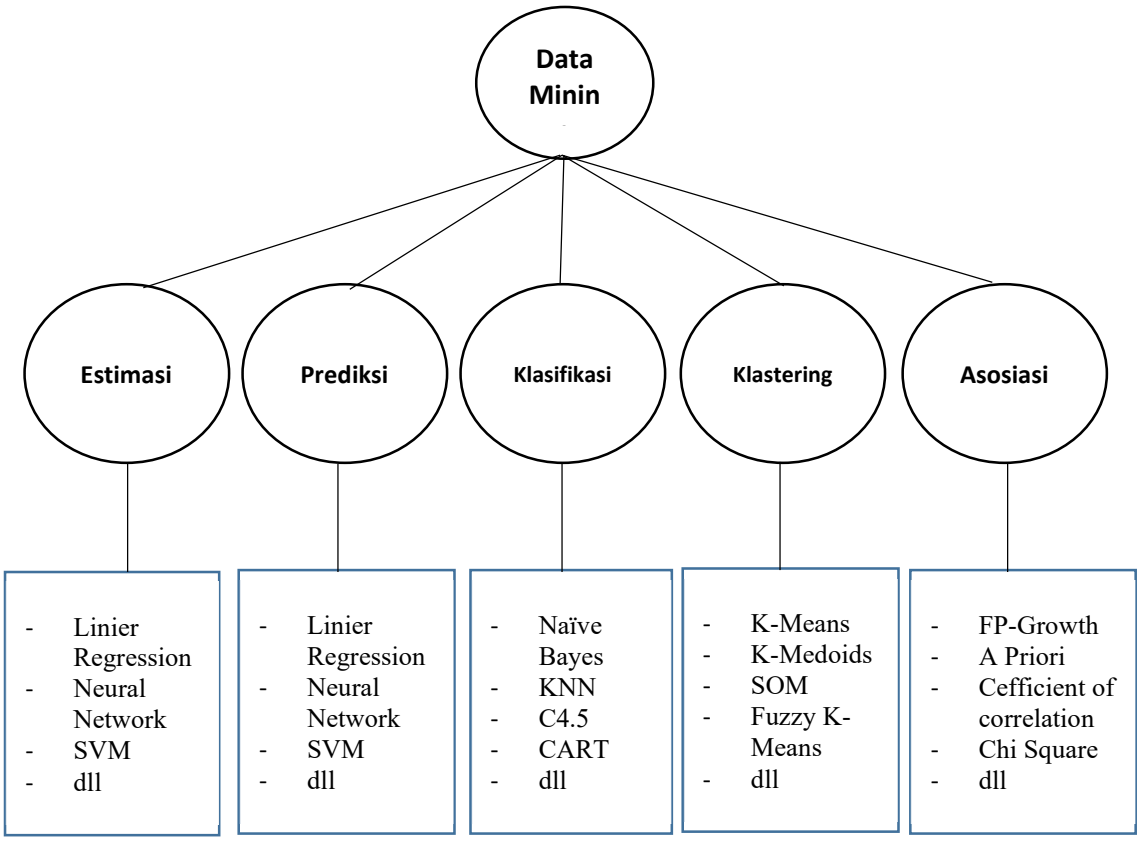
   Klasifikasi merupakan suatu proses mencari suatu himpunan yang dapat mendeskripsikan dan membedakan kelas-kelas data atau konsep dengan tujuan dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui.

5. Pengelompokan (*Clustering*)

   Clustering adalah Pengelompokkan data pada suatu objek data yang ada dalam satu *cluster* yang memiliki tingkat kemiripan antar satu data dengan data lainnya (Frankly Sept Genius Zendrato, Agung Triayudi, Endah Tri E, 2021).

6. Asosiasi (*Association*)

   Asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu.



Gambar II.2. Taksonomi Peranan Data Mining

II.2. Klasifikasi (C*lassification*)

Klasifikasi adalah teknik pengolahan data yang membagi objek menjadi beberapa kelas sesuai dengan jumlah kelas yang diinginkan. Cara kerja dari metode klasifikasi adalah sebuah proses dengan dua langkah. Langkah pertama dari *classification* juga disebut sebagai *learning of mapping atau function*, suatu fungsi pemetaan yang bisa memprediksi class label Y pada suatu tuple X. Pemetaan ini direpresentasikan dalam bentuk *classification rules, decision tree* atau formula matematika. Dari *rules* atau *tree* tersebut dapat digunakan untuk mengklasifikasi *tuple* baru. Langkah kedua adalah klasifikasi yang sudah dibangun akan digunakan

untuk mengklasifikasi data. Pertama, akurasi dari prediksi *classifier* tersebut diperkirakan. Jika menggunakan *training set* untuk mengukur akurasi dari klasifikasi, maka estimasi akan optimis karena data yang digunakan untuk membentuk klasifikasi adalah *training set* juga. Oleh karena itu, digunakan *test set*, yaitu sekumpulan tuple beserta *class* label-nya yang dipilih secara acak dari *dataset*. *Test set* bersifat independen dari *training set* dikarenakan *test set* tidak digunakan untuk membangun klasifikasi (Atika & Priatna, 2020: 89).

II.3. Algoritma Naïve Bayes

Naive Bayes merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris bernama Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Algoritma pengklasifikasi *Naïve Bayes* adalah pengklasifikasi yang berdasarkan probabilitas bersyarat pada *teorema Bayes* (Umar & M. Adnan Nur, 2022).

Naive Bayes merupakan salah satu algoritma klasifikasi yang utama pada data mining yang banyak digunakan dalam masalah klasifikasi di dunia nyata karena metode ini memiliki performa klasifikasi yang tinggi. Beberapa kelebihan dan kekurangan metode *Naïve Bayes* (Hairani et al., 2020) adalah sebagai berikut:

1. Keuntungan:
    a. Mudah diimplementasikan
    b. Memberikan hasil yang baik untuk banyak kasus
    c. Hanya membutuhkan satu kali scan data training

2. Kelemahan:
    a. Harus mengansumsi bahwa antar fitur tidak terkait (*independent*). Namun realitanya keterkaitan itu ada, sebagai contoh: biodata pasien Rumah sakit; umur, riwayat keluarga, dan lain-lain.
    b. Keterkaitan tersebut tidak dapat dimodelkan oleh *Naïve Bayes*

Cara kerja metode pengklasifikasi *Naïve Bayes* dapat diurutkan seperti langkah-langkah berikut (Septiarini et al., 2021) :

1. Diketahui D adalah *dataset training* yang terdiri dari sekumpulan baris data dan label kelasnya. Setiap baris memiliki n dimensi vektor atribut, $X = (x_1, x_2, ... x_n)$, menggambarkan n pengukuran dibuat atas baris dari n atribut, masing-masing sebagai berikut, $A_1, A_2,...A_n$.
2. Terdapat m kelas, $C_1, C_2, ...C_m$ memberikan sampel X, pengklasifikasi akan memprediksi bahwa X termasuk kelas yang memiliki probabilitas *posteriori*, dikondisikan pada X. Dimana X diperkirakan memiliki kelas C_i jika dan hanya jika:

    $P(C_i \mid X) > P(C_j \mid X)$ untuk $1 \leq j \leq m, j \neq i$ (II.1)

    Dengan demikian ditemukan kelas yang maksimal $P(C_i \mid X)$. Kelas C_i untuk setiap $P(C_i \mid X)$ yang dimaksimalkan disebut *hipotesisposteriori* maksimum. Persamaan teorema bayes:

    $$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \quad \text{(II.2)}$$

    Dengan:

    $P(C_i \mid X)$ = Probabilitas hipotesis kelas C_i berdasarkan kondisi X

    $P(X \mid C_i)$ = Probabilitas data X berdasarkan kondisi pada kelas C_i

$P(C_i)$ = Probabilitas awal kelas C_i

$P(X)$ = Probabilitas awal data X

3. P (X) adalah sama untuk semua kelas, hanya $P(X \mid C_i)\ P(C_i)$ yang perlu dimaksimalkan. Jika kelas apriori probabilitas, $P(C_i)$ tidak diketahui, maka umumnya diasumsikan seperti ini $P(C_i) = P(C_2) = ... = P(C_m)$ maka dari itu akan memaksimalkan $P(X \mid C_i)$. Tapi sebaliknya akan memaksimalkan $P(X \mid C_i)\ P(C_i)$. Dapat diperhatikan bahwa kelas probabilitas apriori dapat diperkirakan dengan $P(C_i) = |C_i, D| / |D|$, dimana $|C_i, D|$ merupakan jumlah pelatihan rangkap dari kelas C_i di dalam D.
4. Dataset dengan banyak atribut, akan menjadi perhitungan yang mahal untuk menghitung $P(X \mid C_i)$. Dalam rangka untuk mengurangi perhitungan dalam mengevaluasi $P(X \mid C_i)$. Pada asumsi naive bahwa kelas independen bersyarat dibuat. Ini menganggap bahwa nilai-nilai atribut yang independen bersyarat satu sama lain diberikan pada kelas sampel. Secara matematis berarti bahwa:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

$$= P(x_1 \mid C_i) \times P(X_2 \mid C_i) \times ... \times P(X_n \mid C_i) \qquad \text{(II.3)}$$

5. Untuk memprediksi label kelas X, $P(X \mid C_i)\ P(C_i)$ merupakan evaluasi dari setiap kelas C_i. Pengklasifikasi memprediksi bahwa label kelas X adalah C_i jika dan hanya jika

$$P(X \mid C_i)\ P(C_i) > P(X \mid C_j) \text{ untuk } 1 \leq j \leq m, j \neq i \qquad \text{(II.4)}$$

Contoh: Terdapat sebuah dataset pada table II.1 dibawah ini:

Tabel II.1. Dataset Cuaca

#	Cuaca	Temperatur	Kecepatan Angin	Berolah Raga
1	Cerah	Normal	Pelan	Ya
2	Cerah	Normal	Pelan	Ya
3	Hujan	Tinggi	Pelan	Tidak
4	Cerah	Normal	Kencang	Ya
5	Hujan	Tinggi	Kencang	Tidak
6	Cerah	Normal	Pelan	Ya

Dari Table II.1 diatas di asumsikan seperti dibawah ini :

Y = Berolah-raga

X1 = Cuaca

X2 = Temperatur

X3 = Kecepatan Angin

Fakta :

P (Y = ya) = 4/6

P (Y = tidak) = 2/6

Jika suatu hari:

Cuaca = Cerah

Kecepatan angin = Kencang

Berolahraga = ?

Maka hipotesa yang diambil berdasarkan nilai probabilitas dari kondisi prior yang diketahui:

P (X1 = Cerah, X3 = Kencang, Y = ya)

= { P (X1 = Cerah | Y = ya) . P (X3 = Kencang | Y = ya)} . P (Y = ya)

= { (1) . (1/4) } . (4/6) = 4/24 = 1/6

P (X1 = Cerah, X3 = Kencang, Y = tidak)

= { P (X1 = Cerah | Y = tidak) . P (X3 = Kencang | Y = tidak)} . P (Y = tidak)

= { (0) . (1/2) } . (2/6) = 0

Jadi, prediksi cuaca = cerah, kecepatan angin = kencang adalah berolahraga = ya.

II.4. Algortima C4.5

Algoritma C4.5 merupakan bagian algoritma yang digunakan untuk membentuk pembagian terstruktur atau pengelompokan di dataset. Dasar dari algoritma C4.5 adalah bentuk pohon keputusan *(Decision Tree)* (Ardian Darma *et al.*, 2022).

Tahapan dalam algoritma C4.5 yang pertama adalah menghitung nilai *entropy* total dari setiap label, kemudian menghitung nilai *entropy* dari masing-masing atribut (Saleh et al., 2020). Tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5 adalah :

1. Histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai indeks *entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai indeks *entropy*.

Pohon keputusan dapat digunakan untuk mewakili dan membuat keputusan, dapat dilihat sebagai serangkaian titik yang terdiri dari *node* tunggal atau menyebar ke bagian daun (Gifu, 2021). Untuk memilih attribut sebagai akar, didasarkan pada

nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus sebagai berikut: (Sari Oktapia Ningse et al., 2022).

$$\textbf{Gain S, A = Entropy S - } \sum_{i=1}^{n} \frac{|Si|}{|S|} * \boldsymbol{Entropy\ (Si)} \qquad \text{(II.5)}$$

Keterangan

S = himpunan Kasus

A = attribute

N = jumlah partisi attribute A

|Si| = jumlah kasus pada partisi ke-i

|S| = Jumlah Kasus dalam S

$$\boldsymbol{Entropy\ S = \sum_{i=1}^{n} - pi * log2\ pi}$$

Keterangan

S = himpunan kasus

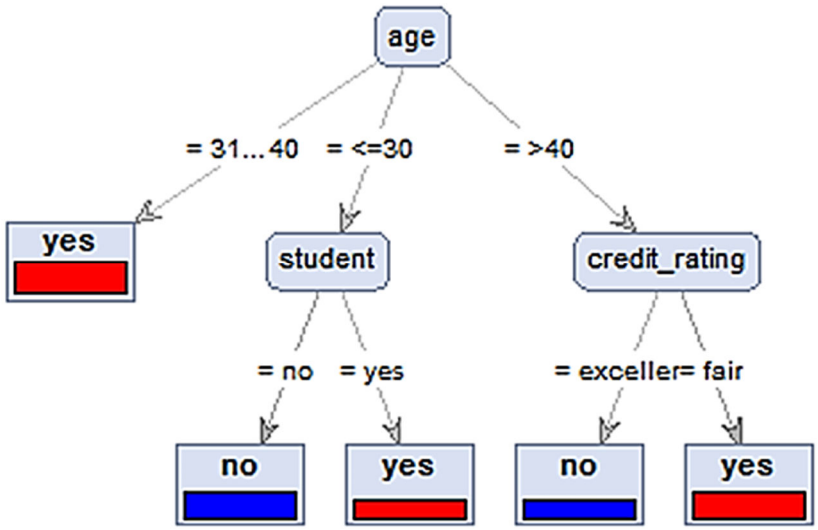
A = fitur

N = jumlah partisi

Pi = perbandingan dari Si terhadap S

Decision tree merupakan model klasifikasi yang berbentuk seperti pohon, dimana *decision tree* mudah untuk dimengerti meskipun oleh pengguna yang belum ahli sekalipun dan lebih efisien dalam menginduksi data (C. Sammut, 2021). Dari akhir tahun 1970 sampai awal 1980-an J.Ross Quinlan, melakukan pengembangan terhadap algoritma *decision tree* yakni ID3 (*Interative Dichotomisemiser*). Kemudian Quinlan juga menghadirkan algoritma C4.5, yang menjadi awal dari

algoritma *supervised learning* yang terbaru. Di tahun 1984 sebuah kelompok statistic (L.Breiman, J. Fridman, R. Olshen dan C.Stone) mempublikasikan *Classification and Regresssion Tree* (CART), yang mengambarkan generasi *binary decision tree* seperti gambar II.3 dibawah ini :



Gambar II.3. Konsep Pohon Keputusan Sederhana

Tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4. 5 yaitu:

1. Mempersiapkan data training, data training biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung total *entropy* sebelum atau dicari masing-masing *entropy class*

   $H(T) = - \sum Pj \log_2 (Pj)$

   Keterangan:

   H = Himpunan kasus

   T = Atribut

Pj = Proporsi dari Hj terhadap H

3. Hitung nilai Gain dengan information gain dengan rata-rata:

   Gain Average = H(T) - Hsaving(T)

   Keterangan:

   H(T) = Total Entropy

   Hsaving(T) = Total Gain information untuk masing-masing atribut

4. Ulangi langkah ke 2 dan ke 3 hingga semua tupel terpartisi.

   Proses partisi pohon keputusan akan berhenti disaat:

   a. Semua tupel dalam node N mendapatkan kelas yang sama.

   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.

   c. Tidak ada tupel di dalam cabang yang kosong

Contoh Penerapan Algoritma C4.5 dari dataset pada table II.2 dibawah ini :

Table II.2. Dataset Contoh Algortima C4.5

No	Outlook	Temperature	Humidity	windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No

Table II.2 Dataset Contoh Algortima C4.5 (Lanjutan)

No	Outlook	Temperature	Humidity	windy	Play
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Penghitungan *entropy* dari dataset yang terlampir pada table II.2 diatas sebagai berikut ini :

Untuk kelas *Outlook*

Entropy (Total) = (- 4/14 x log2 (4/14)) + (- 10/14 x log2 (10/14))

= 0.863

Entropy (Cloudy) = (- 0/4 x log2 (0/4)) + (- 4/4 x log2 (4/4))

= 0.000

Entropy (Rainy) = (- 1/5 x log2 (1/5)) + (- 4/5 x log2 (4/5))

= 0.721

Entropy (Sunny) = (- 3/5 x log2 (3/5)) + (- 2/5 x log2 (2/5))

= 0.970

Untuk kelas *Temperature*

Entropy (Cool) = (- 0/4 x log2 (0/4)) + (- 4/4 x log2 (4/4))

= 0.000

Entropy (Hot) = (- 2/4 x log2 (2/4)) + (- 2/4 x log2 (2/4))

= 1.000

Entropy (Mild) = (- 2/6 x log2 (2/6)) + (- 4/6 x log2 (4/6))

= 0.918

Untuk kelas *Humidity*

Entropy (High) = (- 4/7) x log2 (4/7) + (- 3/7 x log2 (3/7))

= 0.985

Humidity Entropy (Normal) = (- 0/7 x log2 (0/7) + (- 7/7 x log2 (7/7))

= 0.000

Untuk kelas *Windy*

Entropy (False) = (- 2/8 x log2 (2/8) + (- 6/8 x log2 (6/8))

= 0.811

Entropy (True) = (- 4/6 x log2 (4/6) + (- 2/6 x log2 (2/6))

= 0.918

Penghitungan Gain

Gain (Total, *Outlook)* = Entropy(Total-∑ |*Outlook*| / |Total| x *Entropy* (O*utlook*)

= 0.863 – ((4/14x 0.000) + (5/14 x 0.721) + (5/14 x 0.970))

= 0.258

Gain (Total, *Temperature*) = *Entropy* (Total) - ∑ |Temp| / |Total| x *Entropy* (Temp)

= 0.863–((4/14x 0.000) + (4/14 x 1.000) + (6/14 x 0.918))

= 0.183

Gain(Total,*humidity*) = Entropy(Total)-∑ |*Humidity*| / |Total| *xEntropy*(*Humidity*)

= 0.863 – ((7/14 x 0.985) + (7/14 x 0.000))

$= 0.370$

Gain (Total,*Windy*) $= Entropy$ (Total) $-\sum |Windy| / |$ Total $|$ x *Entropy*(*Windy*)

$= 0.863 - ((8/14 \times 0.811) + (6/14 \times 0.918))$

$= 0.005$

Dari hasil penghitungan di atas, dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah *Humidity* yaitu sebesar 0.370, dengan demikian *Humidity* dapat menjadi *node* akar.

II.5. Evaluasi dan Validasi Hasil

Pengukuran model *data mining* secara umum dapat dilakukan dengan 3 kriteria:

a. Akurasi (*Accuracy*)
    1. Ukuran dari seberapa baik model mengkorelasikan antara hasil dengan atribut dalam data yang telah disediakan.
    2. Terdapat berbagai model akurasi, tetapi semua model akurasi bergantung pada data yang digunakan.

b. Kehandalan (*Reliability*)
    a. Ukuran dimana model data mining diterapkan pada dataset yang berbeda
    b. Model data mining dapat diandalkan jika menghasilkan pola umum yang sama terlepas dari data testing yang disediakan

c. Kegunaan (*Usefulness*)

    Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna. Keseimbangan diantara ketiganya diperlukan karena

belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna.

II.5.1 Confusion Matrix

Confusion Matrix adalah alat (*tools*) visualisasi yang biasa digunakan pada *supervised learning*. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya. Sedangkan (Sinaga et al., 2023) menjelaskan bahwa *confusion matrix* adalah tabel yang digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda. *Confusion matrix* merupakan *matrix* 2 dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kenyataan. Tabel II.3 adalah contoh tabel *confusion matrix* yang menunjukan klasifikasi dua kelas.

Tabel II.3 Model Confusion Matrix Prediksi Class

Klasifikasi	Class = Yess	Class = No
Class = Yes	A (True Positive – TP)	B (False Negative – FN)
Class = No	C (False Positive – FP)	D (True Negative – TN)

Keterangan:

TP = proporsi positif dalam data set yang diklasifikasikan positif

TN = proporsi negative dalam data set yang diklasifikasikan negative

FP = proporsi negatif dalam data set yang diklasifikasikan potitif

FN = proporsi negative dalam data set yang diklasifikasikan negatif

Berikut adalah persamaan model confusion matrix:

a. Nilai Accuracy adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{(II.6)}$$

b. Sensitivity digunakan untuk membandingkan proporsi TP terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:

$$\text{Sensitifity} = \frac{\text{TP}}{\text{T}P+FN} \quad \text{(II.7)}$$

c. Specificity digunakan untuk membandingkan proporsi TN terhadap tupel yang yang dihitung dengan menggunakan persamaan:

$$\text{Specificiy} = \frac{\text{TN}}{\text{T}N+FP} \quad \text{(II.8)}$$

d. PPV (Positive Predictive Value) adalah proporsi kasus dengan hasil diagnosa positif, yang dihitung dengan menggunakan persamaan:

$$\text{PPV} = \frac{\text{TP}}{\text{T}P+FP} \quad \text{(II.9)}$$

e. NPV (Negative Predictive Value) adalah proporsi kasus dengan hasil diagnosa negatif, yang dihitung dengan menggunakan persamaan:

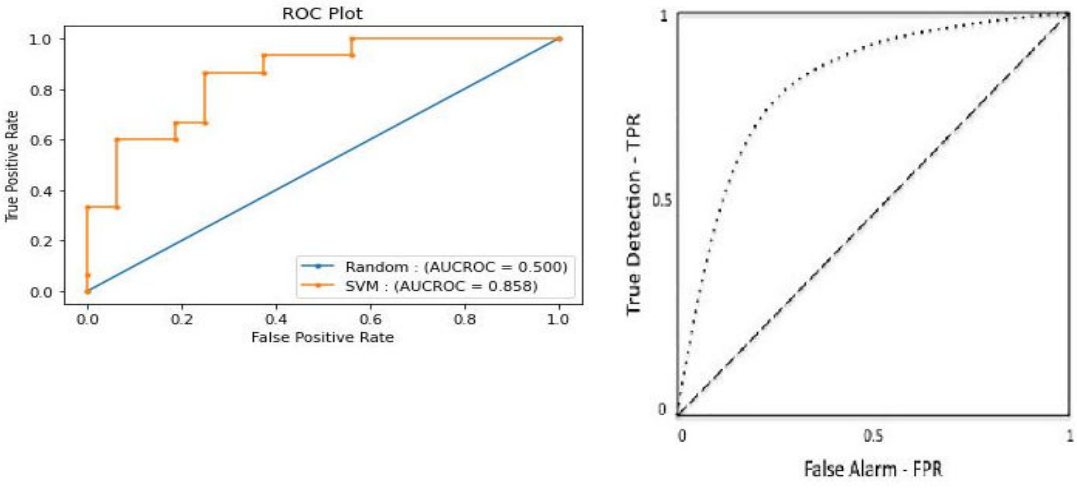
$$\text{NPV} = \frac{\text{TN}}{\text{T}N+FN} \quad \text{(II.10)}$$

II.5.2 ROC (*Receiver Operating Characteristic*)

Kurva ROC (*Receiver Operating Characteristic*) adalah alat visual yang berguna untuk membandingkan dua model klasifikasi hasil ekspresi dari *confusion matrix*. ROC adalah grafik dua dimensi dengan false positives sebagai garis horisontal dan *true positives* sebagai garis vertikal (Yoga Religia & Amali, 2021. ROC adalah ukuran numerik untuk membedakan kinerja

model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif.

Dengan kurva ROC, dapat melihat *trade off* antara tingkat dimana suatu model dapat mengenali tupel positif secara akurat dan tingkat dimana model tersebut salah mengenali tupel negatif sebagai tupel positif. Sebuah grafik ROC adalah plot dua dimensi dengan proporsi positif salah (FP) pada sumbu X dan proporsi positif benar (TP) pada sumbu Y. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus positif dan kasus negatif. nilai positif salah adalah tidak ada (FP = 0) dan nilai positif benar adalah tinggi (TP = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi negatif(0), dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi positif (1). Grafik ROC menggambarkan *trade off* antara manfaat (true positive) dan biaya (false positives). Berikut tampilan dua jenis kurva ROC (*discrete dan continous*) (Qadrini L et al., 2021).



Gambar II.4 Kurva ROC (*Receiver Operating Characteristic*)

Poin di atas garis diagonal merupakan hasil klasifikasi yang baik, sedangkan poin di bawah garis diagonal merupakan hasil klasifikasi yang buruk. Dapat disimpulkan bahwa satu poin pada kurva ROC adalah lebih baik daripada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas di dalam grafik. Untuk tingkat akurasi nilai AUC (*Area Under Curve*) dalam klasifikasi data mining dibagi menjadi lima kelompok yaitu:

a. 0.90 – 1.00 = Excellent Classification
b. 0.80 – 0.90 = Good Classification
c. 0.70 – 0.80 = Fair Classification
d. 0.60 – 0.70 = Poor Classification
e. 0.50 – 0.60 = Failure

II.6 Pelayanan Administrasi

Pelayanan administrasi adalah analisis yang mendalam mengenai konsep, praktik, dan dampak pelayanan administratif dalam berbagai situasi. Dalam tinjauan ini, fokus diberikan pada bagaimana administrasi diintegrasikan dengan pemberian layanan kepada individu, kelompok, atau entitas lain. Konsep dasar administrasi, yang meliputi tahap perencanaan, pengorganisasian, pelaksanaan, dan pengendalian, menjadi dasar dalam memberikan pelayanan yang efektif. Pelayanan administrasi melibatkan berbagai komponen, seperti proses komunikasi yang jelas dan efisien, pencatatan yang akurat, pengarsipan yang teratur, pelaporan berkala, dan penanganan masalah dengan cepat. Kualitas pelayanan administratif menjadi elemen kunci dalam tinjauan ini. Tingkat kepuasan pelanggan ditentukan oleh responsibilitas, kecepatan, dan akurasi dalam memberikan pelayanan. Penggunaan

teknologi modern juga semakin penting, di mana sistem informasi administrasi dapat mempermudah akses dan pengelolaan informasi. Administrasi membantu dalam pengambilan keputusan dengan memberikan data dan informasi yang diperlukan. Selain itu, administrasi memfasilitasi fungsi-fungsi umum seperti pengelolaan keuangan, manajemen sumber daya manusia, dan koordinasi aktivitas organisasi. Namun, ada tantangan dalam pelayanan administratif, termasuk birokrasi yang kompleks dan perubahan regulasi. Oleh karena itu, inovasi dalam proses administrasi diperlukan untuk mengatasi tantangan tersebut. Teknologi dan metode baru membuka peluang untuk memperbaiki efisiensi dan responsibilitas (Along, 2020).

Tinjauan ini juga mengedepankan perspektif pengguna layanan administratif. Dengan memahami harapan dan kebutuhan pelanggan, pelayanan administratif dapat diarahkan untuk memberikan manfaat yang lebih besar. Secara keseluruhan, tinjauan pustaka tentang "Pelayanan Administrasi" adalah penelusuran mendalam mengenai bagaimana administrasi berkontribusi dalam memberikan layanan yang efektif, bagaimana penggunaan teknologi mempengaruhi pelayanan, dan bagaimana tantangan-tantangan dihadapi dan diatasi. Hal ini penting untuk memastikan bahwa pelayanan administratif yang diberikan memiliki dampak positif pada operasional dan pengguna pelayanan (Maulina et al., 2021).

II.7 Kepuasan Pelanggan

Kepuasan Pelanggan adalah kerangka konseptual yang digunakan untuk memahami dan menganalisis persepsi, harapan, dan evaluasi pelanggan terhadap produk, layanan, atau pengalaman yang mereka dapatkan dari suatu perusahaan

atau penyedia layanan. Teori ini mengidentifikasi faktor-faktor yang memengaruhi tingkat kepuasan pelanggan dan bagaimana hal tersebut dapat berdampak pada hubungan jangka panjang antara pelanggan dan perusahaan. Tujuan utama dari Teori Kepuasan Pelanggan adalah untuk meningkatkan kualitas produk atau layanan berdasarkan pandangan dan harapan pelanggan (Astuti & Lutfi, 2020).

Berikut adalah beberapa poin penting dalam Teori Kepuasan Pelanggan:

a. Harapan dan Persepsi: Teori ini mengakui bahwa pelanggan membawa harapan dan ekspektasi tertentu saat menggunakan produk atau layanan. Persepsi pelanggan tentang sejauh mana harapan tersebut terpenuhi atau dilampaui akan mempengaruhi tingkat kepuasan mereka.
b. Faktor-Faktor Pengaruh: Teori ini mengidentifikasi faktor-faktor yang mempengaruhi kepuasan pelanggan. Faktor ini meliputi kualitas produk atau layanan, komunikasi dengan perusahaan, ketersediaan dan aksesibilitas, harga yang wajar, pengalaman pengguna, dan lain-lain.
c. Perbandingan Terhadap Ekspektasi: Kepuasan pelanggan seringkali bergantung pada perbandingan antara harapan awal mereka dengan apa yang sebenarnya mereka alami. Jika kinerja produk atau layanan melebihi harapan, kepuasan cenderung tinggi. Sebaliknya, jika kinerja di bawah ekspektasi, kepuasan cenderung rendah.
d. Dampak Kesetiaan Pelanggan: Teori Kepuasan Pelanggan mengakui hubungan antara kepuasan pelanggan dan tingkat kesetiaan mereka terhadap perusahaan. Pelanggan yang puas cenderung tetap menggunakan produk atau layanan dari

perusahaan yang sama dan bahkan mungkin merekomendasikannya kepada orang lain.

e. Umpan Balik dan Perbaikan: Teori ini mendorong perusahaan untuk mengumpulkan umpan balik dari pelanggan tentang pengalaman mereka. Informasi ini dapat digunakan untuk perbaikan terus-menerus dalam produk, layanan, dan proses.

f. Persepsi Nilai: Kepuasan pelanggan juga bergantung pada persepsi nilai yang diberikan oleh produk atau layanan. Ini mencakup keseimbangan antara manfaat yang diperoleh dan biaya atau usaha yang dikeluarkan oleh pelanggan.

g. Pentingnya Konsistensi: Teori ini menekankan pentingnya konsistensi dalam memberikan kualitas dan layanan yang dijanjikan kepada pelanggan. Perubahan yang tidak konsisten dalam produk atau layanan dapat memengaruhi kepuasan pelanggan.

h. Pentingnya Hubungan: Teori ini mengakui bahwa hubungan jangka panjang antara perusahaan dan pelanggan sangat penting. Kepuasan pelanggan saat ini dapat berdampak pada kesetiaan dan dukungan jangka panjang.

Secara keseluruhan, Teori Kepuasan Pelanggan memberikan kerangka kerja yang komprehensif untuk memahami bagaimana pelanggan menilai pengalaman mereka dengan produk atau layanan. Dengan memahami faktor-faktor yang mempengaruhi kepuasan pelanggan, perusahaan dapat mengambil langkah-langkah untuk meningkatkan kualitas produk, layanan, dan hubungan dengan pelanggan (Subawa & Sulistyawati, 2020)

II.8 Penelitian Sebelumnya

Tabel II.4 Ringkasan Penelitian Sebelumnya

No	Makalah	Metode	Hasil Penelitian
1	Implementasi Algoritma *Decision Tree* Dalam Klasifikasi Kompetensi Siswa (Rifa et al., 2022)	C4.5	Pendekatan penelitian ini menggunakan algoritma *decission tree*, dengan tujuan mendapatkan pola rekomendasi kompeten dan tidak kompeten. Berdasarkan hasil penelitian menjelaskan bahwa akurasi yang didapat yaitu sebesar 76,96 %
2	*Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method* (Irena & Erwin Budi Setiawan, 2020)	C4.5	Hasil akhir menunjukkan bahwa pengujian klasifikasi yang menggunakan pembobotan fitur dan seleksi fitur menghasilkan akurasi terbaik sebesar 72.91% dengan perbandingan data latih 90% dan data uji 10% (90:10) dan jumlah fitur yang digunakan sebanyak 5000 fitur unigram.
3	Klasifikasi Kepuasan Siswa Terhadap Kinerja Guru SMK Satrya Budi 2 Perdagangan Menggunakan Algoritma C4.5 (Kusuma et al., 2021)	C4.5	Nilai akurasi yang ditunjukkan terhadap kesesuaian Algoritma C4.5 dengan kepuasan siswa terhadap kinerja guru adalah 94%.
4	*Classification Based on Decision Tree Algorithm for Machine Learning* (Wilson et al., 2023)	C4.5	Akurasi terbaik yang dicapai untuk algoritma pohon keputusan adalah 99,93% ketika menggunakan repositori pembelajaran mesin sebagai kumpulan data.
5	*Covid Symptom Severity Using Decision Tree* (Wilson et al., 2023)	C4.5	Model prediksi menggunakan struktur hirarki. Konsepnya adalah mengubah data menjadi pohon keputusan atau aturan keputusan. hasil dari J48 sedikit lebih baik dari pohon *Hoeffding* dalam hal akurasi, presisi, dan recall. Sedangkan dari hasil *tree*

			view, Hoeffding Tree lebih sederhana dan jumlah node lebih sedikit dibandingkan J48
6	*Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis* (Umar & M. Adnan Nur, 2022)	Naïve Bayes	Berdasarkan hasil penerapan variasi *Naive Bayes* pada Data set Analisis Umum Indonesia dengan penentuan *min-df* dan *max-df*, didapatkan nilai akurasi tertinggi pada *Naive Bayes* multi nomial. Nilai rata-rata *min-df* yang memiliki akurasi tinggi adalah 0.0005 dan *max-df* 0.5.
7	Deteksi Sarung Samarinda Menggunakan Metode *Naive Bayes* Berbasis Pengolahan Citra (Septiarini et al., 2021)	Naïve Bayes	Nilai akurasi benar dan salah untuk pembagian data menggunakan *percentage split* yang berhasil dicapai adalah 0,987 dan 0,013. Sementara itu, pembagian data menggunakan *cross-validation* mampu dicapai nilai akurasi benar dan salah yaitu 0,984 dan 0,016.
8	Perbandingan Optimasi Feature Selection pada Naïve Bayes untuk Klasifikasi Kepuasan *Airline Passenger* (Yoga Religia & Amali, 2021)	Naïve Bayes	Algoritma Naïve Bayes dengan optimasi PSO, dimana diperoleh nilai akurasi sebesar 86.13%, nilai presisi sebesar 87.90%, nilai recall sebesar 87.29%, dan nilai AUC sebesar 0.923.
9	*A naive Bayes classifier for identifying Class II YSOs* (Wilson et al., 2023)	Naïve Bayes	Kurva ROC meningkat dengan cepat hingga hampir satu dengan area di bawah kurva sekitar 0,998 atau lebih baik, yang mengindikasikan bahwa pengklasifikasi efisien dalam mengidentifikasi kandidat YSO Kelas II.
10	*AI-based smart prediction of clinical disease using random forest classifer and Naive Bayes* (Wilson et al., 2023)	Naïve Bayes dan *Random Forest*	Analisis kinerja data penyakit untuk kedua algoritma dihitung dan dibandingkan. Hasil dari simulasi menunjukkan keefektifan klasifikasi teknik klasifikasi pada dataset, serta sifat dan kompleksitas dataset yang digunakan.

Dalam melakukan penelitian, penulis menggunakan beberapa penelitian yang relevan sebagai tinjauan pustaka dan mencari kelamahan dari penelitian sebelumnya. Dalam Tabel II.4 terlihat dari 10 (sepuluh) penelitian yang dilakukan hanya 1 (satu) yang melakukan komparasi metode. Meskipun dalam penelitian tersebut nilai akurasi dari metode yang diteliti sudah baik, penulis ingin membandingan komparasi dari kedua metode tersebut secara langsung yaitu Naïve Bayes dan C4.5 dengan menggunakan dataset dan data training terbaru dari SMK Swasta siti Banun dengan menggunakan Pyhton dan library sklearn sehingga akurasi dari kedua metode tersebut apakah terdapat perbedaan yang sangat signifikan dari penelitian sebelumnya, tentunya juga dengan melakukan perbaikan-perbaikan yang overfitting.

II.9 Ringkasan Bab

Pada Bab ini telah dijelaskan kajian teoritis dari beberapa istilah yang digunakan yang berhubungan dengan penelitian ini. Sebagai tambahan informasi untuk tambahan referensi, penulis juga menggali informasi tetang penelitian yang terdahulu yang ada hubungannya dengan penelitian ini sehingga ada tambahan informasi untuk melakukan proses penelitian menjadi lebih baik. Berbekal dari referensi maka penulis pada Bab selanjutnya membuat metodologi penelitian sebagai kerangka kerja dalam melakukan penelitian ini.