TUGAS ANALISIS DATA MINING

Mata Kuliah : Data Mining

Topik : Implementasi Algoritma Data Mining Menggunakan Python

Nama Mahasiswa : Ramdhani lutfi

NPM : 231510024

Dosen Pengampu : Erlin Elisa, S.Kom., M.Kom.

Link Colab :

https://colab.research.google.com/drive/15g8vrspBSewglqc-yrqxeuKk15PfFyGO?usp=sharing

Link Dataset : https://drive.google.com/file/d/1DtlFtpqyFWoboywEdeGUpP6WgIjkW9qF/view?usp=drivesdk

1. Klasifikasi Tingkat Kepuasan Penumpang Maskapai Menggunakan Algoritma Random Forest

2. Latar Belakang Masalah

Industri penerbangan merupakan salah satu sektor jasa yang sangat bergantung pada tingkat kepuasan pelanggan. Kepuasan penumpang menjadi indikator penting dalam menilai kualitas layanan maskapai karena berpengaruh terhadap loyalitas pelanggan, citra perusahaan, serta keberlanjutan bisnis maskapai penerbangan (Manurung, 2025). Seiring dengan meningkatnya jumlah penumpang, maskapai penerbangan mengumpulkan data dalam jumlah besar yang mencakup karakteristik penumpang, pengalaman perjalanan, serta penilaian terhadap berbagai aspek layanan yang diberikan.

Namun, data kepuasan pelanggan yang besar dan kompleks tersebut sulit dianalisis secara manual. Tanpa teknik analisis yang tepat, maskapai akan mengalami kesulitan dalam mengidentifikasi faktor-faktor yang memengaruhi kepuasan penumpang serta memprediksi tingkat kepuasan pelanggan secara akurat (Setiono, 2022). Oleh karena itu, diperlukan penerapan data mining untuk mengolah data tersebut secara sistematis dan menghasilkan informasi yang bernilai bagi pengambilan keputusan.

Salah satu algoritma klasifikasi yang banyak digunakan dalam analisis kepuasan pelanggan adalah Random Forest. Algoritma ini mampu menangani data dengan jumlah atribut yang besar, memiliki tingkat akurasi yang baik, serta mengurangi risiko overfitting dibandingkan metode klasifikasi tunggal (Silvana et al., 2025). Dengan menerapkan algoritma Random Forest pada dataset kepuasan penumpang maskapai, diharapkan dapat diperoleh model klasifikasi yang mampu memprediksi tingkat kepuasan penumpang secara akurat dan mendukung pengambilan keputusan berbasis data

3. Rumusan Masalah

1. Bagaimana proses eksplorasi dan preprocessing data pada dataset kepuasan penumpang maskapai?
2. Bagaimana implementasi algoritma Random Forest dalam mengklasifikasikan tingkat kepuasan penumpang maskapai?
3. Bagaimana performa algoritma Random Forest dalam memprediksi kepuasan penumpang berdasarkan metrik evaluasi klasifikasi?

4. Tujuan Penelitian

1. Melakukan eksplorasi dan preprocessing data pada dataset kepuasan penumpang maskapai.
2. Membangun model klasifikasi tingkat kepuasan penumpang menggunakan algoritma Random Forest.
3. Mengevaluasi performa model klasifikasi menggunakan metrik evaluasi seperti accuracy, precision, recall, dan F1-score.

5. Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset kepuasan penumpang maskapai penerbangan yang diperoleh dari platform Kaggle. Dataset ini berisi data hasil survei terhadap penumpang maskapai yang mencakup karakteristik penumpang serta penilaian terhadap berbagai aspek layanan penerbangan. Dataset tersebut banyak digunakan dalam penelitian data mining dan machine learning karena memiliki jumlah data yang besar serta struktur data yang sesuai untuk permasalahan klasifikasi.

Jumlah data pada dataset ini terdiri dari sekitar **129.000 record** dengan **23 atribut**, yang mencakup satu variabel target dan beberapa variabel input. Variabel target pada penelitian ini adalah **satisfaction**, yang merepresentasikan tingkat kepuasan penumpang dan memiliki dua kelas, yaitu *satisfied* dan *neutral or dissatisfied*. Sementara itu, variabel input meliputi karakteristik penumpang seperti jenis pelanggan, tipe perjalanan, kelas penerbangan, serta penilaian terhadap kualitas layanan seperti kenyamanan kursi, layanan dalam penerbangan, ketepatan waktu, dan fasilitas pendukung lainnya.

Dataset ini memiliki kombinasi data numerik dan kategorikal sehingga memerlukan tahapan preprocessing sebelum digunakan dalam pemodelan. Dengan karakteristik tersebut, dataset ini dinilai sesuai untuk diterapkan algoritma klasifikasi Random Forest dalam memprediksi tingkat kepuasan penumpang maskapai penerbangan.

Sumber Dataset : Kaggle – Airline Passenger Satisfaction Dataset
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

tabel Karakteristik Dataset

Tabel berikut menampilkan beberapa contoh atribut utama dalam dataset, sedangkan atribut lainnya tidak ditampilkan secara keseluruhan.

No	Atribut	Tipe Data	Keterangan
1	Gender	Kategorikal	Jenis kelamin penumpang
2	Age	Numerik	Usia penumpang
3	Customer Type	Kategorikal	Jenis pelanggan
4	Type of Travel	Kategorikal	Tujuan perjalanan
5	Class	Kategorikal	Kelas penerbangan
6	Flight Distance	Numerik	Jarak penerbangan
…	…	…	…
23	Satisfaction	Label	Tingkat kepuasan penumpang

6. Metodologi

Metodologi penelitian yang digunakan dalam tugas ini adalah metode **data mining** dengan pendekatan **klasifikasi**. Pendekatan data mining digunakan untuk mengekstraksi pola dan pengetahuan dari kumpulan data berukuran besar guna mendukung proses pengambilan keputusan (Han, Kamber, & Pei, 2012). Pendekatan klasifikasi dipilih karena penelitian ini bertujuan untuk memprediksi tingkat kepuasan penumpang maskapai penerbangan ke dalam kelas tertentu berdasarkan atribut yang tersedia.

Tahapan penelitian disusun secara sistematis mengikuti alur umum proses data mining yang banyak diterapkan dalam penelitian-penelitian di Indonesia, meliputi pemahaman data, preprocessing data, pemodelan, serta evaluasi hasil (Suyanto, 2017).

6.1 Data Understanding

Tahap *data understanding* dilakukan untuk memahami struktur dan karakteristik dataset yang digunakan. Pada tahap ini dilakukan identifikasi jumlah data, jumlah atribut, tipe data setiap atribut, serta penentuan variabel input dan variabel target. Selain itu, dilakukan analisis awal untuk mengetahui distribusi data dan potensi permasalahan dalam dataset, seperti ketidakseimbangan kelas dan keberadaan data kosong. Tahap ini bertujuan untuk memastikan data yang digunakan sesuai dengan tujuan penelitian.

6.2 Data Preprocessing

Tahap preprocessing data bertujuan untuk meningkatkan kualitas data sebelum proses pemodelan dilakukan. Proses preprocessing yang dilakukan meliputi:

1. Pemeriksaan dan penanganan *missing value* pada dataset.
2. Penghapusan atribut yang tidak relevan atau tidak berpengaruh terhadap proses klasifikasi.
3. Transformasi data kategorikal menjadi data numerik menggunakan teknik *encoding*.
4. Normalisasi atau standarisasi data numerik apabila diperlukan.

Tahapan preprocessing merupakan bagian penting dalam data mining karena kualitas data sangat mempengaruhi performa model yang dihasilkan (Prasetyo, 2014).

6.3 Feature Selection

Pada penelitian ini tidak dilakukan proses *feature selection* secara khusus. Seluruh atribut yang relevan digunakan dalam proses pemodelan karena algoritma Random Forest memiliki kemampuan untuk menangani data dengan jumlah fitur yang cukup banyak serta menyediakan mekanisme internal dalam menentukan tingkat kepentingan fitur (*feature importance*).

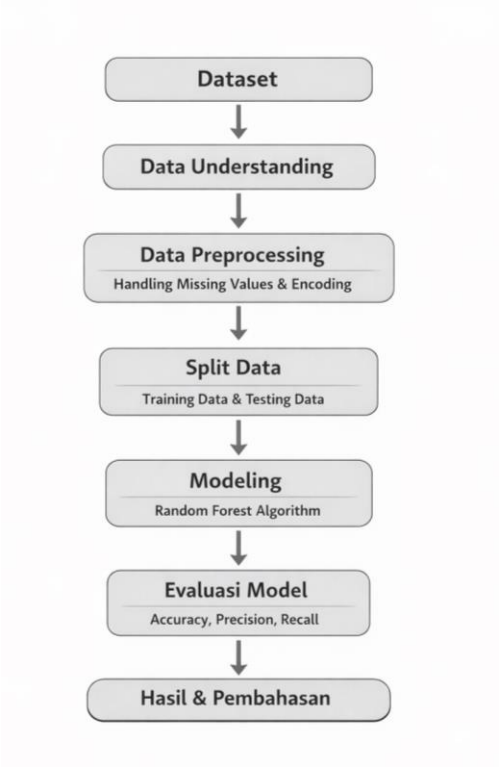
6.4 Pemodelan

Tahap pemodelan dilakukan menggunakan algoritma **Random Forest**, yaitu algoritma klasifikasi berbasis *ensemble learning* yang membangun banyak pohon keputusan dan menggabungkan hasil prediksinya untuk meningkatkan akurasi serta mengurangi risiko *overfitting* (Breiman, 2001). Algoritma Random Forest dipilih karena memiliki performa yang baik dalam menangani data berdimensi tinggi dan data dengan karakteristik kompleks, sebagaimana ditunjukkan dalam berbagai penelitian data mining (Sari & Wibowo, 2020).

6.5 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja model klasifikasi yang telah dibangun. Metrik evaluasi yang digunakan dalam penelitian ini meliputi **akurasi**, **precision**, **recall**, dan **confusion matrix**.

Penggunaan metrik evaluasi ini bertujuan untuk menilai sejauh mana model mampu mengklasifikasikan tingkat kepuasan penumpang secara tepat dan konsisten (Gorunescu, 2011).



7. Implementasi Python

7.1 Lingkungan Pengembangan

Pada penelitian ini, proses implementasi dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan platform Google Colab. Google Colab dipilih karena menyediakan lingkungan komputasi berbasis cloud yang telah terintegrasi dengan berbagai library data science seperti Pandas, NumPy, Scikit-Learn, serta Matplotlib sehingga memudahkan proses pengolahan dan analisis data.

7.2 Import Library

Tahap awal implementasi adalah melakukan import library yang dibutuhkan. Library Pandas dan NumPy digunakan untuk pengolahan data, Matplotlib dan Seaborn untuk visualisasi, sedangkan Scikit-Learn digunakan untuk proses pembagian data, pelatihan model, serta evaluasi performa model.

```
df = pd.read_csv('/content/train.csv.csv')
```

```python
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

7.3 Load Dataset

Dataset yang digunakan pada penelitian ini dimuat dalam format CSV. Dataset tersebut berisi data karakteristik penumpang dan tingkat kepuasan terhadap layanan maskapai penerbangan.

```python
df.head()
```

	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
0	1	0	13	1	2	460	3	4	3	1	...	5	4	3	4	4	5	5	25	18.0
1	1	1	25	0	0	235	3	2	3	3	...	1	1	5	3	1	4	1	1	6.0
2	0	0	26	0	0	1142	2	2	2	2	...	5	4	3	4	4	4	5	0	0.0
3	0	0	25	0	0	562	2	5	5	5	...	2	2	5	3	1	4	2	11	9.0
4	1	0	61	0	0	214	3	3	3	3	...	3	3	4	4	3	3	3	0	0.0

5 rows × 23 columns

7.4 Exploratory Data Analysis (EDA)

Tahap Exploratory Data Analysis (EDA) dilakukan untuk memahami struktur data, tipe data, serta distribusi kelas pada variabel target. EDA juga digunakan untuk mengidentifikasi adanya nilai kosong (missing value) pada dataset.

Data Understanding

a. Informasi Struktur Dataset (df.info())

```
[5]
✓ 0s
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Unnamed: 0                        103904 non-null  int64
 1   id                                103904 non-null  int64
 2   Gender                            103904 non-null  object
 3   Customer Type                     103904 non-null  object
 4   Age                               103904 non-null  int64
 5   Type of Travel                    103904 non-null  object
 6   Class                             103904 non-null  object
 7   Flight Distance                   103904 non-null  int64
 8   Inflight wifi service             103904 non-null  int64
 9   Departure/Arrival time convenient 103904 non-null  int64
 10  Ease of Online booking            103904 non-null  int64
 11  Gate location                     103904 non-null  int64
 12  Food and drink                    103904 non-null  int64
 13  Online boarding                   103904 non-null  int64
 14  Seat comfort                      103904 non-null  int64
 15  Inflight entertainment            103904 non-null  int64
 16  On-board service                  103904 non-null  int64
 17  Leg room service                  103904 non-null  int64
 18  Baggage handling                  103904 non-null  int64
 19  Checkin service                   103904 non-null  int64
 20  Inflight service                  103904 non-null  int64
 21  Cleanliness                       103904 non-null  int64
 22  Departure Delay in Minutes        103904 non-null  int64
 23  Arrival Delay in Minutes          103594 non-null  float64
 24  satisfaction                      103904 non-null  object
dtypes: float64(1), int64(19), object(5)
memory usage: 19.8+ MB
```

Dataset yang digunakan memiliki **103.904 baris data (record)** dan **25 kolom (atribut)**. Berdasarkan hasil df.info(), dataset terdiri dari tiga jenis tipe data, yaitu **integer (int64)**, **floating point (float64)**, dan **object (kategorikal)**.

Atribut numerik didominasi oleh data bertipe **int64**, seperti usia penumpang, jarak penerbangan, serta penilaian layanan maskapai. Sementara itu, atribut bertipe **object** digunakan untuk data kategorikal seperti *Gender*, *Customer Type*, *Type of Travel*, *Class*, dan *Satisfaction* sebagai variabel target.

Selain itu, penggunaan memori dataset tercatat sekitar **19.8 MB**, yang masih tergolong efisien untuk diproses menggunakan Python dan Google Colab.

b. Pemeriksaan Missing Value

```
[6]
✓ 0s
df.isnull().sum()
```

	0
Unnamed: 0	0
id	0
Gender	0
Customer Type	0
Age	0
Type of Travel	0
Class	0
Flight Distance	0
Inflight wifi service	0
Departure/Arrival time convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Online boarding	0
Seat comfort	0
Inflight entertainment	0
On-board service	0
Leg room service	0
Baggage handling	0
Checkin service	0
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	310
satisfaction	0

dtype: int64

Berdasarkan hasil pemeriksaan menggunakan fungsi df.isnull().sum(), diketahui bahwa **sebagian besar atribut tidak memiliki nilai kosong (missing value)**. Namun, terdapat **310 data kosong pada atribut *Arrival Delay in Minutes***.

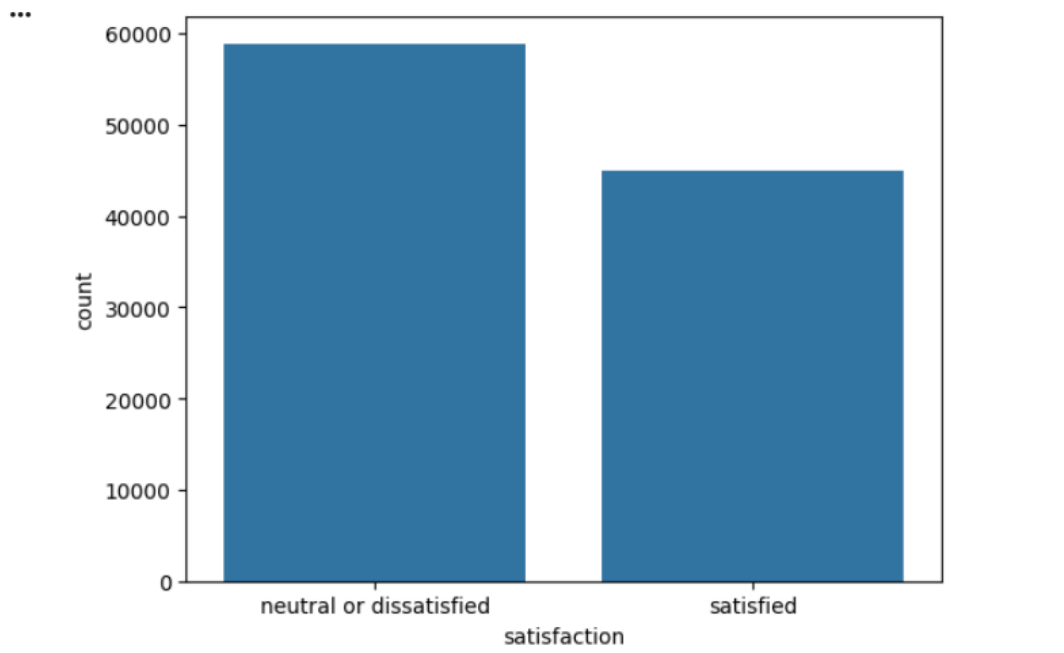
Keberadaan missing value ini perlu ditangani pada tahap preprocessing agar tidak mengganggu proses pelatihan model. Atribut lainnya memiliki nilai lengkap, sehingga tidak memerlukan penanganan khusus terkait missing value.

Exploratory Data Analysis (EDA)

Distribusi Data Kepuasan Penumpang

Visualisasi distribusi variabel target dilakukan menggunakan grafik *countplot* untuk melihat keseimbangan kelas pada data kepuasan penumpang.

```
sns.countplot(x='satisfaction', data=df)
plt.show()
```



Gambar menunjukkan distribusi data kepuasan penumpang yang terdiri dari dua kelas, yaitu *satisfied* dan *neutral or dissatisfied*. Berdasarkan grafik tersebut, jumlah penumpang dengan kategori *neutral or dissatisfied* lebih banyak dibandingkan dengan kategori *satisfied*. Hal ini menunjukkan bahwa dataset memiliki distribusi kelas yang relatif tidak seimbang, meskipun perbedaannya tidak terlalu ekstrem.

7.5 Preprocessing Data

Preprocessing data dilakukan untuk memastikan data siap digunakan dalam proses pemodelan. Tahapan preprocessing meliputi penghapusan kolom yang tidak relevan, penanganan missing value, serta transformasi data kategorikal menjadi numerik menggunakan teknik encoding.

Data Preprocessing

1. Penghapusan Kolom ID

Kolom **id** dan **Unnamed: 0** dihapus dari dataset karena **tidak memiliki kontribusi terhadap proses klasifikasi kepuasan penumpang**. Kolom tersebut hanya berfungsi sebagai identitas atau indeks data dan tidak mengandung informasi yang relevan untuk memprediksi variabel target.

Keberadaan kolom identitas dapat menyebabkan **noise** pada model dan berpotensi memengaruhi hasil pelatihan secara tidak tepat. Oleh karena itu, penghapusan kolom ini dilakukan sebelum tahap pemodelan.

2. Pengisian Missing Value dengan Nilai Rata-rata

Berdasarkan hasil pemeriksaan missing value, ditemukan **310 data kosong pada atribut Arrival Delay in Minutes**. Untuk mengatasi hal tersebut, dilakukan pengisian missing value menggunakan **nilai rata-rata (mean)** dari atribut tersebut.

Metode pengisian dengan rata-rata dipilih karena:

1. Jumlah missing value relatif kecil dibandingkan total data
2. Atribut bertipe numerik
3. Tidak mengubah distribusi data secara signifikan

Pendekatan ini bertujuan untuk menjaga jumlah data tetap utuh tanpa menghapus baris yang berpotensi penting.

3. Label Encoding

Penjelasan

Label Encoding digunakan untuk mengubah data kategorikal menjadi numerik agar dapat diproses oleh algoritma machine learning.

```
[ ]
le = LabelEncoder()

for col in df.select_dtypes(include='object').columns:
    df[col] = le.fit_transform(df[col])
```

7.6 Split Data (Train dan Test)

Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Pembagian ini bertujuan untuk menguji kemampuan model dalam melakukan generalisasi terhadap data baru.

```
[ ]
X = df.drop('satisfaction', axis=1)
y = df['satisfaction']
```

```
[ ]
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

7.7 Pelatihan Model

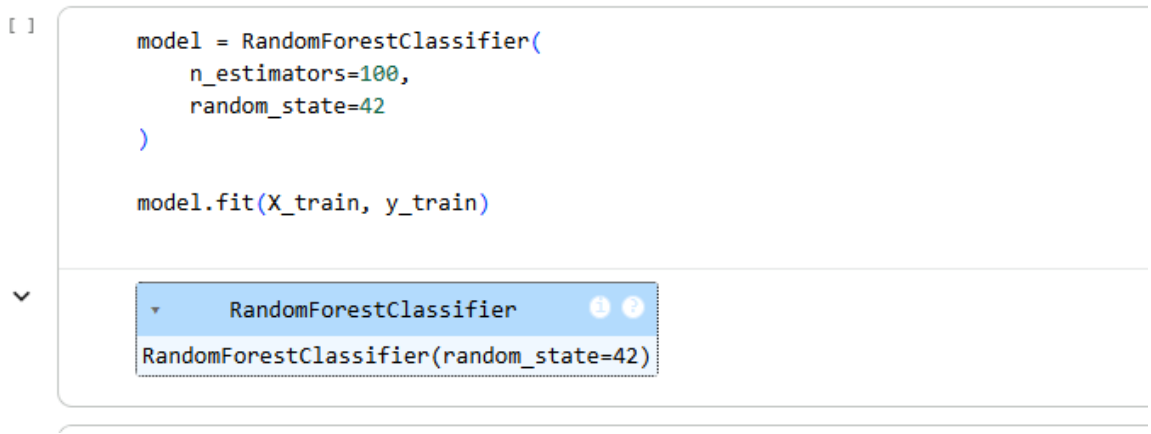
Tahap pelatihan model dilakukan menggunakan algoritma **Random Forest**, yaitu metode klasifikasi berbasis *ensemble learning* yang mengombinasikan sejumlah pohon keputusan (*decision tree*) untuk menghasilkan prediksi yang lebih akurat dan stabil. Pada tahap ini, model dilatih menggunakan data latih (*training data*) yang telah melalui proses preprocessing dan pembagian data sebelumnya.

Algoritma Random Forest bekerja dengan membangun banyak pohon keputusan secara acak berdasarkan subset data dan fitur yang berbeda, kemudian menggabungkan hasil prediksi dari seluruh pohon tersebut melalui mekanisme *majority voting*. Pendekatan ini

bertujuan untuk mengurangi variansi model dan meminimalkan risiko *overfitting* yang sering terjadi pada model pohon keputusan tunggal (Breiman, 2001).

Dalam penelitian ini, parameter utama yang digunakan pada model Random Forest adalah jumlah pohon (*n_estimators*) sebanyak 100 dan nilai *random state* untuk memastikan hasil pelatihan dapat direproduksi. Proses pelatihan model dilakukan menggunakan data latih dengan tujuan agar model mampu mempelajari pola hubungan antara atribut karakteristik dan layanan penerbangan dengan tingkat kepuasan penumpang.

Hasil dari tahap ini berupa model Random Forest terlatih yang selanjutnya digunakan untuk melakukan prediksi pada data uji (*testing data*) serta dievaluasi kinerjanya menggunakan metrik evaluasi klasifikasi.



7.8 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja algoritma **Random Forest** dalam mengklasifikasikan tingkat kepuasan penumpang maskapai penerbangan. Evaluasi dilakukan menggunakan data uji (*testing data*) yang sebelumnya tidak digunakan dalam proses pelatihan model, sehingga hasil evaluasi mencerminkan kemampuan generalisasi model terhadap data baru.

Metrik evaluasi yang digunakan dalam penelitian ini meliputi **Accuracy**, **Precision**, **Recall**, dan **F1-Score**. Nilai accuracy digunakan untuk mengetahui tingkat ketepatan model secara keseluruhan dalam melakukan klasifikasi. Sementara itu, precision dan recall digunakan untuk mengukur ketepatan dan kelengkapan prediksi model pada masing-masing kelas kepuasan penumpang. F1-Score digunakan sebagai metrik gabungan yang merepresentasikan keseimbangan antara precision dan recall.

Perhitungan metrik evaluasi dilakukan menggunakan fungsi accuracy_score() dan classification_report() dari library Scikit-Learn. Hasil evaluasi ini menjadi dasar dalam menilai performa model Random Forest sebelum dilakukan analisis lebih lanjut pada bab hasil dan pembahasan.

```
y_pred = model.predict(X_test)
```

```
accuracy = accuracy_score(y_test, y_pred)
accuracy
```

```
0.9621288677157018
```

```
print(classification_report(y_test, y_pred))
```

```
                precision    recall  f1-score   support

           0       0.95      0.98      0.97     11713
           1       0.97      0.94      0.96      9068

    accuracy                           0.96     20781
   macro avg       0.96      0.96      0.96     20781
weighted avg       0.96      0.96      0.96     20781
```

7.9 Visualisasi Confusion Matrix

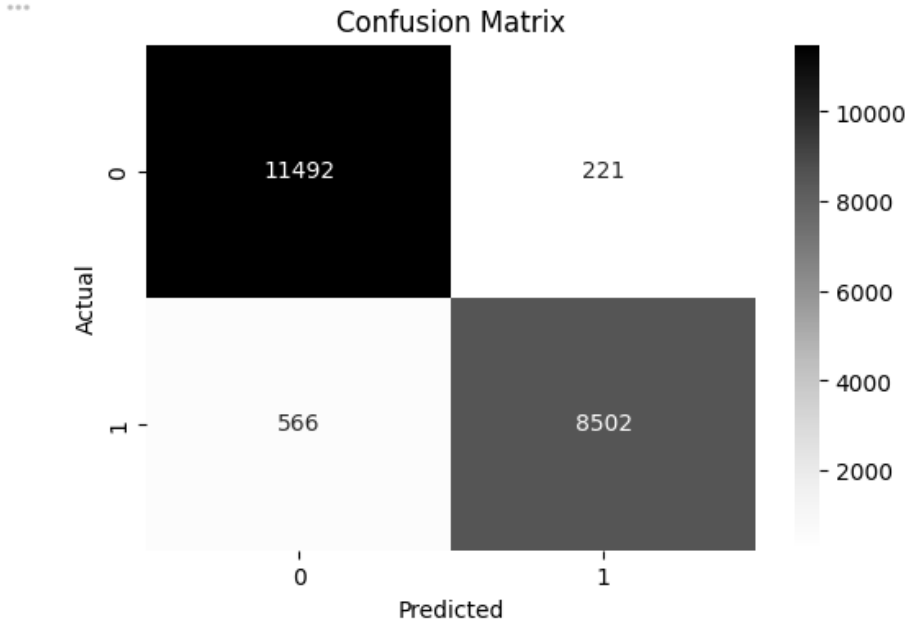
Confusion matrix digunakan untuk memberikan gambaran visual mengenai kinerja model Random Forest dalam mengklasifikasikan tingkat kepuasan penumpang. Confusion matrix menampilkan jumlah prediksi yang benar dan salah untuk setiap kelas, yaitu kelas *satisfied* dan *neutral or dissatisfied*.

Pada penelitian ini, confusion matrix ditampilkan dalam bentuk grafik *heatmap* untuk mempermudah interpretasi hasil klasifikasi. Sumbu horizontal merepresentasikan kelas hasil prediksi model, sedangkan sumbu vertikal menunjukkan kelas aktual dari data uji. Nilai diagonal pada confusion matrix menunjukkan jumlah prediksi yang benar, sedangkan nilai di luar diagonal menunjukkan kesalahan klasifikasi yang dilakukan oleh model.

Visualisasi confusion matrix ini membantu dalam memahami pola kesalahan prediksi yang terjadi serta memberikan informasi tambahan mengenai kekuatan dan kelemahan model Random Forest dalam membedakan tingkat kepuasan penumpang maskapai penerbangan.

```
[ ]
cm = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Greys')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```



8. Hasil dan Pembahasan

8.1 Hasil Evaluasi Model

Berdasarkan hasil pelatihan dan pengujian model klasifikasi menggunakan algoritma **Random Forest**, diperoleh beberapa metrik evaluasi untuk mengukur kinerja model dalam memprediksi tingkat kepuasan penumpang maskapai penerbangan.

Evaluasi dilakukan menggunakan data uji (*testing data*) dengan metrik **Accuracy, Precision, Recall, dan F1-Score**. Metrik-metrik ini digunakan untuk menilai ketepatan model dalam melakukan klasifikasi serta keseimbangan antara prediksi yang benar dan kesalahan prediksi.

[23]
0s

```
from sklearn.metrics import accuracy_score, classification_repor

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.9621288677157018
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     11713
           1       0.97      0.94      0.96      9068

    accuracy                           0.96     20781
   macro avg       0.96      0.96      0.96     20781
weighted avg       0.96      0.96      0.96     20781
```

[24]
0s

```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test, y_pred)
accuracy
```

```
0.9621288677157018
```

Algoritma	Akurasi	Precision	Recall	F1-Score
Random Forest	0.9621	0.96	0.96	0.96

8.2 Interpretasi Akurasi Model

Berdasarkan hasil evaluasi, algoritma **Random Forest** menghasilkan nilai akurasi sebesar **96,21%**. Nilai ini menunjukkan bahwa model mampu mengklasifikasikan tingkat kepuasan penumpang dengan tingkat ketepatan yang sangat tinggi.

Tingginya nilai akurasi mengindikasikan bahwa kombinasi berbagai pohon keputusan dalam Random Forest mampu menangkap pola hubungan antara atribut layanan penerbangan dan tingkat kepuasan penumpang secara efektif. Dengan demikian, model memiliki kemampuan generalisasi yang baik terhadap data baru.

8.3 Interpretasi Precision dan Recall

Nilai **precision sebesar 0,96** menunjukkan bahwa sebagian besar prediksi yang dihasilkan model untuk kelas tertentu adalah benar, sehingga tingkat kesalahan prediksi positif relatif rendah. Hal ini penting untuk memastikan bahwa model tidak memberikan banyak prediksi kepuasan yang keliru.

Sementara itu, nilai **recall sebesar 0,96** menunjukkan bahwa model mampu mengenali hampir seluruh data kepuasan penumpang yang sebenarnya. Nilai recall yang tinggi menandakan bahwa model memiliki sensitivitas yang baik dalam mendeteksi kedua kelas kepuasan.

Keseimbangan antara precision dan recall menunjukkan bahwa model Random Forest tidak hanya akurat, tetapi juga konsisten dalam mengklasifikasikan data.

8.4 Interpretasi F1-Score

Nilai **F1-Score sebesar 0,96** menunjukkan keseimbangan yang sangat baik antara precision dan recall. F1-Score digunakan sebagai indikator performa model secara keseluruhan, terutama pada data yang memiliki distribusi kelas yang tidak sepenuhnya seimbang.

Nilai F1-Score yang tinggi mengindikasikan bahwa model Random Forest memiliki stabilitas performa yang baik dan mampu memberikan hasil klasifikasi yang andal dalam memprediksi tingkat kepuasan penumpang maskapai penerbangan.

8.5 Pembahasan Hasil Secara Umum

Secara keseluruhan, hasil evaluasi menunjukkan bahwa algoritma **Random Forest** memiliki performa yang sangat baik dalam mengklasifikasikan tingkat kepuasan penumpang maskapai penerbangan. Kemampuan Random Forest dalam menggabungkan banyak pohon keputusan memungkinkan model untuk mengurangi kesalahan prediksi dan risiko overfitting.

Atribut-atribut layanan penerbangan seperti kenyamanan, pelayanan selama penerbangan, serta ketepatan waktu terbukti memiliki kontribusi yang signifikan terhadap tingkat kepuasan penumpang. Hasil penelitian ini menunjukkan bahwa pendekatan data mining dengan algoritma Random Forest dapat digunakan sebagai alat bantu pengambilan keputusan bagi pihak maskapai dalam meningkatkan kualitas layanan secara berbasis data.

9. Kesimpulan

1. Proses eksplorasi dan preprocessing data pada dataset kepuasan penumpang maskapai penerbangan berhasil dilakukan dengan baik, sehingga data siap digunakan dalam proses pemodelan klasifikasi.
2. Algoritma Random Forest berhasil diimplementasikan untuk mengklasifikasikan tingkat kepuasan penumpang maskapai penerbangan berdasarkan karakteristik penumpang dan penilaian layanan penerbangan.
3. Hasil evaluasi model menunjukkan bahwa algoritma Random Forest memiliki performa yang sangat baik, dengan nilai akurasi, precision, recall, dan F1-score yang tinggi, sehingga model mampu memprediksi tingkat kepuasan penumpang secara akurat dan konsisten.

10. Saran

1. Penelitian selanjutnya disarankan untuk melakukan analisis **feature importance** guna mengetahui atribut layanan penerbangan yang paling berpengaruh terhadap tingkat kepuasan penumpang.
2. Untuk memperoleh perbandingan performa yang lebih komprehensif, penelitian berikutnya dapat mengimplementasikan dan membandingkan algoritma klasifikasi lain, seperti Decision Tree, Logistic Regression, atau Support Vector Machine (SVM).
3. Penelitian lanjutan dapat menggunakan dataset yang lebih besar atau lebih beragam serta menerapkan teknik penanganan ketidakseimbangan kelas agar kemampuan generalisasi model dapat ditingkatkan.

11. Daftar Pustaka

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Manurung, J. (2025). *Analisis Kepuasan Pelanggan di Industri Jasa Penerbangan*. Jurnal Manajemen Pelayanan, 8(1), 45-58.

Mckinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Andi Offset.

Setiono, B. (2022). *Penerapan Algoritma Random Forest untuk Klasifikasi Data*. Jurnal Teknologi Informasi dan Komputer, 5(2), 112-120.