TUGAS 1
DATA PREPARATION DARI SUMBER OPEN SOURCE

disusun untuk memenuhi
tugas mata kuliah Pembelajaran Mesin A

oleh :

Sadinal Mufti	(2208107010007)
M. Agradika Ridhal Eljatin	(2208107010020)
Jihan Nabilah	(2208107010035)
Firjatullah Afny Abus	(2208107010059)
Athar Rayyan Muhammad	(2208107010074)

DEPARTEMEN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

TAHUN 2025

1. Pendahuluan

1.1. Latar Belakang

Dalam era digital saat ini, data memainkan peran krusial dalam pengambilan keputusan di berbagai sektor. Namun, data yang diperoleh seringkali dalam bentuk mentah dan memerlukan proses persiapan sebelum dapat digunakan untuk analisis atau pelatihan model machine learning. Proses ini dikenal sebagai persiapan data, yang mencakup langkah-langkah seperti pengumpulan, pembersihan, dan transformasi data.

Menurut AWS, "Persiapan data adalah proses menyiapkan data mentah sehingga layak untuk diproses dan dianalisis lebih lanjut." Pentingnya persiapan data juga ditekankan dalam berbagai penelitian. Misalnya, sebuah studi menyatakan bahwa "Tahap data preparation merupakan tahap persiapan data sebelum dipakai guna proses modelling dan evaluation." Hal ini menunjukkan bahwa kualitas data yang baik sangat mempengaruhi kinerja model machine learning yang dibangun.

Pada tugas ini, dataset yang digunakan adalah *Titanic Dataset*, yang berisi informasi tentang penumpang kapal Titanic, termasuk status kelangsungan hidup mereka. Dataset ini memiliki berbagai tantangan, seperti adanya nilai yang hilang pada beberapa kolom (misalnya, *Age* dan *Cabin*), serta fitur kategorikal yang perlu dikonversi agar dapat digunakan dalam analisis machine learning. Oleh karena itu, langkah-langkah preprocessing yang tepat sangat diperlukan untuk memastikan bahwa data siap digunakan dalam analisis lebih lanjut.

1.2. Tujuan Tugas

- Memahami proses persiapan data sebelum analisis.
- Mengaplikasikan teknik preprocessing pada dataset open source.
- Mendokumentasikan hasil dan insight dari proses eksplorasi dan persiapan data.

2. Data Description

2.1. Informasi Dataset

- Nama Dataset: Titanic Dataset
- Sumber: https://www.kaggle.com/datasets/yasserh/titanic-dataset
- Deskripsi Singkat: Dataset Titanic merupakan dataset klasik yang digunakan untuk analisis kelangsungan hidup penumpang kapal Titanic berdasarkan berbagai faktor seperti usia, jenis kelamin, kelas tiket, dan lainnya.

2.2. Struktur Dataset

- Jumlah Sampel: 891 baris
- Jumlah Fitur: 12 kolom
- Label : Kolom Survived (0 = Tidak Selamat, 1 = Selamat)
- Format Data: CSV (Comma-Separated Values)

- Deskripsi Kolom:
    - PassengerId: ID unik untuk setiap penumpang
    - Survived: Status kelangsungan hidup (0 = Tidak Selamat, 1 = Selamat)
    - Pclass: Kelas tiket penumpang (1 = Kelas 1, 2 = Kelas 2, 3 = Kelas 3)
    - Name: Nama penumpang
    - Sex: Jenis kelamin penumpang
    - Age: Usia penumpang
    - SibSp: Jumlah saudara atau pasangan di kapal
    - Parch: Jumlah orang tua atau anak di kapal
    - Ticket: Nomor tiket
    - Fare: Tarif tiket
    - Cabin: Nomor kabin (banyak yang hilang)
    - Embarked: Pelabuhan embarkasi (C = Cherbourg, Q = Queenstown, S = Southampton)

3. Data Loading

3.1. Metode Memuat Data

Untuk memuat dataset, digunakan bahasa pemrograman Python dengan library Pandas. Dataset Titanic dalam format CSV dibaca menggunakan fungsi pd.read_csv().

```python
import pandas as pd

data = pd.read_csv('/content/Titanic-Dataset.csv')
data.head()
```

Dataset berhasil dimuat ke dalam variabel data, dan fungsi head() digunakan untuk menampilkan lima baris pertama dari dataset.

3.2. Tantangan dalam Memuat Data

- Beberapa nilai dalam dataset kosong atau hilang (NaN), yang akan ditangani dalam tahap Data Preparation.

4. Data Understanding

4.1. Analisis Statistik Dasar dan Visualisasi Data

Setelah dataset berhasil dimuat, dilakukan analisis awal untuk memahami struktur dan karakteristik data.

1. Identifikasi Missing Values
   Untuk mengetahui jumlah nilai yang hilang dalam setiap atribut:

```
data.isnull().sum()
```

output:

```
data.isnull().sum()
```

	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

dtype: int64

Hasil menunjukkan bahwa terdapat missing values pada beberapa atribut, terutama pada Age, Cabin, dan Embarked. Kehadiran missing values ini perlu ditangani dalam tahap preprocessing.

2. Distribusi Title Penumpang

Atribut Title ditambahkan dengan mengekstrak gelar dari atribut Name:

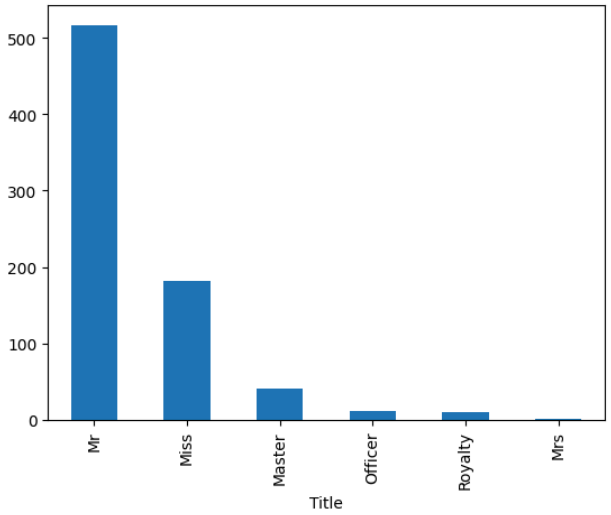
```
data['Title'] = data['Name'].apply(lambda name: name.split(',')
[1].split('.')[0].strip())
data.insert(loc=12, column='Title', value=data.pop('Title'))
```

Selanjutnya, dilakukan transformasi untuk mengelompokkan gelar penumpang:

```
title_map = {
    "Capt": "Officer", "Col": "Officer", "Major": "Officer", "Johkheer": "Royalty",
    "Don": "Royalty", "Sir": "Royalty", "Dr": "Royalty", "Rev": "Officer",
    "The Countess": "Royalty", "Dona": "Royalty", "Mme": "Mrs", "Mile": "Miss",
    "Mr": "Mr", "Miss": "Miss", "Master": "Master", "Lady": "Royalty"
}

data["Title"] = data.Title.map(title_map)
```

Visualisasi distribusi title penumpang:



Hasil visualisasi menunjukkan bahwa gelar Mr, Miss, dan Mrs merupakan gelar yang paling umum digunakan.

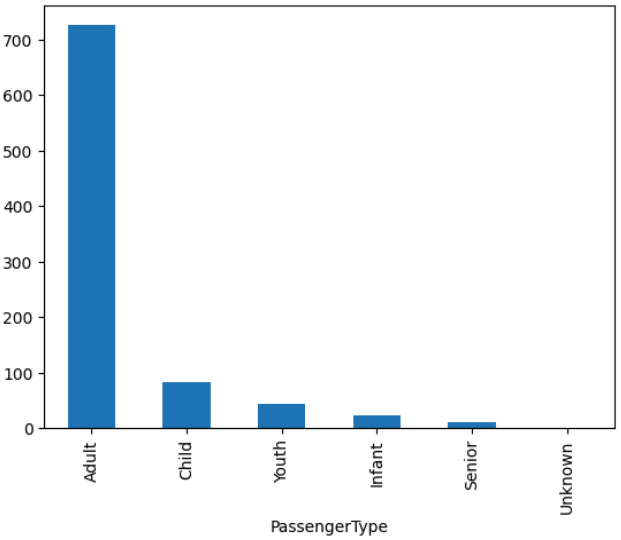
3. Pengelompokan Usia Penumpang
   Kelompok umur didefinisikan sebagai berikut:

   Infant (0-2), Child (3-11)", Youth (12-17), Adult (18-64), Senior (65+).

```python
def passenger_type(row):
    if row['Age'] <= 2:
        return "Infant"
    elif 2 < row['Age'] < 12:
        return "Child"
    elif 12 <= row['Age'] < 18:
        return "Youth"
    elif 18 <= row['Age'] < 65:
        return "Adult"
    elif row['Age'] >= 65:
        return "Senior"
    elif row['Title'] == "Master":
        return "Child"
    elif row['Title'] == "Miss":
        return "Child"
    elif row['Title'] == "Mr" or row['Title'] == "Mrs":
        return "Adult"
    else:
        return "Unknown"

data['PassengerType'] = data.apply(lambda row: passenger_type(row), axis=1)
```

Visualisasi distribusi kelompok umur:

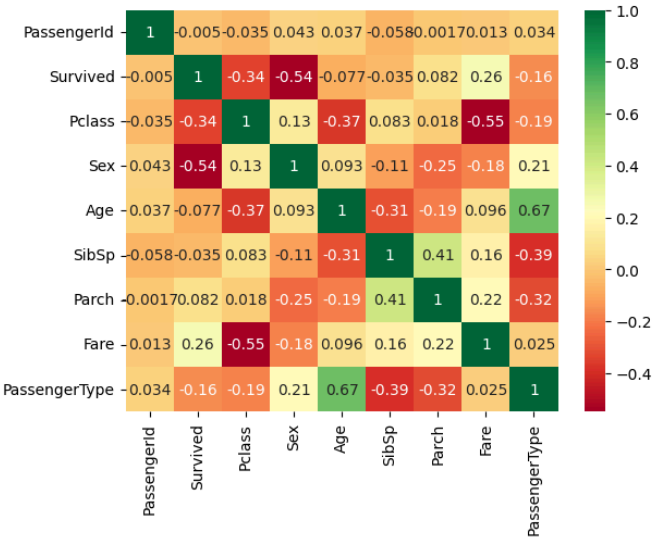


Analisis menunjukkan bahwa sebagian besar penumpang termasuk dalam kategori Adult, diikuti oleh Child dan Youth.

4. Korelasi Antar Variabel

Untuk memahami hubungan antar fitur dalam dataset:

```
correlation = data.select_dtypes(include=['number']).corr()
sns.heatmap(correlation, annot=True, cbar=True, cmap="RdYlGn")
plt.show()
```

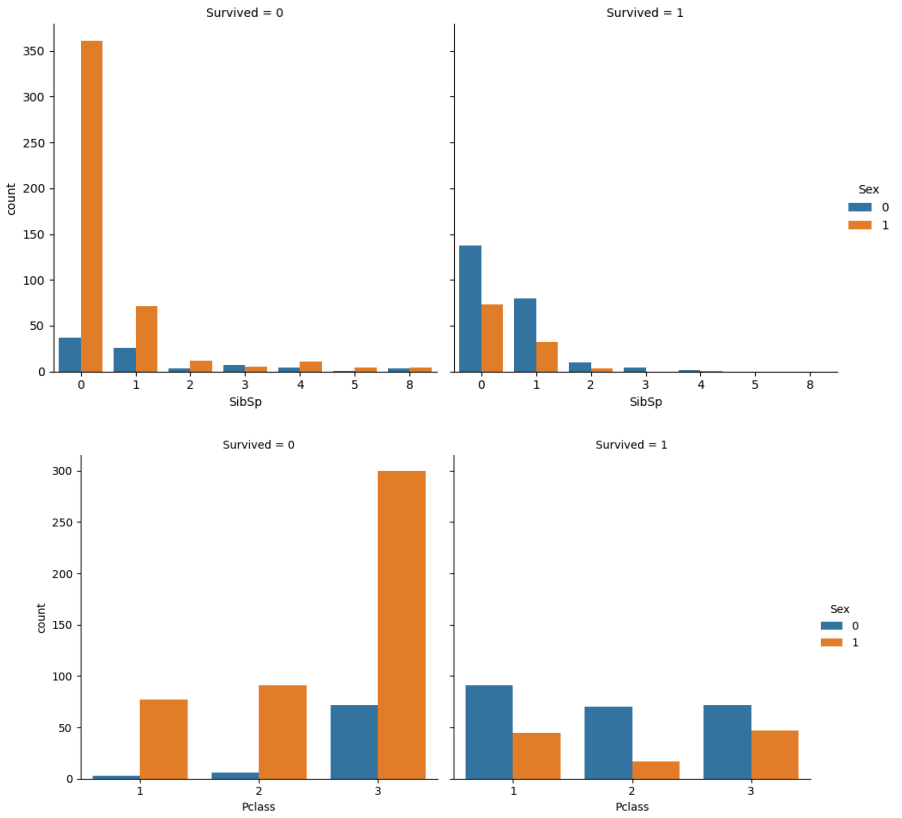


Hasil heatmap menunjukkan korelasi antar variabel numerik, termasuk hubungan antara Pclass, Fare, dan Survived.

Analisis lebih lanjut terhadap fitur yang berhubungan dengan Survived:

```
sns.catplot(x="SibSp", col="Survived", data=data, hue="Sex", kind="count")
plt.show()

sns.catplot(x="Pclass", col="Survived", data=data, hue="Sex", kind="count")
plt.show()
```



Hasil analisis menunjukkan bahwa penumpang dari kelas pertama memiliki tingkat kelangsungan hidup lebih tinggi dibandingkan kelas lainnya.

4.2. Insight dari Data Eksplorasi

Berdasarkan eksplorasi data, ditemukan bahwa atribut Age, Cabin, dan Embarked memiliki missing values, dengan Cabin yang paling banyak sehingga dihapus dari analisis. Distribusi Title menunjukkan mayoritas penumpang adalah "Mr", "Miss", dan "Mrs", sedangkan gelar seperti "Royalty" dan "Officer" jarang ditemukan. Analisis usia mengelompokkan penumpang ke dalam beberapa kategori, di mana Adult (18-64 tahun) mendominasi, sementara bayi dan anak-anak lebih sedikit namun memiliki tingkat keselamatan lebih tinggi.

Korelasi antar variabel menunjukkan bahwa Sex, Pclass, dan Fare berpengaruh terhadap keselamatan penumpang. Perempuan dan penumpang kelas satu memiliki peluang lebih besar untuk selamat dibanding laki-laki dan penumpang kelas bawah. Selain itu, penumpang dengan jumlah keluarga lebih banyak (SibSp tinggi) cenderung memiliki tingkat keselamatan lebih rendah. Faktor utama yang memengaruhi keselamatan adalah jenis kelamin, kelas tiket, usia, dan jumlah keluarga, yang akan menjadi dasar dalam analisis lebih lanjut dan pemodelan prediksi.

5. Data Preparation

5.1. Mengatasi Missing Values

Pada dataset Titanic, terdapat beberapa atribut dengan missing values, terutama pada kolom "Age" dan "Title". Missing values pada "Title" diisi berdasarkan informasi jenis kelamin, di mana jika "Sex" adalah "male", maka "Title" diisi dengan "Mr", sedangkan jika "Sex" adalah "female", maka "Title" diisi dengan "Mrs".

```
 Menampilkan jumlah missing values pada setiap kolom
data.isnull().sum()

# Mengisi missing values pada kolom "Title"
data.loc[data["Title"].isnull() & (data["Sex"] == 'male'), "Title"] = 'Mr'
data.loc[data["Title"].isnull() & (data["Sex"] == 'female'), "Title"] =
'Mrs'
```

Selain itu, atribut "Cabin" memiliki banyak missing values sehingga atribut ini dihapus dari dataset, bersama dengan "PassengerId" dan "Ticket" yang dianggap tidak informatif.

```
# Menghapus atribut yang tidak diperlukan
data.drop(labels=["Cabin", "PassengerId", "Ticket"], axis=1, inplace=True)
```

5.2. Encoding Kategori

Beberapa kolom kategorikal seperti "Sex" dan "PassengerType" dikonversi ke bentuk numerik agar dapat digunakan dalam model pembelajaran mesin.

```
# Encoding atribut "Sex"
sex_map = {"male": 1, "female": 0}
data["Sex"] = data["Sex"].map(sex_map)

# Encoding atribut "PassengerType"
passenger_type_map = {"Unknown": 0, "Infant": 1, "Child": 2, "Youth": 3,
"Adult": 4, "Senior": 5}
data["PassengerType"] = data["PassengerType"].map(passenger_type_map)
```

5.3. Normalisasi dan Standardisasi

Dalam dataset ini, normalisasi tidak diterapkan karena sebagian besar fitur numerik sudah dalam skala yang sebanding. Jika diperlukan, metode seperti Min-Max Scaling atau Standardization dapat digunakan.

5.4. Feature Selection/Extraction

Atribut yang tidak memberikan informasi signifikan terhadap prediksi "Survived" dihapus, seperti "PassengerId", "Ticket", dan "Cabin". Selain itu, dilakukan analisis korelasi untuk menentukan fitur yang memiliki pengaruh signifikan terhadap "Survived".

```
# Menampilkan heatmap korelasi
sns.heatmap(data.corr(), annot=True, cmap="RdYlGn")
plt.show()
```

Selain itu, dilakukan analisis hubungan antara "SibSp", "Pclass", dan "Survived" untuk memahami dampaknya terhadap keselamatan penumpang.

```
# Korelasi jumlah saudara/kakak (SibSp) dengan peluang selamat
sns.catplot(x="SibSp", col="Survived", data=data, hue="Sex", kind="count")
plt.show()

# Korelasi kelas penumpang (Pclass) dengan peluang selamat
sns.catplot(x="Pclass", col="Survived", data=data, hue="Sex", kind="count"
plt.show()
```

5.5. Justifikasi Keputusan Preprocessing

- Missing values pada "Title" diisi berdasarkan jenis kelamin karena umumnya "Mr" digunakan untuk laki-laki dan "Mrs" untuk perempuan.

- "Cabin" dihapus karena terlalu banyak missing values yang sulit untuk diimputasi secara akurat.
- "Sex" dan "PassengerType" dikonversi ke bentuk numerik agar dapat digunakan oleh algoritma machine learning.
- Feature selection dilakukan dengan menghapus atribut yang tidak memiliki pengaruh signifikan terhadap target variabel "Survived".

6. Kesimpulan

Proses preprocessing pada dataset Titanic telah dilakukan secara sistematis untuk memastikan kualitas data yang lebih baik sebelum analisis atau pembuatan model machine learning. Langkah-langkah yang diterapkan mencakup identifikasi dan penanganan missing values, encoding variabel kategorikal, serta pemilihan fitur yang relevan. Hasil eksplorasi menunjukkan bahwa faktor-faktor seperti jenis kelamin, kelas tiket, dan usia berpengaruh signifikan terhadap peluang keselamatan penumpang.

Dengan menghapus fitur yang tidak relevan dan menangani data yang hilang dengan strategi yang sesuai, dataset yang telah diproses kini lebih siap untuk analisis lebih lanjut atau diterapkan dalam model prediksi. Pendekatan preprocessing ini tidak hanya meningkatkan kualitas data tetapi juga membantu dalam memahami pola dan faktor utama yang mempengaruhi keselamatan penumpang Titanic.