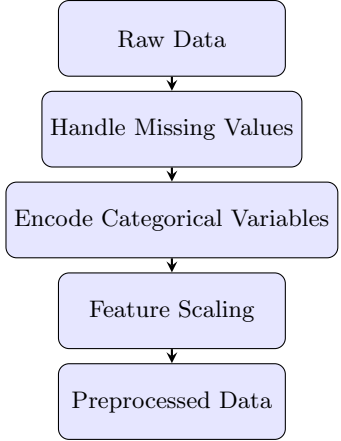

Handling Missing Data and Categorical Features

Bindeshwar Singh Kushwaha

Overview of Data Preprocessing

- Load Titanic dataset from CSV file.

Overview of Data Preprocessing

- Load Titanic dataset from CSV file.
- Handle missing values using various techniques.

Overview of Data Preprocessing

- Load Titanic dataset from CSV file.
- Handle missing values using various techniques.
- Encode categorical data for machine learning.

Overview of Data Preprocessing

- Load Titanic dataset from CSV file.
- Handle missing values using various techniques.
- Encode categorical data for machine learning.
- Save the cleaned dataset to a new CSV file.

Step 1: Load the Titanic Dataset

Python Code

```
import pandas as pd
df = pd.read_csv('titanic.csv')
```

Step 1: Load the Titanic Dataset

Python Code

```
import pandas as pd
df = pd.read_csv('titanic.csv')
```

- We load the Titanic dataset using `pd.read_csv`.

Step 1: Load the Titanic Dataset

Python Code

```
import pandas as pd
df = pd.read_csv('titanic.csv')
```

- We load the Titanic dataset using `pd.read_csv`.
- The dataset is stored in a pandas DataFrame.

Step 2: View the First Few Rows

Python Code

```
print(df.head())
```

Step 2: View the First Few Rows

Python Code

```
print(df.head())
```

- Use `df.head()` to preview the first 5 rows of the dataset.

Step 2: View the First Few Rows

Python Code

```
print(df.head())
```

- Use `df.head()` to preview the first 5 rows of the dataset.
- This helps in understanding the data structure and columns.

Step 3: Checking for Missing Values

Python Code

```
print(df.isnull().sum())
```

Step 3: Checking for Missing Values

Python Code

```
print(df.isnull().sum())
```

- Use `df.isnull().sum()` to check for missing values.

Step 3: Checking for Missing Values

Python Code

```
print(df.isnull().sum())
```

- Use `df.isnull().sum()` to check for missing values.
- Helps in identifying columns with incomplete data.

Step 4: Handle Missing Values

Python Code

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df.dropna(subset=['Embarked'], inplace=True)
```

Step 4: Handle Missing Values

Python Code

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df.dropna(subset=['Embarked'], inplace=True)
```

- Fill missing `Age` values with the median of the column.

Step 4: Handle Missing Values

Python Code

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df.dropna(subset=['Embarked'], inplace=True)
```

- Fill missing **`Age`** values with the median of the column.
- Drop rows with missing values in **`Embarked`**.

Step 5: Verify Missing Values

Python Code

```
print(df.isnull().sum())
```

Step 5: Verify Missing Values

Python Code

```
print(df.isnull().sum())
```

- After handling missing data, verify using `df.isnull().sum()`.

Step 5: Verify Missing Values

Python Code

```
print(df.isnull().sum())
```

- After handling missing data, verify using `df.isnull().sum()`.
- This ensures there are no remaining missing values.

Step 6: Data Overview

Python Code

```
print(df.describe())
```

Step 6: Data Overview

Python Code

```
print(df.describe())
```

- Use `df.describe()` for basic statistical analysis.

Step 6: Data Overview

Python Code

```
print(df.describe())
```

- Use `df.describe()` for basic statistical analysis.
- Helps in understanding the distribution of numerical columns.

Step 7: Encoding Categorical Data

Python Code

```
df['Sex'] = df['Sex'].map('male': 0, 'female': 1)
```

Step 7: Encoding Categorical Data

Python Code

```
df['Sex'] = df['Sex'].map('male': 0, 'female': 1)
```

- Convert the **`Sex`** column into numerical values (0 for male, 1 for female).

Step 7: Encoding Categorical Data

Python Code

```
df['Sex'] = df['Sex'].map('male': 0, 'female': 1)
```

- Convert the **`Sex`** column into numerical values (0 for male, 1 for female).
- This is necessary for machine learning algorithms.

Step 8: Save the Cleaned Dataset

Python Code

```
df.to_csv('cleaned_titanic.csv', index=False)
```

Step 8: Save the Cleaned Dataset

Python Code

```
df.to_csv('cleaned_titanic.csv', index=False)
```

- The cleaned dataset is saved to a new CSV file.

Step 8: Save the Cleaned Dataset

Python Code

```
df.to_csv('cleaned_titanic.csv', index=False)
```

- The cleaned dataset is saved to a new CSV file.
- This ensures that all preprocessing steps are retained for future use.

Reach PostNetwork Academy

Website

www.postnetwork.co

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

GitHub Repositories

www.github.com/postnetworkacademy

Thank You!