Machine Learning Lab - Experiment-1: Visualizing and Handling the Titanic Dataset

Date-15-7-2025

Total Marks: 100

 # Objective:

The objective of this lab is to handle, analyze, and visualize the Titanic dataset using Python. You will perform data cleaning, exploratory data analysis (EDA), and create various visualizations to gain insights from the dataset.

 # Dataset:

The Titanic dataset contains data about the passengers on the Titanic, including whether they survived or not. The dataset has the following attributes:

- `PassengerId`: Unique ID for each passenger
- `Pclass`: Passenger class (1st, 2nd, 3rd)
- `Name`: Passenger's name
- `Sex`: Passenger's gender
- `Age`: Passenger's age
- `SibSp`: Number of siblings or spouses aboard the Titanic
- `Parch`: Number of parents or children aboard the Titanic
- `Ticket`: Ticket number
- `Fare`: Passenger fare
- `Embarked`: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
- `Survived`: Survival status (0 = No, 1 = Yes)

 # Instructions:

1. Data Loading and Inspection (20 Marks)

- Load the attached Titanic dataset into a Pandas DataFrame.
- Display the first few rows of the dataset.
- Print the summary information of the DataFrame, including the number of rows and columns, and the data types of each column.

- Identify and list any columns with missing values and the percentage of missing values in each column.

2. Data Cleaning (20 Marks)

-Identify and list missing values and their percentages

- Handle missing values: Drop columns with more than 50% missing values and fill or drop rows with missing values in critical columns.

- Convert columns to appropriate data types (e.g., categorical columns to category type, `Survived` column to binary).

- Remove duplicates if any.

3. Exploratory Data Analysis (EDA) (30 Marks)

- Descriptive Statistics (10 Marks):

  - Provide summary statistics (mean, median, standard deviation, etc.) for numerical columns.

  - Analyze the distribution of passenger ages.

- Comparative Analysis (10 Marks):

  - Compare the survival rates across different passenger classes (`Pclass`).

  - Analyze survival rates based on gender (`Sex`).

- Correlation Analysis (10 Marks):

  - Calculate the correlation matrix for the numerical features.

  - Create a heatmap to visualize the correlations and identify any strong correlations between features.

4. Data Visualization (20 Marks)

- Create meaningful visualizations to represent the following:

  - The distribution of passenger classes using a bar plot.

  - The distribution of survival status (Survived vs. Not Survived) using a pie chart.

  - The distribution of fares paid by passengers using a histogram.

  - Box plots to compare the fare distribution across different passenger classes (`Pclass`).

5. Conclusion and Insights (10 Marks)

- Write a comprehensive conclusion summarizing the key insights from your analysis.

- Discuss the potential implications of your findings and how they can be useful for understanding survival rates and improving safety measures.

 # Submission:

- Submit a Jupyter Notebook file (.ipynb) with all the code, outputs, and visualizations.

- Ensure the notebook is well-documented with comments and markdown cells explaining each step.

- Include a summary report (1-2 pages) in PDF format summarizing the key insights and findings from your analysis.

 # Evaluation Criteria:

- Correctness and completeness of data loading and inspection (20 Marks)

- Effectiveness and accuracy of data cleaning (20 Marks)

- Depth and relevance of exploratory data analysis (30 Marks)

- Quality and clarity of visualizations (20 Marks)

- Quality of conclusion and insights (10 Marks)

 # Additional Resources:

- [Pandas Documentation](https://pandas.pydata.org/pandas-docs/stable/)

- [Matplotlib Documentation](https://matplotlib.org/stable/contents.html)

- [Seaborn Documentation](https://seaborn.pydata.org/)

- [Titanic Dataset on Kaggle](https://www.kaggle.com/c/titanic/data)