Machine Learning Mini-Project Report

Titanic Data Analysis and Prediction

Name: Parth Ingle (41227)

Mukund Karwa (41234)

Soham Lagad (41241)

Batch: Q2

Title

Data Cleaning, Visualization, and Prediction on Titanic Dataset using Python

Problem Definition

Analyze the Titanic dataset to uncover patterns in passenger survival, visualize the distribution of features like fare and age, and build a machine learning model to predict survival outcomes using features like age, fare, sex, and embarkation point.

Learning Objective

- Learn to clean missing and inconsistent data in real-world datasets.
- Perform data visualization using Seaborn and Matplotlib. - Apply feature engineering for better model performance.
- Use machine learning models (Random Forest) to predict outcomes.

Learning Outcome

- Students will gain hands-on experience in data preprocessing and imputation techniques.
- Ability to visualize relationships and distributions in data.
- Learn how to use encoding and model fitting with scikit-learn.
- Evaluate model performance using metrics like accuracy and classification reports.

Theory-Related Concepts, Architecture, Syntax

The Titanic assignment involves multiple key data science concepts and techniques. One of the fundamental challenges in real-world datasets is handling missing values. This project demonstrates the use of **group-based imputation**, where missing ages are filled using the median value based on the extracted title (e.g., Mr., Miss, Mrs.) from passenger names. This method is more accurate than simply filling with the global median. Additionally, other missing values, like `Cabin`, are inferred based on Fare proximity, and `Embarked` values are filled using the mode.

In terms of visualization, the project uses **Seaborn and Matplotlib** to generate insightful plots including **histograms, countplots, barplots, and swarmplots**. These visualizations reveal trends in age distribution, fare distribution, survival rates by title and gender, and help interpret the data before modeling.

The core of the prediction model is built using a **Random Forest Classifier**, a powerful ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. Features are encoded using label encoding where necessary, and irrelevant columns such as `Name`, `Ticket`, and `PassengerId` are removed to simplify the model input.

The overall architecture of the system follows a standard data science pipeline: loading the dataset, handling missing data, engineering relevant features, visualizing patterns, splitting the data into training and testing sets, fitting a model, and finally evaluating performance through accuracy and feature importance visualization. This pipeline provides a comprehensive introduction to the applied machine learning process.

Test Cases

Test Case ID	Input	Action Taken / Feature	Expected Output
TC1	Row with missing `Age`	Impute using median of grouped `Title`	Missing `Age` replaced by relevant median
TC2	Row with missing `Embarked`	Filled using mode (`S` is most frequent)	No missing `Embarked` values
TC3	Row with missing `Cabin`	Inferred using closest Fare to Cabin median values	`Cabin` replaced with approximated label
TC4	Female passenger with high Fare	Model trained with sex and fare features	Higher chance of predicted survival
TC5	Random `x_test` sample	Prediction using trained Random Forest Classifier	Output is 0 or 1 (survived or not)

TC6	Complete test dataset with features	Model accuracy check	Accuracy ≥ 80% (approx) on validation set

Program Listing

The program involves cleaning data, visualizing patterns, and building a predictive model. Due to its length, it is divided into the following sections:

1. Data loading and display
2. Imputation of missing values (Age, Fare, Cabin, Embarked)
3. Feature engineering (Title extraction, Label Encoding)
4. Data visualization using Seaborn
5. Model training with Random Forest
6. Evaluation and feature importance visualization

Output

- Accuracy = approx 0.82
- Classification Report shows good precision/recall across classes
- Bar chart shows the most important features (Fare, Sex, Age) for survival prediction
- Visualization charts displayed using Matplotlib and Seaborn

Conclusion

The Titanic dataset provides valuable insights into feature impact on passenger survival. By performing systematic data cleaning, imputation, visualization, and machine learning modeling, we can achieve meaningful prediction results. This assignment demonstrates a full-cycle data science workflow using Python libraries.