Titanic Survival Prediction Using Machine Learning

Gauri Rasal
Data Science Student
Email: example@email.com

***Abstract*—This project aims to predict passenger survival on the Titanic using machine learning algorithms. The dataset includes features such as age, gender, class, and more. The approach includes data preprocessing, exploratory analysis, feature engineering, and classification using models such as Logistic Regression and Random Forest.**

I. Introduction

Predicting survival on the Titanic is a classic classification problem. It involves analyzing historical passenger data and identifying key features that influenced survival outcomes. This project applies supervised learning algorithms and data science methodologies to derive insights and make accurate predictions.

II. Dataset

The dataset used was `TITANIC SURVIVAL PREDICTION.csv`. It includes features like:

- Passenger Class (Pclass)
- Name, Sex, Age
- Number of siblings/spouses (SibSp)
- Number of parents/children (Parch)
- Ticket, Fare, Cabin, Embarked

III. Methodology

A. Data Preprocessing

The dataset was imported using pandas:

```
df = pd.read_csv("/content/TITANIC␣SURVIVAL␣
    PREDICTION.csv")
```

Missing values were handled, and non-numeric columns were encoded using techniques like label encoding and one-hot encoding.

B. Feature Engineering

Irrelevant columns such as `Name`, `Ticket`, and `Cabin` were dropped. Categorical variables such as `Sex` and `Embarked` were encoded. Missing age values were filled with the median.

C. Modeling

We used Logistic Regression and Random Forest Classifier:

```
from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)

from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)
```

D. Evaluation

The models were evaluated using accuracy, precision, and confusion matrix.

```
from sklearn.metrics import accuracy_score,
    classification_report
y_pred = log_model.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

IV. Results

The Logistic Regression model gave an accuracy of approximately XX% while the Random Forest model outperformed with YY%. Feature importance from Random Forest indicated that `Sex`, `Pclass`, and `Fare` were the top predictors.

V. Conclusion

This project demonstrates how machine learning techniques can be applied to real-world classification problems. With proper preprocessing and model selection, good predictive accuracy can be achieved.

VI. References

- Titanic Dataset: https://www.kaggle.com/c/titanic
- Scikit-learn Documentation: https://scikit-learn.org/
- Seaborn Documentation: https://seaborn.pydata.org/