Titanic Survival Prediction Using Machine Learning

Gauri Rasal

Deprtment Of Technology

Email: gaurirasal2919@gmail.com

Abstract—**This project aims to predict the survival of the Titanic passengers using machine learning algorithms. The data set includes features such as age, sex, class, and more. The approach includes data pre-processing, 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 data set was imported using pandas:

```python
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. The missing age values were filled with the median.

C. Modeling

We used Logistic Regression and Random Forest Classifier:

```python
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 an accuracy, precision, and confusion matrix.

```python
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

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/