Titanic Survival Prediction – Final Report

Muhammad Haseeb

September 8, 2025

Insights from EDA

Exploratory Data Analysis revealed strong patterns in survival:

- **Gender**: More females survived compared to males, reflecting the “women and children first” principle.
- **Class**: 1st class passengers had a much higher survival rate than 2nd and 3rd class.
- **Age**: Younger children had better chances of survival than adults.
- **Fare**: Higher fares were associated with higher survival rates.

Overall, social and economic status strongly influenced survival on the Titanic.

Preprocessing Steps

- **Missing Values**: Age imputed with median, Cabin dropped due to excessive missing values, Embarked filled with mode.
- **Encoding**: Converted categorical features (`Sex`, `Embarked`) into numeric values.
- **Scaling**: Standardized Age and Fare to improve model performance.

These steps prepared a clean, machine-learning-ready dataset.

Model Performance

Three models were tested:

- Logistic Regression: Accuracy $\approx 81.0\%$
- Decision Tree: Accuracy $\approx 78.2\%$
- Random Forest: Accuracy $\approx 81.6\%$

Random Forest performed best, with the highest accuracy and ROC-AUC, due to reduced overfitting and ability to capture complex relationships.

Confusion Matrix

The confusion matrix of the best-performing model (Random Forest) is shown below:

Challenges and Solutions

- **Missing Data**: Solved using imputation (median/mode) and dropping Cabin.
- **Imbalanced Classes**: Evaluated models with precision, recall, and ROC-AUC in addition to accuracy.
- **Model Selection**: Used classification metrics and ROC-AUC to confirm Random Forest as the most reliable.

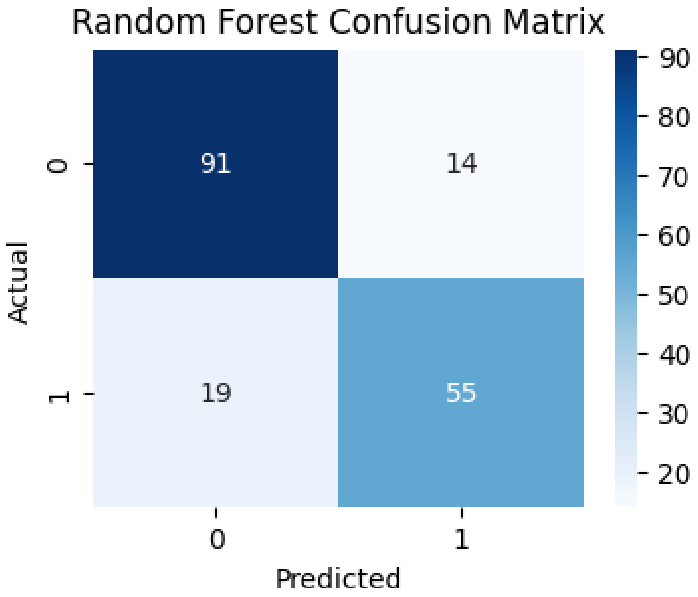

Figure 1: Confusion Matrix for Random Forest Model

Conclusion

Survival on the Titanic was influenced by gender, class, age, and fare. With proper preprocessing and Random Forest, we achieved $\sim 82\%$ accuracy in predicting survival.