Task 5: Titanic Survival Prediction

Soumen Das

November 20, 2025

1 Introduction

The objective of Task 5 is to perform data cleaning and exploratory data analysis (EDA) on the Titanic dataset. The dataset contains demographics and passenger information from 891 passengers (training set) and 418 passengers (test set). The goal is to understand the factors that contributed to survival during the disaster.

2 Data Preprocessing

Before analysis, the data required cleaning to handle missing values and irrelevant features.

2.1 Handling Missing Values

- **Age:** The 'Age' column had approximately 177 missing values. These were imputed using the median age of the passengers to maintain the distribution without introducing outliers.
- **Embarked:** Two missing values in the 'Embarked' column were filled with the mode (most frequent value), which was 'S' (Southampton).
- **Cabin:** The 'Cabin' column was dropped from the analysis as it contained a significant number of missing values (>77%), making it unreliable for direct prediction without complex feature engineering.

3 Exploratory Data Analysis (EDA)

We analyzed the relationship between survival and key features such as Gender and Passenger Class.

3.1 Survival by Gender

Gender was the strongest predictor of survival. Female passengers had a significantly higher chance of survival compared to male passengers.

3.2 Survival by Passenger Class

Socio-economic status, represented by Passenger Class (Pclass), also played a vital role. First-class passengers were prioritized during the rescue.

Table 1: Survival Rates by Gender

Sex	Count	Survival Rate (%)
Female	314	74.20%
Male	577	18.89%

Table 2: Survival Rates by Passenger Class

Pclass	Class Type	Survival Rate (%)
1	Upper	62.96%
2	Middle	47.28%
3	Lower	24.23%

4 Conclusion

The analysis of the Titanic dataset confirms that priority was given to women and passengers of higher socio-economic status.

1. **Gender bias:** Females had a nearly 74% survival rate, whereas males had less than 19%.
2. **Class bias:** First-class passengers were more than twice as likely to survive as third-class passengers.

This preprocessing and analysis form the foundation for building a machine learning model (such as Logistic Regression or Random Forest) to predict passenger survival on the test set.