Exploratory Data Analysis (EDA) Report on Titanic Dataset

Introduction

This report summarizes the key findings from the Exploratory Data Analysis (EDA) performed on the Titanic dataset. The objective of this EDA was to understand the dataset's structure, identify patterns, and prepare it for further analysis or machine learning tasks.

Data Overview and Preprocessing

The dataset contains information about passengers on the Titanic, including their demographics, travel details, and survival status. Initial inspection revealed missing values in 'Age', 'Embarked', and 'Cabin' columns. Missing 'Age' values were imputed with the median, 'Embarked' with the mode, and the 'Cabin' column was dropped due to a high percentage of missing values.

Key Findings from EDA

1. Survival Rate

Approximately 38% of the passengers survived the Titanic disaster.

2. Survival by Gender

Females had a significantly higher survival rate compared to males. This suggests that 'Sex' is a crucial predictor of survival.

3. Survival by Passenger Class (Pclass)

Passengers in higher classes (1st class) had a better chance of survival. This indicates a correlation between socio-economic status and survival.

4. Age Distribution and Survival

The age distribution showed that children (younger ages) had a higher survival rate. There were noticeable differences in age distributions between survivors and non-survivors.

5. Fare Distribution and Survival

Passengers who paid higher fares tended to have a higher survival rate, further supporting the observation that socio-economic status played a role in survival.

6. Survival by Embarked Port

The port of embarkation (Cherbourg, Queenstown, or Southampton) also showed some correlation with survival rates, though further investigation would be needed to understand the underlying reasons.

7. Feature Engineering: Family Size

A new feature, 'FamilySize' (SibSp + Parch + 1), was created. Analysis showed that small to medium-sized families (2-4 members) had a better survival rate compared to individuals traveling alone or very large families.

8. Correlation Matrix

The correlation matrix of numerical features provided insights into the relationships between variables, such as a negative correlation between 'Pclass' and 'Fare', and a positive correlation between 'SibSp' and 'Parch'.

Conclusion

The EDA revealed several important insights into the factors influencing survival on the Titanic. Gender, passenger class, age, and fare paid were identified as strong indicators of survival. These findings are crucial for feature selection and engineering in subsequent machine learning model development.