Titanic Survival Analysis

By OmarM.Taha

Introduction

The Titanic Disaster: A Brief overview of the historical event.

Project Goal: To analyze the Titanic dataset to identify factors that influenced passenger survival and potentially build a predictive model.

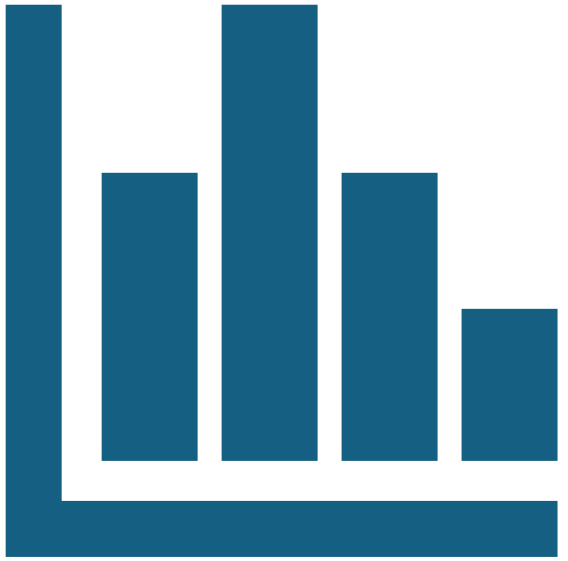
Dataset Overview

Source: Click Here & Click Here

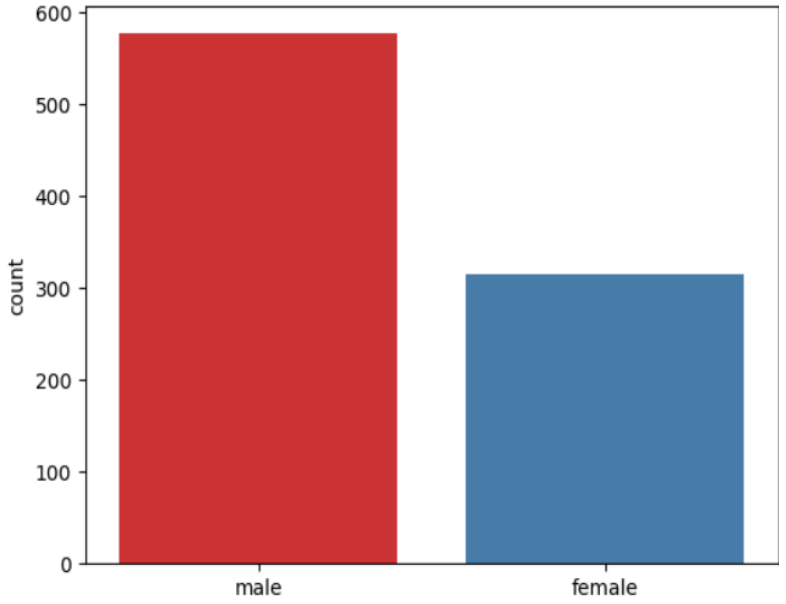
It includes key details such as Survived, passenger class, Sex, Age, and Fare, among other attributes, for 891 entries

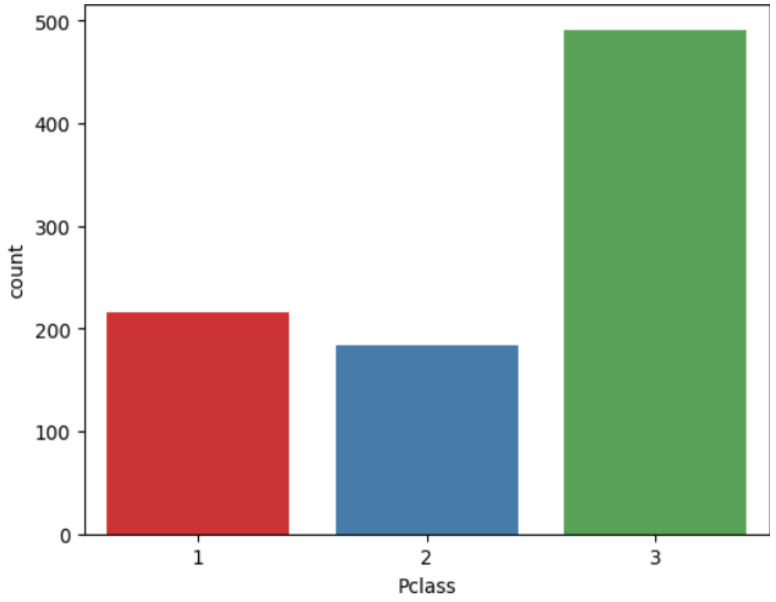
Data Statistics and Information

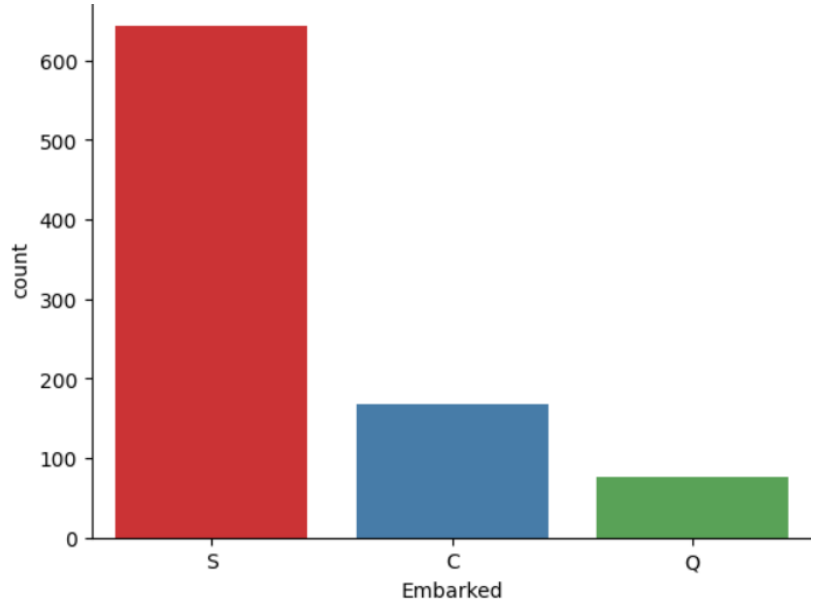
- **Descriptive Statistics:**
  - Average Age: Approximately 29.7 years
  - Survival Rate: Around 38.4% of passengers survived.
  - Passenger Class Distribution: The Majority of passengers were in the 3rd class
  - Average Fare: Approximately 32.2
- **Data Types and Missing Values:**
  - The Age, Cabin, and Embarked columns have missing values
  - Cabin has a significant number of missing values.



Exploratory Data Analysis (EDA) - Categorical Attributes





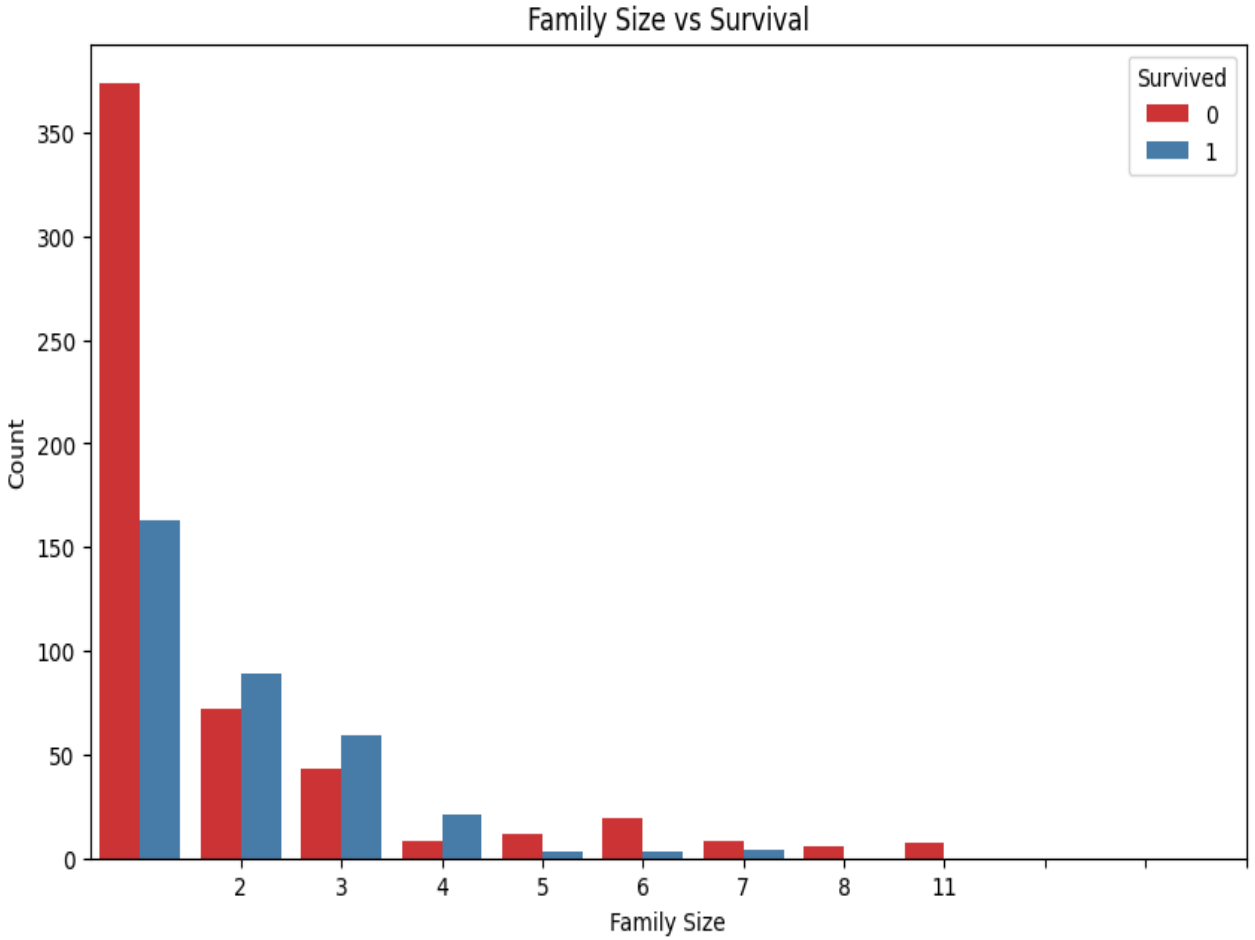


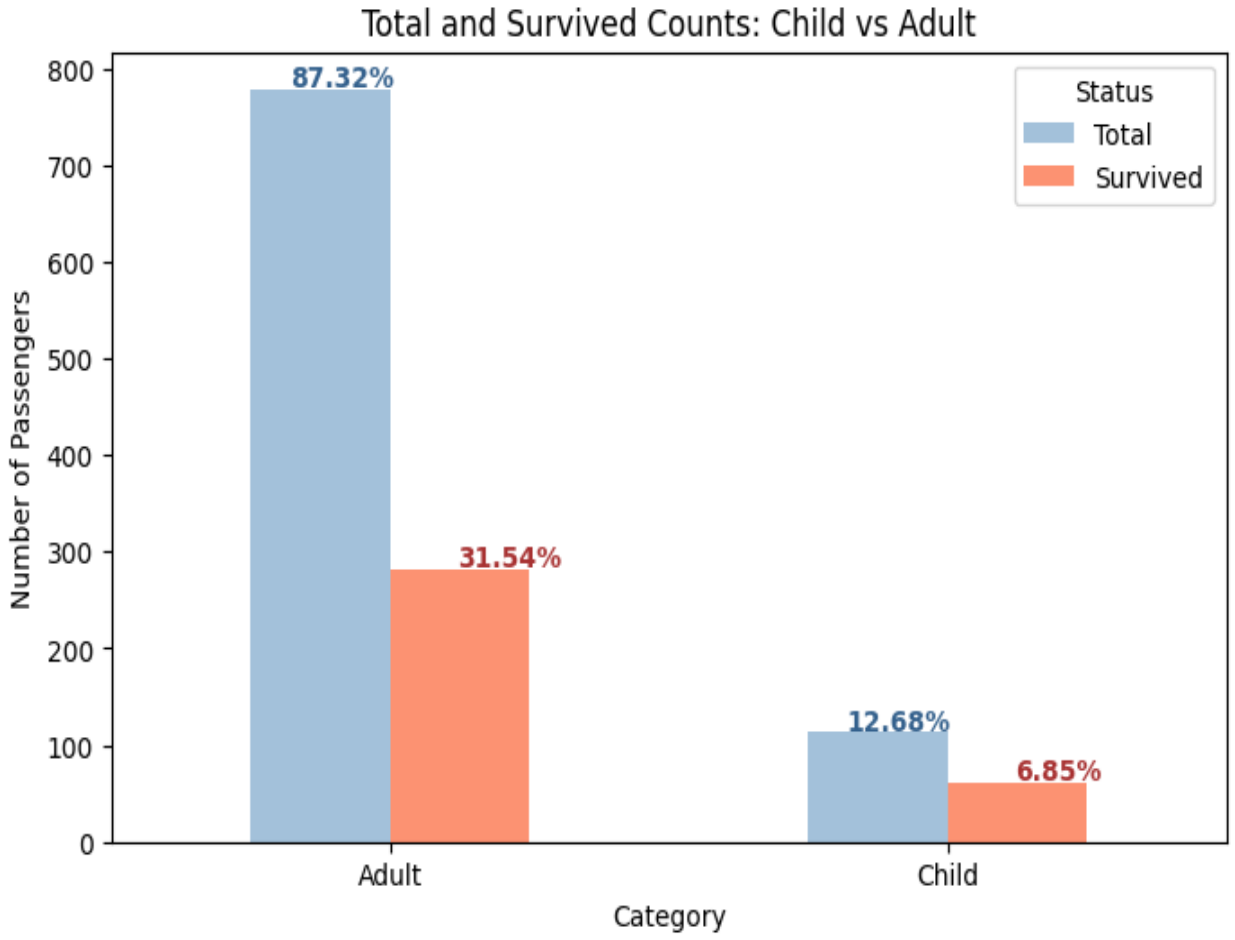
- Women were more likely to survive than men.

Higher passenger classes (1st and 2nd) had better survival rates compared to 3rd class.

Survival rates varied by embarkation point

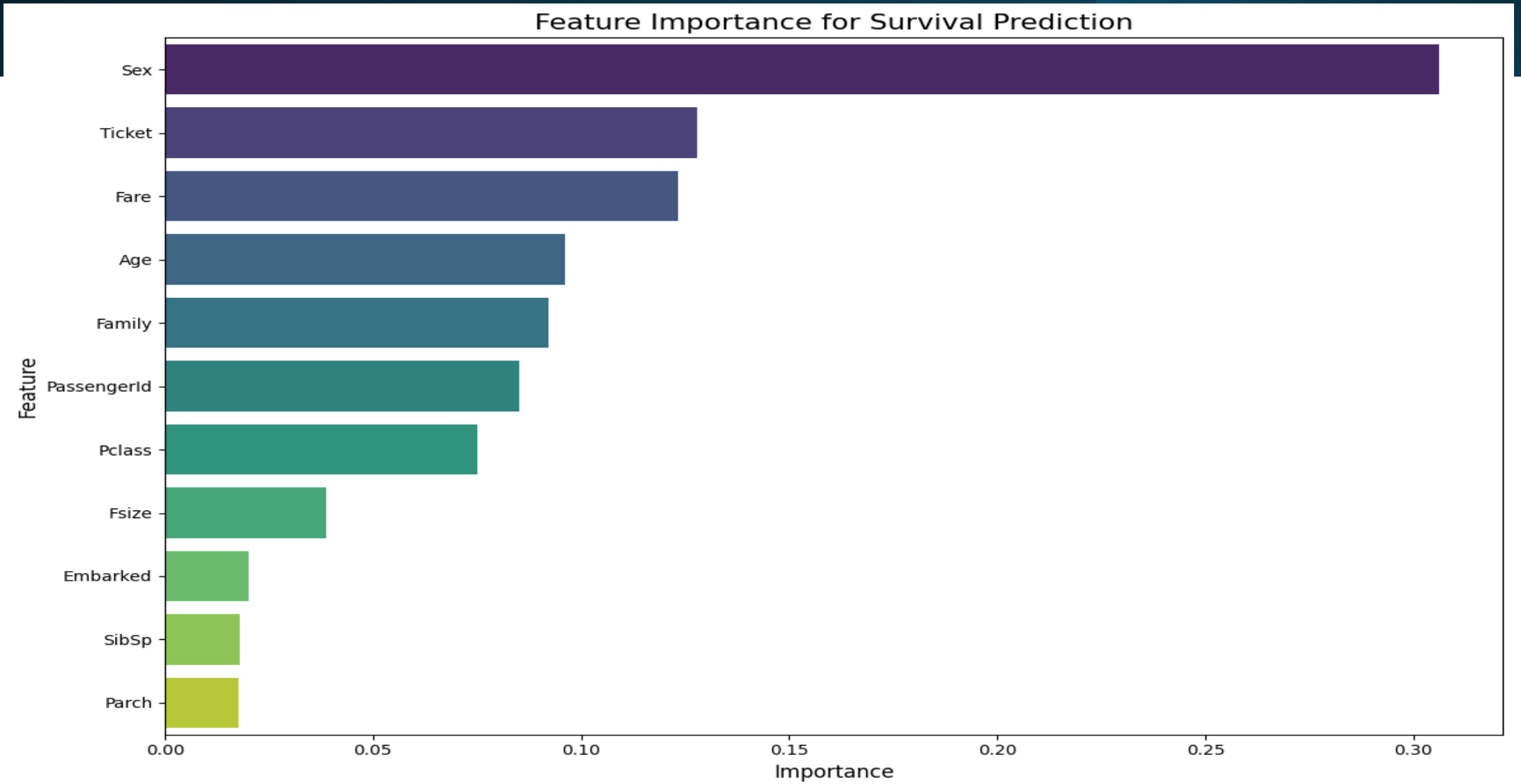
Exploratory Data Analysis (EDA) - Categorical Attributes





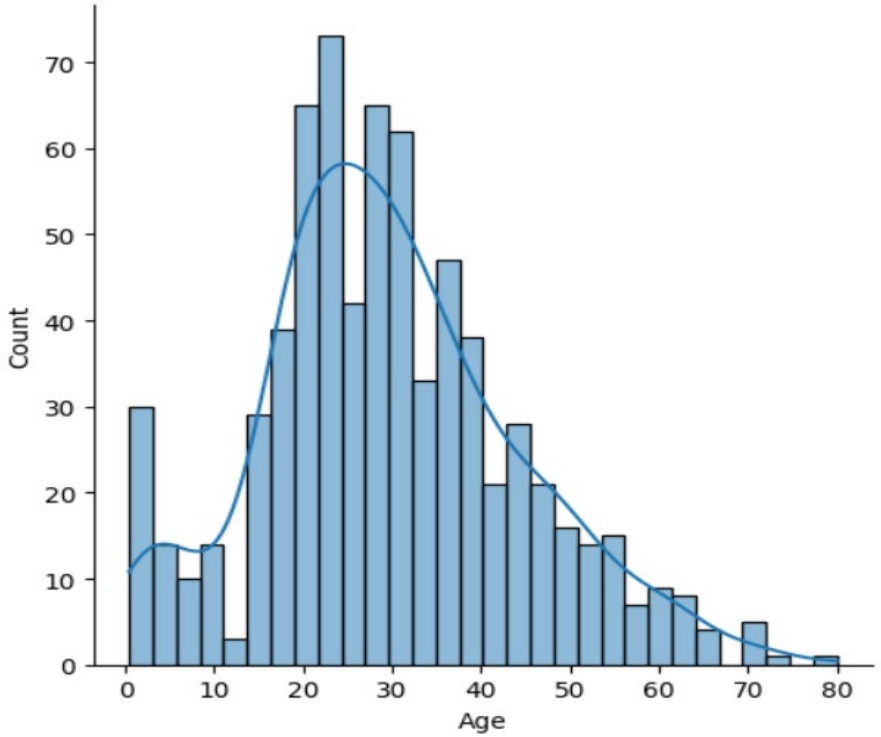
- Do families sink or swim together We Will Create A New Variable Named Family
- Do The Passenger Are Adult Or Child Will Affects Survival Rates

Exploratory Data Analysis (EDA) - Categorical Attributes

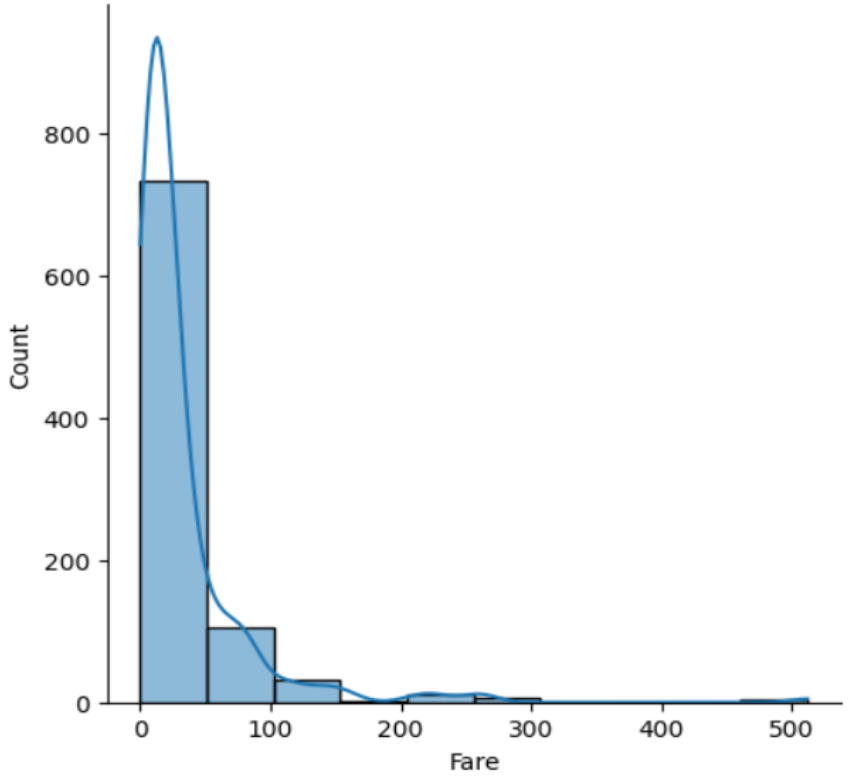


- Let's look at relative variable importance by plotting the mean decrease in Gini calculated across all trees.

Exploratory Data Analysis (EDA) - Numerical Attributes

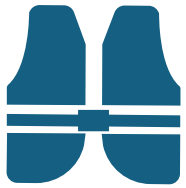


Children and younger passengers might have had higher survival rates.

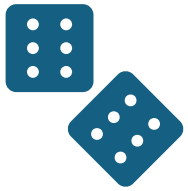


Passengers who paid higher fares tended to have a higher survival rate.

Predictive Modeling - Selection & Training



Goal: Build and evaluate classification models to predict Titanic survival.



Models Explored: Logistic Regression, Decision Tree, and Random Forest Classifiers.

Approach: Data split into 75% training and 25% testing sets.

Model Performance Overview:

Random Forest: Best performance (80.7% CV Score), indicating strong generalization.

Logistic Regression: Solid performance (80.7% Test Accuracy, 78.3% CV Score).

Decision Tree: Lower performance (72.6% Test Accuracy, 76.9% CV Score).

Key Insight: Random Forest emerged as the most robust model for predicting Titanic survival.

Q&A

Thank You!