TITANIC DATASET ANALYSIS REPORT

Analytical Findings and Key Interpretations

891 Total Passengers **38.4% Survival Rate**

Executive Summary:

This report presents a comprehensive analysis of the Titanic passenger dataset to determine the key factors influencing survival outcomes and to develop predictive models for survival prediction. Following rigorous data cleaning and preprocessing, an in-depth exploratory data analysis (EDA) uncovered significant relationships between survival rates and passenger characteristics, including gender, class, age, fare and family relations. Subsequently, multiple machine learning models were trained and evaluated, achieving predictive accuracies of up to 85

Introduction:

The Titanic disaster of 1912 resulted in a significant loss of life. The purpose of this analysis is to understand which passenger characteristics contributed to survival, and to build predictive models using the Kaggle Titanic dataset. This project demonstrates the end-to-end process of data cleaning, exploration, and modeling.

Dataset Overview:

Dataset Source: Kaggle Titanic Dataset

Feature name	Categorical /Numerical	Null Values	Column Dropped (Yes/No)	Information Delivered
Passenger ID	Numerical	0	**Yes**	Unique passenger identifier
Survived	Numerical	0	No	Survival status indicator
Pclass	Numerical	0	No	Passenger travel class
Name	Categorical	0	**Yes**	Full passenger name
Sex	Categorical	0	No	Passenger gender information
Age	Numerical	**177**	No	Passenger age years
SibSp	Numerical	0	No	Siblings/spouses aboard
Parch	Numerical	0	No	Parents/children aboard
Ticket	Categorical	0	**Yes**	Ticket number issued
Fare	Numerical	0	No	Fare paid amount
Cabin	Categorical	**687**	**Yes**	Cabin number/location
Embarked	Categorical	**2**	No	Port of embarkation

Data Quality Assessment:

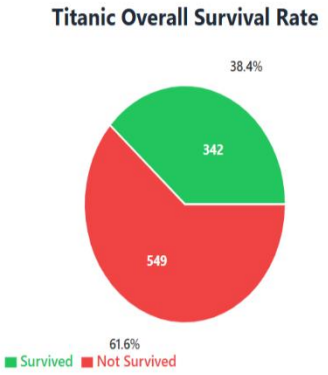
Feature Name	Null Values	Missing %	Solution	Column Dropped (Yes/No)
Age	177	19.86%	Moderate missingness – imputed by median values	No
Embarked	02	0.22	Negligible missingness – imputed by mode value	No
Cabin	687	77.10%	High missingness – column dropped	Yes

- **Overall Data Quality:** Good – except Age column as it was required for calculations.
- **Dropped Columns:** Categorical – Name, Cabin and Ticket, Numerical – Passenger ID

Data Preprocessing:

- **Age Column (19.86% missing values):** Filled with median values.
- **Embarked Column (0.22% missing values):** Filled with 'S' (Mode value)
- **Label/One Hot Encoding:** Converted categorical columns line Sex and Embarked into numeric form.
- **Normalising Data:** Normalised numerical features like Fare and Age.

Exploratory Data Analysis – EDA:

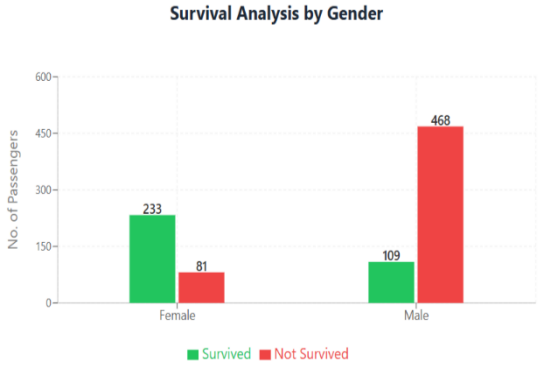


Total passengers: 891

Passengers survived: 342 (38.4%)

Passengers not survived: 549 (61.6%)

Insights: It represents one of the most tragic losses with ~ 62% people losing lives



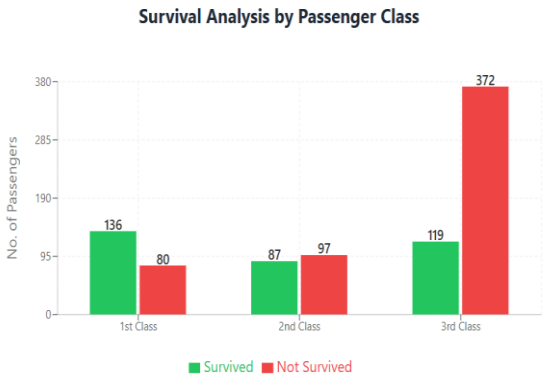
Total females: 314

Females survived: 233 (74.2%)

Total males: 577

Males survived: 109 (18.8%)

Insights: Clearly shows that Female/Children first protocol was implemented and were saved first.

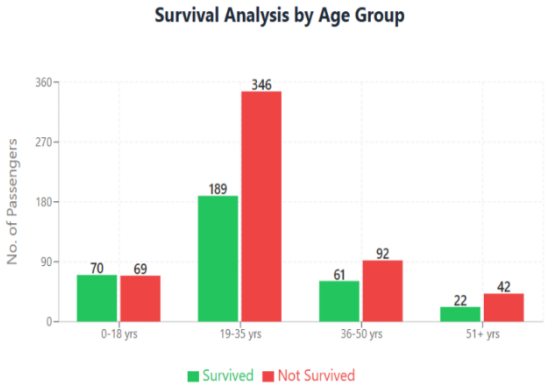


Class 1 people survived: 136 of 216 (62.6%)

Class 2 people survived: 87 of 184 (47.2%)

Class 3 people survived: 119 of 491 (24.2%)

Insights: Socio economic status significantly impacted survival as rate of class 1 people is 2.6 times higher than that of class 3



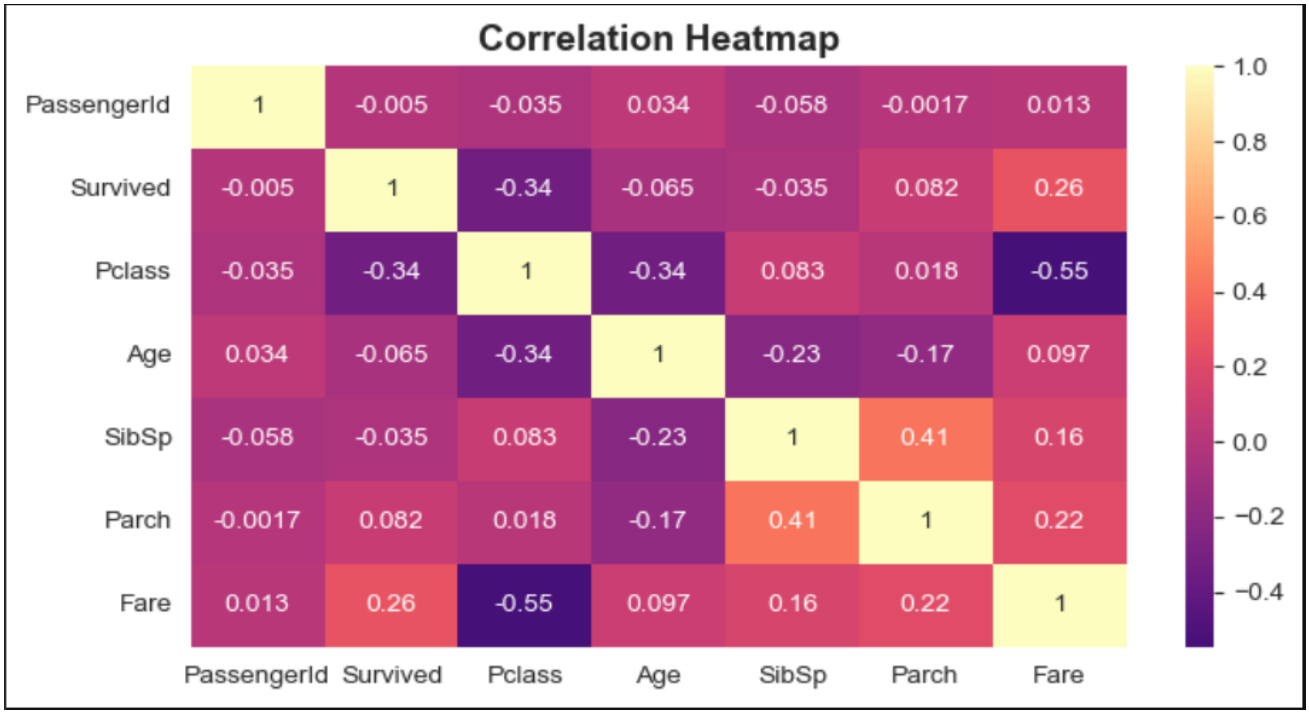
00 - 18 yrs survived: 70 of 139 (50.3%)

19 - 35 yrs survived: 189 of 535 (35.3%)

36 – 50 yrs survived: 61 of 153 (39.8%)

51+ yrs survived: 22 of 64 (34.3%)

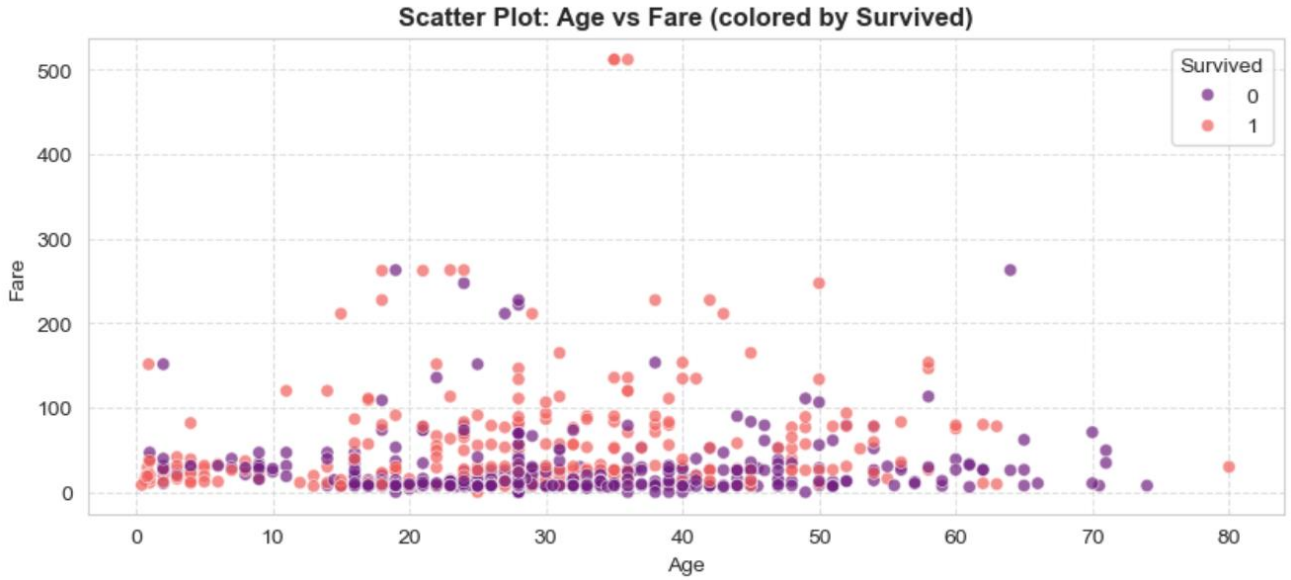
Insights: Children had highest survival rate, supporting prioritization of young passengers in rescue efforts



Key insights from the correlation heatmap:

- **Class was the strongest survival factor (-0.34):** Lower passenger classes had significantly worse survival rates, making it the most important predictor.
- **Wealth improved survival chances (0.26):** Higher fare prices correlated positively with survival, reinforcing that money bought safety.

- **Class and fare were tightly linked (-0.55):** This strong negative correlation shows 3rd class passengers paid much less than 1st class.
- **Family connections helped slightly (0.082):** Having parents/children aboard provided a small survival advantage, likely due to "women and children first" policy.
- **Age and class relationship (-0.34):** Younger passengers were more likely to be in lower classes, suggesting different travel demographics by age group.



Key insights from the scatter plot:

- **High-fare passengers had better survival rates:** Most red dots (survivors) appear in the upper fare ranges above $100, while purple dots (non-survivors) dominate the lower fare areas.
- **Expensive tickets clustered around ages 20-40:** The highest fares ($200-500+) are concentrated in the 20-40 age range, suggesting wealthy middle-aged passengers.
- **Low-fare passengers across all ages had poor survival:** The dense purple cluster at fares below $50 spans all age groups, showing consistent low survival rates for cheaper tickets.
- **Children with high fares survived more:** Red dots appear among younger passengers (under 18) in higher fare ranges, supporting "women and children first" for wealthy families.
- **Most passengers paid under $50:** The majority of data points cluster in the 0-50 fare range regardless of age, indicating most passengers were in lower-cost accommodations with poor survival odds.

Model Building:

- **Training Data:** 80% (~712)
- **Testing Data:** 20% (~179)
- **Models Used:** Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier

- **Before Tuning Model With Hyperparameters:**

Model Used	Accuracy	Precision	F1 - Score	Recall
Logistic Regression	0.8045	0.7931	0.7244	0.6667
Random Forest Classifier	0.8268	0.8065	0.7634	0.7246
Decision Tree Classifier	0.8324	0.8000	0.7761	0.7536
Support Vector Classifier	08212	0.8033	0.7538	0.7101

- **After Tuning Model With Hyperparameters:**

Model Used	Accuracy	Precision	F1 - Score	Recall
Logistic Regression	0.8045	0.7297	0.7552	0.7826
Random Forest Classifier	0.8156	0.8214	0.7360	0.6667
Decision Tree Classifier	0.7765	0.9143	0.6154	0.4638
Support Vector Classifier	08268	0.8167	0.7597	0.7101

Conclusions:

- Survival of passengers on Titanic was strongly associated with **gender, passenger class** and **age.**
- Machine learning models achieved **up to 80%** accuracy in predicting the survival of the passengers.
- Highlights the importance of **data preprocessing, feature selection, and model comparison** in predictive analysis.

Recommendations & Next Steps:

- Perform **feature engineering**. E.g. extracting the titles from the names of the passengers.
- Apply **cross-validation** for more robust performance estimation.
- Create a **dashboard** for real-time visualisation of survival factors.

Limitations:

- Dataset size is relatively **small** ~ 891 rows only.
- Missing key variables, e.g. exact location on the ship ~ Cabin (687 rows null)
- **Imputation of Age** may introduce bias results as it is an important feature for predicting results.

Another Project: Bangalore House Price Prediction Model (use of Flask Server) at my Github.

Link for the same: [House Price Prediction Model Project]