Titanic Survival Prediction Report

Introduction

The Titanic dataset, obtained from Kaggle, provides insights into the factors influencing passenger survival during the tragic sinking of the RMS Titanic. In this report, we will detail the steps taken to predict passenger survival, encompassing data preparation, exploratory data analysis, feature engineering, and modeling techniques employed throughout the study.

Data Overview

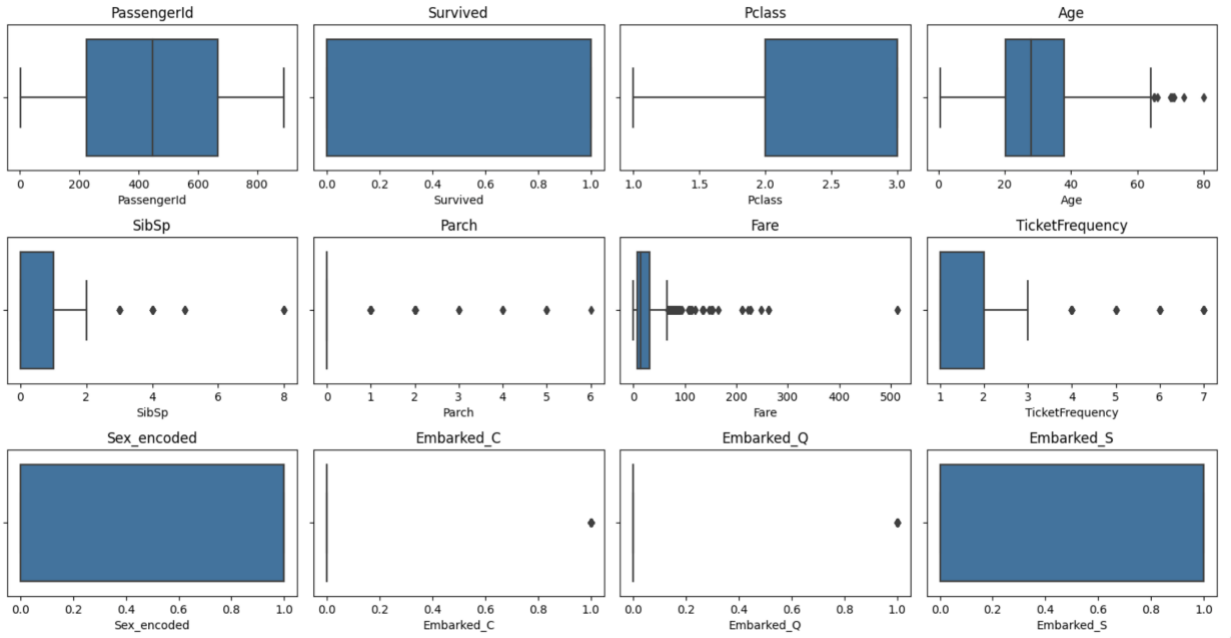
This dataset contains information on 891 passengers and provides a wide range of details that can be used to explore and analyze the factors affecting their survival on the Titanic:

1. **PassengerId:** A unique identifier assigned to each passenger.
2. **Survived:** Indicates whether the passenger survived (1) or did not survive (0) the Titanic disaster.
3. **Pclass:** Represents the passenger's ticket class, with three categories: 1st class, 2nd class, and 3rd class.
4. **Name:** The name of the passenger.
5. **Sex:** The gender of the passenger, either male or female.
6. **Age:** The age of the passenger in years.
7. **SibSp:** The count of siblings or spouses traveling with the passenger.
8. **Parch:** The count of parents or children traveling with the passenger.
9. **Ticket:** The ticket number associated with the passenger.
10. **Fare:** The fare or price paid for the ticket.
11. **Cabin:** The cabin number where the passenger stayed (if available).
12. **Embarked:** The port of embarkation, with three options: C (Cherbourg), Q (Queenstown), or S (Southampton).

Analyzing more in-depth we have the following statistics ì about the missing values, the quantity of outliers, and useful insights:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200



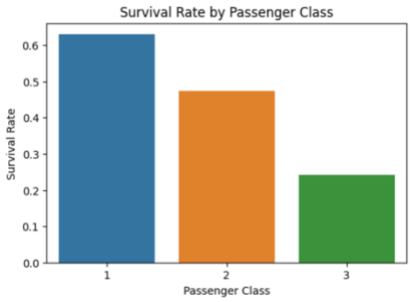
Data Preprocessing

Given the statistics, was decided to drop:

- the 'Cabin' feature because has too many missing values;
- the 'Name' feature because does not give useful information;
- the 'Ticket', after being used for the feature engineering process, was dropped because not give useful information.

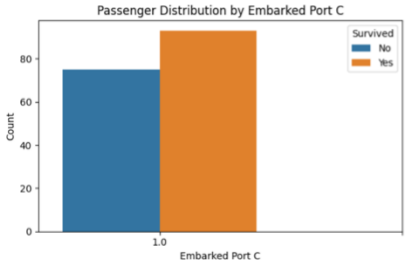
Another preprocessing applied was the encoding with the Label Encoder for the 'Sex' feature (female = 0, male = 1) and of the 'Embarked' feature with the One Hot Encoder.

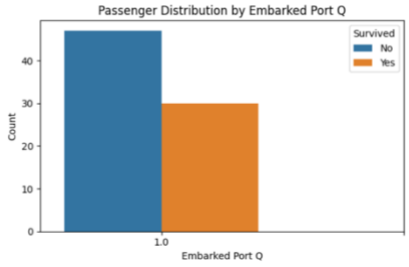
Eventually, the missing values of the feature 'Age' were filled with the mean because the presence of the outlier did not influence the distribution.

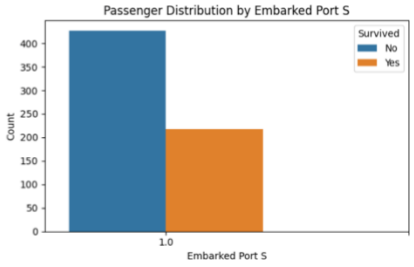


Exploratory Data Analysis (EDA)

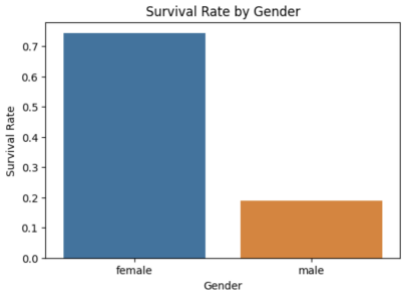
Firstly, is possible to notice that the survival rate increases as the passenger class goes toward the first one and similarly the rate is greater for the people embarked from port C, then from Q, and lastly to S.

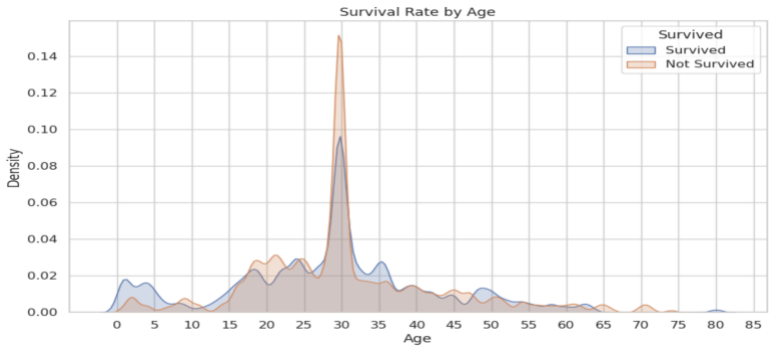




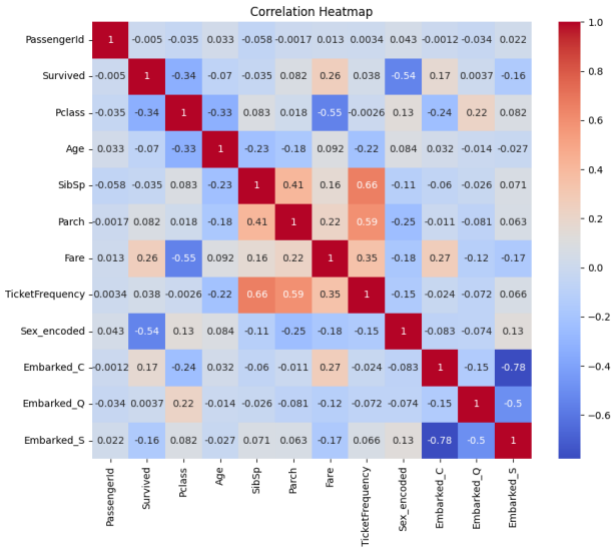


Other important features for the survival are the gender and the age. From the graph below, is possible to notice that females and young people, under 10 years old, have more probability of surviving.





In addition, a feature correlation analysis was applied to understand whether any feature engineering methods needed to be applied among the highly correlated features.



Feature Engineering

In this process 2 new feature was added to the dataset:

1. ‘FamilySize’ was computed as the sum of 'SibSp' and 'Parch'.
2. ‘TicketFrequency’ that counts how many times each ticket number appears in the dataset. Passengers with the same ticket number might have traveled together.

Modeling and Evaluation

Once the dataset was ready some classification algorithms were applied obtaining different scores:

DecisionTreeClassifier Results:
Accuracy: 0.7374301675977654
Classification Report:

	precision	recall	f1-score	support
0	0.77	0.78	0.78	105
1	0.68	0.68	0.68	74
accuracy			0.74	179
macro avg	0.73	0.73	0.73	179
weighted avg	0.74	0.74	0.74	179

RandomForestClassifier Results:
Accuracy: 0.8044692737430168
Classification Report:

	precision	recall	f1-score	support
0	0.81	0.88	0.84	105
1	0.80	0.70	0.75	74
accuracy			0.80	179
macro avg	0.80	0.79	0.79	179
weighted avg	0.80	0.80	0.80	179

SVC Results:
Accuracy: 0.5977653631284916
Classification Report:

	precision	recall	f1-score	support
0	0.60	0.98	0.74	105
1	0.67	0.05	0.10	74
accuracy			0.60	179
macro avg	0.63	0.52	0.42	179
weighted avg	0.62	0.60	0.48	179

KNeighborsClassifier Results:
Accuracy: 0.659217877094972
Classification Report:

	precision	recall	f1-score	support
0	0.66	0.85	0.74	105
1	0.64	0.39	0.49	74
accuracy			0.66	179
macro avg	0.65	0.62	0.62	179
weighted avg	0.66	0.66	0.64	179

GaussianNB Results:
Accuracy: 0.776536312849162
Classification Report:

	precision	recall	f1-score	support
0	0.84	0.76	0.80	105
1	0.70	0.80	0.75	74
accuracy			0.78	179
macro avg	0.77	0.78	0.77	179
weighted avg	0.78	0.78	0.78	179

Hyperparameter tuning

Finally, the best models (DecisionTree and RandomForest) were selected and parameter tuning, with Optuna framework, was applied to increase the performance. Unluckily the final scores were not improved due to a few trials of tuning being completed and the necessity of more time.