# REPORT ON ANALYSIS AND PREDICTION OF SURVIVORS ON THE TITANIC

**Name:** **Deepika V**

## Problem statement:

The RMS Titanic, a luxury steamship, sank in the early hours of April 15, 1912 off the coast of Newfoundland in the North Atlantic after sideswiping an iceberg during its maiden voyage. Of the 2,240 passengers and crew on board, more than 1,500 lost their lives in the disaster.

The goal of this project is to perform an analysis of the survival of the passengers based on the features available in the titanic dataset.

## Scope:

- The titanic dataset contains data about 418 passengers and 12 features.
- Since this dataset contains information of only a limited number of passengers, all the analysis and predictions made are only for the group of passengers given in the dataset.
- This project can be extended by increasing the number of input parameters and by using several advanced machine learning algorithms and perform more effective predictions.

## Dataset Description:

- PassengerId
- Survived – (Yes=1, No=0)
- Pclass – Passenger Class (Class 1 = 1, Class 2 = 2, Class 3 = 3)
- Name – Name of the Passenger
- Sex – (male or female)
- Age – Age of the passenger in years
- Sibsp – Number of siblings or spouses aboard
- Parch – Number of parents or children aboard
- Ticket – Ticket number
- Fare
- Cabin
- Embarked – Port of embarkation (Q-Queenstown, C-Cherbourg, S-Southampton)

Dataset link: https://www.kaggle.com/brendan45774/test-file

## Proposed Solution:

Data visualization based on some of the features in the dataset is done to analyse the relationships between them.

Data pre-processing is done for the following purposes.

- ❖ Some of the Age values are found to be null and hence they are replaced with the mean value of Age.
- ❖ One value of Fare is found to be null and hence it is assigned to be zero.
- ❖ The Cabin feature is dropped as it is not of much importance in this analysis.

Machine Learning models used are:

- ✓ Logistic Regression
- ✓ Support Vector Machine (SVM)

These models are implemented and their performance is evaluated to compare and decide which algorithm produces better results for the available titanic dataset.

## Tools used:

- ❑ Numpy
- ❑ Pandas
- ❑ Matplotlib
- ❑ Seaborn
- ❑ Scikit-learn

## Other similar tools:

- ❑ Scipy
- ❑ Plotly
- ❑ Altair

## Performance metrics:

Performance evaluation of the models helps us to choose the best model which predicts the most probable outcome.

The various performance metrics used to evaluate the models are as follows:

- **Confusion matrix:** a specific table layout that allows visualization of the performance of an algorithm.
  - **True Positives** − both actual class and predicted class of data point is 1.
  - **True Negatives** − both actual class and predicted class of data point is 0.
  - **False Positives** − actual class of data point is 0 and predicted class of data point is 1.
  - **False Negatives** − actual class of data point is 1 and predicted class of data point is 0.
- **Classification report:** This displays the precision, recall, F1, and support scores for the model.
- **Accuracy:** The number of correct predictions made as a ratio of all predictions made.
- **Precision:** The ratio of total number of true positives to the total number of data points with ground values as positive.
- **Recall:** The ratio of total number of true positives to the total number of data points with predicted values as positive.
- **F1-score:** This score will give the harmonic mean of precision and recall. It is the weighted average of precision and recall.

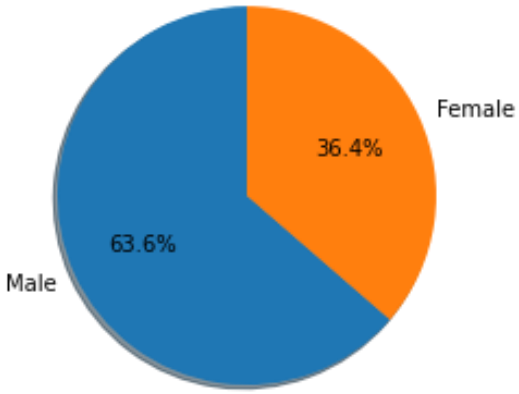
## Data visualization:

First five rows from the dataset:

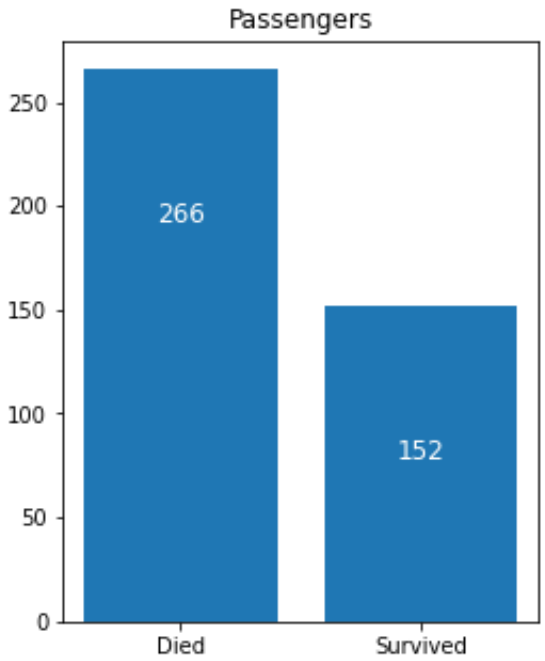
| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 892 | 0 | 3 | Kelly, Mr. James | male | 34.5 | 0 | 0 | 330911 | 7.8292 | NaN | Q |
| 1 | 893 | 1 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47.0 | 1 | 0 | 363272 | 7.0000 | NaN | S |
| 2 | 894 | 0 | 2 | Myles, Mr. Thomas Francis | male | 62.0 | 0 | 0 | 240276 | 9.6875 | NaN | Q |
| 3 | 895 | 0 | 3 | Wirz, Mr. Albert | male | 27.0 | 0 | 0 | 315154 | 8.6625 | NaN | S |
| 4 | 896 | 1 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22.0 | 1 | 1 | 3101298 | 12.2875 | NaN | S |

After data pre-processing:

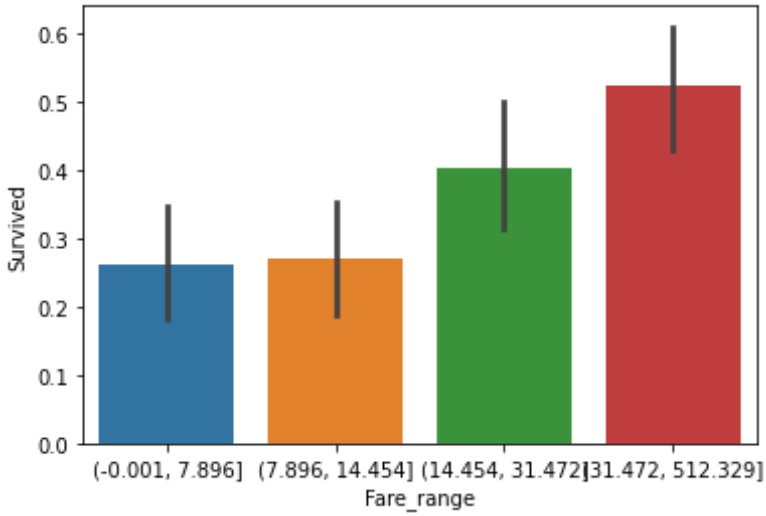
| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 892 | 0 | 3 | Kelly, Mr. James | male | 34.50000 | 0 | 0 | 330911 | 7.8292 | Q |
| **1** | 893 | 1 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47.00000 | 1 | 0 | 363272 | 7.0000 | S |
| **2** | 894 | 0 | 2 | Myles, Mr. Thomas Francis | male | 62.00000 | 0 | 0 | 240276 | 9.6875 | Q |
| **3** | 895 | 0 | 3 | Wirz, Mr. Albert | male | 27.00000 | 0 | 0 | 315154 | 8.6625 | S |
| **4** | 896 | 1 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22.00000 | 1 | 1 | 3101298 | 12.2875 | S |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **413** | 1305 | 0 | 3 | Spector, Mr. Woolf | male | 30.27259 | 0 | 0 | A.5. 3236 | 8.0500 | S |
| **414** | 1306 | 1 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00000 | 0 | 0 | PC 17758 | 108.9000 | C |
| **415** | 1307 | 0 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50000 | 0 | 0 | SOTON/O.Q. 3101262 | 7.2500 | S |
| **416** | 1308 | 0 | 3 | Ware, Mr. Frederick | male | 30.27259 | 0 | 0 | 359309 | 8.0500 | S |
| **417** | 1309 | 0 | 3 | Peter, Master. Michael J | male | 30.27259 | 1 | 1 | 2668 | 22.3583 | C |



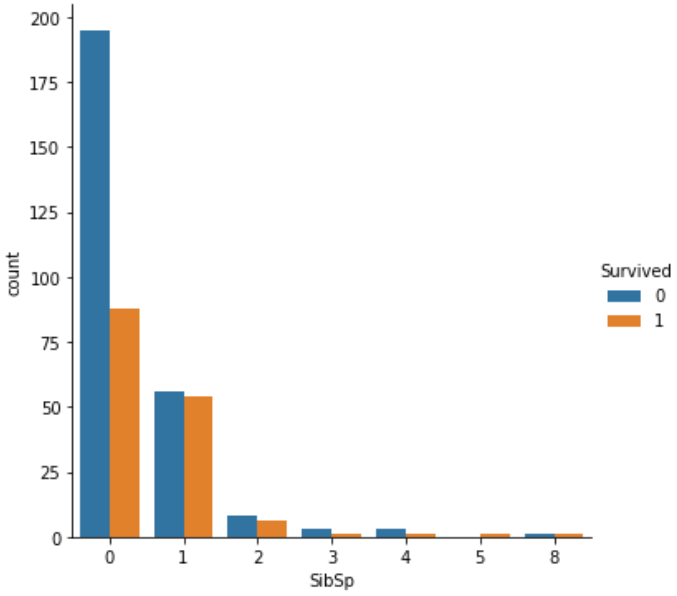

From the values in the dataset, it is found that the number of male passengers is higher than the number of female passengers.



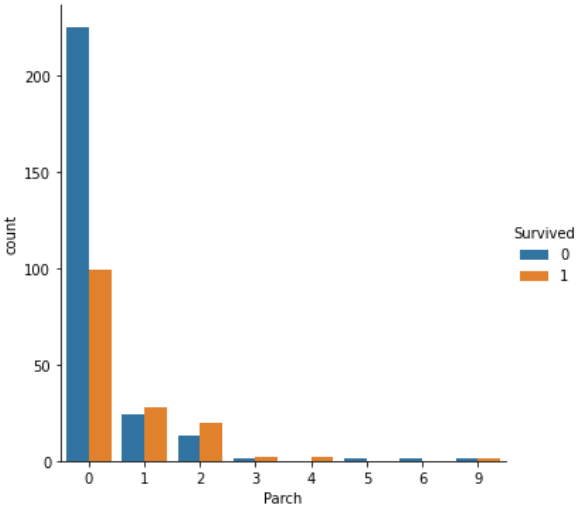

Out of the 418 passengers, 152 passengers survived and 266 didn't survive.



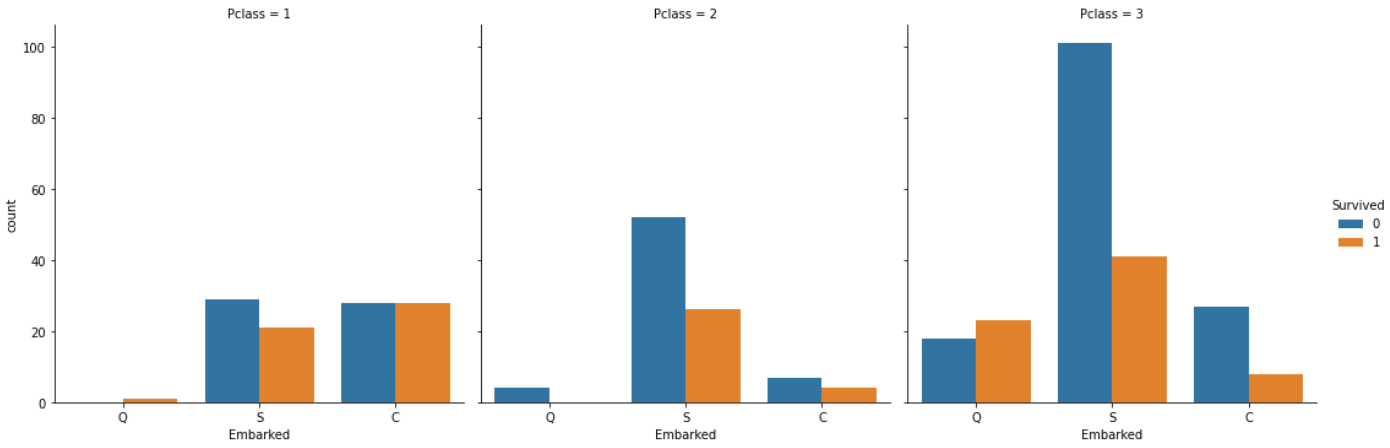

The survival of the passengers who paid more for the travel (more than 31 dollars) is higher when compared with the passengers who paid less than 31 dollars.




From the plot it is found that the count of survivors is highest in the case where the passengers didn't have siblings or spouse aboard the Titanic and the number of deaths is also the highest in this category.




Similar to the previous observation, it is found that the count of survivors is highest in the case where the passengers didn't have parents or children aboard the Titanic and the number of deaths is also the highest in this category.



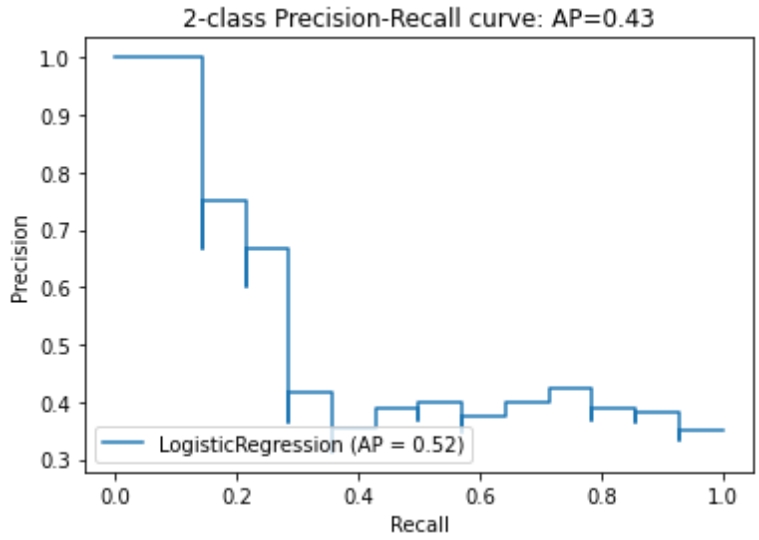

From the above plot it can be inferred that the group of passengers who travelled in the third class and departed from the Southampton port have the highest count of deaths and survivors when compared with the passengers of other passenger classes.

## Results:

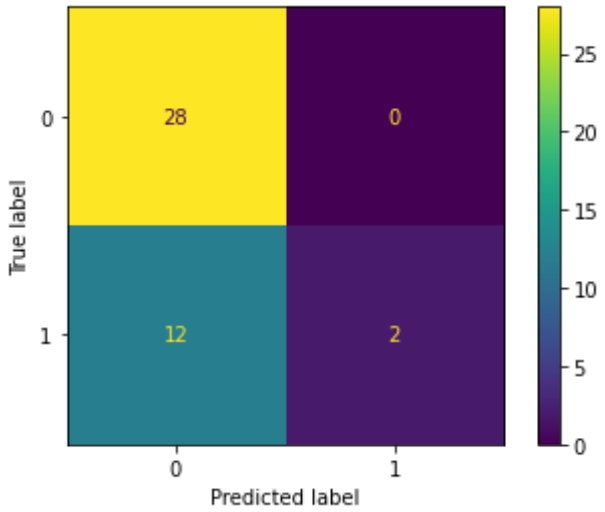
### Logistic Regression:

| Performance Metrics | |
|---|---|
| **Accuracy** | **0.7142857142857143** |
| **Precision score** | **0.7999999999999998** |
| **Recall score** | **0.7142857142857143** |
| **F1 score** | **0.25** |

**Average precision-recall score: 0.43**




## Confusion Matrix:




## Classification Report:

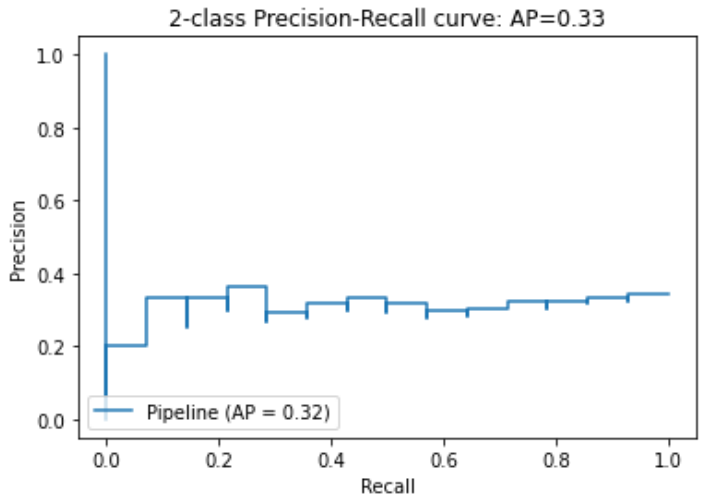
```
from sklearn.metrics import classification_report
target_names = ['Survived', 'Died']
print(classification_report(y_test, y_predict, target_names=target_names))
```

```
              precision    recall  f1-score   support

    Survived       0.70      1.00      0.82        28
        Died       1.00      0.14      0.25        14

    accuracy                           0.71        42
   macro avg       0.85      0.57      0.54        42
weighted avg       0.80      0.71      0.63        42
```

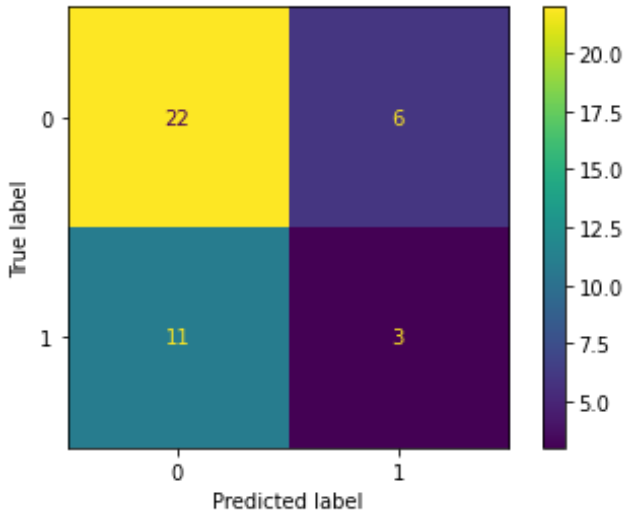
## Support Vector Machine:

| Performance Metrics | |
|---|---|
| **Accuracy** | **0.5952380952380952** |
| **Precision score** | **0.5555555555555555** |
| **Recall score** | **0.5952380952380952** |
| **F1 score** | **0.2608695652173913** |

**Average precision-recall score: 0.33**




**Confusion Matrix:**



**Classification report:**

```
target_names = ['Survived', 'Died']
print(classification_report(y_test, y_prediction, target_names=target_names))
```

```
              precision    recall  f1-score   support

    Survived       0.67      0.79      0.72        28
        Died       0.33      0.21      0.26        14

    accuracy                           0.60        42
   macro avg       0.50      0.50      0.49        42
weighted avg       0.56      0.60      0.57        42
```

## Conclusion:

Hence the titanic dataset has been analysed by data pre-processing and data visualization and the Machine Learning algorithms have been implemented.

Considering the analysis and performance of the two models it can be inferred that the Logistic Regression algorithm is more efficient. Hence the Logistic Regression algorithm can be used to predict the survival of the passengers on the Titanic according to this particular dataset.

## References:

Dataset taken from: https://www.kaggle.com/brendan45774/test-file