TITANIC SURVIVAL PREDICTIONS
PROJECT REPORT

Attila Péter Lőrincz

Budapest

2021

INTRODUCTION

1. Project description

The aim of this project was to predict the probability of survival of passengers based on different features like *Age, Sex, #Parents/Children, #Siblings/Spouses, Cabin code, Fare, Embarkation, Ticket.* These features are the predictor variables, and *Survived* (binary, 0 = no, 1 = yes) is the target variable.

The goal is to come up with a probability score of Survived = 0 for each passenger in the test dataset.

2. Project plan

2.1 Data preparation (training set)

The first step is ***data exploration and visualization*** in order to get an insight about how the predictor variables correlate to the target variable. One can use *countplots, distribution plots, boxplots, pointplots etc.* to visualize one by one the relationship of predictor variable and target variable. After that to get an overall information about the correlation of predictor values and the target value, one can creat a *heatmap* with a colorbar which would reflect if there is no correlation, there is a positive/negative correlation and the shade of the color would give us information about how strong the correlation is. This will be helpful to select features we want to move further with.

The second step is to ***encode features*** if it is necessary (in the case of categorical features) because machine learning model can work only with numerical features.

The third step is to handle **missing values and outliers**. It is a crucial step to handle missing values in such a way like *drop rows with missing values, fill missing value of a column with the mean/mode/median of that column, generate random values from the same distribution with the same mean and std value as the original distribution (to reduce bias), fill missing value with 0 or -1 etc.*

In the case of *outliers* one can *drop them, replace with the 2nd highest value of that column etc.*

The fourth step is ***feature engineering*.** Based on the data explorations and visualizations we could notice deeper relationship between predictor values which could have a significant effect (for example fare_per_person, age categories etc.) on the performance of the statistical model.

The last step of data preparation is to create a ***heatmap*** to visualize the correlation between the original+new features and target variable so we could get an idea about the importance of each predictor feature.

2.1. Data preparation (test set)

Perform the same feature engineering, missing value and outlier handling, feature encoding steps on the test set as on the training set.

2.2. Build Machine Learning models and choose the best

Choose different machine learning models which could be a good choice for this task and evalue the performance (accuracy) of each of them on the test set and then move further with the model with the best performance.

2.3. Build the final model and evaluate its performance

Build the model which performed the best. Apply it on the test set and then perfom ***hyperparameter tuning with gridsearch*** (to search for the best model parameters), ***validate*** your result ***with*** an n-fold ***cross validation***.

After that one has to ***evaluate the performance*** of the model. It is crucial to choose the best metric in order to avoid misinterpreting the performance of the model. For example if one has to predict a binary variable (0 or 1) then it is not enough to know that the accuracy of the model is 92 %. One needs to get a deeper insight into the performance. In the case of binary classification it is a smart decision to use *confusion matrix, precision and recall* to discover that if the model predicts something positive or negative then in what proportion of cases is true.

METHODS

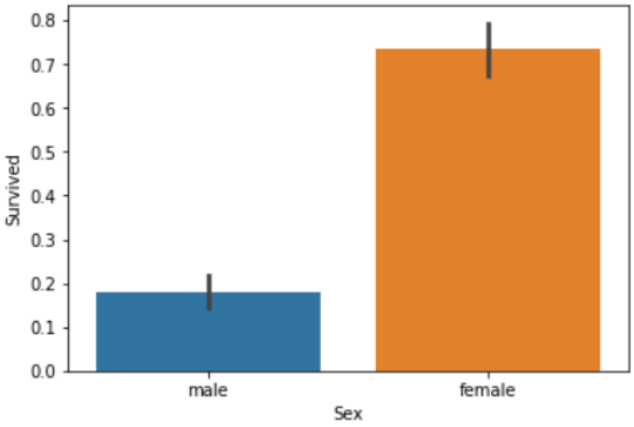
I have completed this project in ***python programming language*** beacuse it has the most userfriendly libraries for data science. The environment I was working with was ***Jupyter Notebook*** beacuse it is interactive, easy to use and can be used for presentations.

In the following sections I will present the data exploration results.

3. Data Exploration and Visualization Results

3.2 Sex

I have investigated the relationship between sex and survival rate with a barplot. The result can be seen on figure 1.

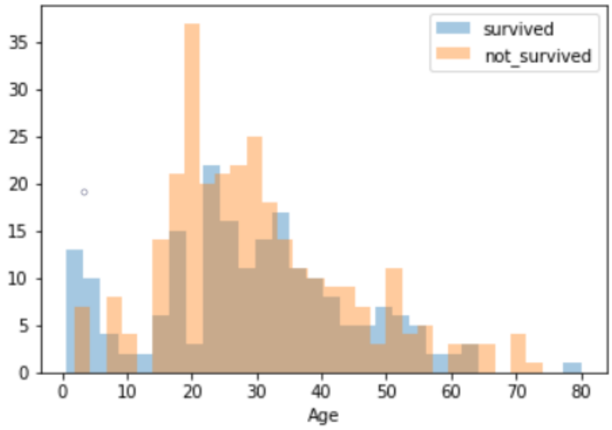


1. ábra Survival rate by Sex

Conclusion: Females had a greater chance of survival.

3.2 Age

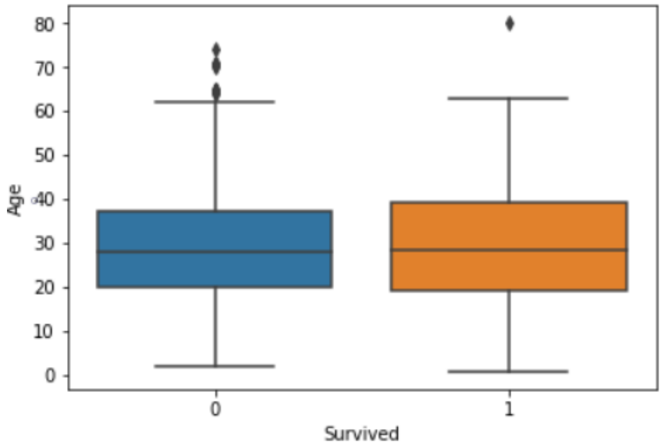
I have investigated the relationship between age and survival rate with a barplot. The result can be seen on figure 2.



2. ábra Survival by Age

Conclusions:

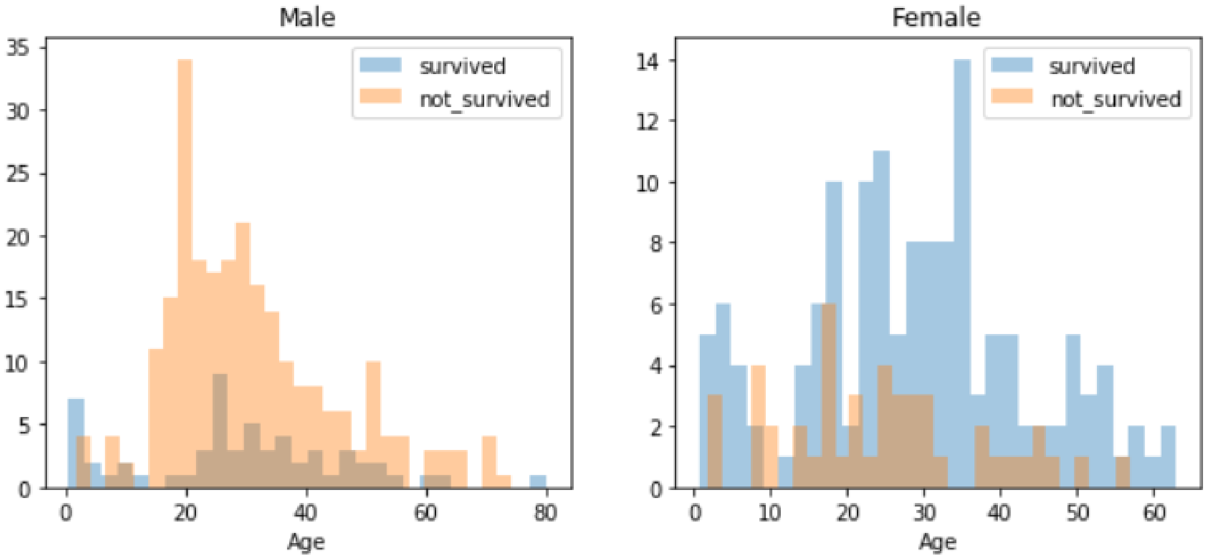
Figure 2 looks a bit crowded but one can notice a pattern in the age distribution. Most of the passengers who surived are aged between 20 and 40 years. But the same is true for the majority of dead passengers. I have made a boxplot which confirms my conclusion.



3. ábra Boxplot of age and survived

It can be seen on figure 3 that my conclusion from figure 2 is true. That is **most of the passengers who surived/died are aged between 20 and 40 years.**

I also investigated the ages by sex. The result can be seen on figure 4.



4. ábra Survival by Ages and Sex

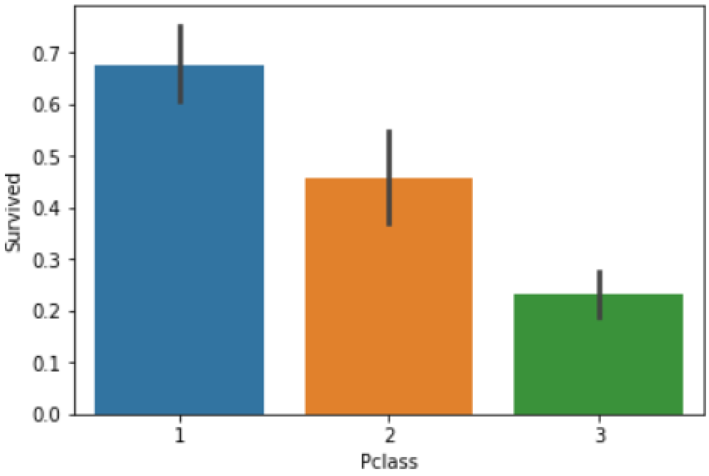
Figure 4 reflects that

- **females and children had a greater chance of survival**
- **most of the passengers who surived/died are aged between 20 and 40 years**

It would be a smart choice to **divide Age into 6 bins.** Which I did.

3.3 Pclass

I have investigated the relationship between pclass and survival rate with a barplot. The result can be seen on figure 5.

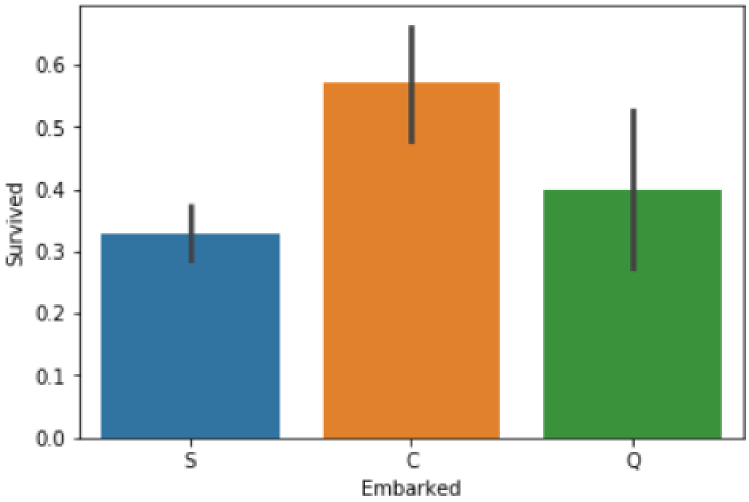


5. ábra Survival by Socieconomic Class

Conclusion: People with a higher socioeconomic class are more likely to survive. There is a much higher probability that the person in pclass 3 will not survive.

3.4 Embarked

I have investigated the relationship between port of embarkation and survival rate with a barplot. The result can be seen on figure 6.



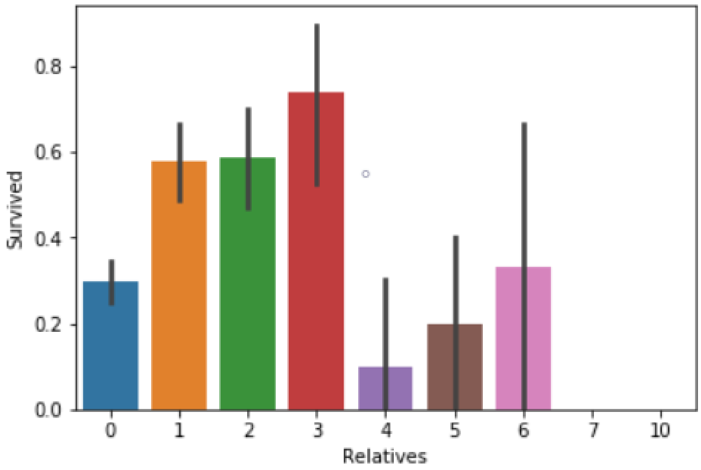
6. ábra Survival Rate by Port of Embarkation

Conclusion: Most of the passengers embarkedked at port S and most of the survived passengers embarked port C.

3.5 SibSp + Parch = Relatives

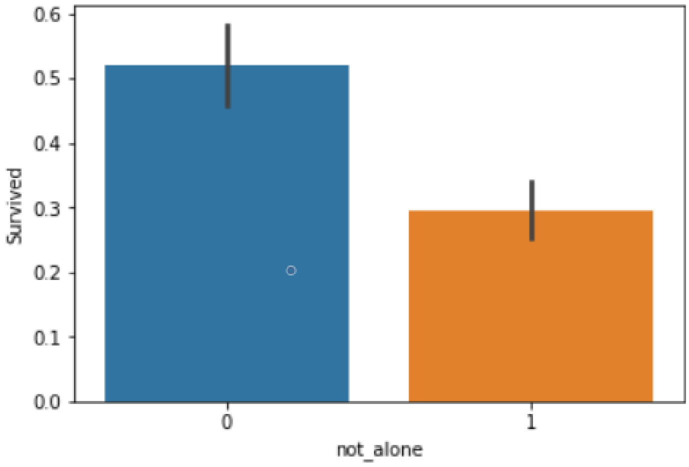
I have added together number of sibling/spours and number of parents/children and named the feature „Relatives”.

I have investigated the relationship between number of relatives and survival rate with a barplot. The result can be seen on figure 7.



7. ábra Survival Rate by Number of Relatives

Conclusion: We can see from the figure 7 that **those with 1-3 number of relatives has a higher probability to survive while those who are alone are less likely to survive. -->** We could introduce a **new feature** called **" Not Alone".**



***8**. ábra Survival Rate by alone or not alone*

I have investigated the relationship between number of relatives and survival rate with a barplot. The result can be seen on figure 8.

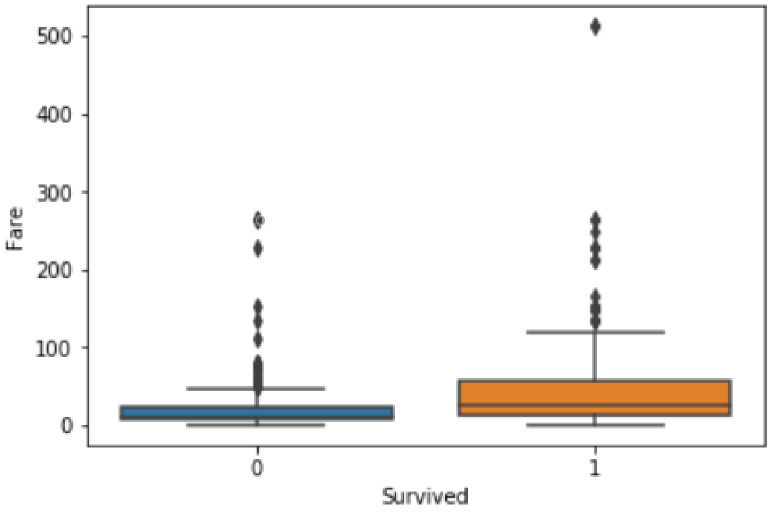
Conclusion: It confirms the hypothesis that those who are alone, are less likely to survive.

3.6 Cabin

79.5% of the cabin values are missing so I decided to drop this column.

3.7 Fare

I have investigated the relationship between fare and survival rate with a barplot. The result can be seen on figure 9.



9. ábra Survival Rate by Fare

Fare column contains an outlier. I decided to remove it.

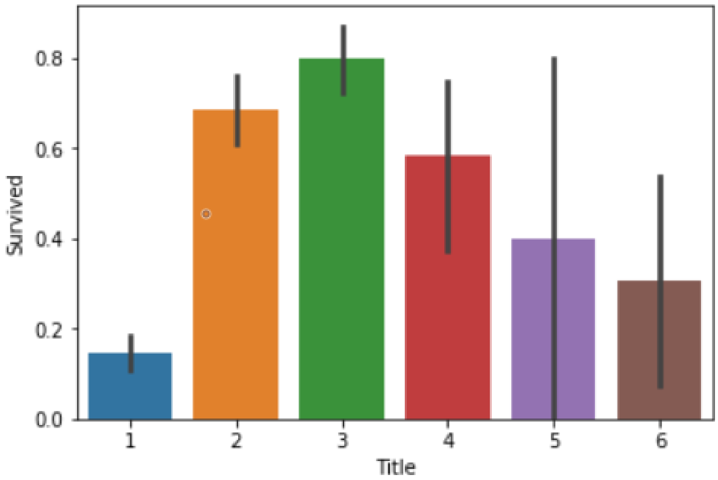
Conclusions:

- **Those who survived payed a higher fare.**
- **Most of them who payed a higher fare were not alone and not alone means higher chance to survive so it would be a smart decision to introduce a new feature "Fare per person**
- **Fare contains outliers. We have to handle them (for example replace it with the 2nd maximum value, median value or just drop them)**

3.8 Name / Titles

I have extracted titles from names because I thought it would matter wheter or not a passenger is a doctor or miss etc. These titles are in relationship with the financial state of the passenger. So if someone is a doctor then maybe he/she is richer than a mister and since he/she is richer he/she could payed a higher fare which mean higher chance of survival.

The result can be seen on figure 10.



10. ábra Survival Rate by Titles

Conclusion: This plot almost confirms my theory about the relationship of survival and titles.

3.9 Data Manipulation

3.9.1 Missing values

- **Embarked:** I filled the missing embarked value with the most common value of that column
- **Age:** I dropped rows with missing age values

3.9.2 Outliers

- **Fare:** I dropped outlier (> 263) values

3.9.3 Age bins

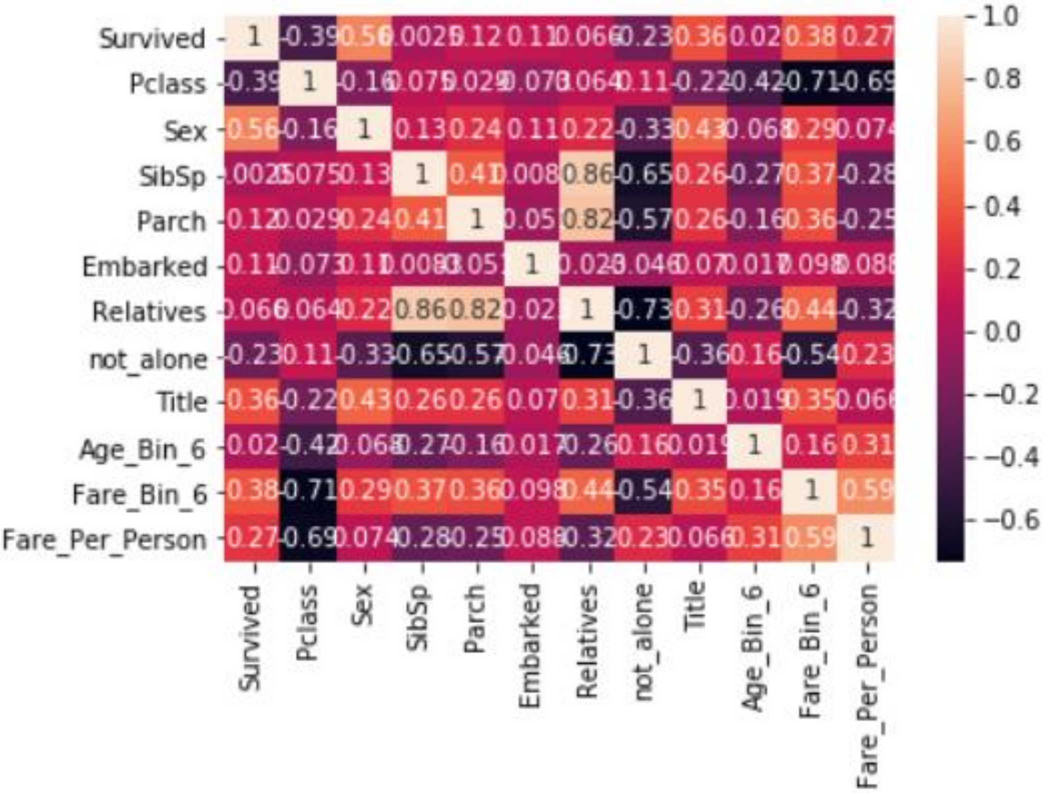
I have divided age values into 6 bins so that age values are roughly at the same scale as other feature values.

3.9.4 Fare bins

I did the same with the fare values as with the age values.

3.9.4 Heatmap

I have created a heatmap to visualize the overall correlation between features.



Conclusions: Pclass, Sex, Title, Fare_Bin_6, Fare_Per_Person are the most important predictors of the target variable.

4. Choose the Best Machine Learning model

Before testing several machine learning models, I have divided the training set into **X_train** and **Y_train** dataframes containing the predictor and target variables, respectively. I also created an **X_test** dataframe containing every variable except *PassengerId.*

I have tested several regression and classifier models (*logistic regression, randomforest classifier, decision tree classifier, stochastic gradient classifier)* and evaluated their accuracy and based on the accuracy values I have chosen the best one.

1. táblázat Model performances

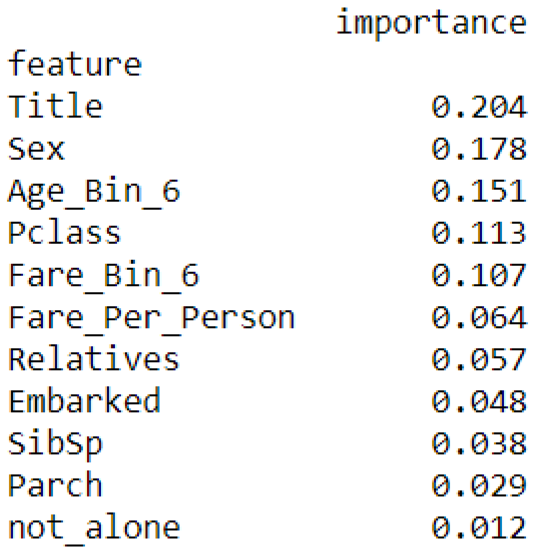
MODEL	ACCURACY
LOGISTIC REGRESSION	80.89 %
RANDOM FOREST CLASSIFIER	93.09 %
DECISION TREE CLASSIFIER	93.09 %
STOCHASTIC GRADIENT CLASSIFIER	68.5 %

In table 1 one can see that randomforest classifier and decision tree classifier performed the best. I have **chosen random forest classifier** because it is a widely used model for machine learning tasks.

5. Final Machine learning model

5.1. Feature importance

At first I investigated the feature importance of each feature for randomforest classifier. The result can be seen on the following figure.



```
                 importance
feature
Title                 0.204
Sex                   0.178
Age_Bin_6             0.151
Pclass                0.113
Fare_Bin_6            0.107
Fare_Per_Person       0.064
Relatives             0.057
Embarked              0.048
SibSp                 0.038
Parch                 0.029
not_alone             0.012
```

11. ábra Feature importance

As we have seen ont he heatmap, ***title, sex, age_bin_6, pclass, fare_bin_6*** has significant effect on the accuracy of our machine learning model.

5.2 Data Manipulations on The Splits

Based on feature importance values I modified the **X_train** and **Y_train** dataframes to contain only these top 5 features and I repeated the model fitting.

After fitting I performed **gridsearch** to find the best parameters which with the model perfomes the best. Then I did refit the model and validated the results with a **10-fold cross validation.**

I will discuss the results in chapter 6.

6. PERFORMANCE

- **Before top 5 features:** 80.52 % accuracy with a standard deviation of 6.61 %.
- **After grid search: 83.76 % accuracy**

After selecting the 5 most important features, the model accuracy has been improved with 3.24 % which I think is a great improvement and my model can be used to predict probability of death.

6.1. Precision and Recall

- **Precision**: Proportion of positives that are correctly identified (if my model predicts a passenger survived, it will be correct 80.41% of the time).
- **Recall:** Proportion of negatives that are correctly identified (if my model predicts a passenger dead, it will be correct 78.78% of the time).

Precision: 80.41 %

Recall: 78.78 %

A good model has the same precision and recall (or almost the same). For my model, the difference between the two is 2.03 % which I think is a good result and my model can be used to complete the probability prediction of death.

7. SUMMARY:

I have build a random forest classifier machine learning model to predict the probability of death for each passenger. After hyperparameter tuning my model perfromed with **83.76 % of accuracy**, **80.41 % of recall** & **78.78 % of precision.**

Since the precision and recall values are almost the same (which is true for a good model), my model is trained well enough to predict the probability of death.