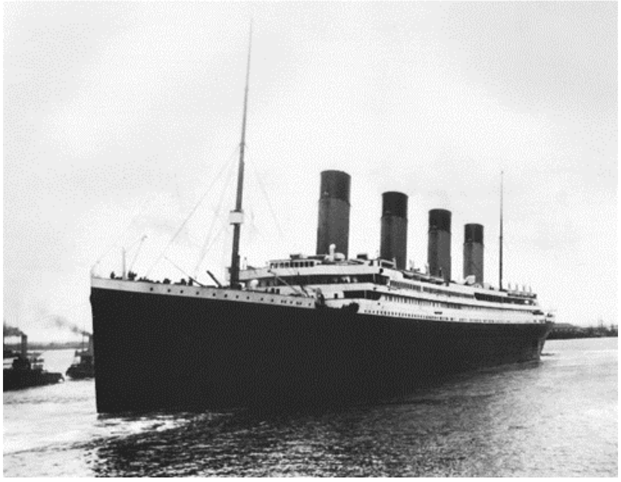
MACHINE LEARNING PROJECT REPORT

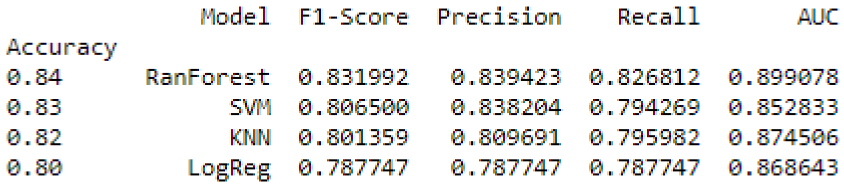
Machine learning combined with Titanic, the most famous ship of all the time. Using different classification models on the Titanic data set, we will compare the models according to their performance.

Since there are so many algorithms, it can be difficult to decide which one to choose. By comparing the models we use, we can determine the most suitable model for our data set.



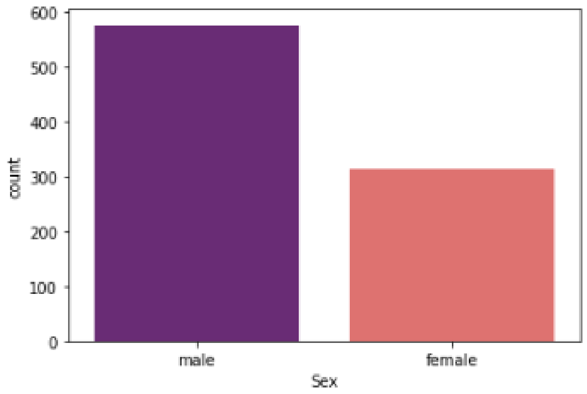
Let's examine and interpret the comparison table at the end of the code. There are several performance metrics to evaluate classification models:

Accuracy, F1-Score, Precision, Recall and AUC.



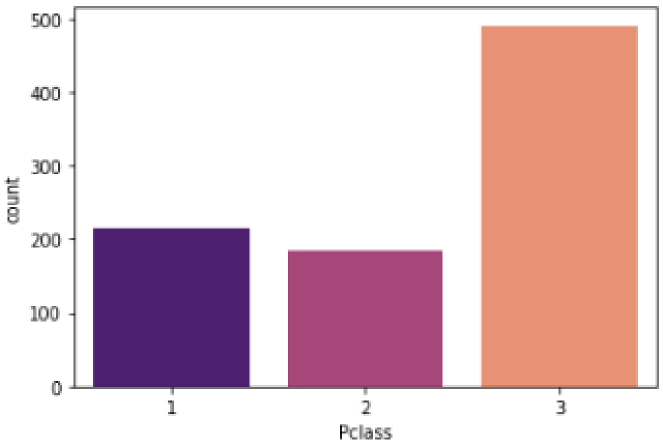
```
              Model  F1-Score  Precision    Recall       AUC
Accuracy
0.84      RanForest  0.831992   0.839423  0.826812  0.899078
0.83            SVM  0.806500   0.838204  0.794269  0.852833
0.82            KNN  0.801359   0.809691  0.795982  0.874506
0.80         LogReg  0.787747   0.787747  0.787747  0.868643
```

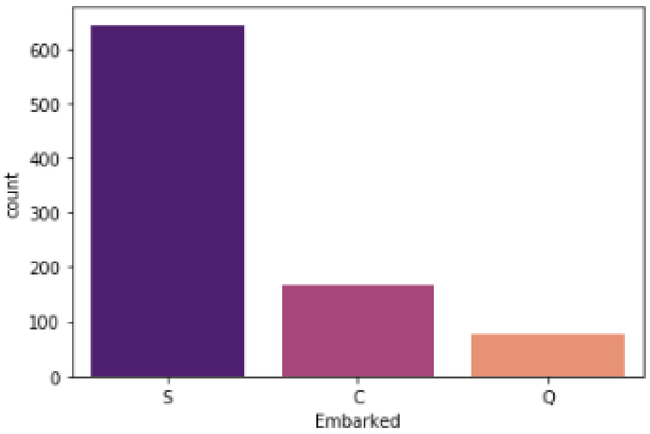
Here I made a sorting according to accuracy. Accuracy is the best metric to measure performance relatively. However, every metric makes a unique comment about our model. So every metric has something to tell us about our model. Therefore, instead of interpreting the performance of the model by looking at a single metric, we can reach a more accurate decision by examining other metrics. We determine the most appropriate performance metric by looking at our data set. Whether our data set is balanced or imbalanced changes our preference. The Titanic dataset is an imbalanced data set. I can show this with examples. I will give two examples showing that the **data set is imbalanced.**



In the graphic on the left, we see that the number of men on the ship is twice the number of women.

In the example on the right, we see that the class P3 is more than the sum of the classes P1 and P2.





We also see a significant disproportion in the number of preferred ports. Sex, Pclass and Embarked are relevant features for the survival.

As a result our data set is skewed.So it is imbalanced data set. Therefore, we will give more importance to F1-Score and accuracy. Because these are more suitable metrics to evaluate imbalanced data sets. Anyway, F1-Score combined precision and recall.

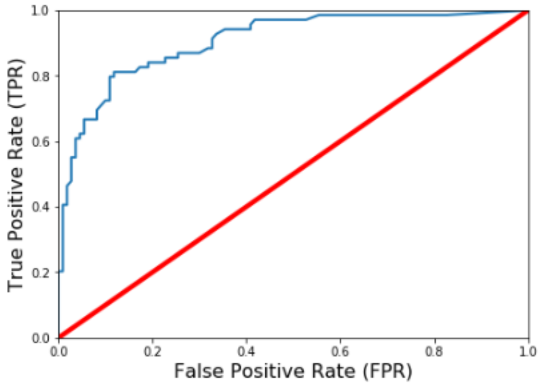
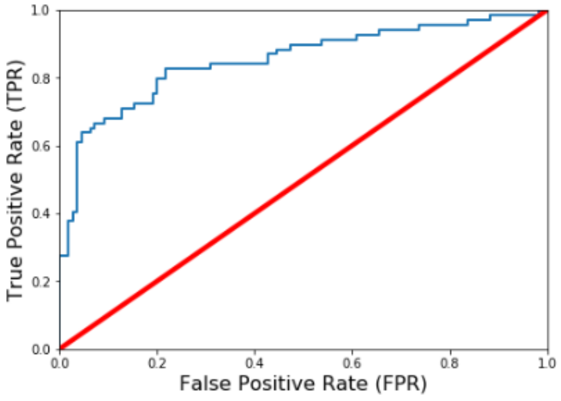
In summary, we are interested in predicting the survival status of all passengers in the test set correctly and accuracy is the most vivid metric for us(Accuracy is used when the TruePositive and TrueNegatives are more important). Besides accuracy, we'll look at the F1-Score to avoid choosing the wrong model. In addition, we will examine the ROC curves and AUC values. However, ROC curve could be misleading with imbalanced data set. We will also consider this. If we consider all of this on the table, we can clearly see the performance sorting. So we can show the performance sorting as follows:

Random Forest > SVM > KNN > Logistic Regression

Random Forest is the best performing model according to the data set.

Let's examine a little more.

The AUC value is pretty good on every model. Let's compare the ROC curves of the two best models.

	RANDOM FOREST	SVM
ROC CURVE		
AUC	0.899078	0.852833

As clearly seen, the AUC area is larger on the chart of the Random Forest. Also, the ROC curve of Random Forest tends to the upper left corner. It is directly proportional to the success of this model. In addition, the performance metrics of Random Forest are higher than other models.

The results and observations show that Random Forest are a more reliable more of classifiers according to our data set.

Of course, this does not mean that Random Forest is a better model than SVM. This is a result we get only according to the data set we have and the hyperparameters we have determined.