North South University

Project Report

Submitted To:

Intisar Tahmid Naheen (ITN)
North South University

Submitted By :

Name : Md Habibullah
ID : 1712220642
Sec : 04
Course Title : CSE445

Submission Date: 23-09-2021

Abstract

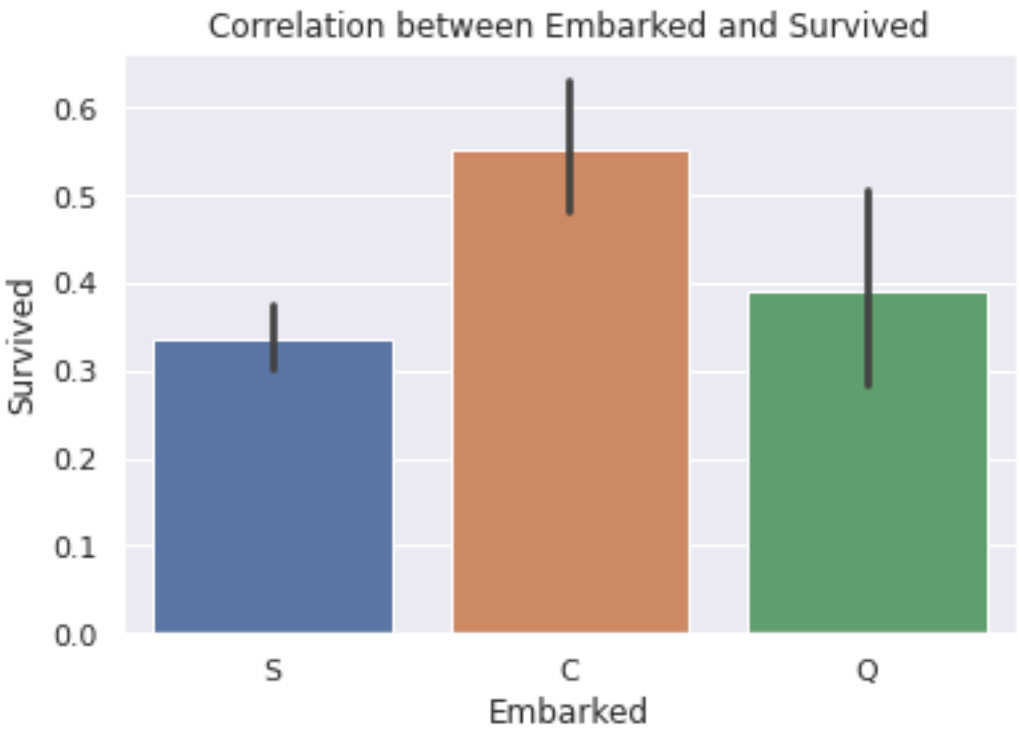
The Titanic was a British passenger ship that sank after striking with an iceberg in the North Atlantic Ocean. Although there was some element of luck involved in surviving the sinking, particular groups of people, such as women, children, and the upper-class, were more likely to survive than others. The Titanic's sinking is one of history's most famous catastrophes. The international society was shocked by the catastrophe, and the ship's safety standards were raised as a result. The absence of lifeboats for both passengers and crew was one of the reasons for the high death toll. We will utilize machine learning approaches to determine which passengers survived in this report, as well as perform a predictive study of the categories of persons who were likely to survive the accident.

Introduction

The Titanic was lost at sea. One of the worst accidents in history, with thousands of passengers and crew members killed. The majority of the deaths were caused by a lack of lifeboats. Observations that will leave you speechless. It was discovered that certain people, such as children and women, were more prone to being overweight than others. Who had a higher priority to save? Technology's inexorable growth has made our lives easier while also posing certain obstacles. One of the technology's advantages is that it makes it simple to access a large range of data when required. However, finding appropriate information is not always possible. Raw data from online sources does not make sense on its own and must be processed before being used as a source of information. Feature engineering methods and machine learning algorithms are critical in this instance. The goal of this paper is to combine machine learning and feature engineering to produce as accurate outcomes from raw and missing data as possible. As a result, one of the most important datasets in information science, the Titanic, was used. This dataset provides data on Titanic passengers, including who survived and who did not. A number of missing and non-linear variables were revealed to have hindered the prediction's efficiency. For a thorough data analysis, the impact of characteristics was investigated. Some new features have been added to the dataset as a consequence, while others have been removed.

Data Pre-processing

Some of the data in dataset accessible for prediction and also missing values or they are null. The missing data was reduce the overall model accuracy result. Data preprocessing is the solution transforming raw data into an understandable data. Missing values are replaced by mean value which is average of that column. So, the missing and unknown data of the passengers which is easily predictable is filled up by this step.



Dataset

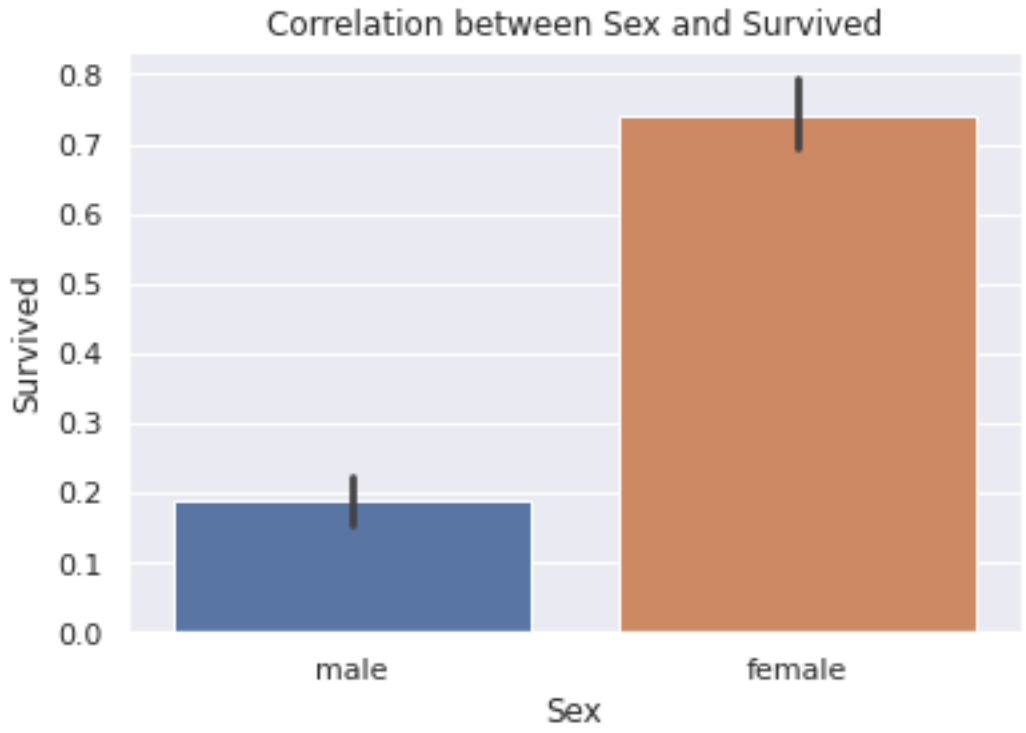
The training dataset (70%) and the test dataset (30 %) were separated from the original data. The training set is used to create our machine learning models. The training set includes our objective variable, passenger survival status, as well as other independent variables like gender, class, fare, and Pclass. The test set should be used to see how well our model performs with based on train model. We are not able to use anything from the test set. The probability of passengers surviving We'll use our model to predict if a passenger will make survive or not. The test set should be used to assess our knowledge of this particular data.

Attributes in Data Set:

	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Passenger class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age of the passenger	

sibsp	Number of siblings or spouse on the ship	
parch	Number of parents or children on the ship	
ticket	Ticket number	
fare	Price of the ticket	
cabin	Cabin number of the passenger	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Plot of Survive Male and Female:



Methodology

The training data from the Kaggle titanic dataset contains 892 rows and 12 columns. The first row of the dataset describes the different parameters for a passenger. The first column in the dataset gives the Passenger Id of a passenger and the second column of the dataset gives whether the person survived or not. The PClass attribute defines the class in which the passenger was travelling in the ill-fated ship.

Algorithm

Prediction models are generated using four machine learning algorithms Logistic Regression, Naïve Bayas, Support Vector machine (SVM), Decision tree and Random Forest. Those algorithms are compared to another on the basis of the validity percentage. The attributes used in the test and train dataset for implementing these algorithms are- Pclass, Sex, Age, SibSp, Parch and Fare respectable.

Logistic Regression

Logistic Regression is a type of classification algorithm which target variable is categorical and binary. This algorithm founded on idea that independent variables can predict the value of a dependent variable. Similar to the other regression analyses, LR is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and other are nominal, interval, ordinal or ratio-level independent variables. It uses a method of using regression line between dependent and independent variable to predict the value of the dependent variable.

Naive Bayes:

Naive Bayes algorithm is a linear-classifiers depend on Bayes theorem. The model generated is probabilistic. It estimates conditional probability which is the probability that something will happen, given that something else has already occurred. Missing values are much easier to deal with this Naive Bayes model. NB performs effectively on datasets with high dimensional and complexity.

Support Vector Machine:

SVM - Support Vector Machine is a supervised machine learning algorithm, which is also use both classification and regression algorithm. In this algorithm each data item represent point in a “n” dimensional (ie, 1D,2D,3D..nD) space with the value of each feature being the value of a particular identical, where “n” = “number of features”. Each SV represents the co-ordinates of

an individual observation. Support Vector Machine itself is a boundary, which is soft and hard boundary. SVM using this kernel trick make best fit line and gave higher accuracy.

Decision Tree

One of the most commonly used classifiers is decision trees. Decision tree algorithm give a graphical representation of all the possible solutions to decision make on perticular conditions. Decisions made can be easily explained. Branching is done with decision nodes, and class labels are specified via prediction nodes. In this way decision tree algorithm work: Begin the tree with the root node, says S, which contains the complete dataset.--> Find the best attribute in the dataset using Attribute Selection Measure (ASM). -->Divide the S into subsets that contain possible values for the best attributes.-->Generate the decision tree node, which contains the best attribute.-->Recursively make new decision trees using the subsets of the dataset created in 3rd step. Repeating this procedure until a stage reached where further classify are not reach the nodes and mention the final node as a leaf node.

Random Forest

Random forest(RF) algorithm can be used for both classification and regression problems. RF is a supervised learning algorithm. It creates a forest to evaluate results. Random Forest builds multiple decision trees by picking 'K' number of data points from the dataset and merges them together to get a more accurate and stable prediction. For each 'K' data points decision tree we have many predictions and then we take the average of all the predictions. Random forest is an Ensemble learning Algorithm. Ensemble learning is the act of combining numerous models to predict a single outcome.

Results

All algorithms are run to assess the probability of passengers and crew surviving, as well as what characteristics are linked to passenger and crew survival. We observed that certain model parameter modifications were required to correct the approach when applying the algorithm to the Titanic dataset. For comparing the five techniques used in this project two metrics are used. First metric is accuracy and the second metric is false discovery rate. Both these metrics are computed using the confusion matrix. The structure of the confusion matrix is shown in below. Accuracy is the measure of how a model best predicts. Higher the accuracy of better. Accuracy is calculated using the formula TN+TP/ Total number of test set rows*100. False discovery rate are the false positive measures of confusion matrix where the model predicts that the passenger would survive but in reality, it doesn't. This would prove dangerous as the prediction may go wrong and hampers the accuracy of the results. The attempts are being made to increase the accuracy rate and reduce the false discovery rates.

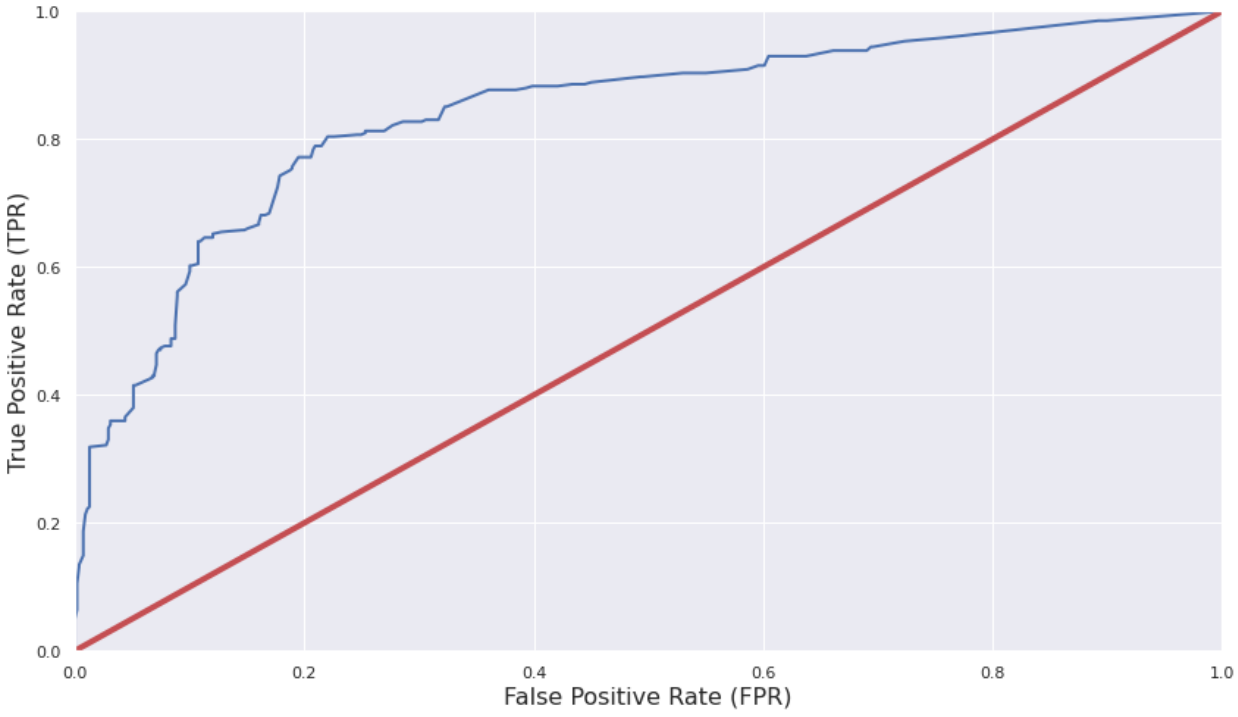
False discovery rate is calculated using the formula FP/FP+TP *100. Hence lower the false discovery rate the better. The accuracy and false discovery rate for each of the algorithm.

Accuracy Results

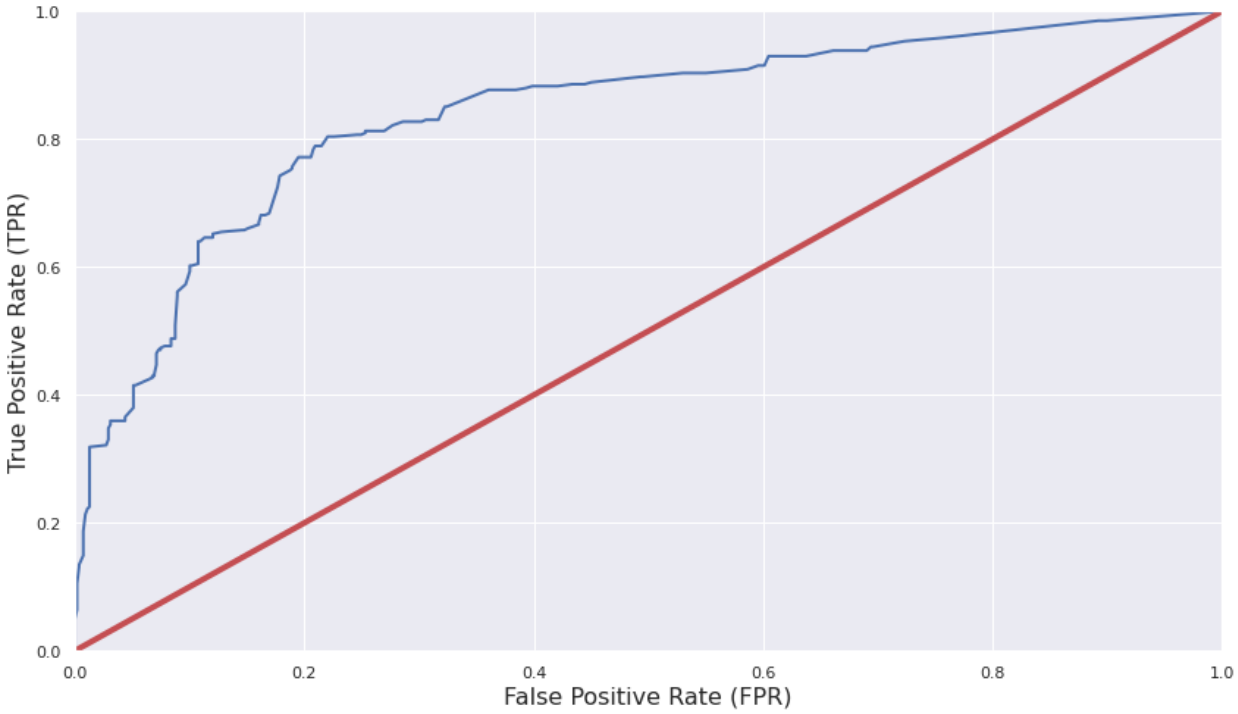
Score	Model
78.90	Naive Bayes
80.25	Logistic Regression
82.04	Support Vector Machines
87.21	Decision Tree
87.21	Random Forest

ROC Curves:

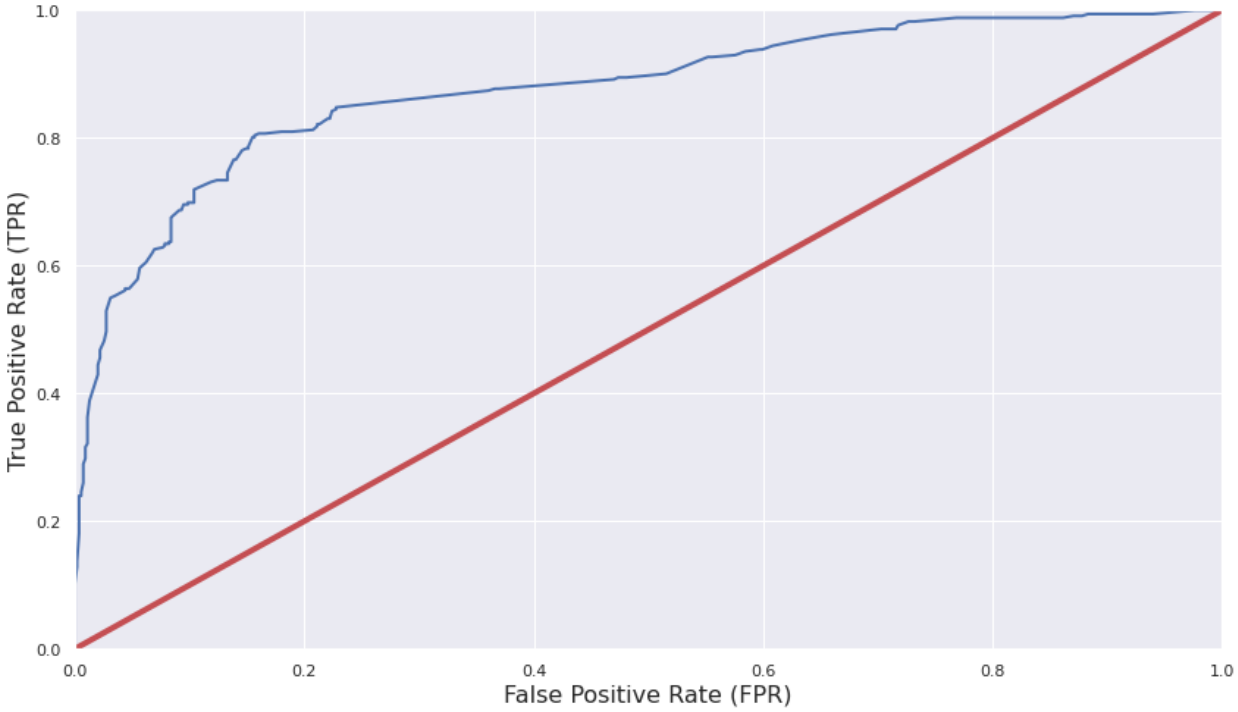
1. Logistic Regression



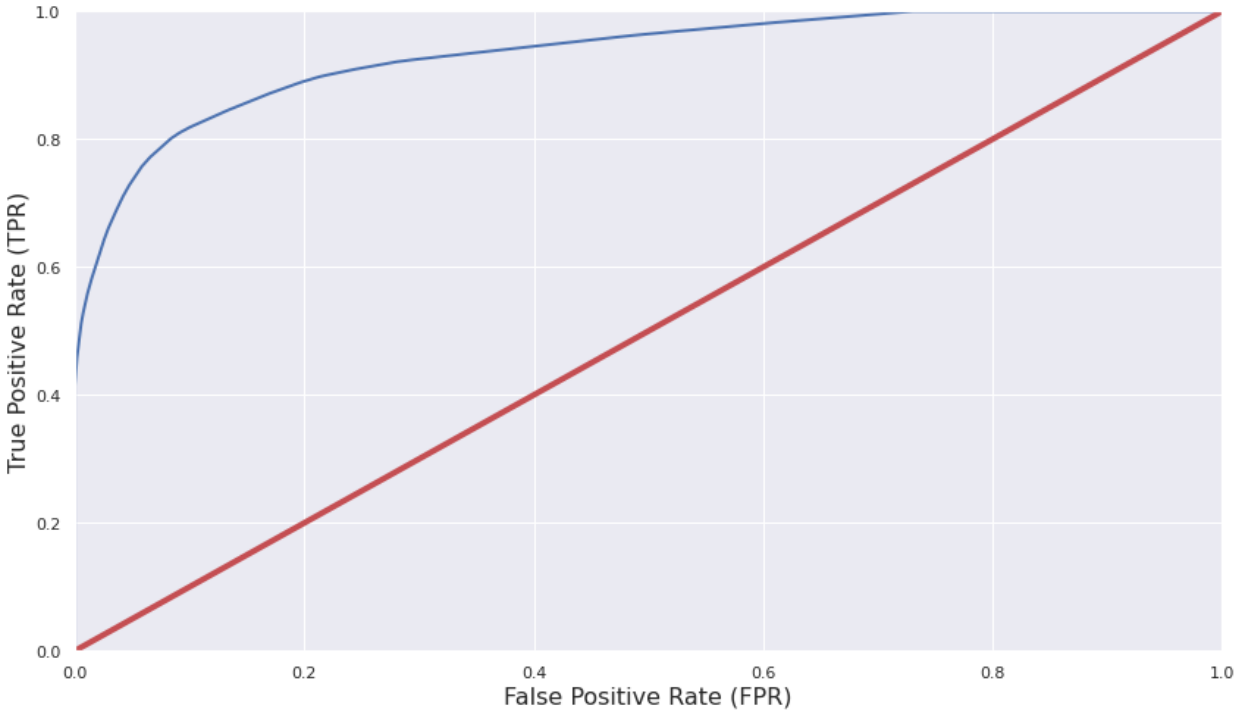
2. Naive Bayes



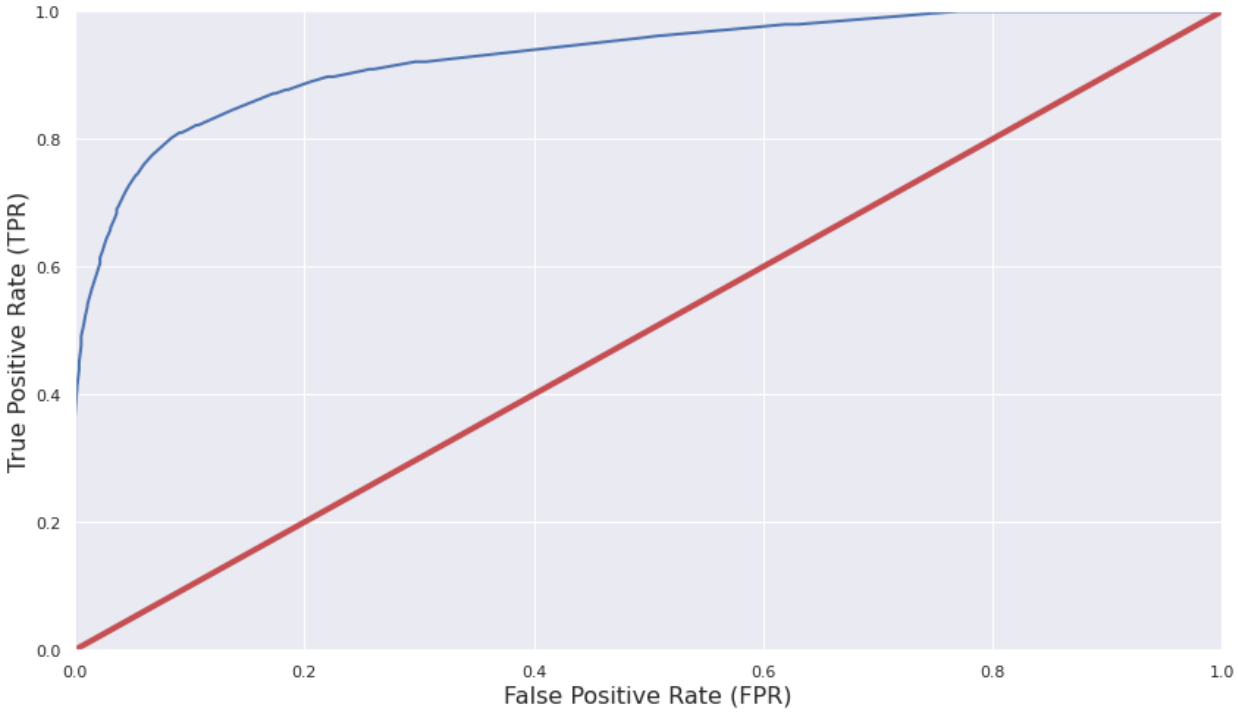
3.SVM



4.Decision Tree

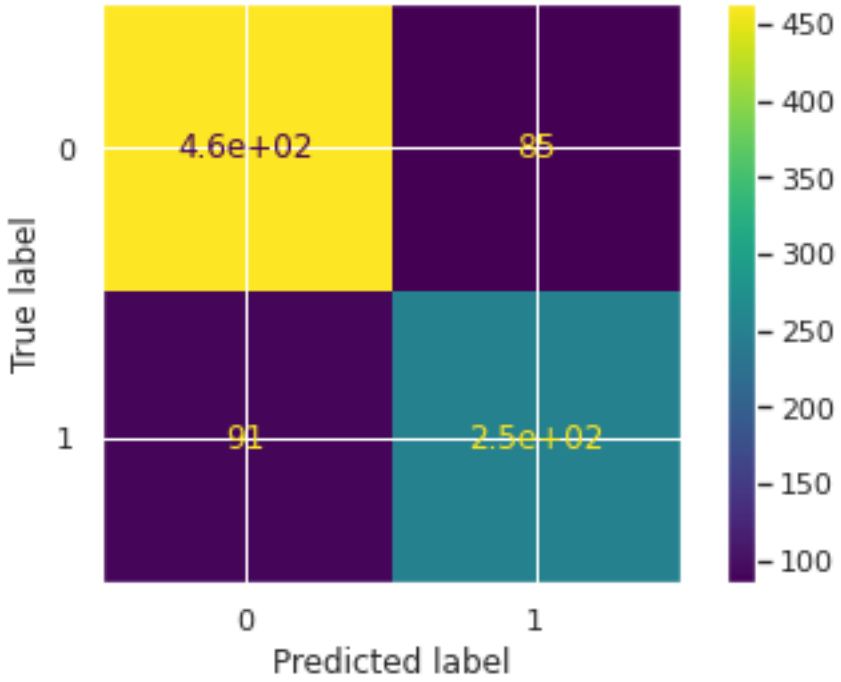


5.Random Forest

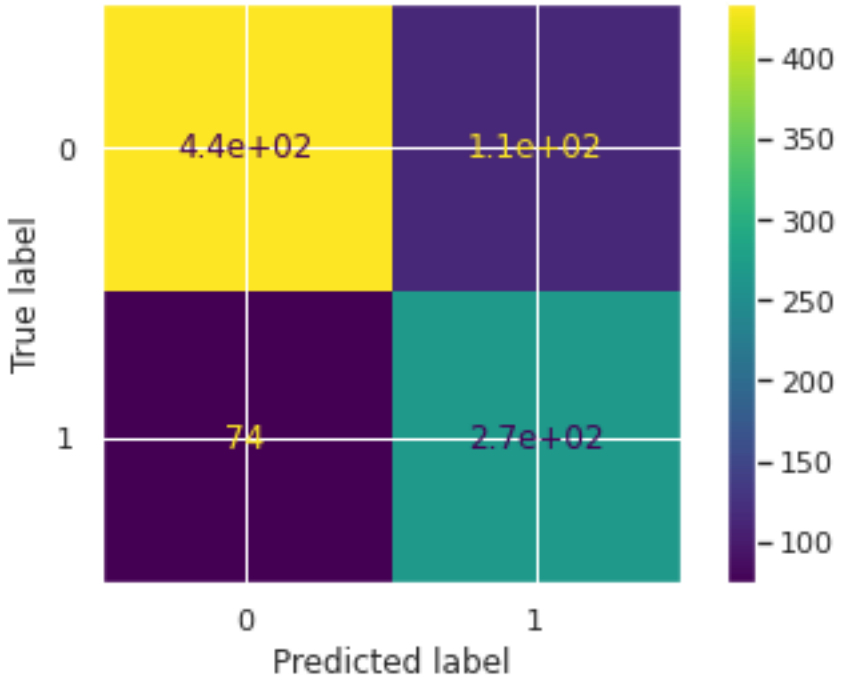


Confusion Matrices

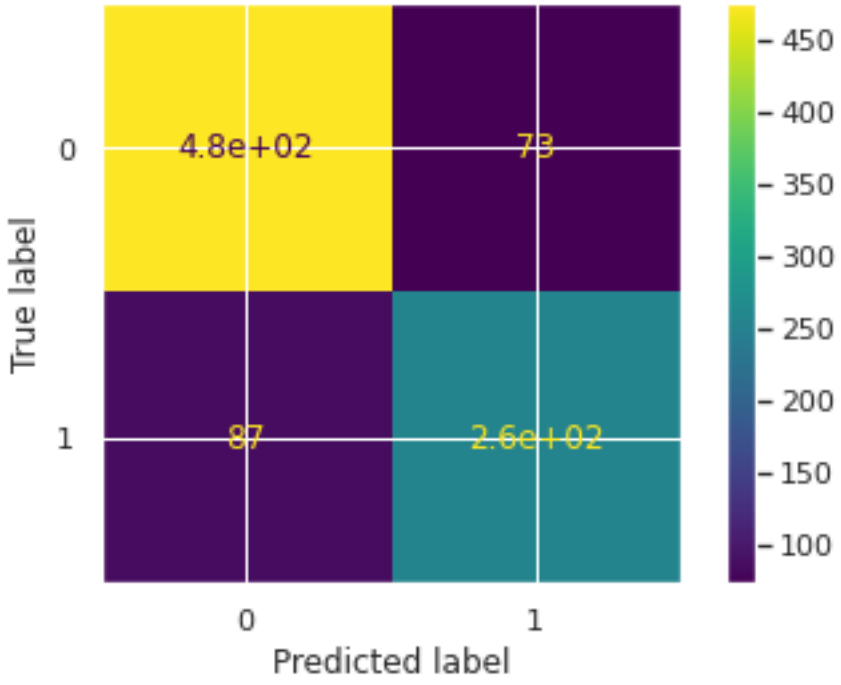
1.Logistic Regression



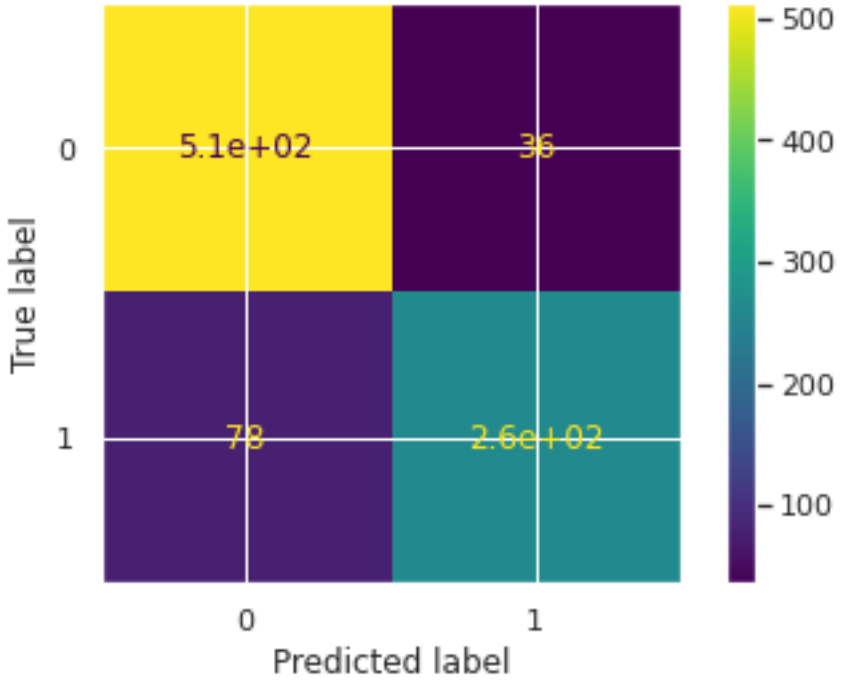
2.Naïve Bayes



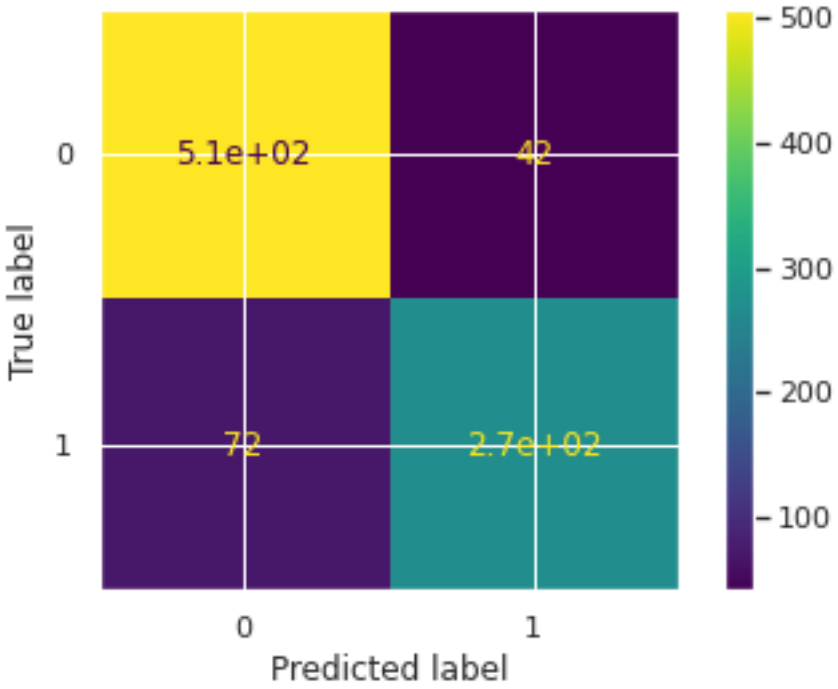
3.SVM



4.Decision Tree



5.Random Forest



Conclusion

The accuracy of the five techniques we evaluated many significantly different. Random Forest and Decision Tree proved to be the best algorithm for the Titanic classification problem since the accuracy of Random Forest and decision tree is the highest and the false discovery rate is the lowest as compared to all other implemented algorithms. This project also determined the features that are the most significant for the prediction. Decision Tree, as well as Random Forest, suggested that Pclass, sex, age, children, and SibSp are the features that are correlated to the survival of the passengers. Future work might include more different algorithms for better accuracy achieved. Include also cross-validation that is calculating accuracy based on different combinations train and test data. It would be interesting to work more with dataset and introducing more attributes that might lead to good results.