NORTHEASTERN UNIVERSITY
Department of Electrical & Computer Engineering

EECE 5644 - Machine Learning / Pattern Recognition
Fall 2022

Final Report
Titanic - Machine Learning from Disaster

Instructor: Professor David Brady

Group 1
Jianing Chen
Tianqi Huang
Guangping Liu
Juncong Zhu

Semester: 2022 Fall

Abstract

Titanic was a passenger liner which sank in the North Atlantic Ocean in 1912 after striking an iceberg. A good model which analyzes the critical elements in survival can benefit companies and other institutes in many aspects. Shipbuilding companies who enhance their security level based on information given by our model would hold a strong lead in local markets, and earn more profits. Insurance companies may also have a good reference for their security evaluating insurance policy against all kinds of similar accidents. This research would also increase knowledgement to this infamous accident of human history .
In this project, we are trying to perform the Exploratory Data Analysis (EDA) and using various machine learning methods based on the passengers training and testing datasets of Titanic and predict the survival rate based on the analysis.
Since analyzing the relevance of samples in one category and the samples of the survival rate could be insufficient to predict the survival model when a similar disaster occurs. Therefore, if time allows, more sample data on different categories should be analyzed together to obtain classifiers with more precision.

Data Overview & Correction

Variable Name	Meaning
PassengerID	Random id assigned by the database
Survived	1=survived; 0 = not survived
Pclass	1=upper class; 2 = middle class; 3 = lower class
Name, Sex & age	Basic information of passengers
SibSp & Parch	# of siblings / spouses & parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation: S=Southampton; C=Cherbourg; Q=Queenstown

Correction
Missing value
    Age:median()
    Embarked: mode()
    Fare: median()
Modification
    “male”: 1 , “female” : 2
    Embarked: “S”:1, “C”:2, “Q”:3
    Selection
Drop: “PassengerID” , “Cabin”, “Ticket

Beginning Stage

In the beginning of the project, it is proposed that the survival rate should be directly associated with the money a passenger spent and the class of his/her ticket. Therefore, using the training data set, creating a histogram will be the best way to visualize the proportion of people on different amounts of money he/she spent on boarding. This could be done using the “histogram” function of MATLAB.

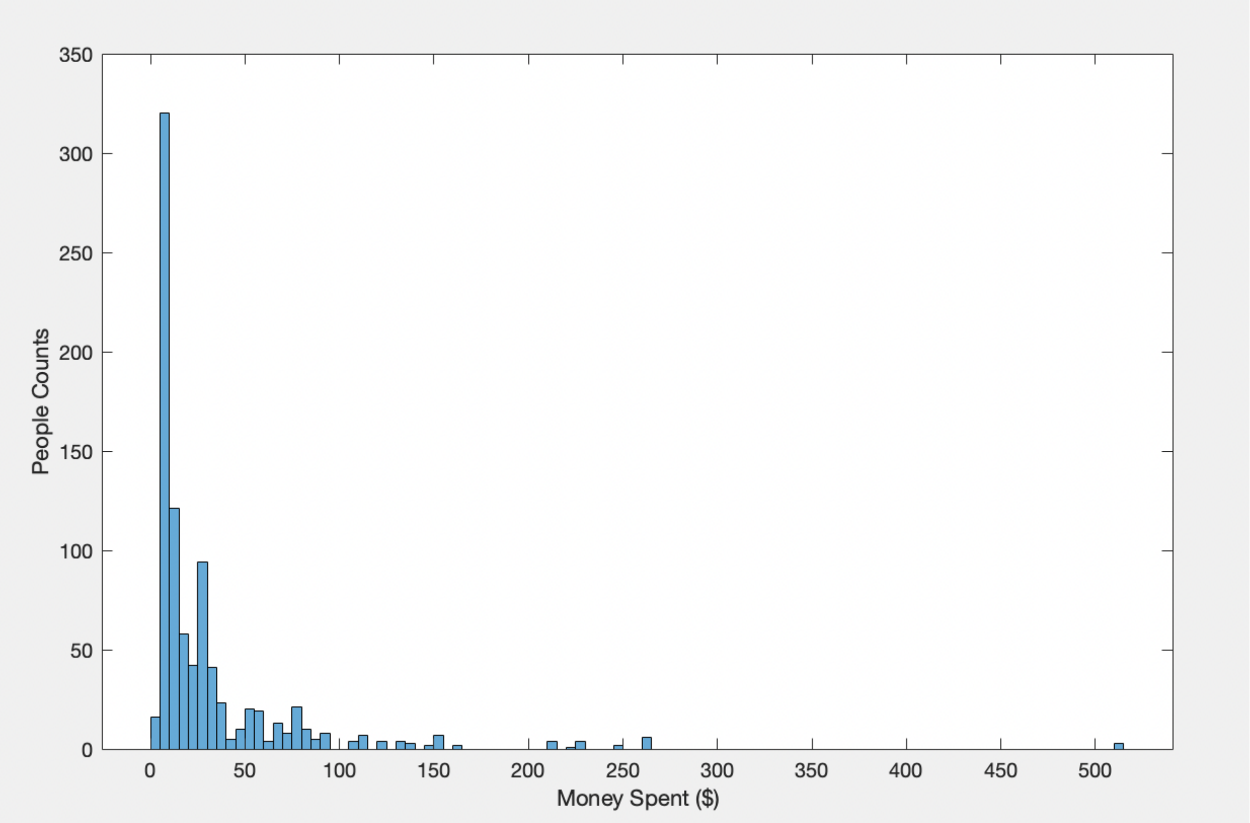

Figure 1: The histogram of Number of People vs. The money he/she spent

From the histogram, it shows that the majority of passengers spent between 0 dollars and 100 dollars to get on board, meanwhile, there are some outliers. Before analyzing, to make the estimation more precise, the outliers have to be removed. After removing the outliers, the first model being implemented is the logistic regression model. The reason for applying logistic regression is because a binary class set is being analyzed (Died vs. Survived), and logistic regression often provides a clear understanding for the relationship between how one feature of a training set relates to the class of the corresponding data.

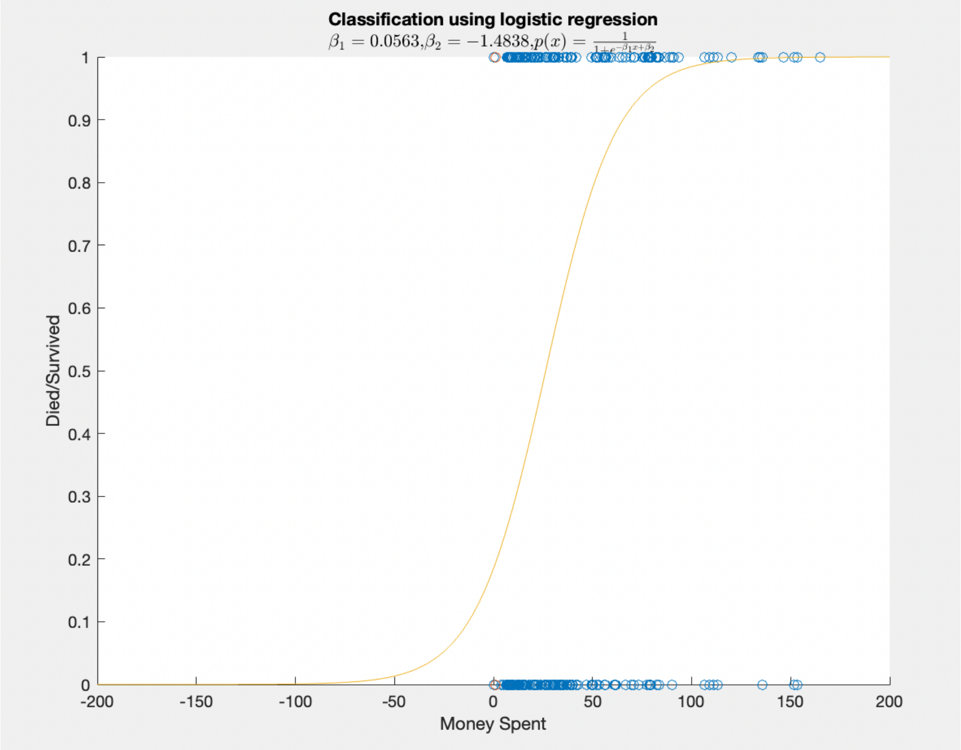

Figure 2: The logistic regression model for the feature “Fare”

The logistic regression is calculated based on maximum likelihood estimation. Since the logistic regression does not provide statistical classification, it only plays a basic role for binary response. Therefore, it is not sufficient to use it to show the relationship between the money spent by a passenger with his/her survival rate. From the figure shown above, the logistic regression does not provide a clear distinguishing between 2 classes, mainly because the data are compacted, a similar amount of people died and survived with the same amount of money spent. That suggests more data should be involved, and more classification tools should be implemented in order to find the optimized model for the dataset.

After a thorough examination of the dataset, it is decided that the “Age” data should also be used for classification. The reason for using “Age” is because the range of the data set does not vary too much (from 0 to 80), and it is also believed that the survival rate of seniors and youth should be higher than adults. After combining the “Age” dataset with the “Fare” dataset, removing the outliers and the empty cells, a complete feature set of the training data is obtained.

Naive Bayes Classification

The first classifier implemented is the Naive Bayes Classifier, this classifier applies the Bayes’ theorem on independent probabilities with a strong assumption of independence among features. That means, given a set of feature samples, the classifier will connect each feature value with its corresponding class, and calculate the posterior probabilities, more training data should mean a more accurate prediction that the classifier can make. After removing the outliers and the empty cells, there are 584 groups of training data remaining, which can also mean a reduced accuracy for the classifier.

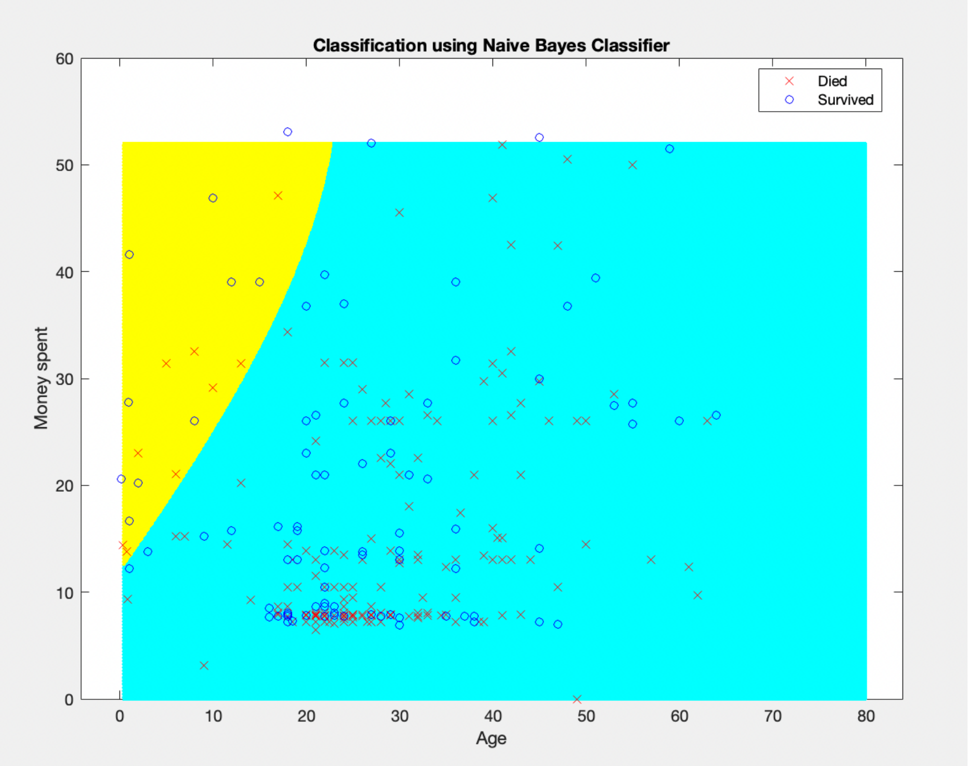

Figure 3: Classification of the data using Naive Bayes Classification
(Blue = Died Region, Yellow = Survive Region)

As shown in the decision region above, Naive Bayes does not provide a successful classification, because the data are stacked together and Naive Bayes will only be able to make a curve to distinguish between 2 regions that will be undesired for this set of data. A more straightforward method to visualize the accuracy of the classifier is by putting the predicted labels and the testing class together, create a confusion matrix, plot its ROC curve, and show its accuracy score.

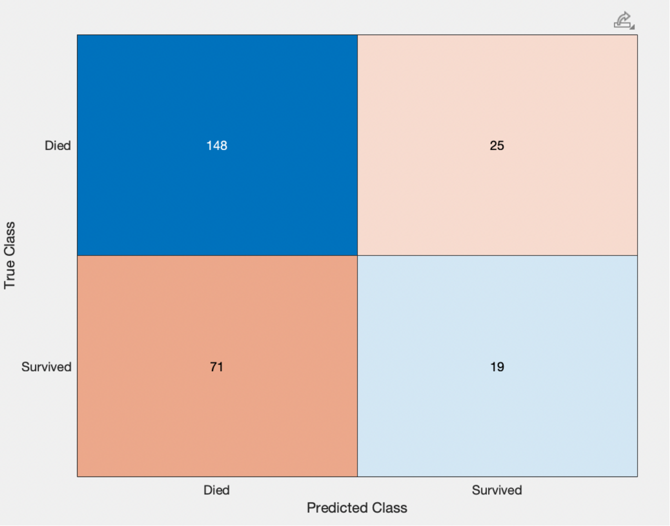

Figure 4: Confusion Matrix

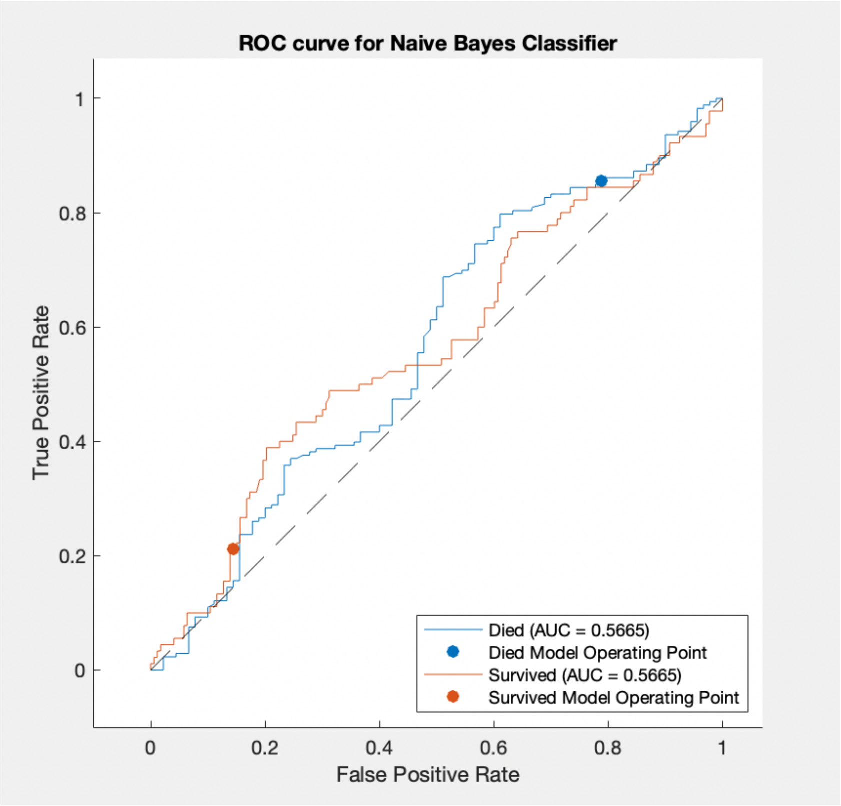

Figure 5: ROC Curve

The accuracy score obtained from cross validation reveals that the accuracy rate of the Naive Bayes Classifier is 56.65%, which is far from a desired estimation. In general, Naive Bayes is a strong method for prediction, but will be undesired if the training data is not large enough or when the data points are stacked together.

Support Vector Machine

To explore more on data classification, and to try to make a decision region for 2 classes, implementing the support vector machine is decided to be the next step. Support vector machine is used widely for data classification and outlier detection. Similar to regressions, the objective of the support vector machine is to construct a hyperplane that distinguishes data points in different classes. In the project, the support vector machine will be implemented in a 2-dimensional space, and the 2 features going to be trained will be “Age” and “Fare”. There are 2 advantages for doing this: the SVM will train specifically on the relationship between “Age” and “Train” to the survival rate, that makes the significance of the age and the money spent by the passenger much more clear to be observed; also, the 2D support vectors will be easier to implement comparing to the 3D support hyperplanes.

The figure below is the linear SVM implementation on the 2D space of “Age” (x-axis) and “Fare” (y-axis). As shown in the figure, there are no linear vectors that can be plotted. That is because the data points are clustered together and the data points in different classes are irregularly distributed. In fact, MATLAB has plotted over 385 vectors on this 2D space. More vectors mean more uncertainties on the prediction the SVM can predict. That is shown in the ROC curve on the right, which shows the AUC of the ROC curve to be only around 0.51. The ROC curve is obtained based on the support vectors obtained from the training data and implemented on the features of the test data for prediction, and it is showing the prediction had a bad performance.

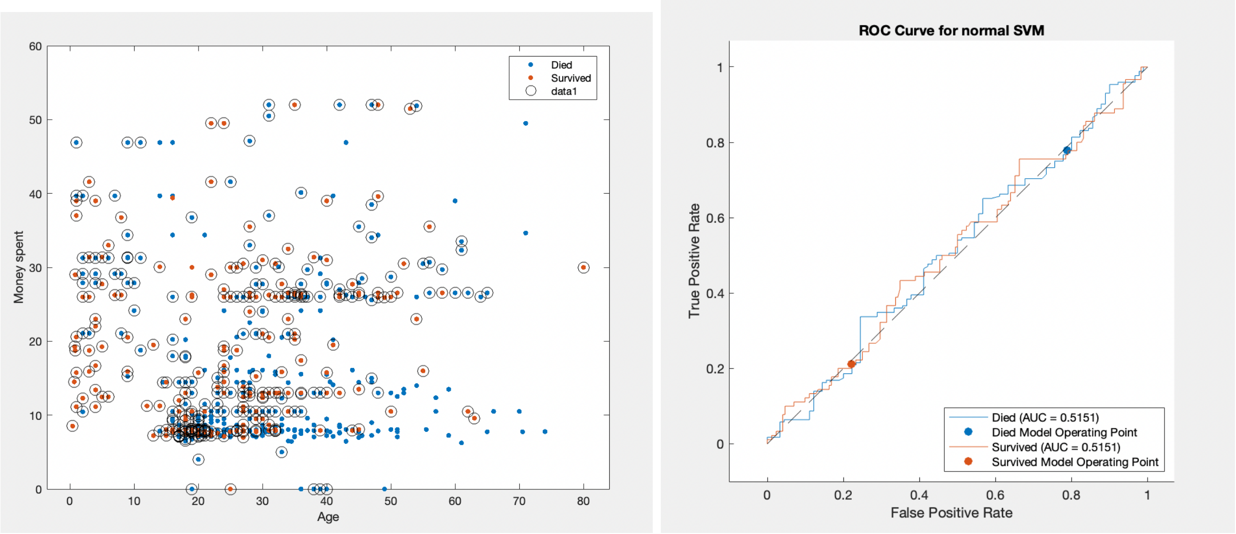

Figure 6: Classification of the data using Linear SVM and its ROC curve

Realizing that the data is clustered together, a linear model will fail for most of the trails. Therefore, a kernel trick is decided to be used for the SVM. The kernel trick implemented here is the gaussian kernel, and as shown in the plot below, after the implementation of the gaussian kernel, the support vectors can be plotted in the graph as a circle and a curve. Even though the prediction for the gaussian kernel SVM is still not precise, the performance has improved by a great amount: from 51.5% to 57.8%. The gaussian kernel is separating the “Died” region with the ‘Survived’ region by circling the area where the majority of the red dots (data that belongs to the ‘Survived’ class) are present.

The reason for the gaussian kernel SVM still to be imprecise is because there are too many passengers who spent between 10 dollars to 40 dollars and are aged between 7 years old to 20 years old. That suggests: the combination of “Age” and “Fare” might not be an optimized feature group for classification, more features might need to be considered to make a more accurate prediction, and the previous classification methods might not be perfect choices for this dataset. At the end of the basic classification stage, the group has decided to investigate more algorithms like random forest and decision trees, and try to add more attributions to the feature group for training to see if a difference or a better performance can be achieved with this change.

The missed class for the Naive Bayes Classifier is 31.34%, the missed class for the Linear SVM is 30.31%, and the missed class for the Gaussian Kernel SVM is 29.97%. All data is obtained from cross validation and the “kfoldLoss” function in MATLAB. Also, the Naive Bayes Classification model is obtained through the “fitcnb” function in MATLAB, and the SVM model is obtained through the “fitcsvm” function in MATLAB.

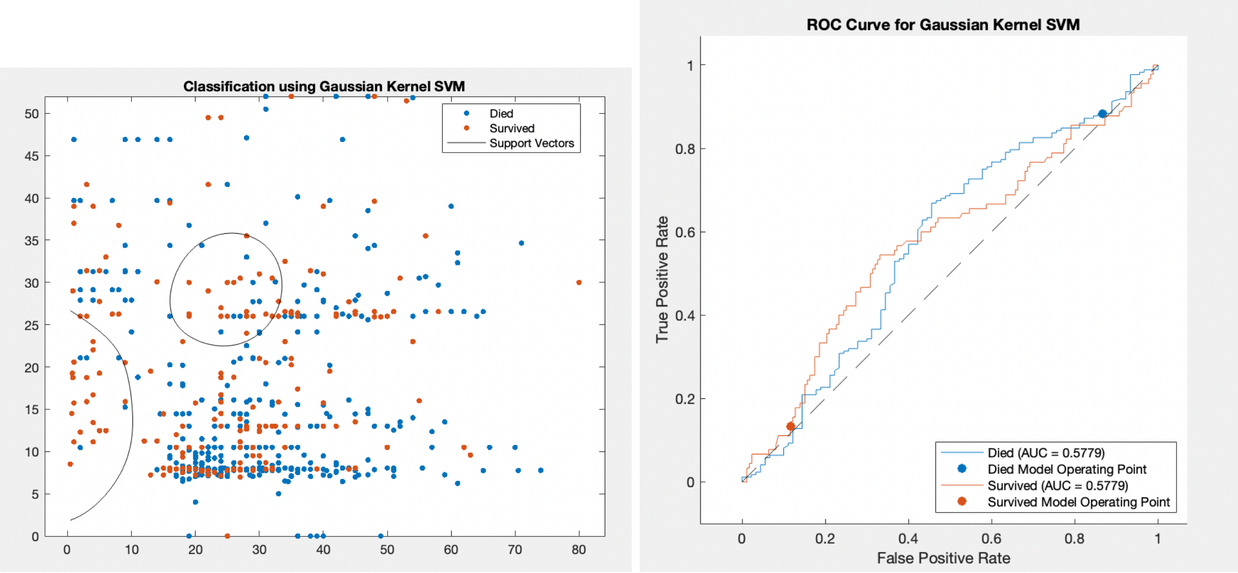

Figure 7: Classification of the data using Gaussian kernel SVM and its ROC curve
The region surrounded by the support vectors are the 'Survived' regions

Random Forest Classification

In the next stage, the group is trying to find out how 'Age' relates to survival. The group assume younger passengers survived more than other passengers.

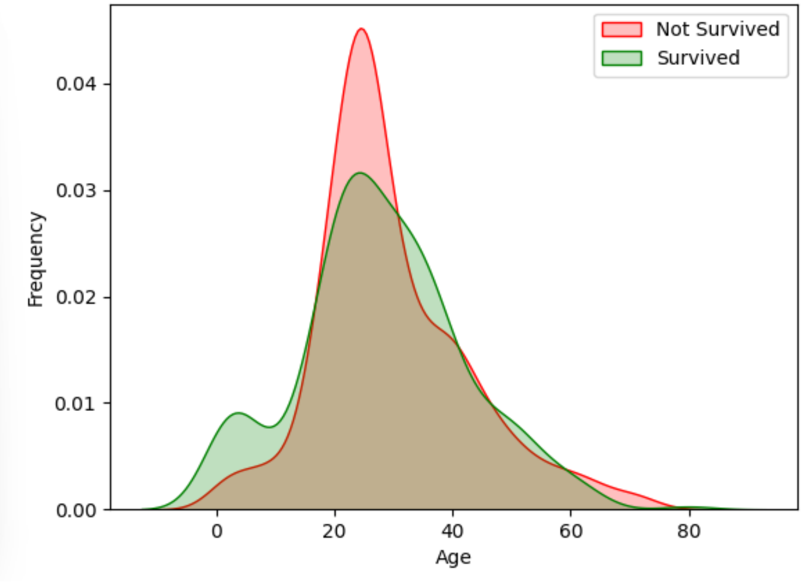

Figure 8: Survived by Age

From the distributions above, it is able to see that passengers around age of 25 have less survived, and barely no one survived over age of 60. The group also noticed that there is a peek on the surviving plot that means children have survived more. Although some patterns of age distribution can be observed, it is still not persuasive to tell if ‘Age’ has a strong correlation with ‘Survived’. The single feature can not determine whether the passenger survived or not. The group is going to discover more than two feature correlations in a single graph.

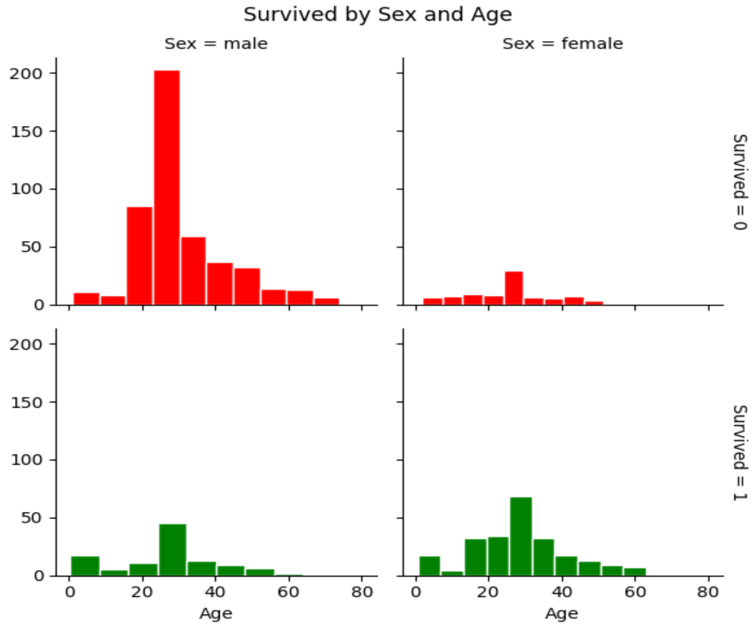

Figure 9: Survived by Sex and Age

From the plot above, there is an intuition that female passengers had better priority than male passengers.

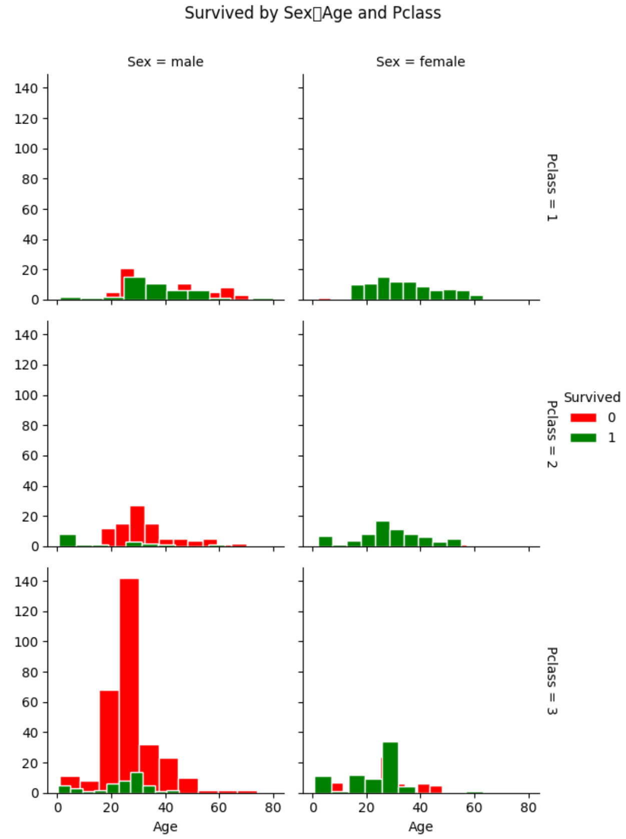

Figure 10: Survived by Sex, Age, Pclass

From the plot above, most passengers were class 3 which represent the low class. By vertical comparison, lower class passengers have a higher death rate than higher class. By horizontal comparison, barely all male passengers from class 3 and class 2 did not survive, and the majority of female passengers from all classes survived. Within multi-dimensional features analysis, it is able to say that female passengers and higher class passengers had higher survival rate.

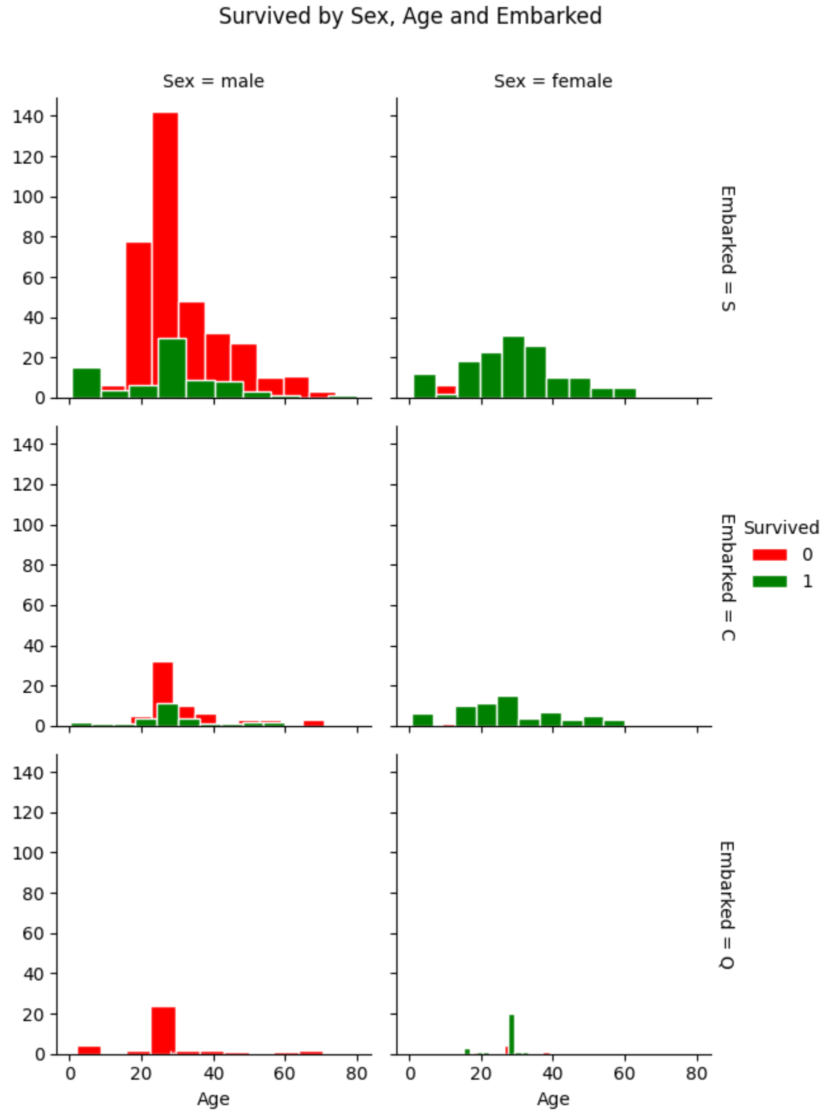

Figure 11: Survived by Sex, Age, Embarked

From the plot above, most passengers were boarded on Southampton(S), which also reveals that the death rate is higher for passengers who boarded there.

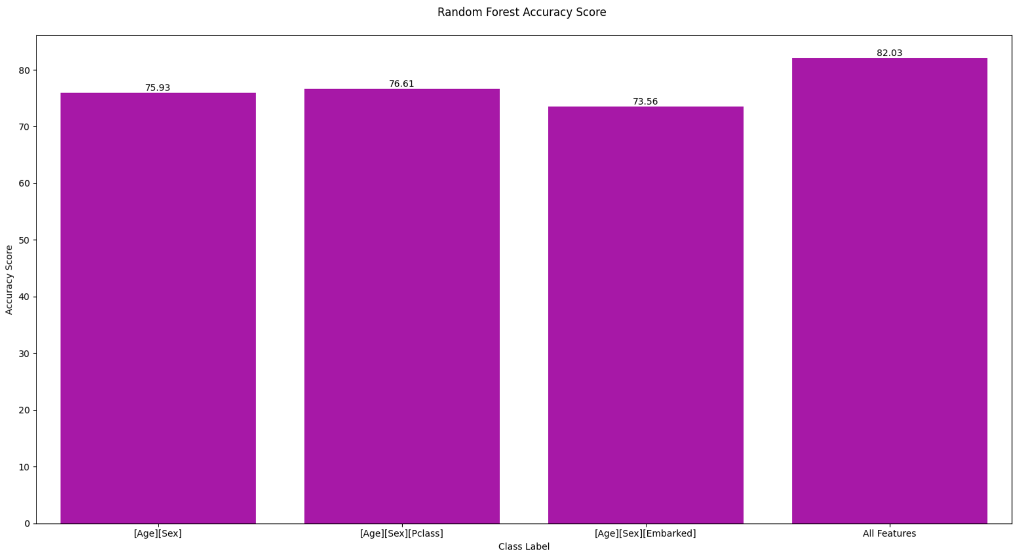

Figure 12: Random forest accuracy score of different feature combinations

The Random Forest classification for each feature combination was implemeneted. From the above bar chart, the ‘Age’, ‘Sex’ and ‘Embarked’ feature combination scores 75.56% accuracy which is the lowest accuracy score. ‘Age’, ‘Sex’ and ‘Pclass’ feature combination scores 76.61% accuracy which is the highest accuracy score.’Age’ and ‘Sex’ feature combination scores higher than ‘Age’, ‘Sex’ and ‘Embarked’ feature combination. That means the feature ‘Embarked’ does not have a strong correlation with survival rate. I also implement Random Forest classification for all important features including ‘Sex’, ‘Age’, ‘Fare’, ‘Pclass’, ‘SibSp’, ‘Parch’ and ‘Embarked’. The final accuracy score is 82.03%. Random Forest classification has great performance on predicting the survival rate.

Decision Tree Classification

At the beginning of the project, the group focused on the impact of a single feature in the data set, and that resulted in poor feedback in the model evaluation. The group first tried to adapt what was learned in the lectures, Fisher's LDA, and is expected to see a visual representation of the classification. However, the group found that it was far from satisfactory for the dataset to use any single one of the features to determine the fate of a passenger.

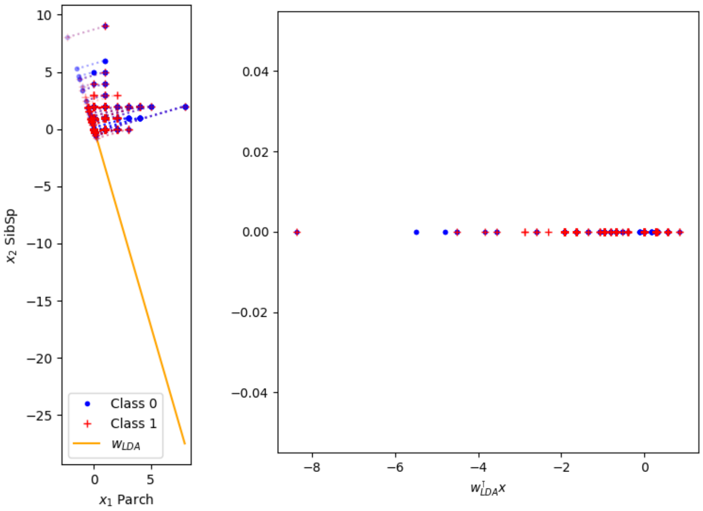

Figure 13: Fisher's LDA on SibSp (Sibling = brother, sister, stepbrother, stepsister)

Therefore, the group moved forward with the analysis across multiple features. PCA (Principal Component Analysis) on several features is tried to be implemented as a trial. However, the group found out that finding proper features by manual selection would be a struggle. For instance, most of our data are binary (0 & 1), instead of various value distributions. After some undesired results were obtained (shown below), it turned out that more features in the model have to be added to have a suitable result. Therefore, lightGBM library in Python is introduced, which would be considered as a high-performance open source library built with decision-tree based algorithms. After the data set is fully expanded, the data with a feature importance concern is analyzed.

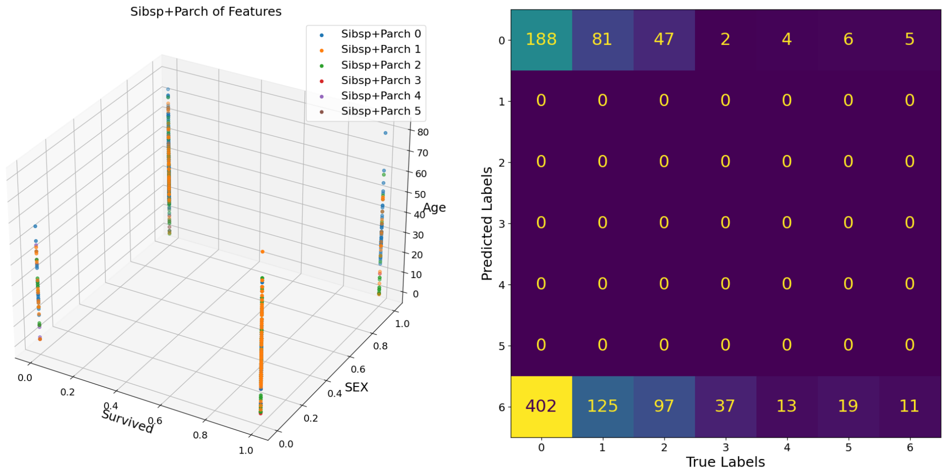

Figure 14: Unsuccessful PCA Analysis

From the bar chart shown below, “Fare” came to be the most important feature against the rest of other features in our dataset.

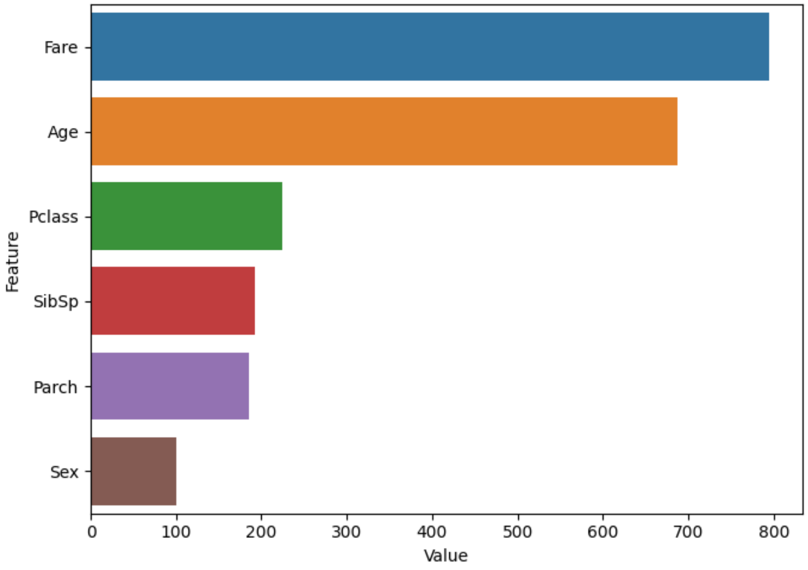

Figure 15: Feature Importance

Decision trees have been widely adopted in many different fields of study, from finance and economics to biology and medicine. They are particularly useful in scenarios where the relationships between the features and the target variable are complex and non-linear, and where the data may have missing or noisy values. Furthermore, decision trees can handle both numerical and categorical data, making them versatile and adaptable to a wide range of data types and distributions. Additionally, the hierarchical structure of decision trees allows for easy pruning and regularization, which can improve the performance and prevent overfitting. Overall, our decision tree implementation offers a reliable, efficient, and interpretable solution for solving real-world problems in machine learning.

We started our idea with a simple decision tree (shown below), and we were planning to have a vertical growing decision tree across the features depending on their importance scores.

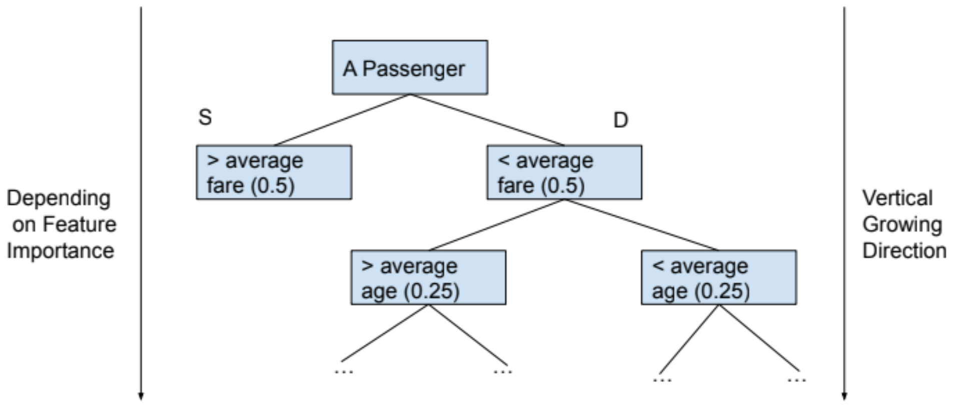

Figure 16: Scratch Note of Decision Tree

We tried to add as many features in our model as possible. After we had put all of the concerned features in the tree decision classifier, we reached 82.183% predicted accuracy on the training set, and surprisingly found 77.612% prediction accuracy on the testing set. Meanwhile, we randomly selected 30% samples from the training set, and the accuracy came to 89.234%.

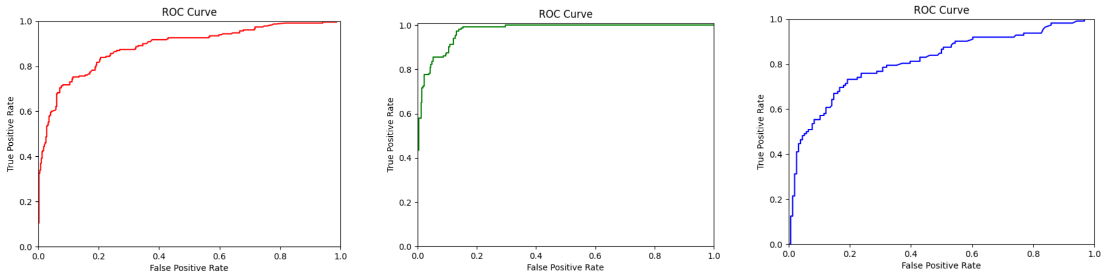

Figure 17: Trained ROC (shown as red), Valid ROC (shown as green), Test ROC (shown as blue)

Model Evaluation:

	Training	Validation	Testing
Accuracy Rate %	82.183	89.234	77.612
Recall Rate %	75.652	86.184	66.964
Precision Rate %	75.983	84.516	76.531
F1 - Score	75.817	85.342	71.429
Roc Accuracy %	88.581	97.440	81.313

Confusion Matrix:

Training Set	Predicted 0	Predicted 1
True Label 0	338	55
True Label 1	56	174

Valid Set	Predicted 0	Predicted 1
True Label 0	133	23
True Label 1	37	75

Testing Set	Predicted 0	Predicted 1
True Label 0	242	21
True Label 1	24	131

Our decision tree classifier provided us a good accuracy in determining a survivor in this disaster, and also having a really good computing speed against the normal models we had dealt with in the previous assignments. This implementation is really helping us to understand and exercising the decision tree concept we discussed previously in this semester.

Feature Attribution

We have made many classifiers with different features, including Naive Bayes, linear regression and something else. It is reasonable that complex classifiers to which many features are involved get a better prediction and higher accuracy. For those simple classifiers, which get bad predictions, they only have one or two features and this is the reason why these classifiers perform badly. In the milestone report, only one feature "gender" is used to make a Naive Bayes classification and the result is easy to imagine. This classifier told me all females will survive. When we only use the other two features, like sibsp and parch, the prediction also fails. That makes me interested in a question, which feature is the most important, and the easiest to get a good prediction.

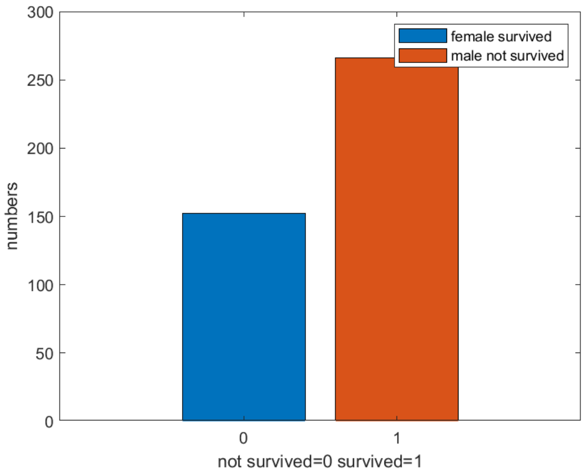

Figure 18: Classifier based on gender

This is related to feature attribution which is an analysis of relationships among features. There are many ways of feature attribution, including integrated gradiences, lime and something else. Here we chose integrated gradiences to calculate the importance of each feature.

As before, we cleared the data and deleted some features like passenger's id and name because characters will increase the complexity of classification. Then we built a four layer neural network model and its accuracy is about 0.8 which is good to see. Under this model, we predict the survival rate of the test data and the result is about 0.32.

In the next step, we chose integrated gradience to analyze the importance of features. And the result is like below.

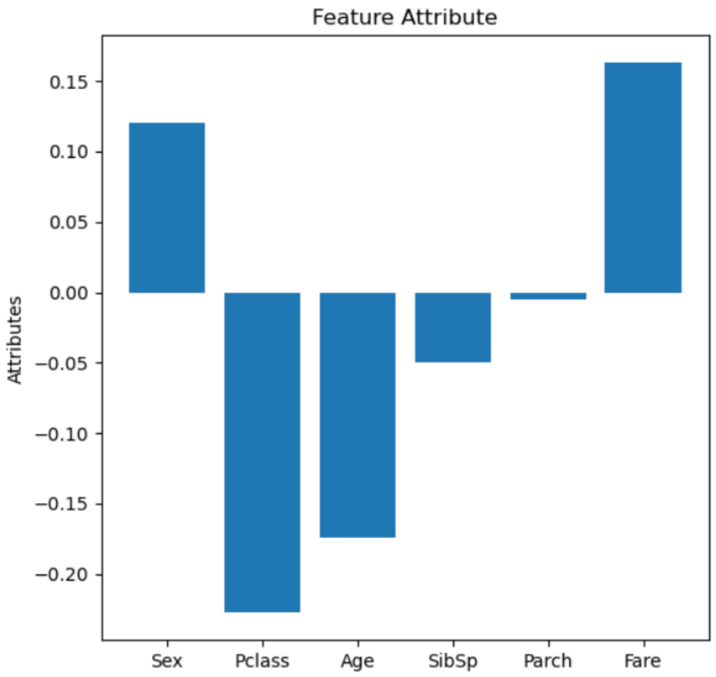

Figure 19: Feature attribution

From the plot above, we can get a lot of information. Pclass and age are strongly and negatively correlated with survival, which means the number of classes and age have great importance to survival. From class1 to 3, the difficulty of survival is increasing and it is harder for the old to be alive in this catastrophe. Fare also contributes much, which is corresponding to the importance of survival because the fare of class 1 is the highest whereas the class 3 is the cheapest, causing low rate of survival. Parents and children attribute least and that means there is not a close tie between this feature and survival.

Conclusion

Analyzing the data obtained from the Titanic tragedy is helpful for estimating the survival rate in a lot of catastrophic accidents. By choosing the proper classification tool, predictions can be made given a passenger's different traits.

In the beginning of the project, the implementation of the Naive Bayes Classifier and the SVM is proved to be unsuccessful, since Naive Bayes Classifier is able to make a great estimation when a large training dataset is available and the features should have a strong independence with each other, and the SVM is unable to separate the data because the irregular distribution of data points on the 2D space and the cluster of data that belongs to both classes. The Randomforest model has high accuracy scores on each feature combination. Based on the random forest accuracy scores, we can conclude that when all important features are selected, the prediction of passengers survival rate will be more reliable. After first our time of implementation with the decision tree, we met the basic expectation with having a functional classifier against multiple features. The tree optimized our decision from various concerns, and it was definitely an improvement from our basic classifiers targeting any single one of the features.

To predict the survival rate, compared to the previous failed classification, the neural network model works well and shows good accuracy. Feature attribution gives us a direct instruction on the relationship of features. Pclass and age are the most important features, and then are fare and sex. If we would like to build a good classification of the Titanic data, we need to at least include these four features to train our model.

Appendix

Source codes are available on:
https://github.com/LiLJN/EECE-5644-Final-Project-Group-1
Zoom Recording (Presentation) :
https://northeastern.zoom.us/rec/share/DVioBWsitl1S-cTIsyUEXkzmHESsQ7Wy6aEfWhGXOi2D65DYXGF2mb7wSyP0k2-9.r4w-O697rxajaOzR
Passcode:9t28x$nE
Google Drive Recording:
https://drive.google.com/file/d/1bmdDVcZc6WwnkWHpyGZLXXHo8EEDD_hU/view?usp=sharing