Analyzing Titanic Disaster using Machine Learning Algorithms

Aakriti Singh
Department of CSE
Amity University, Uttar Pradesh
India
aakriti1495@gmail.com

Shipra Saraswat
Department of CSE
Amity University, Uttar Pradesh
India
sshipra1510@gmail.com

Neetu Faujdar
Department of CSE
Amity University, Uttar Pradesh
India
neetu.faujdar@gmail.com

Abstract—Titanic disaster occurred 100 years ago on April 15, 1912, killing about 1500 passengers and crew members. The fateful incident still compel the researchers and analysts to understand what can have led to the survival of some passengers and demise of the others. With the use of machine learning methods and a dataset consisting of 891 rows in the train set and 418 rows in the test set, the research attempts to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. These factors may or may not have impacted the survival rates of the passengers. In this research paper, various machine learning algorithms namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest have been implemented to predict the survival of passengers. In particular, this research work compares the algorithm on the basis of the percentage of accuracy on a test dataset.

Index Terms—Titanic; Prediction; Classification; Data mining; R; Python; Logistic Regression; Random Forest; Decision Tree; Nave Bayes

I. Introduction

The field of machine learning has allowed analysts to uncover insights from historical data and past events. Titanic disaster is one of the most famous shipwrecks in the world history. Titanic is a British cruise liner that sank in the North Atlantic Ocean a few hours after colliding with an iceberg. While there are facts available to support the cause of the shipwreck, there are various speculations regarding the survival rate of passengers in the Titanic disaster. Over the years, data of survived as well as deceased passengers has been collected. The dataset is publically available on a website called Kaggle.com [1].

This dataset has been studied and analyzed using various machine learning algorithms like Random Forest, SVM etc. Various languages and tools are used to implement these algorithms including Weka, Python, R, Java etc. The approach of the research paper is centered on R and Python for executing algorithms- Nave Bayes, Logistic Regression, Decision Tree, and Random Forest. The prime objective of the research is to analyze Titanic disaster to determine a correlation between the survival of passengers and characteristics of the passengers using various machine learning algorithms. In particular, this research work compares the algorithms on the basis of the percentage of accuracy on a test dataset.

II. Dataset

The dataset used for the paper is provided by the Kaggle website. The data consists of 891 rows in the train set which is a passenger sample with their associated labels [1]. For each passenger, the name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation were provided. The data is in the form of a CSV (Comma Separated Value) file. For the test data, the website provided a sample of 418 passengers in the same CSV format. The structure of the dataset with a sample row has been listed in Table I. The attributes of the training set and their description have been mentioned in Table II.

Before building a model, data exploration is done to determine what all factors or attributes can prove beneficial while creating the classifier for prediction. To start the exploration, few X-Y generic plots are made to get an overall idea for each attribute. Some of the generic plots have been shown below. The age plot in Fig 1 suggested that maximum or majority of the passengers belonged to the age group of 20-40.

Similarly, a graph Fig 2 is plotted and some calculations are performed for the sex attribute and the results suggested that the survival rate of the female is 25.67% higher to that of the male. Similarly, each of the attribute are explored to extract those attributes or features which would be used later for prediction.

A survival histogram is generated to determine the number of people survived vs. number of people who can not survive. From the histogram it is clear that the number of people who survived is less than the number of people who could not survive. The survival histogram is shown in Fig 3.

In order to deal with the missing values data cleaning is done. While observation it has been found that the dataset is not complete. There are various rows for which one or more fields are marked empty (especially age and cabin). But the age is an important attribute to predict the survival of passengers. Hence a technique to replace the NAs in the age column has been used. The gender column has been changed to 0 and 1

(0 for male and 1 for female) to fit the prediction model in a better manner. Some new variables are introduced into the dataset to predict the survival more closely.

TABLE II
ATTRIBUTES IN THE TRAINING DATA SET

Attributes	Description
PassengerID	Identification no. of the passengers.
Pclass	Passenger class (1, 2 or 3)
Name	Name of the passengers
Sex	Gender of the passengers (male or female)
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 (Cherbourg, Queenstown or Southampton)
Survived	Target variable (values 0 for perished and 1 for survived)

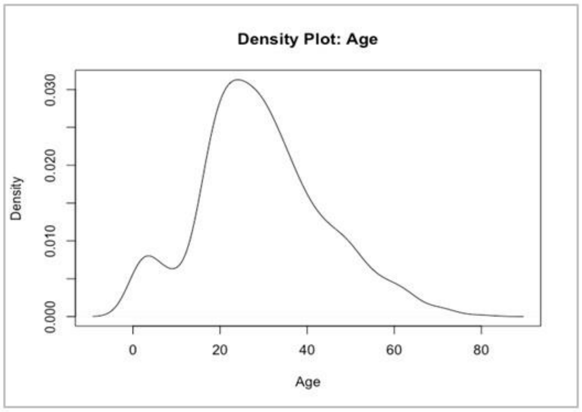


Fig. 1. Age plot

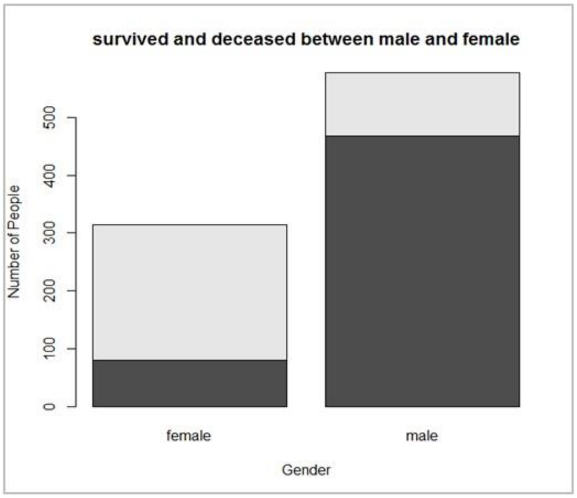


Fig. 2. Sex bar plot

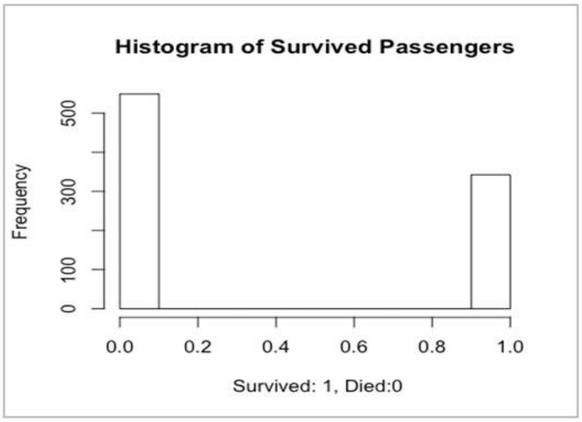


Fig. 3. Survival Histogram

III. Related Work

Eric Lam and Tang used the Titanic problem to compare and contrast between three algorithms- Naive Bayes, Decision tree analysis and SVM. They concluded that sex was the most dominant feature in accurately predicting the survival. They also suggested that choosing important features for obtaining better results is important. There are no significant differences in accuracy between the three methods they used [2].

Shawn Cicoria, John Sherlock, Lauren Clarke and Manoj Muniswamaiah performed Decision tree classification and Cluster analysis to suggest that sex is the most important feature as compared to other features in determining the likelihood of the survival of passengers [3].

Kunal Vyas, Lin Li and Zeshi Zhengsuggested that dimensionality reduction and playing more with the dataset could improve the accuracy of the algorithms. The most important conclusion provided by them is that more features utilized in the models do not necessarily make results better [4].

Mikhael Elinder in his post, analyzed the direct relationship of social norms and sex with survival. He concluded that on the Titanic, the survival rate of women is more than three times higher than the survival rate of men [5].

Bruno S. Frey, David A. Savage, and Benno Torgler concluded that people in their prime age died less often than older people. Passengers with high financial stability, traveling in first class, are better able to save themselves as are passengers in second class as compared to third class [6].

Trevor Stephens has carried out the prediction using Random forest and decision tree algorithms. He has used the following parameters- Title, Fare, Pclass, FamilyID, Family Size, SibSp, Parch, Sex, Age and Embarked. He has not mentioned the accuracy percentages of the implemented algorithms [7-8].

Rex Morgan in a post, submitted to www.beacuse.uk suggested that human behavior also determines the survival rates of the passengers. He also mentioned that lifeboats were less and most of them were not filled up to their capacities [9].

TABLE I
KAGGLE DATA SET

PassengerID	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr.	Male	22	1	0	A/521171	7.25	NA	S
2	1	1	Cunnings, Mrs.	Female	38	1	0	PC17599	71.2833	C85	C

IV. DATA CLEANING AND FEATURE EXTRACTION

The analysis started with exploring different attributes for NA values. Age and cabin column had NA values in them. Age column had 177 rows with NA values and cabin column had 687 rows with NA values. As most of the data in the cabin column is missing, cabin column was dropped from the analysis. Since age can have been a very important attribute. Hence, the age column is kept for the analysis.

A relationship between the title of the passengers and their age is established. It is assumed that Ms. A will be younger than Mrs. A and that the people having same titles will be closer in age. Titles (Mr, Mrs, Miss, Master etc.) of the passengers are extracted from the name of the passengers and the names in the name column is replaced with the extracted titles. The NA values in the age column are replaced by the calculated average age of the particular title-group i.e. if there is a NA age value for a woman with title Mrs, the NA value is replaced with the average age of all the women with title Mrs.

In the past marine disasters, the policy of Women Children First (WCF) is used by the crew members giving women and children survival advantage over men [5]. Based on this social norm some new attributes have been introduced (Children and Mother) to strengthen the dataset and improve the analysis. A research study suggested that titles such as Dr, Col, etc. in the Name column can be an important part of the analysis since it shows that these people are influential and respectable which might increase their survival rates [4]. Based on this an attribute called respectable is introduced. The newly introduced attributes and their descriptions are listed in the Table III.

Dataset after the addition of these attributes have been shown in the above tables Table IV. Name, Ticket, Cabin and Embarked are dropped out of the analysis as it is believed that these variables are not relevant to the analysis. Similarly, due to large variation in values, fare attribute is also removed.

V. ALGORITHM USED

Prediction models are generated using four machine learning algorithms namely- Naive Bayes, Logistic Regression, Decision tree and Random forest. Each of these algorithms are compared to one another on the basis of the accuracy percentage.

The attributes used in the test and train dataset for implementing these algorithms are- Pclass, Sex, Age, SibSp, Parch, Mother, Children, Family and Respectable. Naive Bayes prediction is performed using Python and rest of the algorithms are executed using R.

A. Naive Bayes

Nave Bayes is a classification algorithm that applies Bayes theorem to build a prediction model. Naive Bayes is a classification algorithm that applies Bayes theorem to build a prediction model. It is based on some naive assumptions about the features. The assumption is that all the features are independent of each other.

In other terms, the probability of value of one feature belonging to a class is independent of all other features. The probability of each class value for a given value of a feature is called the conditional probability.

The probability of a class value is obtained by multiplying all the conditional probabilities. The class with the highest probability is the assigned class of a given instance. There are different types of Naive Bayes algorithm. In this analysis, Gaussian Naive Bayes algorithm is used. The workflow for Naive Bayes Algorithm used in this research paper is shown in Fig. 4.

The model is built using Pclass, Sex, Age, SibSp, Parch, Mother, Children and Respectable. All features are categorical (or numeric). The target variable is a categorical variable having values 0 and 1(1 for survival and 0 for demise). All the other parameters or variables are categorical as well.

The parameter such as mother, children and respectable are categorical while the parameter such as age and family are numerical. After loading the training and test data in python, the data is summarized.

Summary data included calculating mean and the standard deviation for each attribute, by class value. The prepared summaries are used for prediction. Prediction involved the use of Gaussian function to estimate the probability of a given attribute value of belonging to a particular class.

The probabilities of all the attribute values for a data instance have been combined and a probability of the entire data instance belonging to the class is generated. The class with the largest probability value is considered as the predicted class. The predicted results using the Naive Bayes model is shown in Fig. 5.

A classification accuracy is obtained by comparing the predictions to the class values of the test data [10]. A classification accuracy of 91.38755% was achieved.

TABLE III
NEW ATTRIBUTES AND DESCRIPTION

New Attributes	Description
Mother	Column value is 1 only if the title is Mrs. and value of parch is greater than 0. Otherwise, 2 is assigned.
Children	Column value is 1 only if age is less than or equal to 14. Otherwise, 2 is assigned.
Family	Z= X+Y+1 where X is the value of SibSp and Y is the value of Parch.
Respectable	Column value is 1 is the title is Dr, Rev, Capt., Col., Don. Or Lady. Otherwise 2 is assigned.

TABLE IV
DATASET WITH NEW ATTRIBUTES

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Mother	Children	Family	Respectable
0	3	Mr	male	22	1	0	2	2	2	2
1	1	Mrs	female	38	1	0	2	2	2	2

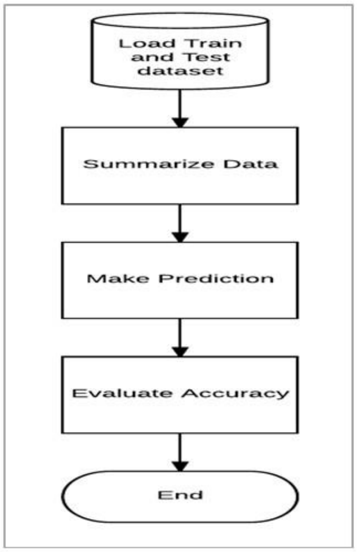
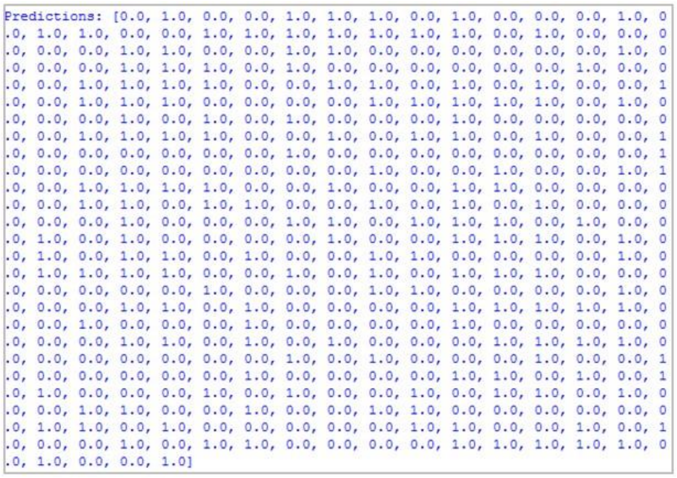


Fig. 4. Naive Bayes workflow



```
Predictions: [0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0
.0, 1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0
.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0
.0, 0.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1
.0, 0.0, 1.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0
.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1
.0, 0.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 1.0, 1.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0
.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0
.0, 1.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0
.0, 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 1.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0
.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0
.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1
.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0
.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1
.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0
.0, 1.0, 0.0, 0.0, 1.0]
```

Fig. 5. Prediction using Naive Bayes

B. Logistic Regression

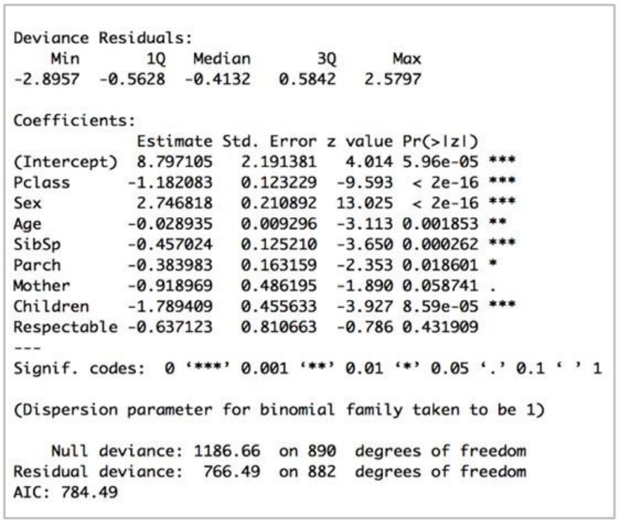
After Naive Bayes classification Logistic Regression is implemented. Logistic Regression is a type of classification algorithm in which the target variable is categorical and binary. In the dataset survived column is the dependent variable which is both binary and categorical (1 for survival and 0 for demise) [11].

The prediction model is built by including the features i.e. Pclass, Sex, Age, SibSp, Parch, Mother, Children, Family and Respectable. After running the model, it is observed that family is the least significant variable. Hence, family is dropped from the dataset and the model is built again as shown in Fig. 6.

From the above results, it is concluded that mother, parch and respectable are not stastically significant and they have high P-value. Pclass, sex and children are the most significant values as they have low P-values. Parch, mother and respectable are from the dataset to build the logistic model again. The summary of the improved model is shown in Fig. 7.

The low P-values of Pclass and sex suggested that they have high correlation with the probability of having survived the disaster. The negative coefficient of Pclass implies that if all other variables are kept constant, people with higher Pclass value are less likely to survive. This basically means people with Pclass value 1 are more likely to survive than Pclass value 2 and people with Pclass value 2 are more likely to survive than Pclass value 3.

Similarly, the positive coefficient of sex implies that if all other variables are kept constant people with higher sex value are more likely to survive. This translates to females (sex value 1) are more likely to survive the disaster than males (sex value 0). Lastly, the accuracy of the model using the test data is determined. An accuracy of 94.26% has been achived. The decision boundary is taken to be 0.5. The class for which the probability is greater than 0.5, is the predicted class.



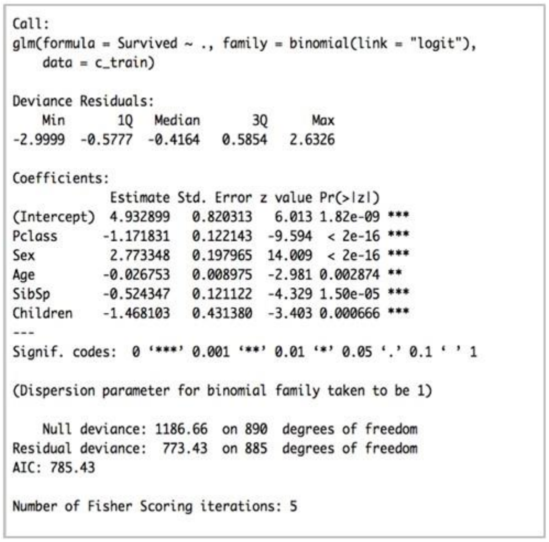
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8957  -0.5628  -0.4132   0.5842   2.5797

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.797105   2.191381   4.014 5.96e-05 ***
Pclass      -1.182083   0.123229  -9.593  < 2e-16 ***
Sex          2.746818   0.210892  13.025  < 2e-16 ***
Age         -0.028935   0.009296  -3.113 0.001853 **
SibSp       -0.457024   0.125210  -3.650 0.000262 ***
Parch       -0.383983   0.163159  -2.353 0.018601 *
Mother      -0.918969   0.486195  -1.890 0.058741 .
Children    -1.789409   0.455633  -3.927 8.59e-05 ***
Respectable -0.637123   0.810663  -0.786 0.431909
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  766.49  on 882  degrees of freedom
AIC: 784.49
```

Fig. 6. Summary of logistic model



```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = c_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9999  -0.5777  -0.4164   0.5854   2.6326

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.932899   0.820313   6.013 1.82e-09 ***
Pclass      -1.171831   0.122143  -9.594  < 2e-16 ***
Sex          2.773348   0.197965  14.009  < 2e-16 ***
Age         -0.026753   0.008975  -2.981 0.002874 **
SibSp       -0.524347   0.121122  -4.329 1.50e-05 ***
Children    -1.468103   0.431380  -3.403 0.000666 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  773.43  on 885  degrees of freedom
AIC: 785.43

Number of Fisher Scoring iterations: 5
```

Fig. 7. Summary of the improved model

C. Decision Tree

Next, the research analysis is carried by implementing Decision tree algorithm [8]. Decision tree learning is the method of construction of a decision tree from class-labeled training tuples. A decision tree can be considered as a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label.

The node at the top is called as the root node. In the research the generated decision tree gave some useful insights. Some of the insights are that if a passenger is female and she belonged to a passenger class of either 1 or 2, then the probability of survival is 0.95 and if a passenger is male and greater than or equal to an age of 13, then the probability of his survival is 0.16 as shown in Fig. 7.

This proves that the survival chances of females are greater to that of males and also people belonging to passenger class 1 0r 2 had greater chances of survival than the passengers belonging to Pclass 3. Decision rule is similar to the ones that developed from other algorithms. The decision tree has been shown in Fig. 8. From the confusion matrix in Table V, It has been observed that out of 418 predictions, the prediction model using decision tree algorithm made 389 correct predictions, giving an accuracy of 93.06%.

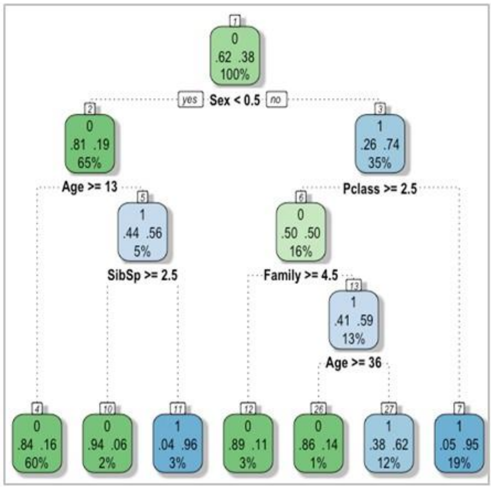


Fig. 8. Decision Tree

TABLE V
CONFUSION MATRIX FOR DECISION TREE

	Actual	Actual
Predicted	Survived: NO	Survived: YES
Survived: NO	252	15
Survived: YES	14	137

D. Random Forest

Random forest algorithm is implemented for improving the accuracy of the classification model even further and determining the most significant features for survival. Random forest algorithm is a classification algorithm that constructs a multitude of decision trees at the time of training and it outputs the class which is the mode of the individual trees [7].

The model has been built with all the variables of the cleaned train dataset, that are Pclass, sex, Age, Family, Children, SibSp, Mother, Parch and Respectable. In order to understand the significance of all these different variables in the classification process, an argument importance while building our model is used.

From Fig. 9, it is clear that Sex and Pclass play the most significant role in the classification model while Mother, Parch and Respectable are the least significant variables. This is in alignment with our analysis using logistic regression algorithm. The accuracy of random forest algorithm has been checked on the test data. After executing the Random forest analysis on test cases the model generated a confusion matrix as shown in Table VI. Using the confusion matrix we determined that out of total 418 predictions, the model made 384 correct predictions, giving an accuracy of 91.8%.

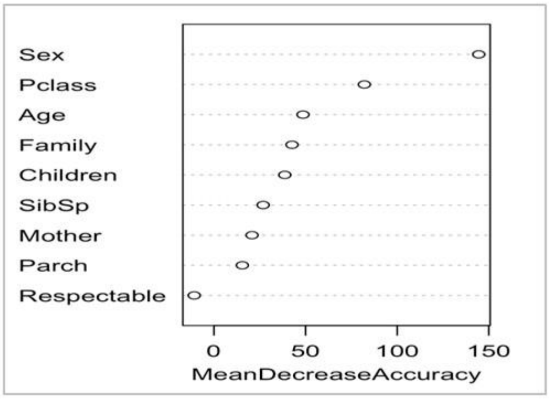


Fig. 9. Variable importance

TABLE VI
CONFUSION MATRIX OF RANDOM FOREST

	Actual	Actual
Predicted	Survived: NO	Survived: YES
Survived: NO	250	18
Survived: YES	16	134

VI. RESULTS

For comparing the four techniques used in this research work 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 table XI. Accuracy is the measure of how well a model predicts. Higher the accuracy the better. Accuracy is calculated using the formula TN+TP/ Total number of test set rows*100.

False discovery rate is a method of conceptualizing the rate of type I (false positive) errors in null hypothesis testing when conducting multiple comparisons. For the problem used in the research paper, false discovery rate is an important metric as it would be dangerous if the system predicts a passenger would survive but in reality he does not survive. 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 is listed in the Table VII .

TABLE VII
STRUCTURE OF CONFUSION MATRIX

	Actual	Actual
Predicted	Survived: NO	Survived: YES
Survived: NO	True Negative (TN)	False Negative (FN)
Survived: YES	False Positive (FP)	True Positive (TP)

TABLE VIII
COMPARISON OF ALGORITHMS

Algorithms	Accuracy	False discovery Rate
Nave Bayes	91.3%	15.47%
Logistic Regression	94.26%	7.90%
Decision Tree	93.06%	9.26%
Random Forest	91.86%	10.66%

VII. CONCLUSION AND FUTURE WORK

Logistic Regression proved to be the best algorithm for the Titanic classification problem since the accuracy of Logistic Regression is the highest and the false discovery rate is the lowest as compared to all other implemented algorithms. The research also determined the features that are the most significant for the prediction. Logistic regression 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 potentially validating more using pruning techniques that is to see if a shallower tree with same or improved accuracy can be achieved. Cross validation could also be used that is calculating accuracy based on different combinations of training and test data. It would be interesting to play more with dataset and introducing more attributes which might lead to good results. Various other machine learning techniques like SVM, K-NN classification can be used to solve the problem.

REFERENCES

[1] Kaggle.com, ‘Titanic:Machine Learning form Disaster’,[Online]. Available: http://www.kaggle.com/. [Accessed: 10- Feb- 2017].

[2] Eric Lam, Chongxuan Tang, “Titanic Machine Learning From Disaster”, LamTang-TitanicMachineLearningFromDisaster, 2012.

[3] Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, “Classification of Titanic Passenger Data and Chances of Surviving the Disaster”, Proceedings of Student-Faculty Research Day, CSIS, pp. 1-6, May 2014.

[4] Vyas, Kunal, Zeshi Zheng, and Lin Li, “Titanic-Machine Learning From Disaster”, Machine Learning Final Project, UMass Lowell, pp. 1-7, 2015.

[5] Mikhael Elinder.(2012). ‘Gender, social norms, and survival in maritime disasters’, [Online]. Available: http://www.pnas.org/content/109/33/13220.full. [Accessed: 8- March - 2017].

[6] Frey, B. S., Savage, D. A., and Torgler, B, “Behavior under extreme conditions: The Titanic disaster”, The Journal of Economic Perspectives, 25(1), pp. 209-221, 2011.

[7] Trevor Stephens. (2014), ‘Titanic: Getting Started With R - Part 3: Decision Trees’, [Online]. Available: http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/. [Accessed: 11- March- 2017].

[8] Trevor Stephens. (2014). ‘Titanic: Getting Started With R - Part 3: Decision Trees’, [Online]. Available: http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/. [Accessed: 8- March - 2017].

[9] Rex Morgan. (2016). Titanic [Online]. Available:http://www.because.uk.com/wp-content/uploads/Because-2016-03w.pdf. [Accessed: 9- March - 2017].

[10] Jason Brownlee. (2014). How to implement Nave Bayes in Python from scratch [Online]. Available: http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/. [Accessed: 9- March - 2017].

[11] Santos, K.C.P, Barrios, E.B, “Improving Predictive accuracy of logistic regression model using ranked set sample,” Communication in statistic simulation and computation, 46(1),pp. 78-90, 2017.