A Novel Approach for Titanic Survival Prediction Using Machine Learning

Dr. Meghna Utmal

HOD-MCA, GGITS, Jabalpur, (M.P.), India

Abstract: The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. Our paper proposes a predictive model to predict which of the passengers survived using various machine learning techniques namely decision tree, logistic regression and linear SVM. In particular, the response variable Survived will be modeled given ten possible predictors. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Index Terms – Titanic Dataset, Decision Tree, Logistic Regression, Linear Support Vector Machine (SVM), Accuracy.

I. INTRODUCTION

Machine learning[1] means the application of any computer-enabled algorithm that can be applied against a data set to find a pattern in the data. This encompasses basically all types of data science algorithms, supervised, unsupervised, segmentation, classification, or regression". few important areas where machine learning can be applied are Handwriting Recognition, Language Translation, Speech Recognition, Image Classification, Autonomous Driving. Some features of machine learning algorithms can be observations that are used to form predictions for image classification, the pixels are the features, For voice recognition, the pitch and volume of the sound samples are the features and for autonomous cars, data from the cameras, range sensors, and GPS. The area of machine learning has enabled experts to reveal bits of knowledge from the useful information and past occasions. One of the familiar histories in the world is Titanic disaster. The main aim is to anticipate the passengers who have survived using the machine learning techniques. To make the correct predictions about the disaster various parameters are included such as Name, Sex, Age, PassengerID, Embarked etc. Initially the dataset has collected. The dataset has been contemplated and deselected utilizing different machine learning calculations like SVM, Random forest and so forth. The methods are used in this are decision tree, linear SVM, and logistic regression. Evaluating the Titanic disaster to decide a relationship between the survival of passengers and attributes of the travelers utilizing different machine learning calculations is the main goal of this project. Hence, various algorithms can be compared based on the accuracy of a test dataset [2].

Figure 1: Snapshot of Famous Titanic Ship

II. LITERATURE REVIEW

Shikha Chourasia [3] proposed a various technique of classification of the ID3 decision tree. In the developing region of data mining, the supreme classification method is Decision tree. In many fields Decision tree classifiers (DTC) are found. For example, in expert system, various types of recognition, in the fields of medical. For building the decision tree, the primary algorithm developed is Induced Decision Tree(ID3). So, in this the variety of techniques that is improved version of ID3 that are fixed induced decision tree(FID3) and variable precision rough set fixed induced decision tree(VPRSFID3) are explained. By comparing all the methods for any data sets Accuracy is always high in the case of VPRSFID3.The disadvantages are present in the FID3 algorithm is solved by VPRSFID3.So they concluded that (VPRSFID3) is considered as the best method.

Pea-Lei Tu & Jen- Yao Chung [4] proposed a new algorithm to overcome the problem of dependency relation of the ID3 algorithm which can degrade the overall performance of the classification. Therefore, they presented a new decision tree classification algorithm, IDA As against the ID3, which accounts the local dependency, IDA counts the global dependency of the variables and it leads to better classification algorithm by selecting the helpful attributes. Here, the comparison is carried out against ID3 and IDA in terms of time analysis. The experimental studies have shown that IDA algorithm outperforms in terms of efficiency and effectiveness.

Baigal tugsSanjaa & Erdenebat Chuluun [5] proposed an approach for detecting the malicious software and performed the investigation on the malicious detection with the help of linear SVM algorithm. The basic principle behind the detection is that, this algorithm learns from the malicious software's dataset and creates a model for detection. It is observed that the rate of detection can be raised by discarding the less weighed features. The experiment is conducted on 297003 features and the study has shown that, detection rate of linear SVM is 75% for unknown malware samples.

Yue Zhou & JinyaoYan [6] proposed an approach for Software Test Management. For the academic and industry purpose software test management is one of the major area in the field of software engineering. Many experts are concentrated on the quality of the software instead of test quality. So, this can be achieved by Software Test Management Consequently, the goal is to set up a calculated relapse-based approach for programming test administration to assess test quality. In this paper, system with manufacture measurements structure for test administration, and count the definition, sort and scope of every metric. Additionally demonstrate a few aftereffects of our investigations utilizing a few information tests from an enormous informational collection.

Akriti Singh et.al [7] proposed the overall view of the debacle of the titanic. But they confirmed that the analysis is still going on for determining the survival of travelers. In this paper they have done the comparison of accuracy of survival rate based on the four major approaches namely logistic regression, Random forest, Naive Bayes and decision tree. Here they have considered a few features that are sex, Pclass, age and children to compute the survival rate. After all the computation they have decided that for finding the survival rate of accuracy, logistic regression will perform better because false rate is also less compare to all other algorithms.

III. METHODOLOGY

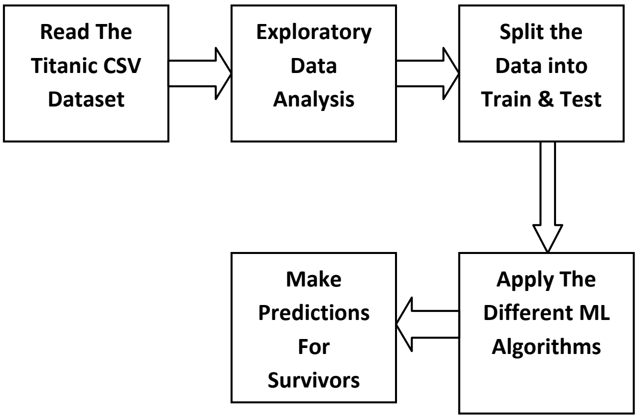

Read The Titanic CSV Dataset
Exploratory Data Analysis
Split the Data into Train & Test
Apply The Different ML Algorithms
Make Predictions For Survivors

Figure 2: Flowchart for the proposed Methodology

The proposed work is based on the concept that firstly we read the titanic dataset obtained by well known kaggle data repository. Next we perform the exploratory data analysis on the dataset. In the third level we break our dataset into two parts training and testing. Further we apply the different well known machine learning algorithm to get the predictions for the expected survivors of titanic accident.

3.1. Dataset

Kaggle website provides the dataset for this work [7]. The data comprises of 891 rows in the prepare set which is a traveller test with their related names. The Passenger class, Ticket number, Age, Sex, name of the passenger, Embarkations, Cabin are provided to each passenger. So here all the provided data are stored in the format of CSV (comma separated value) file. For the test data, the website provided a sample of 418 passengers in the same CSV format. Attributes in the training data set is shown in Table 1:

Table 1: Attributes and their Description in Titanic Dataset

Passenger ID	Identification number of passengers
Pclass	Passenger class (1,2or3)
Name	Name of Passenger
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 passenger
Embarked	Port of embarkation (Cherbourg, Queenstown or Southampton)
Survived	Target Variable (values 0 for perished and 1 for survived)

3.2 Different variables present in the datasets : There are four variables present in the given dataset like Numerical features like age, fare,SibSp and Parch. Apart from this categorical features as sex, embarked, survived and Pclass. Alphanumeric features like ticket and cabin and finally text feature name. All such attributes need to be addressed before performing analytics.

3.2 Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

From the above correlation matrix we see that a high correlation can be seen between Fare and Survived that is around 0.3 which shows that the passengers who paid more amount as a part of their fare are given top priority for rescue operation. There is a negative correlation seen between Fare and Pclass that is -0.55 which shows there is a least correlation between them. Gender too have no correlation with survival whose value is -0.54.

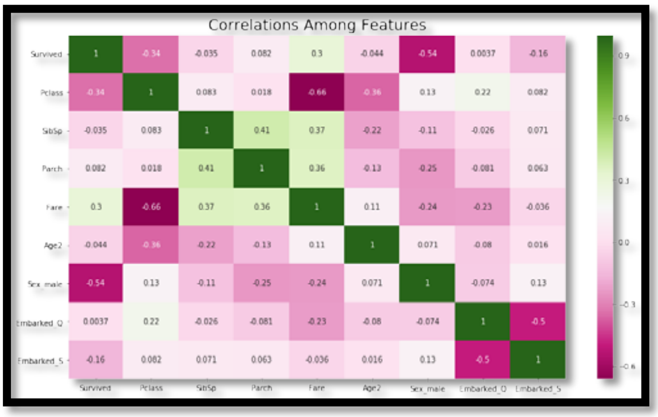


Figure 3: Correlation Matrix of Features of Titanic Dataset

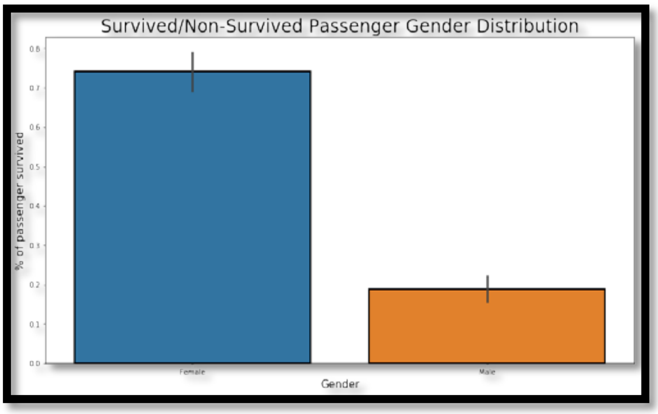


Figure 4: Gender wise Distribution of Survived/Non-Survived

This bar plot above shows the distribution of female and male survived. The x_label shows gender and the y_label shows % of passenger survived. This bar plot shows that 74% female passenger survived while only ~19% male passenger survived.

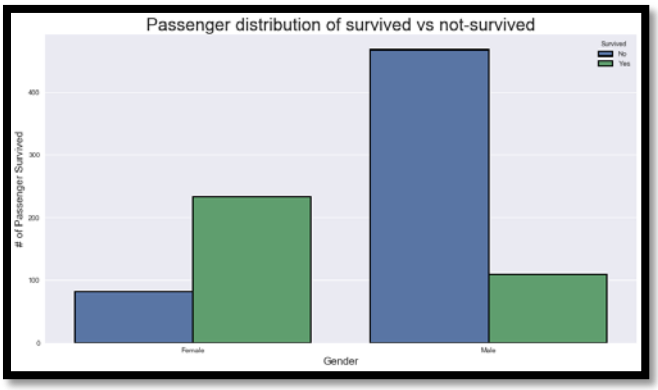


Figure 5: Passenger wise Distribution of Survived/Non-Survived

This count plot shows the actual distribution of male and female passengers that survived and did not survive. It shows that among all the females ~ 230 survived and ~ 70 did not survive. While among male passengers ~110 survived and ~480 did not survive.

As a part of summary to the exploratory data analytics we can claim that female passengers survived much better with respect to male as females carried children with them and therefore they were given the priority in rescue operation.

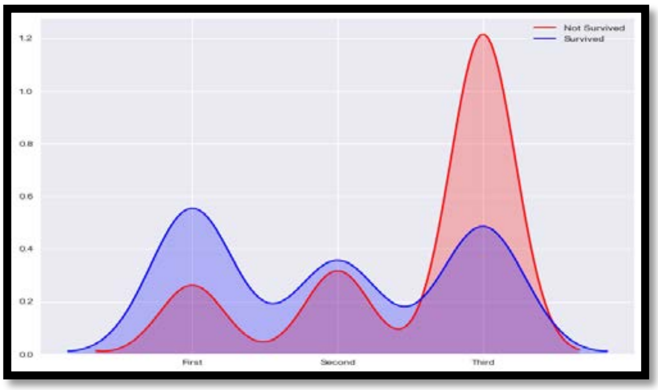


Figure 6: Kde Plot of Survived/Non-Survived

This kde plot is pretty self explanatory with all the labels and colors. Something I have noticed that some readers might find questionable is that in, the plot; the third class passengers have survived more than second class passengers. It is true since there were a lot more third class passengers than first and second.

3.4 Modeling the Data : After EDA activities we have applied several machine learning algorithms like Logistic Regression, Gaussian Naive Bayes, AdaBoost, Support Vector Machines, Decision Tree Classifier, Random Forest Classifier and many more algorithms to train our model and after training we have shown the accuracy bar graph for a comparative study and claim that Support Vector Classifier gives the best accuracy of all the algorithms.

IV. RESULTS & DISCUSSION

In this section, the analysis is done for the following categories: Gender by Survival, Age group by Survival, Survival and Fare relationship, Passenger class by Survival, Survival rate of Gender based on Passenger class. From the above barplot, we can clearly see that the accuracy of the SVC classifier is best out of all other classifiers.

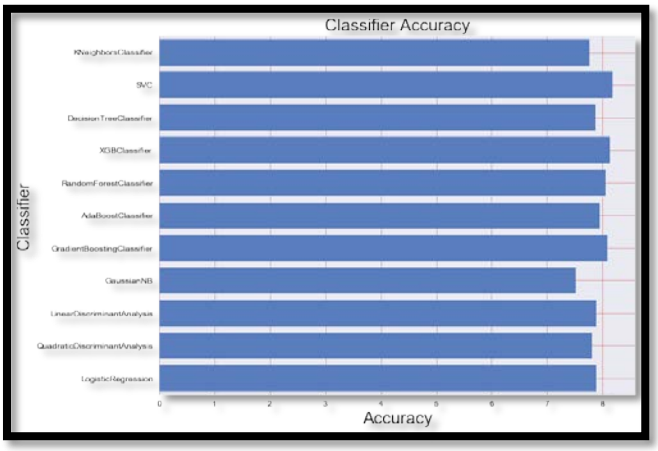
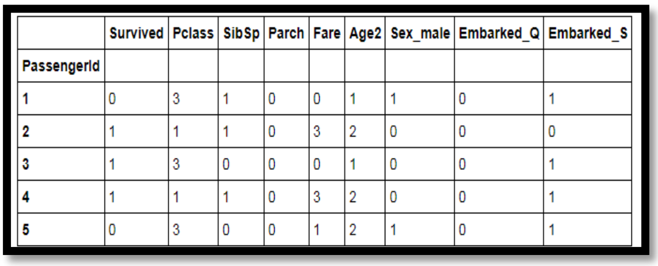


Figure 7: Accuracy Comparison Bar Graph for Different Algorithms

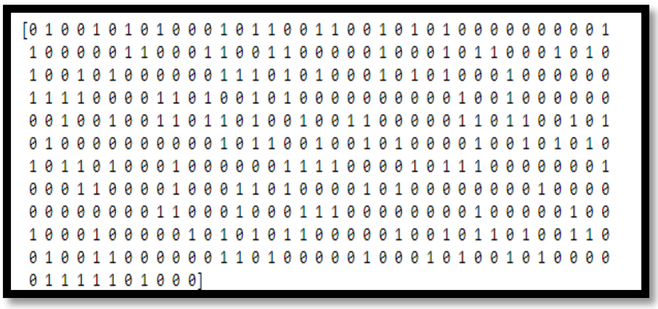
Prediction : Finally on applying the SVC Classifier we obtain the following table

Table 2: Prediction table after applying SVC Classifier



	Survived	Pclass	SibSp	Parch	Fare	Age2	Sex_male	Embarked_Q	Embarked_S
PassengerId									
1	0	3	1	0	0	1	1	0	1
2	1	1	1	0	3	2	0	0	0
3	1	3	0	0	0	1	0	0	1
4	1	1	1	0	3	2	0	0	1
5	0	3	0	0	1	2	1	0	1

The final result is given by 0's and 1's for the entire passenger dataset where 0 stands for not survived and 1 stands for survived.



```
[0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1
 1 0 0 0 0 0 1 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 1 0 1 0
 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0
 1 1 1 1 0 0 0 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0
 0 0 1 0 0 1 0 0 1 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 0 0 1 1 0 1 1 0 0 1 0 1
 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 1
 0 0 0 1 1 0 0 0 0 1 0 0 0 1 1 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0
 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0
 1 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 0 1 1 0
 0 1 0 0 1 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0
 0 1 1 1 1 1 0 1 0 0 0]
```

Figure 8: Final Prediction Using SVC Classifier

V. CONCLUSION & FUTURE WORK

The proposed paper aimed at improving the prediction accuracy of the survivors of titanic accident using the machine learning technique. our model applied different ML classifiers to see for the accuracy and on the basis of all we see that Support Vector Classifier gives the best accuracy of all the models and hence we have also shown the result in terms of 0 and 1 where 0 means not survived and 1 means survived. In future we can also make use of different cross validation techniques as well as deep learning so as to enhance the prediction accuracy of our model.

REFERENCES

[1] Eric Lam, Chongxuan Tang (2012), “Titanic Machine Learning from Disaster”, LamTang-TitanicMachineLearningFromDisaster, 2012.

[2] Haifley, T. (2002, October). Linear logistic regression: An introduction. In Integrated Reliability Workshop Final Report, 2002. IEEE International (pp. 184-187). IEEE.

[3] Chourasia, S. (2013). Survey paper on improved methods of ID3 decision tree classification. International Journal of Scientific and Research Publications, 3(12), 1-2.

[4] Tu, P. L., & Chung, J. Y. (1992, November). A new decision-tree classification algorithm for machine learning. In Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on (pp. 370-377). IEEE. B. Simpson, et al, “Title of paper goes here if known,” unpublished.

[5] Sanjaa, B., & Chuluun, E. (2013, June). Malware detection using linear SVM. In Strategic Technology (IFOST), 2013 8th International Forum on (Vol. 2, pp. 136-138). IEEE.

[6] Zhou, Y., & Yan, J. (2016, October). A Logistic Regression Based Approach for Software Test Management. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2016 International Conference on (pp. 268-271). IEEE.

[7] Singh, A., Saraswat, S., & Faujdar, N. (2017, May). Analyzing Titanic disaster using machine learning algorithms. In Computing, Communication and Automation (ICCCA), 2017 International Conference on (pp. 406-411). IEEE. H. Simpson, Dumb Robots, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.