2018 4th International Conference for Convergence in Technology (I2CT)
SDMIT Ujire, Mangalore, India. Oct 27-28, 2018

Predicting the Survival Rate of Titanic Disaster Using Machine Learning Approaches

Jyothi Shetty
Department of CSE
NMAM Institute of Technology
Nitte – 574110
jyothi_shetty@nitte.edu.in

Pallavi S
Department of CSE
NMAM Institute of Technology
Nitte – 574110
pallavisajangadde@gmail.com

Ramyashree
Department of ISE
NMAM Institute of Technology
Nitte – 574110
ramyashreebhat1994@gmail.com

***Abstract* - The Titanic incident has led the scientist and investigators to comprehend what can have prompted the survival of a few travelers and death of the rest. Many machine learning algorithms contributed in predicting the survival rate of passengers. In addition to the this, a dataset of 891 rows which includes the attributes namely Age, PassengerID, Sex, Name, Embarked, Fare etc. has been used. In this paper, survival of passengers is figured out using various machine learning techniques namely decision tree, logistic regression and linear SVM. The main focus of this work is to differentiate between the three different machine learning algorithms to analyze the survival rate of traveller based on the accuracy.**

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

I. Introduction

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 [1].

The overall accuracy can be calculated by undergoing several stages as depicted by the below Fig. 1 using aforesaid machine learning approaches.

A. Dataset

Kaggle website provides the dataset for this work [10]. 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 I.

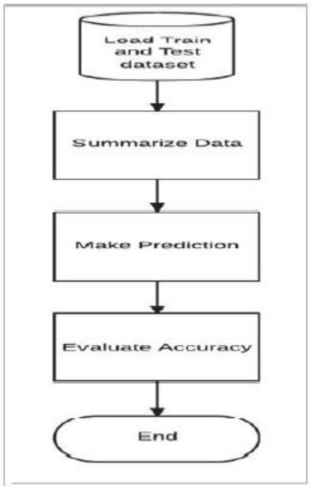


Fig. 1. Overall accuracy calculation

TABLE I Attributes of training dataset

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 the 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, Southampton)
Survived	Target variable (values 0 for perished and 1 for survived)

The data cleaning technique is also applied to manage the missing elements. While observing the csv format of dataset, it is understood that the dataset is incomplete. Because, some of the attribute fields are empty (particularly cabin and age). In

978-1-5386-5232-9/18/$31.00 ©2018 IEEE

order to analyze the survival of passengers age is the central parameter. Therefore, to fill empty values with some numeric values a technique has been employed. To achieve a better prediction model, Sex column has been modified to 0 to 1. (1 for female and 0 for male) [1].

II. LITERATURE SURVEY

Akriti Singh et.al [1] 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.

Pea-Lei Tu & Jen- Yao Chung [2] 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.

Priyank Pandey & Amita Jain [3] proposed FSVM (Fuzzy Support Vector Machine) for the purpose of assigning the multiple points to one particular class. This paper provides the comparative analysis of three classification approaches where they are compared on the basis of input and the obtained output. The study has shown that, the results obtained for FSVM and SVM is different for the same dataset. Hence, they concluded that SVM performs better than FSVM since it cannot always give the better result because it relies upon the problem domain. In addition, the result of decision tree depends on the type of attribute and the criteria being chosen.

Baigal tugsSanjaa & Erdenebat Chuluun [4] 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.

Stephan Dreiseitl & Lucila Ohno-Machado [5] summarized the similarities and dissimilarities of the Logistic regression as well as Artificial Neural Network (ANN) in the field of medical literature. These models are compared with the rest of the machine learning classification algorithms. They outlined the process of building the models, evaluation of the quality and performance factors to report. They found out that, the information that is suitable for calculating the goodness of paper is high for logistic regression because of the easier model-building process. They concluded that, all the algorithms perform differently on any datasets and have their own novelty in terms of results and application areas.

Shikha Chourasia [6] 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.

Satish Kumar et.al [7] explained about the mapping function that are used for Linear SVM. Here it explains that linearity of SVM and the dataset are linearly proportional to each other. The linear SVM is extents undeviating with the extent of the dataset. If the non-linearity is present in the dataset then classification is the challenging task. So, some mapping function are needed to improve the dimensionality. Sometimes obscenity of dimensionality can be appeared. To solve this some mapping functions used widely is kernel function. While doing this function the optimization of parameters is one of the challenging task. So, for this type of case, some replacement of kernel function is needed. So, in this they proposed one intelligent approach that is co- evolutionary approach. Based upon the communication CA is classified as cooperative co-evolution and competitive co- evolution. Here in the mapping function various combination of features are taken. So, it overcomes many disadvantages' too.

Tim Haifley [8] provides the detailed description about the linear logistic regression, value and analysis of reliability. The trustworthy neighbourhood has used well in fitting of survival distributions and the use of design of experiments(DOE). In this paper they give the example of model of human body(HBM) electrostatic discharge(ESD). The purpose of this is to identify the failure at the different stages of voltages. So, in this statistical method are applied. Therefore, this is applicable to variety of linear logistic model.

Yue Zhou & JinyaoYan [9] 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.

Eric Lam &Tang [10] utilized the Titanic issue to think about what's more, differentiate between three calculations Naive Bayes, Decision tree examination and SVM. They presumed that sex was the most prevailing element in precisely anticipating the survival. They additionally proposed that picking vital highlights for getting better outcomes is vital. There are no huge contrasts in exactness between the three techniques they utilized.

III. IMPLEMENTATION DETAILS

Learning models are created using three machine learning methods-Logistic Regression, Linear SVM and Decision tree. A comparison of these algorithms is done based on the accuracy of the result. For developing these algorithms various attributes are used in the train dataset as well as test dataset. All the algorithms are executed using scikit-learn based on Python.

A. Extraction of feature and cleaning of data

The prediction is initiated by dealing with the various parameters of Not Applicable values. In this data set, Age and Cabin column contains the Not Applicable values. Age section had 177 lines with Not Applicable esteems and Cabin segment had 687 lines with Not Applicable esteems. The column Cabin can be dropped from the prediction because there is not relevant feature to predict the survival rate. Since age is a critical trait, the age section is kept for the investigation. Some of the variables are not useful for prediction so such attributes can be dropped from the dataset [1] [10].

B. Logistic Regression

One of the famous classification algorithms is logistic regression to analyze the target feature. It is a nonlinear function which uses the sigmoid function as hypothesis which is given by p=1/(1+e-y). Here categorical and binary are taken as the target variable. In the given dataset the survived attribute is the supported variable (0 for death and 1 for survival) [8].

Steps to evaluate the accuracy using Logistic regression is given as follows:

Step 1: Initially read the dataset using the function panda's read_csv ().

Step 2: The target will be "survivable"' variable from the titanic data frame. To ensure that its factor, utilize the count plot () function.

Step 3: Check the missing value. It can be calculated using is null () function. So once the missing values are identified then the attributes which are not relevant to the decisions can be dropped. In the titanic dataset Ticket, Cabin, Name, Passenger id are not use full for analyzing the survivability. So, drop these attributes.

Step 4: Based on the respective classes, approximate the age of the travelers. This essentially implies individuals with Class esteem 1 will probably make due than Class esteem 2 and individuals with Class esteem 2 will probably survive than Class esteem 3.

Step 5: Changing over unmitigated factors to fake pointers.

Step 6: Evaluate the model. From the confusion matrix it can be analyzed that there are 137 and 69 number of right forecasts and 34 and 27 are the number of wrong predictions.

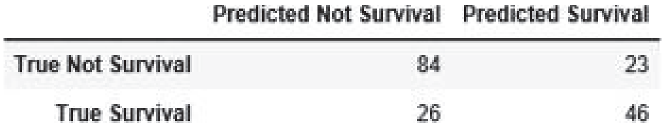
Step 7: Finally calculate the accuracy. For our dataset the accuracy obtained is 80%.

C. Decision Tree

Decision tree characterization procedure is a standout amongst the most prevalent systems in the developing field of information mining. A method of building a decision tree from the set of samples is the method involved in the implementing decision tree algorithm. It is the form of flow chart where every non-terminal node represents the test on a particular attribute and class labels are held with the terminal node [2]. Here, the chance of survival can be calculated basically using Sex. Therefore, initially divide the given data into males and females. Using this field, the accuracy is achieved up to 73.74%. The below Table II shows the matrix that describes survival rate.

D. Linear SVM

TABLE II. CONFUSION MATRIX



	Predicted Not Survival	Predicted Survival
True Not Survival	84	23
True Survival	26	46

It is a supervised learning algorithm which is applicable for classification, regression etc. It performs effectively even if the number of dimensions higher than the samples. It encourages both dense and sparse vector of input. Linear Support Vector Classification (SVC), SVC are the classes based on the kernel which can be used for the multi-class classification. The following illustrates the steps for calculating the accuracy by employing this learning model.

Step 1: Initially read the given dataset

Step 2: Filter out the columns such as Sex, Name, Pclass, Fare which leads to the survival prediction.

Step 3: Preprocessing the data involves the removal of improper data like Cabin, Embarked.

Step 4: The attribute which contributes to the prediction which are null must be filled with the appropriate values using median such as age.

Step 5: Parse the categorical value to the integer type such as sex.

Step 6: Split the data. Step 7: Select the model. Step 8: Train the model.

Step 9: Make predictions for the given training elements.

Step 10: Finally check the accuracy. For our dataset the accuracy obtained is 80.33%.

IV. Results and 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.

A. Gender by Survival

Here the statistical analysis is performed between the gender and survival. The significant result produced by the analysis is t-value is -19.28 and p-value is 0.0, which indicates that the rate of survival of women is more compared to men.

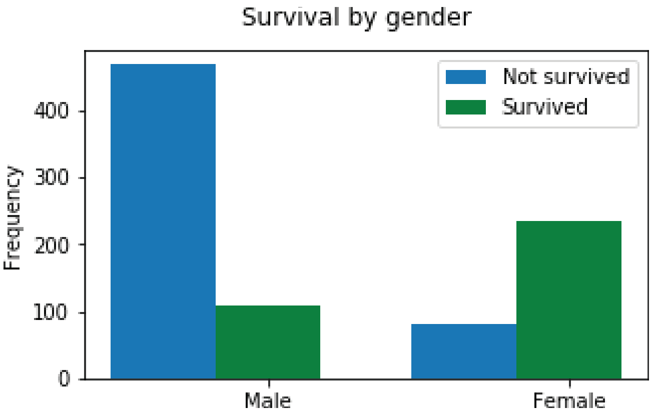


Fig. 2. Survival based on the gender

B. Age group by Survival

Here the statistical analysis is performed between the age group and survival. The t-test is conducted to calculate the mean age difference between the survival and non-survival. The significant result produced by the analysis is for t-value is -2.067 and p-value is 0.039, which indicates that the average age of no survivor is more compared to survivals.

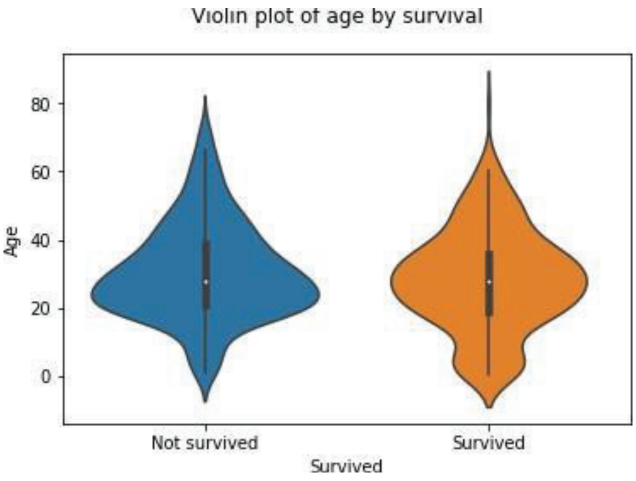


Fig. 3. Survival based on the age group

C. Survival and Fare relationship

A t-test is directed to analyze the distinction in the mean of charge between non-survivors and survivors. The significant result produced by the analysis for t-value is -7.939 and p-value is 0.000, which indicates that the amount paid by the survivor is greater compared to non-survivor.

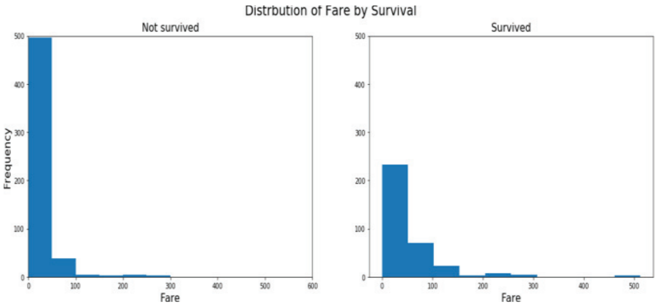
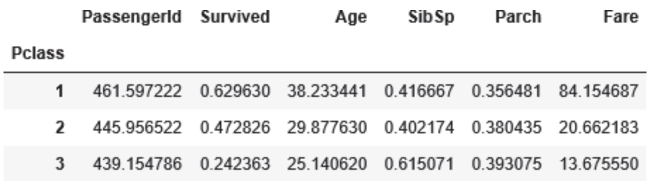


Fig. 4. Survival based on the fare

D. Passenger class by Survival

A test is performed to check the survival rate based on the type of the class.

TABLE III



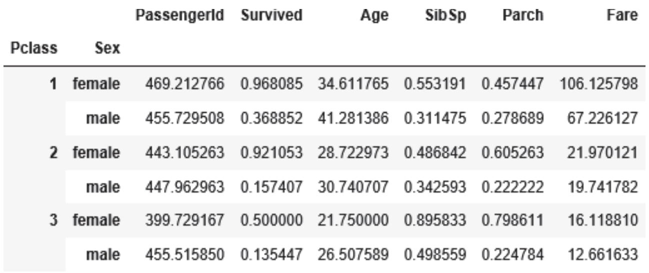
	PassengerId	Survived	Age	SibSp	Parch	Fare
Pclass						
1	461.597222	0.629630	38.233441	0.416667	0.356481	84.154687
2	445.956522	0.472826	29.877630	0.402174	0.380435	20.662183
3	439.154786	0.242363	25.140620	0.615071	0.393075	13.675550

From the Table III, it is observed that, for the 1st class passengers, the rate of survival is 62.96% whereas for 3rd class passengers are 24.2%.

E. Survival rate of Gender based on Pclass

A test is performed to check the survival rate of gender present in the different passenger classes.

TABLE IV. Survival rate of gender based on the Pclass



		PassengerId	Survived	Age	SibSp	Parch	Fare
Pclass	**Sex**						
1	**female**	469.212766	0.968085	34.611765	0.553191	0.457447	106.125798
	male	455.729508	0.368852	41.281386	0.311475	0.278689	67.226127
2	**female**	443.105263	0.921053	28.722973	0.486842	0.605263	21.970121
	male	447.962963	0.157407	30.740707	0.342593	0.222222	19.741782
3	**female**	399.729167	0.500000	21.750000	0.895833	0.798611	16.118810
	male	455.515850	0.135447	26.507589	0.498559	0.224784	12.661633

The above Table IV depicts that the survival rate of female in 1st class is 96% whereas male is 36%. But, it is also observed that the chance of survival of females who belong to the 3rd class is 50% which is less than the 1st class females.

V. Conclusion and future work

This paper aims at providing the accuracy for the titanic dataset using three popular machine learning methods. Out of three techniques, Linear SVM turned out to be the best compared to other two aforesaid techniques with the accuracy of 80.33%. It is also observed that, the features like Pclass, Sex, Age are the most contributing terms towards the survival rate. Table V provides overall accuracy obtained using Linear SVM, Decision tree and Logistic regression.

TABLE V. COMPARISON OF MODELS BASED ON ACCURACY

Models	Accuracy (%)
Logistic Regression	80
Decision tree	73.74
Linear SVM	80.33

Future work includes calculating the accuracy of the train set as well as test set using the cross-validation technique. Different approaches of machine learning such as K-NN classification, clustering can also be developed for finding the survival rate. The accuracy can also be calculated using MAPE (Mean Absolute Percent Error) where it compares the predicted values with the target value.

REFERENCES

[1] 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.

[2] 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.

[3] Pandey, P., & Jain, A. (2016, March). A comparative study of classification techniques: Support vector machine, fuzzy support vector machine & decision trees. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* (pp. 3620-3624). IEEE.

[4] 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.

[5] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352-359.

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

[7] Jaiswal, S. K., & Iba, H. (2017, June). Coevolution of mapping functions for linear SVM. In *Evolutionary Computation (CEC), 2017 IEEE Congress on* (pp. 2225-2232). IEEE.

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

[9] 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.

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

[11] Eric Lam, Chongxuan Tang (2012), "Titanic Machine Learning from Disaster", LamTang-TitanicMachineLearningFromDisaster, 2012

[12] Cicoria, S., Sherlock, J., Muniswamaiah, M., & Clarke, L. Classification of Titanic Passenger Data and Chances of Surviving the Disaster.

[13] Santos, K.C.P., Barrios, E.B. (2017). Improving predictive analysis of logistic regression model using ranked set samples. *Communications in Statistics-Simulation and Computation, 46(1),78-90.*

[14] Whitley, M. A. (2015). Using statistical learning to predict survival of passengers on the RMS Titanic

[15] Russel, S., & Norvig, P. (2015). "Artificial Intelligence–A Modern Approach", Pearson Education, 2003.