TITANIC SURVIVAL ANALYSIS USING LOGISTIC REGRESSION

Kalsalingam Academy of Research and Education

**J.Jayanthi**  
CSE-Dept(Associate professor)  
Kalsalingam Academy of Research and educationVirudh nagar,India  
[jayanthi.j@klu.ac.in](mailto:jayanthi.j@klu.ac.in)

**Dhanekula Giri VenkataManohar**  
Computer science engineering  
Kalsalingam Academy of Research and educationGuntur,India  
[manojdhanekula2002@gmail.com](mailto:manojdhanekula2002@gmail.com)

**Pamulapati Vamsi Krishna**   
Computer science engineering  
Kalsalingam Academy of Research and educationGuntur,India  
[vamsipamulapati999@gmail.com](mailto:vamsipamulapati999@gmail.com)

**Mirza Adnan Baig**  
Computer science engineering  
Kalsalingam Academy of Research and educationAdilabad,India  
[mirzaadnan93817@gmail.com](mailto:mirzaadnan93817@gmail.com)

**Maraka Rajesh Kumar**  
Computer science engineering  
Kalsalingam Academy of Research and educationKothacheruvu,India  
[marakarajeshkumarreddy@gmail.com](mailto:marakarajeshkumarreddy@gmail.com)

**Abstract :** A logistic regression analysis of an extensive data set on the Titanic passengers is presented which tests the likelihood that a Titanic passenger survived the accident--based upon passenger characteristics. The main finding is that underneath the strong overt preference afforded in the rescue by the authorities to women and children over men, there was a complex class determination of survival rates among men, on the one hand, and women and children, on the other. We hypothesize that the statistical interactions of gender and class are explained by two crucial decisions made by the ship's authorities:

1. to encourage, and perhaps direct, some three to four hundred 3rd Class men quartered in the lower decks at the forward end back to the afferent of the stern.

2. to restrict the access of men at the after- end of the ship to the lifeboats launched from the after Boat and Promenade Decks.

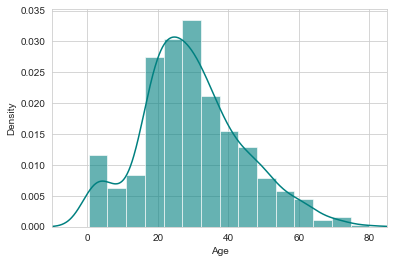
**Key Words**: Logistic Regression, Data Analysis, Kaggle Titanic Dataset, Data pre-processing , Cross validation, Confusion Matrix.

**1.INTRODUCTION**

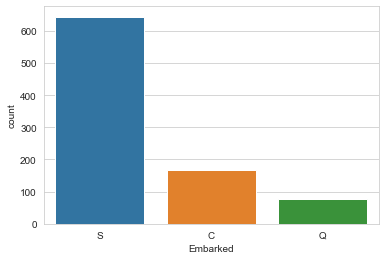
The disaster that occurred over a century ago ripped off Many parts of the Titanic during the fateful night. Regrettably, There were not enough lifeboats present to rescue all the 2224 Passengers on-board. The dead included many men whose Place were given to the children and women. The goal of this research paper is to correctly predict who Would survive the Titanic given a set of demographic Information. A predictive model using passenger data was Built so people’s genders, their ages, what class of ticket they Belonged from, and their socio-economic class, all contributed To whether they would be lucky enough to survive or Tragically perish on the Titanic. Predictive analysis is a Method of determining important and useful patterns in broad Data sets by combining statistical approaches. To determine Significant and useful trends in large data. Survival is Predicted using machine learning algorithms based on Different feature combinations. Aim of this research is thereby to conduct exploratory data Analytics to excavate different knowledge existing in available Data set and to perceive the impact of every field with respect To the passengers’ survival by the use of “Survival” field Analytics in between each field of the data set. Data analysis On applied algorithms was performed and likewise the Accuracy was tested. Based on this, different algorithms are Compared, and the best performing model was selected .After analysing the Titanic dataset, two predictions were Generated. The first was to see what the lucky passengers had In common that helped them survive the shipwreck, while the Second was to see if I would have survived had I been aboard. The fateful ship by applying the tools of machine learning.

**2. ALGORITHM**

Data Pre-processing In the dataset available for the prediction some of the data values are missing or unknown. This missing data was resulting in reducing the accuracy of the overall prediction model and also reduces the size of pure training data which in turn reduces accuracy. Data preprocessing is a technique that involves transforming raw data into an understandable format.



Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Missing values are replaced by average of that column. So, the missing and unknown data of the passengers which is easily predictable is filled up by this step.



**3. Classification**

**Logistic Regression:**

Second step of the algorithm is using a classifier to classify the available information. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous .

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It uses a method of using he regression line between dependent and independent variable to predict the value of the dependent variable.

**4.Creating our training model and predict.**

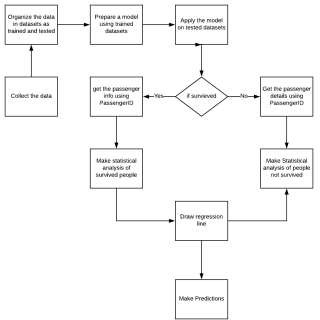
* Now we are ready to create & train a model and also predict the required solution.
* There are lot of predictive modelling algorithms to choose from.
* Here Our problem is a classification and regression problem.
* We want to check relationship between output (Survived or NOT Survived) with other variables or features like (Gender, Age, Class etc).
* We train our model by using Logistic Regression.
* We can also use these models:
* KNN
* Ada Boost Classifier
* XGBoost
* Random Forrest etc.

**5. Cross validation**

Dataset is divided into two main parts namely Train and Test data. Training data will be considered for the training of the machine. Test data will be used for validating the machine. Cross validation technique used here is K-Fold. The method has only one parameter called k that refers to the number of groups into which a given data sample is to be split. As such, the method is also called k-fold crossvalidation. When a particular value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

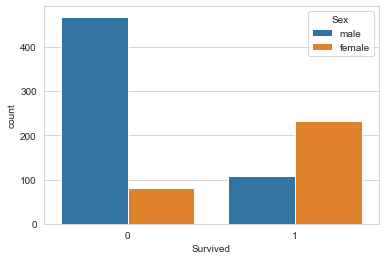
**6. Analysis of confusion matrix**

Confusion matrix is used to show the performance of the algorithm. Accuracy of the model can be predicted using the confusion matrix. It is a plotting of relation between real and predicted outputs. It allows us to check the accuracy and performance of the algorithm. In this case we are using two attributes at a time for the confusion matrix plotting. Test case data is used to build the confusion matrix. The values shown in the confusion matrix are the probability of survival of the individual considering only those parameters. The cell on first column and is of age and the 7th row is sex\_male i.e. the probability of surviving the individual is depending on the age and the gender as if he is male is 0.081. As it is positive there is a possibility that the person with this attribute survives.



**7. Total Number of male & female Survived**

The survival rate for the female passengers is very high because the strict maritime tradition of evacuating women and children first.



Women have a survival rate of 78%, while men have a survival rate of about 19%.

According to Passenger Class (i.e; Pclass) Pclass = 2 has high rate of survival.

**8. RESULTS**

The logistic regression gives the accuracy of **78%** which is based on the confusion matrix. The parameters used here are accuracy and false discovery rate. Accuracy is a measure of the correctness of the prediction of the model. Higher accuracy is always better and is calculated by (TN + TP)/Total number of rows \*100 False discovery rate are the false positive measures of confusion matrix where the model predicts that the passenger would survive but in reality, it doesn’t. This would prove dangerous as the prediction may go wrong and hampers the accuracy of the results. The attempts are being made to increase the accuracy rate and reduce the false discovery rates.

**9. CONCLUSIONS**

The logistic regression provides a better accuracy i.e. almost of about **78%**. It works better with binary dependent variable which means the variable has a binary value as its output like yes or no, true or false.

****

The ROC curve is the plotting of the output based on the false positive rate and the true positive rate plotted along x and the y-axes. The Curve depicts the performance of various algorithms on the same data which helps to compare the performance, accuracy and efficiency of the algorithm. It helps to decide the best algorithm which is suitable for user’s requirement.

**REFERENCES**

Singh, A., Saraswat, S., & Faujdar, N. (2017, May). Analyzing Titanic Disaster using machine learning algorithms. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 406-411). IEEE.

Whitley, M. A. (2015). Using statistical learning to predict survival of Passengers on the RMS Titanic.

Lam, E., & Tang, C. (2012). CS229 Titanic–Machine Learning From Disaster.

Mishra, V. P., Shukla, B., & Bansal, A. (2018). Intelligent Intrusion Detection System with Innovative Data Cleaning Algorithm and Efficient Unique User Identification Algorithm. Jour of Adv Research in Dynamical & Control Systems, 10(01), 398-412.

Mishra, V.P., Shukla, B. and Bansal, A., 2019. Analysis of alarms to Prevent the organizations network in real-time using process mining Approach. Cluster Computing, 22(3), pp.7023-7030.

Vyas, K., Zheng, Z., & Li, L. (2015). Titanic-Machine Learning From

Disaster. Machine Learning Final Project, UMass Lowell, 1-7.

Kaggle, Titanic: Machine Learning form Disaster[Online].Avail

<http://www.kaggle.com/>