NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 – 281263, Fax: 08258 – 281265
**Department of Computer Science and Engineering**
B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

# Mini Project Synopsis

Project Title:

## TITANIC SURVIVAL PREDICTION

Submitted by:

4NM16CS107 Preethi M Shenoy

4NM16CS109 Prithvi A

4NM16CS131 Sanjana Nambiar

Submitted to:

**Ms. Divya Jennifer D'Souza**

Assistant Professor Gd. I

Department of Computer Science and Engineering

# Table of Contents

## Abstract:

The goal of the project was to predict the survival of passengers based off a set of data. We used Kaggle competition "Titanic: Machine Learning from Disaster" to retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not a passenger survived). We used this set to build our model to generate predictions for the test set. For each passenger in the test set, we had to predict whether or not they survived the sinking. Our score was the percentage of correctly predictions.


## Literature Survey:

Yi-han Wang ,Yang Ou ,Xu-dong Deng , Lu-ran Zhao, Chao-yu Zhang in their article [1] proposed a system to predict ship collision accidents based on Logistic Regression and Big Data. In this paper, the formation of a collision accident, especially the human factors, is partially refined into the more detailed factors causing the collision, and the maximum impact factors are obtained by using Logistic analysis. The Logistic Ordered Multiple Regression in Regression Analysis is applied to the study of the causes of ship collision.

Darwin Prasetio , Dra. Harlili presented a paper [2] where a logistic regression model is built to predict matches results of Barclays' Premier League season 2015/2016 for home win or away win and to determine what are the significant variable to win matches.

Tao Lu ,Zhu Dunyao ,Yan Lixin ,Zhang Pan [3] built a the prediction model of accident hotspot was established. The results show that the location of car in road transects, the road safety grade, the road surface condition, the visual condition, the vehicle condition and the driver state are the most significant factors which may lead to traffic accident.

Yong Han , Muyun Yang ,Haoliang Qi ,Xiaoning He ,Sheng Li presented a paper [4] that presents an improved logistic regression model which reduces the impact of the features appearing in both spam messages and ham ones. Byte level n-grams are employed to extract the features from messages, and TONE (train on or near error) is adopted, which are proved effective in state-of-the-art spam filtering system.

Weiguo Li ,Cuiying Li ,Xiaoping Du ,Kun Qian ,Hanjie Zhang ,Dezao Hou this paper [5] formalizes traffic states based on the average vehicle speed. Though applying ordered logistic regression approach to specify the factors affecting traffic flow, a high accurate prediction model was established.

# Methodology:

## Logistic Regression (One-Vs-One and One-Vs-All)

We have used Logistic Regression algorithm. Both One-Vs-One and One-Vs-All logistic regression multi-class classification are based on binary classification. In the logistic regression function, set one class to be positive (yi = 1) and other class to be negative (yi = 0).

In this Logistic Regression Algorithm, logistic function I used is

g(z) = 1/(1+e^-z) formula 1

We apply the Gradient Ascent for the logistic Regression . Firstly, create "w" and initialize w = [1,….,1]

T . Before run the the algorithm, I normalized all the features (IWe have used Z- Score normalization, which is a general requirement for logistic regression), because different features have different number range. The larger value in a feature may mislead the model.

The Z-Score normalization function is:

z = (x-mean(x))/std(x) formula 2

After normalized the data, Repeat formula 1 until convergence:

w = w + a/numrow(x) X'(Y - g(X * w)) formula 3

where w is the weight, X is the input, Y is the output, and $\alpha$ is the hyper-parameter.

One-Vs-One and One-Vs-All are two different approaches to multiclass classification. One-Vs-All strategy is training a single classifier for each class (positive for each class and negative for all others), 9 classifiers for this problem. One-Vs-All pick the class with highest score to be the prediction, the score function 6 is:

$i$ = arg (max $f$i($x$)) formula 4

One-Vs-One is training a classifier for each pair of class, total $K(K-1)/2$ classifiers, 36 classifiers for this problem. Each classifier votes for the wining class in a pair, and the class with most votes wins.

## Dataset

Training and Test data come in CSV file and contains the following fields:

- Passenger ID

- Passenger Class
- Name
- Sex
- Age
- Number of passenger's siblings and spouses on board
- Number of passenger's parents and children on board
- Ticket
- Fare
- City where passenger embarked

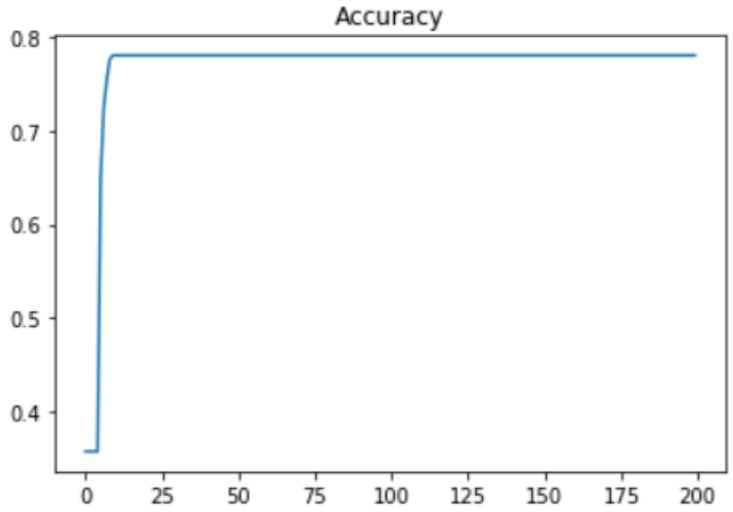
## Procedure

The titanic dataset mainly involves predicting if the passenger with the given details survives the wreck or not. Thus, logistic regression has been used to solve this problem.

- Firstly, the dataset undergoes data preprocessing as there as many inconsistencies in the dataset. There are many attributes that have null values. This needs to be corrected as we can get inaccurate results otherwise.
- All string entries are converted into integer values so that they can be evaluated and used for prediction.
- Once the data has been preprocessed, we split the dataset into training and testing data. First, we train the model using the training data several times so that the model learns the required patterns.
- Then we use the model on the testing dataset to obtain the respective predictions. The obtained values are then compared with the actual values and the error is calculated. This error is further used to update weights and the model is trained accordingly.
- After a reasonable number of epochs, we get the value of weights, which gives us the required accuracy.

# Observation and Discussion:

The plot of the accuracy obtained using training dataset is shown below:




```
Weight [-1.91410393  2.61445987 -0.62648951  0.58425376]
Validate {'loss': 0.45833908745142343, 'accuracy': 0.7798507462686567}
```

Figure 1: Accuracy obtained using training dataset

## Result:

The following shows the figure accuracy obtained while classifying our test data into our implementation Logistic Regression Algorithm.

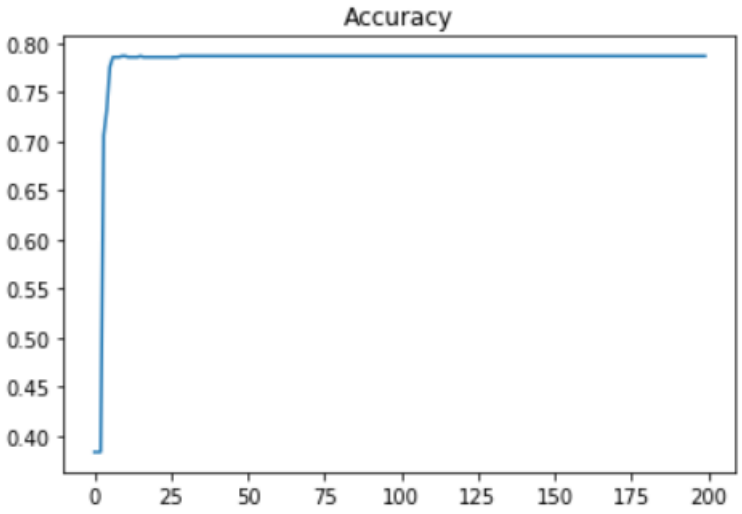


Figure 2: Accuracy obtained using testing dataset

## Conclusion:

In this project we have implemented logistic regression for titanic dataset successfully, and have obtained an accuracy of 78%. Using details of passengers like their age, gender etc. we have predicted if they would survive the ship wreck or not.

## References:


[1] Yi-han Wang ,Yang Ou ,Xu-dong Deng , Lu-ran Zhao The Ship Collision Accidents Based on Logistic Regression and Big Data. 2019 Chinese Control And Decision Conference (CCDC), 3-5 June 2019.

[2] Darwin Prasetio , Dra. Harlili Predicting football match results with logistic regression, 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 16-19 Aug. 2016.

[3] Tao Lu ,Zhu Dunyao ,Yan Lixin ,Zhang Pan The traffic accident hotspot prediction: Based on the logistic regression method ,2015 International Conference on Transportation Information and Safety (ICTIS) ,Year: 2015.

[4] Yong Han , Muyun Yang ,Haoliang Qi ,Xiaoning He ,Sheng Li The Improved Logistic Regression Models for Spam Filtering , 2009 International Conference on Asian Language Processing , Year: 2009.

[5] Weiguo Li ; Cuiying Li ; Xiaoping Du ; Kun Qian ; Hanjie Zhang A traffic flow prediction model based on ordered logistic regression, 6th International Conference on Digital Content, Multimedia Technology and its Applications, 16-18 Aug. 2010.