Survival On Titanic Using Logistic Regression

Axit Dhola
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901004@daiict.ac.in

Bhadrayu Bhalodia
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901049@daiict.ac.in

Dhruvil Gorasiya
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901061@daiict.ac.in

Dhyey Vaghani
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901262@daiict.ac.in

Denish Hirpara
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901424@daiict.ac.in

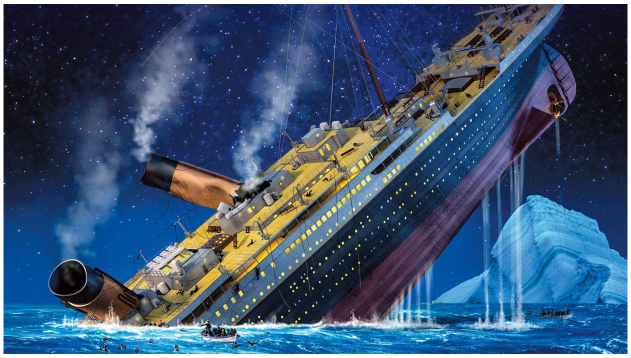
Parin Jasoliya
Student, B.Tech - 3rd year, DAIICT
Gandhinagar, India
201901451@daiict.ac.i

ABSTRACT

The sinking of the Titanic ship resulted in the deaths of approximately thousand passengers and the crew is one of the deaths in history .The loss of life was mainly caused by the shortage of lifeboats.the incident is that some people were more durable to carry than many others ,such a children ,women were the ones with the highest priority to be rescued .The primary goal of the algorithm is first to find predictable or previously unknown data by implementing exploratory data analysis on the available training data and then applying different machine learning models and classifiers to complete the scan,this will help predict which people are most likely to survive.

PROBLEM STATEMENT:

The sinking of the Titanic is one of the most infamous shipwrecks in history.On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this challenge, we ask you to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (ie name, age, gender, socio-economic class, etc).



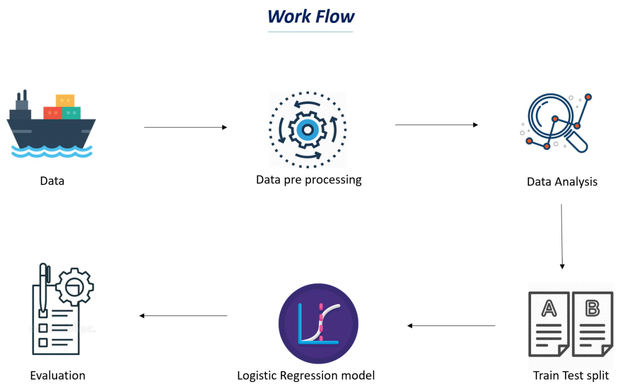
INTRODUCTION

The most infamous disaster which occurred over a century ago on April 15, 1912, is well known as the sinking of "The Titanic". The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender were present on that fateful night, but the bad luck was that there were only a few lifeboats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine.

 Machine learning algorithms are applied to make a prediction about which passengers survived at the time of the sinking of the Titanic. Features like ticket fare, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform exploratory data analytics to mine various information in the dataset available and to know the effect of each field on survival of passengers by applying analytics between every field of the dataset with the "Survival" field. The predictions are done for newer data sets by applying machine learning algorithms. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions.

This dataset has been studied and analyzed using various machine learning algorithms like Logistic Regression, Random Forest, SVM etc. Various languages and tools are used to implement these algorithms including Matlab ,Weka, Python, R, Java etc. The approach of the research paper is centered on Matlab for executing algorithm Logistic Regression. The prime objective of the research is to analyze the 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.



3.LOGISTIC REGRESSION:

- Logistic regression is machine learning algorithm that is used to predict the probability of categorical
- dependent variable.
- In, logistic regression dependent variable is a binary variable that contains data coded as 1 (yes or success) or 0 (no or failure).
- In other words, the logistic regression model predicts p(Y=1) as a function of x.
- Logistic regression just has transformation based on linear regression hypothesis.
- For Logistic regression focusing on binary classification here, we have class 0 and class 1.
- To compare with the target, we want to

constrain prediction to some values between 0 and 1.

CLASSIFICATION:

We are explaining logistic regression with an example of a brain tumor that is benign or malignant.
Tumor: malignant or benign?

here, Y = {0,1} 0:“Negative Class” (e.g., benign tumor)

1:”Positive Class” (e.g. malignant tumor)

x = tumor size

Here Y = 0 or Y = 1

So our hypothesis must be between 0 and 1

so, 0 <= h(θ) <=1

In logistic regression we use the sigmoid function, the benefit of using this function is that the output of this function is all between 0 and 1.

$$h_{\theta}(x) = g(\theta^{T}x)$$

$$g(z) = 1/(1 + e^{-Z})$$

Now, we are explaining each step of logistic regression implemented in matlab.

COST FUNCTION FOR LOGISTIC REGRESSION

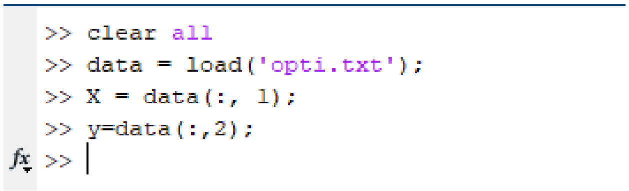
Training set: { (x1 , y1) , (x2 , y2) , (x3 , y3) , ……….. , (xm , ym) }.

m examples: $x \in [x1\ ,\ x2\ ,\ x3\ ,\ ....\ ,\ xn]^{T}$

our training set is:
{(0.5213,1),(0.5632,1),(0.6254,1),(0.5789,1),(0.5164,1),(0.4536,0),(0.4987,0),(0.4632,0),(0.4123,0),

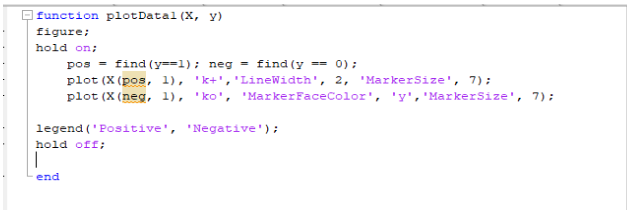
(0.4639,0)}

Now we load data in matlab.

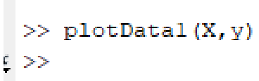


```
>> clear all
>> data = load('opti.txt');
>> X = data(:, 1);
>> y=data(:,2);
>> |
```

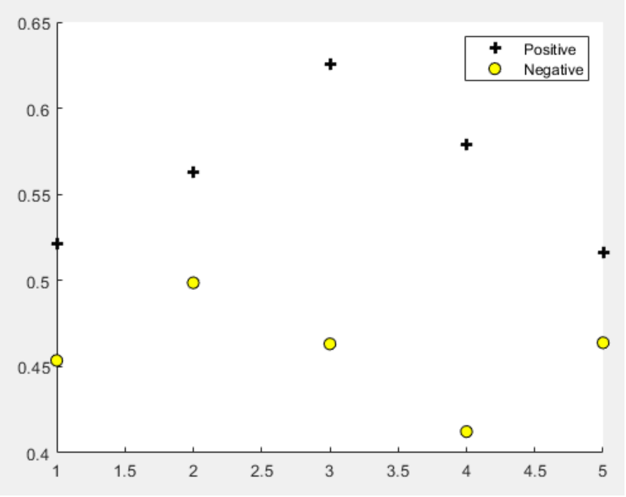
Now we plot our data,



```
function plotData1(X, y)
figure;
hold on;
    pos = find(y==1); neg = find(y == 0);
    plot(X(pos, 1), 'k+','LineWidth', 2, 'MarkerSize', 7);
    plot(X(neg, 1), 'ko', 'MarkerFaceColor', 'y','MarkerSize', 7);

legend('Positive', 'Negative');
hold off;

end
```



```
>> plotData1(X,y)
>>
```



COST FUNCTION:

$J(\theta) = 1/m * \sum Cost(h_{\theta}(x(i)), y(i))$ (where i from 1 to m)

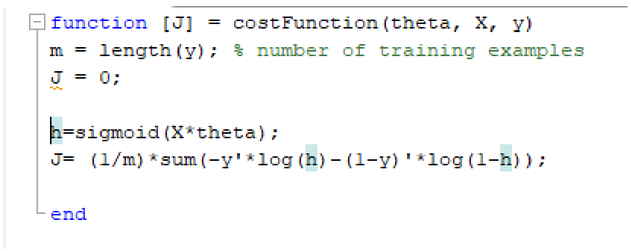
$Cost(h_{\theta}(x),,y) = \quad -log\ (h_{\theta}(x),) \quad if\ \ y = 1$

$\qquad\qquad\qquad\qquad -log(\ 1\ -\ h_{\theta}(x),) \quad if\ \ y = 0$

here,

$J(\theta) = 1/m * \sum Cost(h_{\theta}(x(i)), y(i))$

$J(\theta) =$
$- 1/m * [\sum y(i) * log(h_{\theta}(x(i)) + (1 - y(i)) * log(1 - h_{\theta}(x(i))$
]



```
function [J] = costFunction(theta, X, y)
m = length(y); % number of training examples
J = 0;

h=sigmoid(X*theta);
J= (1/m)*sum(-y'*log(h)-(1-y)'*log(1-h));

end
```

Now, we want to minimize our cost function, so we use an optimization method called gradient descent. Using gradient descent we can minimize our function, in gradient descent we use a differential method to minimize the function.

GRADIENT DESCENT:

$J(\theta) =$
$- 1/m * [\sum y(i) * log(h_{\theta}(x(i)) + (1 - y(i)) * log(1 - h$
]

here we want to minimize J(θ) respects to θ.

so, $min_{\theta} \quad J(\theta)$:

repeat {

$$\theta_j = \theta_j - \alpha \frac{d\, j(\theta)}{d\, \theta_j}$$

simultaneously update all θ_j

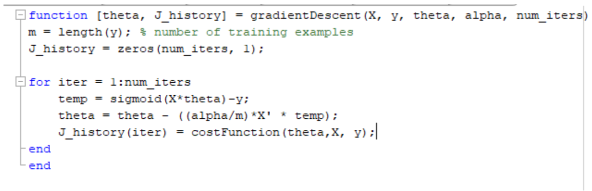
}

so, $min_{\theta} \quad J(\theta)$:

repeat {

$$\theta_j = \theta_j - \frac{\alpha}{m} \sum (h_{\theta}(x(i)) - y(i))\ x_j(i)$$

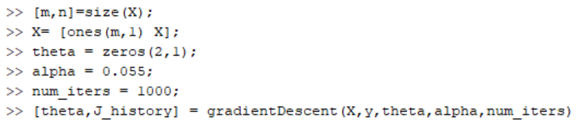
simultaneously update all θ_j



```
function [theta, J_history] = gradientDescent(X, y, theta, alpha, num_iters)
m = length(y); % number of training examples
J_history = zeros(num_iters, 1);

for iter = 1:num_iters
    temp = sigmoid(X*theta)-y;
    theta = theta - ((alpha/m)*X' * temp);
    J_history(iter) = costFunction(theta,X, y);
end
end
```

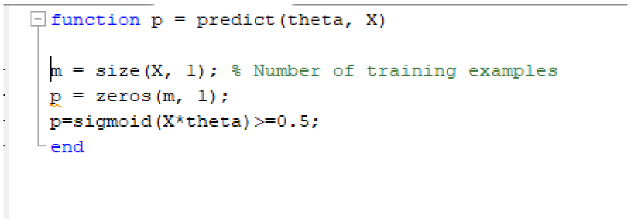
}

After loading data and visualizing the plot of data, we apply a logistic model on data.

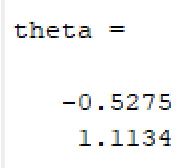


```
>> [m,n]=size(X);
>> X= [ones(m,1) X];
>> theta = zeros(2,1);
>> alpha = 0.055;
>> num_iters = 1000;
>> [theta,J_history] = gradientDescent(X,y,theta,alpha,num_iters)
```

After run this code we can find optimal theta of our data and theta is following:



```
function p = predict(theta, X)

m = size(X, 1); % Number of training examples
p = zeros(m, 1);
p=sigmoid(X*theta)>=0.5;
end
```



```
theta =

   -0.5275
    1.1134
```

hence we find minimal θ for our hypothesis, so now we can predict our output:

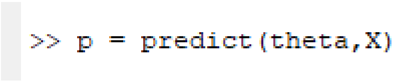
$$Y = h_{\theta}(x) = 1/(1 + e^{-\theta^T x})$$

Interpretation of our output:

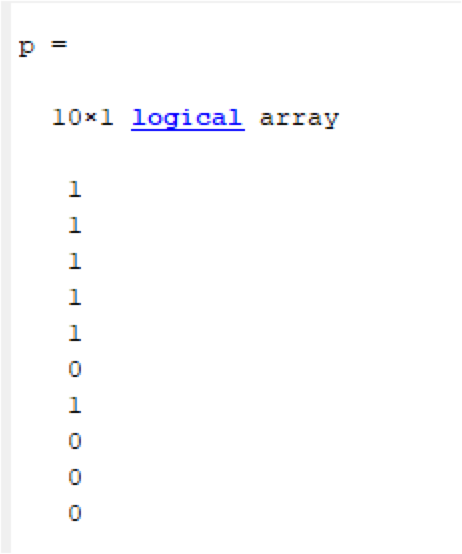
$h_{\theta}(x)$ = estimated probability that Y=1 on input x

so we can predict that if $h_{\theta}(x)$>=0.5 then Y=1

$h_{\theta}(x)$<0.5 then Y=0



```
>> p = predict(theta,X)
```



```
p =

  10×1 logical array

   1
   1
   1
   1
   1
   0
   1
   0
   0
   0
```

So, this way we can predict output of any given data. We also can use this model to predict the output of data which are not available.

LITERATURE REVIEW:

A. Predicting the likelihood of survival of titanic by machine learning

The sinking of RMS Titanic is presumably one of the most infamous and disastrous shipwrecks in history.An eye-opening observation that came forth from the sinking of Titanic is the fact that some individuals had a better chance at surviving than the others. Kids and women had been given foremost priority .Social classes were heavily stratified in the early twentieth century, this was especially implemented on the Titanic Firstly, the aim is use and apply exploratory data analytics (EDA) to uncover previously unknown or hidden facts in the data set available.

B. Logistic Regression by andrew ng (Ng)

This video explains the logistic regression which is the core part of the project. It is a machine learning algorithm. It is used in prediction of output whether it is success or failure. We also practice different examples and techniques of logistic regression.

C. Titanic survival Prediction, Analytics Vidhya

This article contains a brief description about logistic regression. Also explains workflow and machine learning(python) parts. Mainly we used this article to get knowledge about logistic regression.

D. Titanic survival applying data analytic and machine learning techniques

This article explains the flow of work, how we can use data analysis to solve the problem. This article saws that using data features we can fill data where there are nun values. And after applying the model on the dataset this explains how to compare and analyze the basis of accuracy.

PLAN OF WORK:

1) *Goal 1 (completed)*
- **Objective**:

  Finding good quantity and quality of Research Paper/Reference Materials. Formulating the Problem Statement
- **Timeline**:

  15th Sept - 20th Sept
- **Strategy**:

  We decided that each member will find some good research papers. We discussed the papers in online meetings. Then we formulated the Problem Statement. The references we have used for this are mentioned in the Reference section

2) *Goal 2 (completed)*

- **Objective:**

  Learning theory that are involved in project (titanic survival)

- **Timeline**:

  23rd Sept - 29th Sept

- **Strategy** :

  We see some videos (ML by andrew) on youtube.we also read some articles and books.(Ng)

3) *Goal 3 (completed)*

- **Objective**:

  Compute Logistic regeneration and implement MATLAB code for logistic regression.

- **Timeline**:

  1st Oct - 8th Oct

- **Strategy**:

  We discussed our views and tried to solve the queries we had in understanding the theory. Then we computed logistic regression involved in our problem and tried to implement it in MATLAB. For this, we referred to a YouTube video mentioned in the Reference Section.

4) *Goal 4 : (completed)*

- **Objective**:

  Applying logistic regression on titanic survival dataset. And do data features
  before applying the model.

- **Timeline**:

  26th Oct - 30th Oct

- **Strategy**:

  Using our knowledge in logistic regression we applied this on our project and predicted the output.

5) *Goal 5 (completed)*

- **Objective**:

  Implementing the final version of our solution in MATLAB using logistic regression and compute the accuracy of our model.

- **Timeline**:

  1st Nov - 10th Nov

- **Strategy**:

  In the final version of MATLAB code, we will implement logistic regression and using the same method we compute the accuracy of our model.

OBJECTIVE OF PROBLEM

Reading the datasetThe objective of this Kaggle challenge is to create a Machine Learning model which is able to predict the survival of a passenger on the Titanic, given their features like age, sex, fare, ticket class etc.

The outline of this tutorial is as follows:

1)Data Analysis

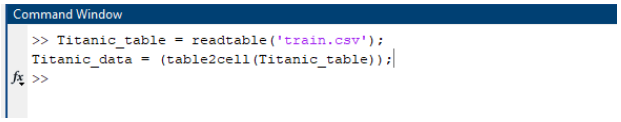
2)Feature Engineering

3)Model Fitting

4)Prediction on the test set

1. Exploratory Data Analysis

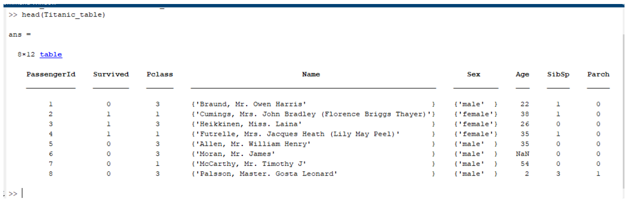
Reading the dataset



```
>> Titanic_table = readtable('train.csv');
Titanic_data = (table2cell(Titanic_table));
>>
```

The train set has 891 passenger entries and 12 columns.

Now, let's take a look into our data. The head function displays the top rows of a table, similar to that in Pandas.



```
>> head(Titanic_table)

ans =

  8×12 table

    PassengerId    Survived    Pclass                            Name                               Sex        Age    SibSp    Parch
    ___________    ________    ______    _____________________________________________________    __________    ___    _____    _____

         1            0          3       {'Braund, Mr. Owen Harris'                           }    {'male'  }    22       1        0
         2            1          1       {'Cumings, Mrs. John Bradley (Florence Briggs Thayer)'}    {'female'}    38       1        0
         3            1          3       {'Heikkinen, Miss. Laina'                            }    {'female'}    26       0        0
         4            1          1       {'Futrelle, Mrs. Jacques Heath (Lily May Peel)'      }    {'female'}    35       1        0
         5            0          3       {'Allen, Mr. William Henry'                          }    {'male'  }    35       0        0
         6            0          3       {'Moran, Mr. James'                                  }    {'male'  }   NaN       0        0
         7            0          1       {'McCarthy, Mr. Timothy J'                           }    {'male'  }    54       0        0
         8            0          3       {'Palsson, Master. Gosta Leonard'                    }    {'male'  }     2       3        1

>>
```

The ‘Survived’ column is our binary target variable which we need to predict; where 0- Not Survived, 1- Survived.

The predictor variables i.e. features are as follows-

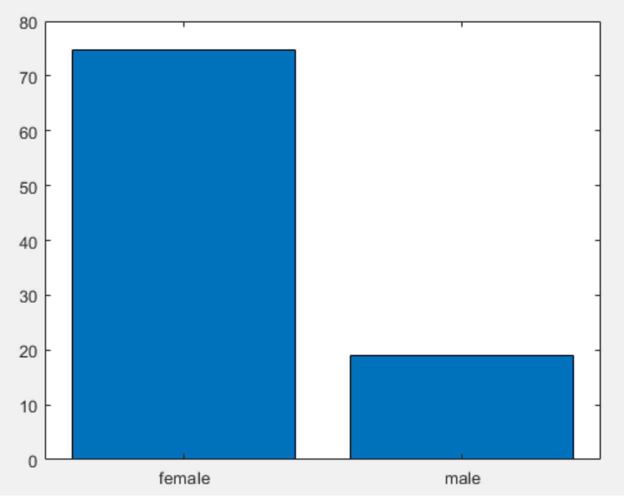
- Pclass: Ticket class
- Sex
- Age
- SibSp: number of siblings, spouses along with the passenger
- Parch: number of parents, children along with the passenger
- Ticket: Ticket number
- Fare
- Cabin: Cabin number
- Embarked: Port of Embarkation

‘Pclass’ variable is indicative of socio-economic status. It has 3 possible values ‘1’, ‘2’ , ‘3’ representing the Upper, Middle and Lower class respectively.

The probability of survival for passengers in the training set is around 38 %

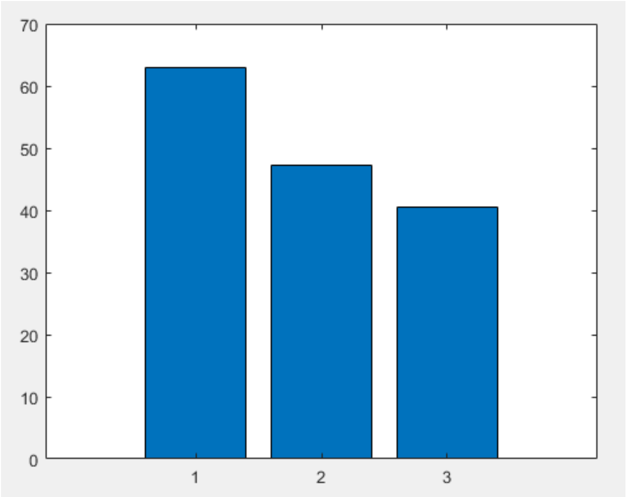
Now, Let’s visualize survival based on different features-

Survival probability by sex



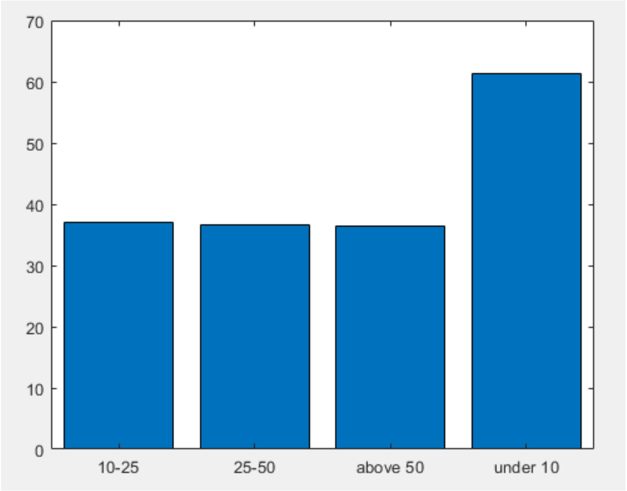
- There are 577 males and 314 female passengers.
- Females have a much higher survival probability of 74.20 % as compared to males which is 18.90 %

Survival probability by pclass



We see that the passengers with a higher ticket class are more likely to have survived.

Survival probability by age



Children have a very high survival rate

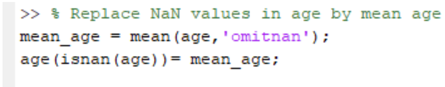
The initial statistics are in lines with our intuition that the strategy must have been to save the lives of women and children first

2. Feature Engineering

To transform the 'sex' feature, I have used grp2idx to convert the categorical values 'male' and 'female' to numerical indices '1' and '2' respectively.

The transformation of the 'age' feature has 2 steps-

Filling missing entries-

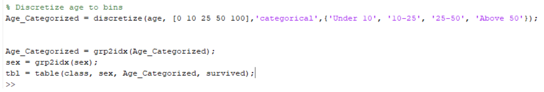


```
>> % Replace NaN values in age by mean age
mean_age = mean(age,'omitnan');
age(isnan(age))= mean_age;
```

We will impute the missing entries in the 'age' feature with the mean age of all other passengers. i.e. 29.61 years. There are some other ways in which we can handle these missing entries, discussed in section 6.

Categorizing into bins-

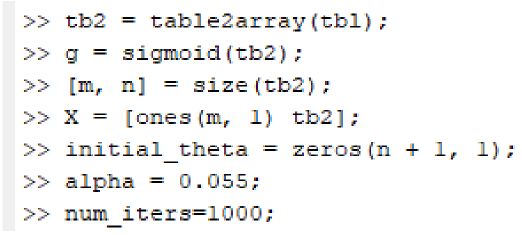
MATLAB provides a discretized method which can be used to group data into bins or categories. I have used this method for categorizing age into the following 4 age groups — 'Under 10', '10–25', '25–50' and 'Above 50'.



```
% Discretize age to bins
Age_Categorized = discretize(age, [0 10 25 50 100],'categorical',{'Under 10', '10-25', '25-50', 'Above 50'});


Age_Categorized = grp2idx(Age_Categorized);
sex = grp2idx(sex);
tbl = table(class, sex, Age_Categorized, survived);
>>
```

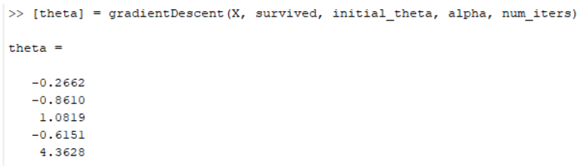
3. Model Fitting

After loading the data,we apply a logistic model on data.



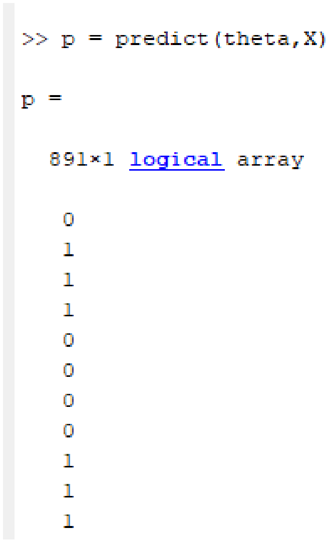
```
>> tb2 = table2array(tbl);
>> g = sigmoid(tb2);
>> [m, n] = size(tb2);
>> X = [ones(m, 1) tb2];
>> initial_theta = zeros(n + 1, 1);
>> alpha = 0.055;
>> num_iters=1000;
```

Now apply gradient descent



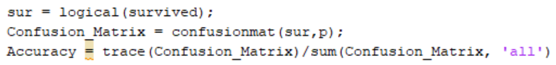
```
>> [theta] = gradientDescent(X, survived, initial_theta, alpha, num_iters)

theta =

   -0.2662
   -0.8610
    1.0819
   -0.6151
    4.3628
```

hence we find minimal θ for our hypothesis, so now we can predict our output:

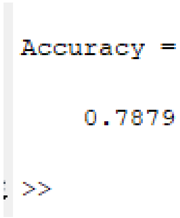


```
>> p = predict(theta,X)

p =

  891×1 logical array

   0
   1
   1
   1
   0
   0
   0
   0
   1
   1
   1
```

Now, we find the accuracy of our model using a confusion matrix.



```
sur = logical(survived);
Confusion_Matrix = confusionmat(sur,p);
Accuracy = trace(Confusion_Matrix)/sum(Confusion_Matrix, 'all')
```

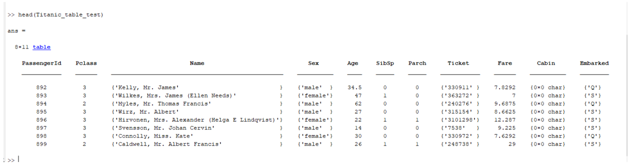


```
Accuracy =

    0.7879

>>
```

Our model has 78.79% of accuracy.

4. Generate Test Predictions

The test set comprises 418 passenger entries and 11 columns without the ‘Survived’ variable, which we will predict using the trained models.



```
>> head(Titanic_table_test)

ans =

  8×11 table

    PassengerId    Pclass                       Name                          Sex        Age    SibSp    Parch      Ticket        Fare        Cabin       Embarked
    ___________    ______    ____________________________________________    __________    ____    _____    _____    ___________    ______    __________    ________

        892          3       {'Kelly, Mr. James'                          }    {'male'  }    34.5      0        0      {'330911' }    7.8292    {0×0 char}     {'Q'}
        893          3       {'Wilkes, Mrs. James (Ellen Needs)'          }    {'female'}      47      1        0      {'363272' }         7    {0×0 char}     {'S'}
        894          2       {'Myles, Mr. Thomas Francis'                 }    {'male'  }      62      0        0      {'240276' }    9.6875    {0×0 char}     {'Q'}
        895          3       {'Wirz, Mr. Albert'                          }    {'male'  }      27      0        0      {'315154' }    8.6625    {0×0 char}     {'S'}
        896          3       {'Hirvonen, Mrs. Alexander (Helga E Lindqvist)'}    {'female'}      22      1        1      {'3101298'}    12.287    {0×0 char}     {'S'}
        897          3       {'Svensson, Mr. Johan Cervin'                }    {'male'  }      14      0        0      {'7538'   }     9.225    {0×0 char}     {'S'}
        898          3       {'Connolly, Miss. Kate'                      }    {'female'}      30      0        0      {'330972' }    7.6292    {0×0 char}     {'Q'}
        899          2       {'Caldwell, Mr. Albert Francis'              }    {'male'  }      26      1        1      {'248738' }        29    {0×0 char}     {'S'}

>>
```

We apply the same feature engineering steps on the test dataset and feed it through the models described above to generate the predictions.

CONCLUSION:

We have seen a step-by-step process in MATLAB to solve the Machine Learning task for visualizing patterns in databases, selection and engineering features, training multiple class dividers to make guesses using trained models.

Other comments-

The increase in the number of drums in the ‘age’ aspect has led to overcrowding.

Adding a new feature calculated as (fare / ticket frequency) did not provide a significant improvement in test accuracy.

ACKNOWLEDGEMENT

We are thankful to Prof. Manish Kumar for providing us the opportunity to research on such topics and have the experience of applying what we study to real-life scenarios.

We are Also thankful to our TAs for guiding us throughout the semester to carry out our research in the correct direction.

REFERENCES

1) Ekinci, E., Omurca, S., & Acun, N. (2018). A comparative study on machine learning techniques using Titanic dataset. In the 7th *international conference on advanced technologies* (pp. 411–416).

2) Kshirsagar, V., & Phalke, N. (2019). Titanic Survival Analysis using Logistic Regression. *International Research Journal of Engineering and Technology*, *6*(8), 89-91.

3) Kakde, Y., & Agrawal, S. (2018). Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. *Int J Comput Appl*, *179*(44), 32-38.

4) Balakumar, B., Raviraj, P., & Sivaranjani, K. (2019). Prediction of survivors in Titanic dataset: a comparative study using machine learning algorithms. *Sajrest Arch*, *4*(4).

5) Ng, Andreq, director. *machine learning - Andrew Ng*. 2017. *youtube.com*, https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN.

ANNOTATIONS :

Following Files can be found in the google drive Folder :-

costFunction.m

gradientDescent.m

predict.m

sigmoid.m

test predict.csv

titanic.zip

titanic1.m

Link :-

Drive Folder