**Project Report**

On

**EXPLORATORY DATA ANALYSIS OF TITANIC DATASET**

**By**

**ANUPREET SINGH 2017A7PS0955G**

**RITURAJ ROY 2017A7PS0957G**

**At**

**IDS INFOTECH, IT PARK,CHANDIGARH**

**A Practice School-I Station Of**

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**(July,2019)**

#### Acknowledgement

We would like to extend our sincere gratitude towards Mr Pratap K Aggarwal,Managing Director of IDS Infotech Pvt Ltd for giving us the opportunity of interning at his prestigious company as a part of the Birla Institute of Science and Technology, Pilani, PS programme.

We are extremely grateful for the support of our Project Mentor Anirudh Munj, without whose guidance and support we could not have achieved the progress we have in data analysis on titanic disaster data set project. He familiarize us with python libraries and methods of data analysis.

We wish to acknowledge the help from our PS Faculty Incharge, Dr. Rajeev Taliyan for guiding us through the Practice School Course.

**Abstract**

Explanatory data analysis is a major part of any machine learning project.The determination of the correlation of attributes with one another is very important in order to figure out the variables via which we need to train our model.Titanic Disaster dataset comprises of details about various passengers who boarded the titanic and whether they survived the accident or not.Data analysis on this dataset points toward the features which determined the possibility of survival of any passenger.

**Introduction**

The Titanic dataset contains different attributes that help in predicting the survival of a particular passenger. Some of the attributes that are important are:-

* Survived: ‘0’ indicating death, ‘1’ indicating survival
* Pclass (Ticket Class): ‘1’=1st ,’2’=2nd,’3’=3rd
* SibSp: No of Siblings/Spouses for a person aboard the Titanic
* Parch: No of Parents/Children for a particular person aboard the Titanic
* Ticket: Ticket Number
* Cabin: Cabin Number
* Embarked: Place of Boarding the Titanic
* Sex: ‘0’: Male, ‘1’:Female

A number of these attributes contribute to the survival of a particular person where as some are irrelevant like Ticket Number, Place of Embark etc.

Further, We use Feature Engineering to convert the descriptive attributes like Pclass etc to machine readable variables.

Now we use Python Data Analytic libraries such as Numpy, Pandas, Matplotlib and Seaborn. These Libraries a number of different features for data analysis and visualization.

Our final goal is to predict the survival of a person based on only the given data with high accuracy and precision.

**Current Progress**

Our final goal is to train a model which correctly predicts whether a passenger survives or not depending on various attributes of a particular passenger.We have already done data analysis on the dataset available for the titanic disaster which is available on Kaggle.com.After pictorial representation of various features vs the probability of survival few features came out to be more dominant in predicting the survival possibility of a passenger than the others.The sex of a passenger proved to be a major factor,females had a higher rate of survival in contrast to their male counterparts.Passengers travelling in first and second class had a better chance of surviving than those travelling in third class.Similarly passengers with age less than 16 years were given more priority when it came to saving lives.On the other hand attributes like embarked I.e. from which port the ship was boarded did not have any remarkable effect on the chances of survival of the passenger.Same goes with passenger name and whether he or she was travelling with family or not.

We used these results which we obtained from data analysis to train a model on the principles of logistic regression where we used sex as the independent variable.We kept the size of test data set as 20% of the training data set and got an accuracy of about 81%.After this we used two attributes as independent parameters namely sex and passenger class by giving them equal weights and the accuracy increased to 83.5%.Our main goal is to increase this accuracy as much as possible by using the best suitable combination of various attributes.We have used python libraries such as numpy,matplotlib and pandas for pictorial representation of data which helped us in determining these attributes.

**Future Plans**

The next step is to excel in data analysis by taking into account many other data sets.We have already started working on a data set containing housing prices in California.This data set requires us to find various attributes via which we can predict the price of houses in California.This requires application of Linear Regression on various attributes to plot a function which gives us the pricing house as result in which weights of the various dependent parameters are determined by the machine via learning in order to reduce the mean squared error and root mean squared error.We have already trained the model and are working on it in order to improve its accuracy.

**Bibliography**

1. Numpy : https://www.numpy.org/devdocs/
2. Pandas : http://pandas.pydata.org/pandas-docs/stable/
3. Jupyter Notebook
4. Seaborn : https://seaborn.pydata.org/
5. Matplotlib: https://matplotlib.org/users/index.html
6. Dataset: www.kaggle.com/c/titanic