Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset

Karman Singh
Department of CSE
Amity University, Uttar Pradesh
India
Karmanminocha@gmail.com

Renuka Nagpal
Department of CSE
Amity University, Uttar Pradesh
India
Rnagpal1@amity.edu

Rajni Sehgal
Department of CSE
Amity University, Uttar Pradesh
India
rsehgal@@amity.edu

***Abstract*—RMS Titanic was a British cruise ship said to be the largest cruise ever made in the history of world. It collided with an iceberg during its maiden journey across the pacific ocean from Southampton to New York City. With more than 2200 passengers on board, nearly half of them died after the unprecedented mishap. The infamous incident compels researchers to dig into the dataset. This research is aimed at achieving an exploratory data analysis and understand the effect or parameters key to the survival of a person had they been on the ship. The survival prediction has been done by applying various algorithms like Logistic Regression, K – nearest neighbours, Support vector machines, Decision Tree. Towards the end, accuracies of the algorithms based on features fed to them has been compared in a tabular form.**

Keywords— RMS Titanic, prediction, Logistic regression, K nearest neighbours, Support Vector Machines, Decision Tree, EDA

I. INTRODUCTION

The trend of machine learning has been quite evident with researchers and enthusiasts falling short of data to experiment on. To assuage the curiosity, the RMS Titanic dataset has been experimented on with different algorithms. Only 712 of the 2456 on board survived the shipwreck. The ship crashed in the North Atlantic Ocean. We will try to answer the following questions:

- Reasons behind people who survived and who they are.
- Did the economic class have to do with the chances of survival?
- How much does the gender of the person influence the chances of survival?
- Is age of someone a crucial parameter in determining the survival?

The article will mainly focus on the features imperative for the survival on Titanic. Understanding data is a key competence to companies and analysts. Exploratory data analysis has been done using python on Jupyter Notebook. Python's readily available libraries and Jupyter's interactive UI facilitated an efficient and comprehensive design of the project. The main idea of the article is focused around the exploration of data and predicting survival rate using various algorithms.

II. DATASET

The dataset is available publically on Kaggle.com in CSV (Comma Separated Values) format. As mentioned before the dataset has 891 rows with attributes - name of the passenger, number of siblings, number of parents or children, cabin, ticket number, fare of ticket and the place where the person has embarked from.[1] The raw dataset has metadata and incomplete or missing entries which have been filtered in preprocessing. Preprocessing includes assigning the median of available values to missing values and converting string values to numeric. For example, converting sex of the person to numeric; assigning 0 to male and 1 to female. Further, dataset has been split into test and train set to predict how efficiently the algorithm works. Before the algorithm is built for this specific model, a few data exploration graphs have been made to analyze which features could be detrimental to the model and which could help us ameliorate our result.

Table 1 gives us a brief outlook of the name of the features and what they depict. The features have been listed below

Table 1: Attributes and their description

Attributes	Description
Passenger ID	Identification Number of Passenger
Pclass	Passenger class(1,2,3)
Name	Name of the passenger
Sex	Gender of the passenger(Male,Female)
Age	Age of the passenger
SibSp	Number of sibling or spouse on the ship
Parch	Number of the children or parent on the ship
Ticket	Ticket Number
Fare	Price of the ticket
Cabin	Cabin number of the passenger
Embarked	Port of embarkation
Survived	Target Variable(value 0 for perished , 1 for survived)

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

The Figure 1 gives us an overview of majority of passengers belonging to the age group of 20-40. Moreover, Figure 2 suggests that the majority of passengers occupied the third class. This helped determine that age and class are germane as features to the model building.

Similarly, Figure 3 helps us establish the relation between the two and give us a rather concrete relationship.

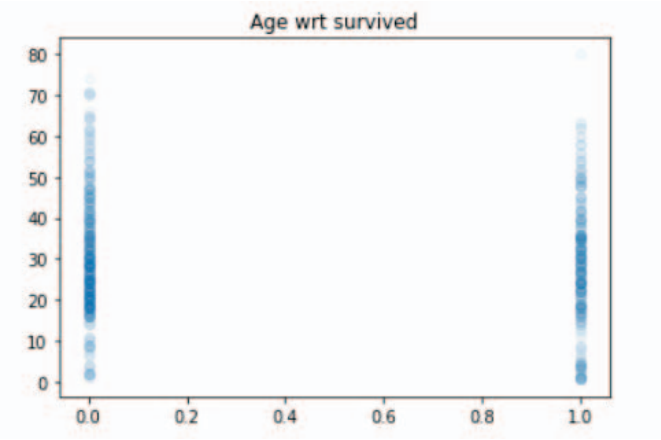

Fig. 1: Age of survived persons

Based on the given data class of occupied passenger is detremined as shown in figure 2

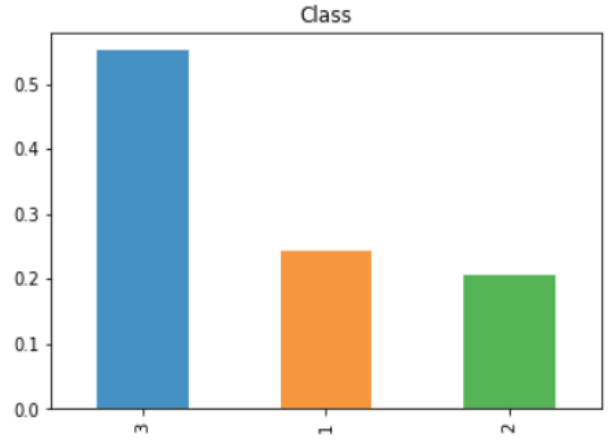

Fig. 2: Class occupied by the passenger

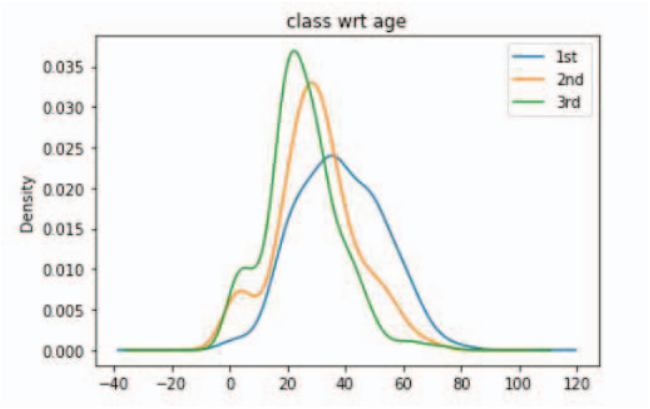

Fig. 3: class w.r.t age

Furthermore, Figure 3 has helped us extrapolate that the majority of passengers i.e between the age gap of 20-30, belonged to the third class.

Moreover, the relation between class and survival rate was made clearer in figure 4 where we can clearly infer that the people who survived (x axis = 1) were maorly of the first class.

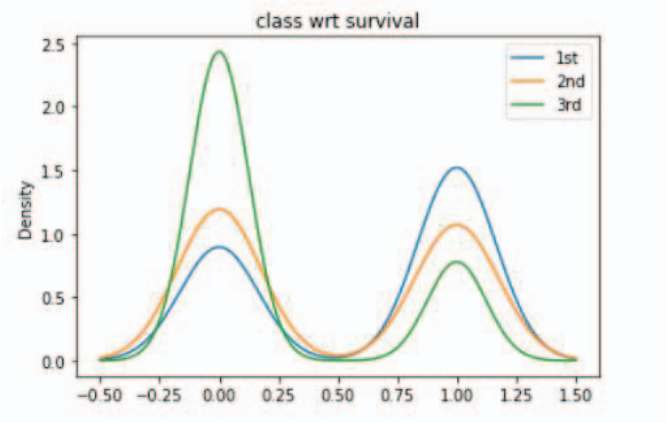

Fig. 4:Number of paggenger survived w.r.t class

A histogram for the people who have survived in the entirety of passengers on the boat has been illustrated in Figure 5. Before drawing anny other circumstantial evidence on these two features. Exploratory Data Analysis has been further done for the feature 'Sex' amongst the parameters.

The parameter sex contains string values namely 'male' and 'female' when further inspected which was rectified using a simple line of code.

df.loc[df["Sex"] == "male", "Sex"] = 0

df.loc[df["Sex"] == "female", "Sex"] = 1

and the missing values with it's median using the following

df["Age"] =df["Age"].fillna(df["Age"].dropna().median())

df["Fare"] =df["Fare"].fillna(df["Fare"].dropna().median())

The data ready to be analysed was further used to plot a relation between the survival of men versus that of women.

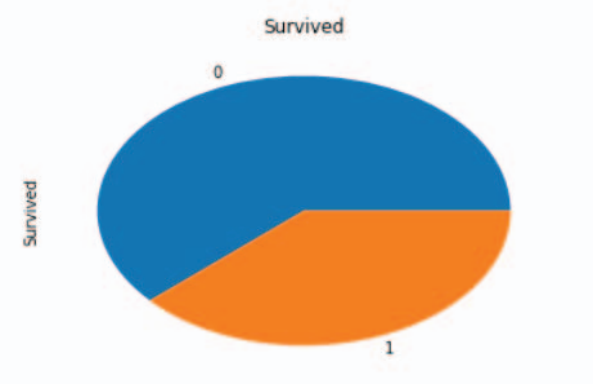

Fig.6: Total number of passenger survived

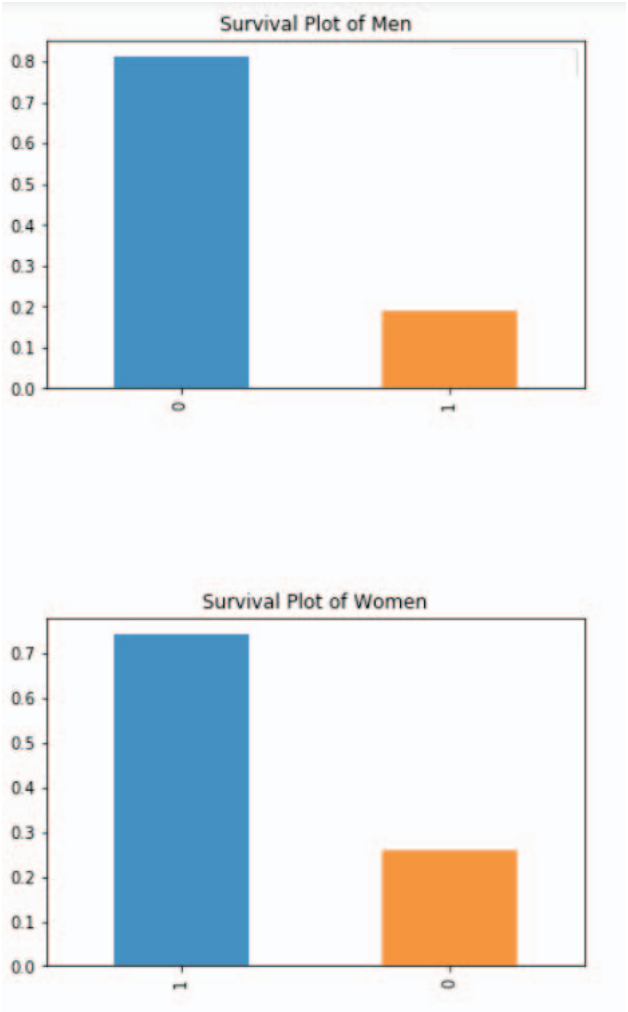

Fig. 6: Survived man Versus women

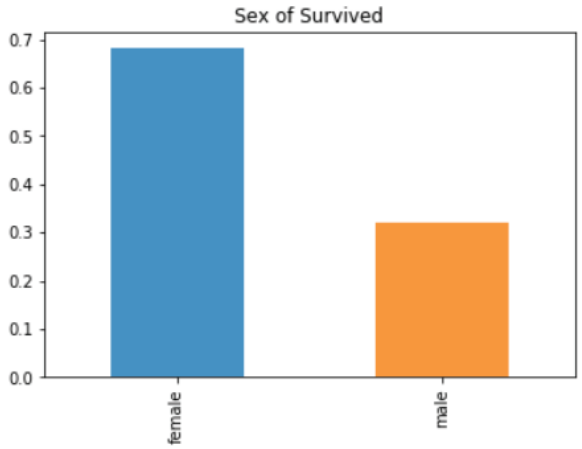

Fig. 7: Percentage of man and women survived.

Figure 6 gave us an exciting revelation into the passengers who survived where the survival rate of men was clearly in contrast with that of women. The two graphs gave us definite evidence linking sex as one of the fundamental factors behind designing a model.

Fig. 7 then gives us a clear insight that around 70% of the women survived the mishap and almost half of that with a staggering approximation of 30% made it out.

III. RELATED WORK

Data mining that we have done is used to create one of the following either a predictive or a descriptive model. The predictive model's purpose is to allow the programmer to predict a value that may be unknown or predict something that might happen in the future. A descriptive model is essential for the aforementioned model as it gives us an idea by evaluating a summary of all the data points therefore giving us an idea about which features may be conducive to the model.[2]

Supervised Learning or predictive modeling uses an atomic or group of columns to predict a variable. The target variable in this case is 'survived' which is a binary value. This value is also called a categorical variable where we will use classification techniques. Had the variable been continuous, we would have used regression techniques to find a particular value for a given unknown value. The nature of the target variable is what led to the first predictive model to be Logistic regression. [3]

No significant difference between the models applied namely Naïve Bayes, Decision Tree and Support Vector Machines suggested how the black box approach leads to answers that could be improved by better preprocessing data or selecting different features. One important inference by Lam and Tang was the importance of the feature 'sex'. This has been

confirmed in the circumstantial evidence that we have surmised using our EDA. [4]

Ju liu has helped illustrate the need for data analysis in a segmented fashion of establishing relations and finding correlations along with applying predictions to it [5].

Lin, Kunal and Zeshi give us an insight that it isn't necessary that the model performs better when fed with more features. Dimensionality reduction on the other hand plays a vital role in the process. [6]

The paper by Saraswat, Faujdar adds evidence that
Sex, Age, Sibsp, Pclass, children are the features that are correlated to the survival of the
passengers. [7]

The feature importance being the main highlight has then been evaluated by plotting a correlation heatmap using seaborn illustrated below in Figure 8.

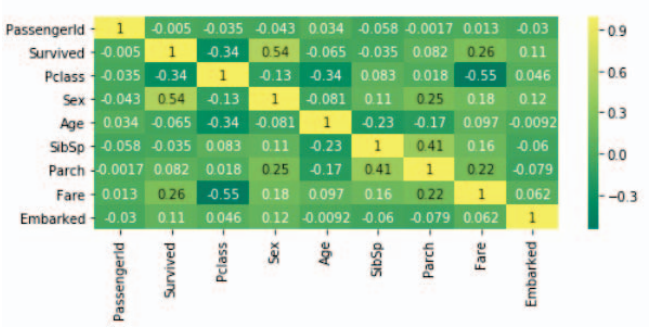

Fig. 8: correlation between the attributes

IV. ALGORITHMS USED

Prediction Models used in this article include the following :

1 Logistic Regression

2 Decision Tree

3 Decision Tree with Hypertuning

4 K – Nearest Neighbors and

5 Support Vector Machines

The prediction as mentioned before has been performed on python on Jupyter notebook. After analyzing the heatmap of correlation of survival with other features, we have assumed Age, Sex, Pclass, SibSP, Parch and embarked as essential features.

(a) LOGISTIC REGRESSION

Logistic regression is a classification based algorithm which works on discrete data instead of continuous. After learning from the data, it implies or classifies what the result might hold for us; true(1) or false(0).

The main contention behind the algorithm is to bring about a result to classify data on a very simple yet effective knowledge. If the result of the data is anywhere above 0.5 it is classified as 1 and if less than that, it is classified as 0.

The hypothesis function used here is called the sigmoid function which is given by

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1} \qquad (8)$$

Equation 1 attempts classification by obtaining an S shaped curve. The adopted method is to map all predictions greater than 0.5 as 1 and all less than 0.5 as 0. Currently, we will focus on the **binary classification problem** in which y can take on only two values, 0 and 1.

On applying the algorithm we get a score of 80.24% with 214 correct predictions. Refer Figure 9

Predicted	0	1
Real data		
0	149	27
1	27	65

Fig. 9

Logistic regression worked on categorical values of Embarked, Sex and Pclass but also with numeric values of greater range like that of Age, SibSP. Probability of all attribute values crucial for prediction of survival have been calculated.

intercept : 1.767342313071837
coefficients : [[-1.08329341 -0.02527854 0.32862081 2.56424716 -0.33539811 0.10419273]]

the positive coefficients of Sex, Embarked and Parch imply that keeping other variables constant, passenger with value of sex tending towards 1 (i.e for female), who have boarded from Queens(3) and have more number of parents or children have better chances of surviving the disaster.

(b) DECISION TREE

The research is then done by implementing Decision Tree Algorithm. Going by the literal meaning, the concept of trees is used to evaluate a decision where the results are 'branched' according to a particular condition. The basic decision comprises of very basic 'if , else' situation which are used to evaluate an end result. [9]
The aforementioned algorithm has been used in the program as an application of supervised machine learning. The forest is somehow random in the sense where the

algorithm itself evaluates particular decision's importance which is called 'feature classifier'. One of the disadvantage of Random Forest is that a huge number of trees are made in the process which makes the algorithm inefficient and sluggish. Albeit, these algorithms are fast to train, but quite slow to produce the desired results. The node at top is called the root node. The branches are called child and the nodes at the most bottom level are called leaf nodes. We have used decision tree imported from Sci-Kit Learn.

On applying the model with same parameters we get an accuracy of 93.6 % with 250 correct predictions. Refer to figure 10.

Predicted	0	1
Real data		
0	171	5
1	17	75

Fig. 10

Further the feature importance has been highlighted in Figure 11 below

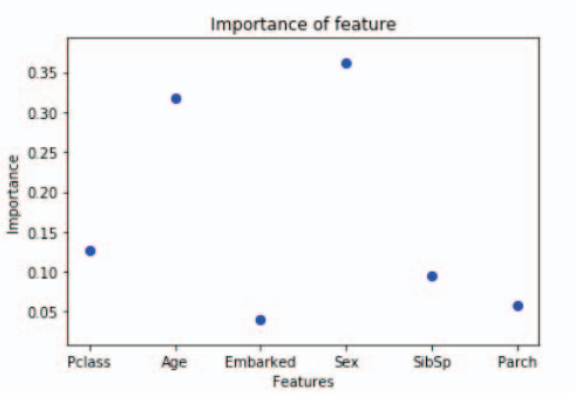

Fig.11

(c) DECISION TREE HYPERTUNING

The decision tree algorithm has further been hypertuned where random state = 1, max depth = 7 and min_sample_split = 2. That is, tree pruning upto 7th level with max 2 children nodes (binary tree) which gave us an 86.76% accuracy. The tree has been visualised as follows in Fig. 12

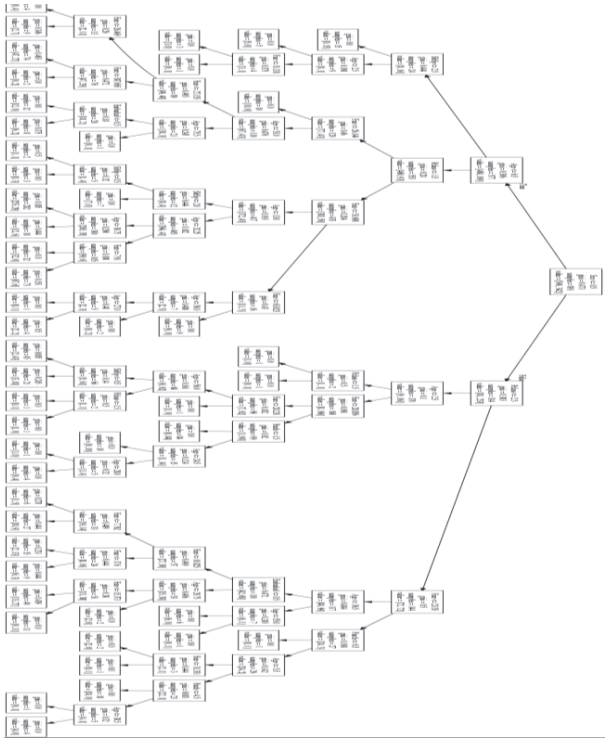

Fig. 12

It is easier to visualise a decision tree with only 7 levels and has been converted by grphviz to a dot file. Feature importance has automatically been assigned highest to 'sex' which supports the results found in the past 3 algorithms.

The feature importance now is changed, illustrated in figure 13 below

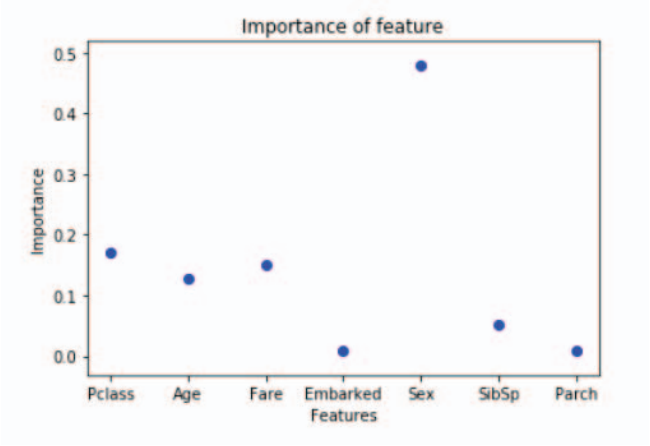

Fig. 13

(d) K NEAREST NEIGHBORS

K nearest neighbors is a classification algorithm highly useful in classifying an unknown dataset primarily on the similarity of the neighboring results. It is used for classification and regression. It is among the simplest of the algorithms and it's result depends on whether it is used for classification or regression. [10]

- When used for classification, the output is a membership of a particular class. Any object is assigned this membership by majority vote of its k nearest neighbors. The degree of 'closeness' is calculated by distance of a said 'object' to its k neighbors. If k is 1, then the object is simply assigned to its closest neighbor.
- When used for regression, the output is an average of the distances from a said object to its k closest neighbors. A commonly used metric for distance is Euclidean distance.

In the classification phase that we have used, the neighbors closest are defined by the user itself i.e the value of 'k'. The model has a training phase where it is trained using some feature vectors. The model is trained with the value of k equal to 3 which yielded 83.95% with 232 correct predictions. Refer to Figure 14.

Predicted	0	1
Real data		
0	165	11
1	25	67

Fig. 14

(e) SUPPORT VECTOR MACHINES

Support vector machines algorithm is like a knife used to separate data, works efficiently on smaller
datasets. It is a supervised machine learning algorithm which can be used both for classification

and regression. Each data item is plotted as a point in n-dimensional space with value of each feature being the value of a particular coordinate. Then,
classification is performed by finding a hyper-plane that discerns the two classes very well. The points closest to the hyper plane are called support vector which are fundamental to deciding the hyperplane
that has maximum margin. It has a unique feature to ignore the outliers, hence it is robust to it.
The algorithm yielded us an accuracy of 79.12 %

The accuracies have further been collated into one table in a decreasing order of accuracies with decision tree having the highest accuracy and support vector, the lowest (Refer to Figure 15). The metric used for comparison is the accuracy itself, percentage of correct predictions.

Score	Model
93.600000	Decision Tree
86.76	Generalised (hypertuned)
83.95	KNN
83.63	Logistic Regression
79.12	Support Vector Machines

Fig. 15

V. CONCLUSION

The comprehensive research gives us a result with decision tree having the highest score with 93.6% correct predictions and lowest false discovery rate. The research also made us aware of the features that are highly relevant to the prediction of survival of a passenger, with Sex being a feature with highest importance. The correlation between factors first evaluated using a basic formula was also justified in some cases and defied in the others.

Future work may include using other algorithms like K means, gradient boosting, adaboost, further hyper tuning the decision tree algorithm and even using advanced neural networks. Validating other techniques like assigning feature importance, introducing a new feature altogether that is, a more robust preprocessing could improve the accuracies and may yield different results for different algorithms.

REFERENCES

[1] Kaggle.com, 'Titanic:Machine Learning form Disaster',[Online]. Available:
http://www.kaggle.com/. [Accessed: 29-October-2019].

[2] Jain, Nikita, and Vishal Srivastava. "Data mining techniques: a survey paper." IJRET: International Journal of Research in Engineering and Technology 2.11 (2013): 2319-1163.

[3] Zhao, Zheng, and Huan Liu. "Spectral feature selection for supervised and unsupervised learning." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[4] Farag, Nadine, and Ghada Hassan. Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. ICSIE '18 Proceedings of the 7th International Conference on Software and Information Engineering , May 2018, dl.acm.org/citation.cfm?id=3220282.

[5] Disaster, CS229 Titanic–Machine Learning From. "Eric Lam Stanford University."

[6] Liu, Ju. “Arkham/Jack-Dies.” GitHub, 30 Aug. 2017, github.com/Arkham/jack-dies.

[7] Singh, Aakriti, Shipra Saraswat, and Neetu Faujdar. "Analyzing Titanic disaster using machine learning algorithms." 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017.

[8] Han, Jun; Morag, Claudio (1995). "The influence of the sigmoid function parameters on the speed of backpropagation learning". In Mira, José; Sandoval, Francisco (eds.). From Natural to Artificial Neural Computation. Lecture Notes in Computer Science. 930. pp. 195–201. doi:10.1007/3-540-59497-3_175. ISBN 978-3-540-59497-0.

[9] Peng, Wei, Juhua Chen, and Haiping Zhou. "An implementation of ID3-decision tree learning algorithm." From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May 13 (2009).

[10[Ekinci, E. Omurca, and N. Acun. "A Comparative Study on Machine Learning Techniques using Titanic Dataset." 7th International Conference on Advanced Technologies. 2018.

[11] Xiao, Yingchao, et al. "Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection." Knowledge-Based Systems 59 (2014): 75-84.