ZKG INTERNATIONAL

ISSN: 2366-1313

# Prediction Of Titanic Survival Using Machine Learning Techniques

[1]Mr.K.Vamshee Krishna, [2] Jampally Chandu Kumar Yadav,[3]Boddu Dayakar,[4] Palagiri Anu Priyanka, [5]paramshetty Satya Sai

[1] Assistant Professor, Dept. Of CSE, Samskruti College of Engineering & Technology, TS.

[2,3,4,5]B. Tech Student, Dept. Of CSE, Samskruti College of Engineering & Technology, TS.

**Abstract**: *It can be very important to uncover the root causes of past human suffering so that future crises can be eliminated. The incident of April 15, 1912 is an example of a human tragedy in which approximately 1,500 passengers and crew members lost their lives. Today, continuing research shows that if the right steps are taken, it is possible to reduce human harm. Nowadays there are many new and effective technologies, with the help of data analysis the truth can be established. In these studies, reviewed, Titanic survivors were studied primarily as a tool to learn techniques. In monitoring, out of all the organizations, 891 organizations were used for learning and 418 organizations were used for testing and comparisons were examined between different learning systems, which gave importance to this research.*

***Keywords***—Machine Learning, Prediction, Pattern Recognition, Statistical Analysis.

## I INTRODUCTION

Machine learning, a great sub field of artificial intelligence, evaluates important tasks that include prediction using the truth or, it can be said, specifically using mathematics to discover hidden patterns in data. It is important to note that where the traditional methods require appropriate methods, machine learning presents better decisions in the output.[1] Look at 1.

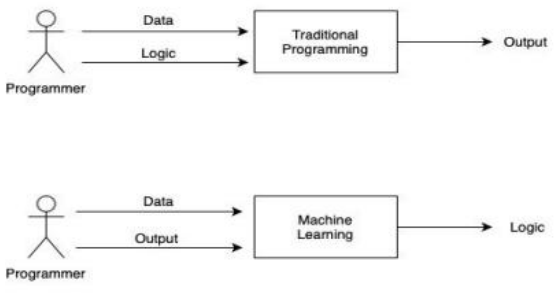


Figure 1: Difference between Traditional Programming and Machine Learning

Fig1: In the traditional system, the programmer feeds this system with the facts, logic and results of the program. But,

ISSN: 2366-1313

with the knowledge acquired by the equipment, the programmer provides the version with data and output, and the model then generates logic or software. In this research work, a

The Titanic data set is used to evaluate the survival of the Titanic based on all observational analyzes of observational learning techniques such as logistic regression, random forest, decision making, K - nearest neighbor, etc.

According to the information, the aliens are the ship's gift, so all the predictions are passed on by the lovers so that it can maximize its survival. Before starting, to train the model, it is important to first record the data in all aspects such as missing values, similar formatting, outliers, etc. [2] [3]. For more information, a clear picture of how to make a cartoon is given in parent 2.


Figure 2: Workflow of Data analysis

Figure 2 describes the process of data analysis and how the author arrived at the belief. It is best to calculate all statistical methods properly

so it is easy to know the truth of the real way.

The purpose of this paper is to explore and analyze the Titanic data using learning tools to predict passenger survival. Special articles in the value of monitoring acquisition methods such as logistic regression, random forest, stochastic gradient descent, decision tree and nearest neighbor Divide the passengers into groups, survivors or not. The purpose of the evaluation is to evaluate the overall performance of the algorithms based on specific evaluation measures, including

ISSN: 2366-1313

accuracy, F1 score, recall, and accuracy. In addition, the article discusses data preparation techniques, feature engineering, and data visualization to better understand data. The results obtained from this study can be useful for the development of prevention methods and emergency strategies of maritime transport in the future.

## II LITERATURE REVIEW

The titanic DATA-SET has been widely used in many studies to explore modeling strategies, task selection strategies, and machine learning algorithms. The DATA SET served as a benchmark for competition and training, and its evaluation confirmed the importance of predicting survival. The studies reviewed as part of this data assessment demonstrate the simplicity and importance of the titanic data set in advancing reconnaissance and predictive modeling. From the study of data mining techniques to the study of the sigma trait Parameters in synthetic neural networks

Table 1: Summarized view of Literature Review

| Ref. | Year | Method Used | Assessment |
|---|---|---|---|
| 1 | 2019 | Data Set Provider | Kaggle.com provides the Titanic dataset and platform for the Machine Learning from Disaster competition, which serves as a popular benchmark for predictive modeling. |
| 2 | 2013 | Data Mining | This paper provides a comprehensive survey of data mining techniques, including supervised and unsupervised learning, and their applications in various fields. |
| 3 | 2007 | Feature Selection | The paper proposes a spectral feature selection method for both supervised and unsupervised learning tasks, which can improve the accuracy and efficiency of predictive modeling. |
| 4 | 2018 | Predictive Modeling | This study uses the Titanic dataset to predict the survivors of the disaster and compares the performance of various machine learning algorithms. |
| 5 | 2012 | Predictive Mmodeling | It proposed a predictive modelling approach using the Titanic data set and offers a comprehensive review of related concepts and methods. |
| 6 | 2017 | Predictive Mmodeling | This GitHub repository contains a predictive model for the Jack Dies competition, which is based on the Titanic data set and serves as a similar benchmark for machine learning. |
| 7 | 2017 | Predictive Mmodeling | This study uses various machine learning algorithms to analyze the Titanic disaster and identifies the most important factors for survival prediction. |
| 8 | 1995 | Neural Networks | The paper investigates the impact of sigma function parameters on the [illegible] learning algorithm in artificial neural networks. |
| 9 | 2009 | Decision Trees | This paper presents an implementation of the ID3 decision tree learning algorithm and provides a tutorial on how to apply it to predictive modeling tasks. |
| 10 | 2018 | Predictive Mmodeling | This study compares the performance of different machine learning techniques on the Titanic data set and identifies the most accurate |

## III DESCRIPTION OF DATA AND EXPERIMENTAL SET UP

The data series consists of two columns and eleven columns. Parch ticket, P-

ISSN: 2366-1313

elegance, call, passenger ID, life expectancy, gender, age, Sib Sp, price,

cabin, and onboard are lines. Besides ability, our goal is survival. The data thus categorizes certain family ties. Blood brothers, followers, spouses and children (wives and fiances are not considered spouses). This is how the data set defines family relationships. Mother and father are parents. A child refers to a son, daughter, granddaughter, or half-brother. Because some young people are better off traveling with a babysitter, the value of their parchment has gone down.

Before any form of analysis of the data is done, the author

should smooth the data set. Some important missing points are also found in the literature and should be addressed. Missing values in the variables such as Embarked, Cabin and Age are filled with the option selected by the main age. In this state, the Cabin column is deleted and updated with the value type of the Onboard Exclusion Value column. Along the middle line, fill in the missing value in the age column.

**Research and evaluation of data**

First, we can do an analysis of the information about our problem. The data were analyzed by clinical analysis of the data to identify the factors that affect the survival rate. By correlating each behavior with survival, the data is analyzed carefully. Figure 3 shows how sex affects survival rates.

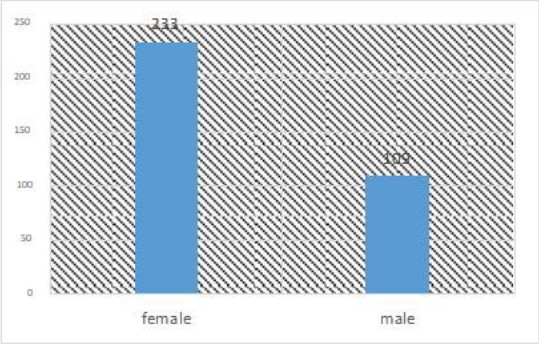


Figure 3: SUM of Survived by Sex

Figure 3 shows that women have a better survival rate than older men, as illustrated in Figure 6. The survival rates of women and men were calculated as 74.20382% and 18.89081%, respectively. Other attributes, including car price, cabin, title, family, P-perfect, auto, and survival, have a similar relationship. The name was created using the call property.' Sibbs and Parch were united. We are able to determine the importance of each character in the passenger's survival on this journey.

## IV MACHINE LEARNING MODELS

Many learning algorithms are used to be accurate and hopefully survive.

ISSN: 2366-1313

a. logistic regression is a version of the distribution in place of a version of regression and is a simple distribution model that produces very good results with linearly separated classes [4] [5]. we use the logistic regression version here because our results are both survival and non-survival. logistic regression is a reliable and useful method for binary and linear class problems.

b.random forest is the owner manager. it can be used in learning gadgets to get answers in regression and distribution mode. "a random forest is a distributed system that has multiple tree selections of specific properties of the given data and uses the average to improve the expected results of that data" [6] according to as the name suggests. as part of the main data, instead of relying on a single selected tree, random forest collects the predictions from each tree and predicts the final result using that prediction. which received the most votes.

c. stochastic gradient descent (sgd), the author uses massive data and a method that optimizes the descent for the duration of each search as quickly as possible when selecting the weight vector. gradual descent is a method of searching in continuous or infinite space where 1) the assumptions are constant and 2) the error varies depending on the parameters. the weights are initialized in sgd from the given data (titanic DATA-SET) and the code modifies the load vector with a statistic. when a calculation error is made, gradient descent gradually adjusts it to improve convergence.

d. decision tree is a method to obtain supervised knowledge that can be used to solve problems of type and regression such as titanic datasets, however, it is mainly used to solving classroom problems [7]. it has a tree structure, with internals for data set locations, branches for input options, and leaves for results. the order node and the leaf node are 2 nodes that form a selection tree. while leaf nodes are the result of such choices and do not have branches, decision nodes are used as choices and include more branches. the examination or assessment is usually based on the characteristics of the given data.

e. k-neighbors neighbors (knn) is a monitoring system that can be used for all types and returns. by calculating the distance between the test data and all the training points present in the big data, knn tries to predict the best classes for the test data. then determine where k is the maximum, just like the test [8] [9]. the knn algorithm determines which classes have the best results by counting the number of

ISSN: 2366-1313

times that control data is available for each "k" statistical class. the value in the regression condition is the average of the "k" decisions of the training program.

## V RESULT

The first step in engaging in a survey is data collection. Exploratory statistics analysis makes understanding data and relationships between capabilities less problematic. Use various graphic techniques. A reference above uses histograms and ggplot. Some conclusions were drawn and data was discovered using a case study. Based on the research data analysis process, the need to construct schools and forecast versions is recognized in engineering works. The mastery of modes by machines presupposes the good quality of passengers who survive. To make predictions in class problems, the Random Forest method is used. With a precision of zero.827261504, a return of 0.813453456, an F1 score of zero.8237261504, and a precision of 0.827261504 in line with the confusion matrix, Random Forest appears to be the true version. This shows that Random Forest has an overall overestimation of the prediction skill in this data set using the selection function. For a detailed picture of the statistical analysis, see Table 2.

Table 2. Performance Matrix Representation

| Algorithm | Accuracy | F1-Score | Recall | Precision |
|---|---|---|---|---|
| Logistic Regression | 78% | 0.78 | 0.78 | 0.79 |
| Random Forest | 82% | 0.82 | 0.81 | 0.82 |
| Stochastic Gradient Descent | 58% | 0.45 | 0.58 | 0.63 |
| Decision Tree | 79% | 0.79 | 0.79 | 0.79 |
| K-nearest neighbor | 66% | 0.64 | 0.66 | 0.67 |

It is very obvious that when using a specific design process, the accuracy of the model can also be affected. The perfect models for this type of problem are Random Forest and Decision Tree because they provide a high level of accuracy. The results of our experiment, as evidenced in Figure 4, show the performance of many machine learning algorithms used to estimate the survival of the Titanic. We evaluate the performance of the algorithms using accuracy, F1 score, recall, and precision. The Random Forest algorithm came out on top with an accuracy of 82%, an F1 index of zero.82, a return of 0.8 one, and a precision of 0.82. The logistic regression and decision tree algorithms also performed well with 78% and 79% accuracy, respectively. However, the stochastic gradient descent algorithm guarantees a perfect result with an accuracy of only 58%. The K-nearest neighbor rule performed slightly better

ISSN: 2366-1313

with sixty-six percent accuracy. These results suggest that Random Forest rules are the most suitable for predicting the survival of Titanic passengers using learning techniques.

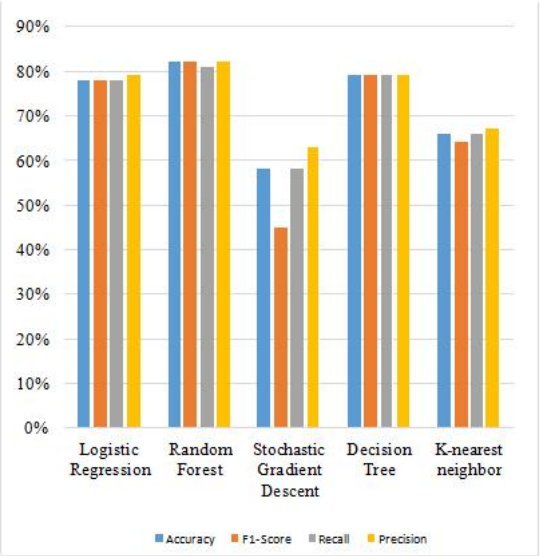


Figure 3: Performance Measures

Fig3: Display the results of the algorithms. This graph demonstrates the algorithm's performance in relation to accuracy and other factors.

## VI CONCLUSION

The design using the system brought awareness to the perceived value of passengers who survived. The random forest method is used to make predictions about target species. The accuracy of each model is determined using the confusion matrix, and the Random Forest model comes out on top with an accuracy of zero.82. This shows that the Random Woodland estimator performs in this data with very good selection ability. It is clear that when using the best modeling methods, the accuracy of the models can also change. The modes that provide the best level of accuracy for classification problems are random forest modes. Machine learning and statistical analysis were used in this work. This diagram can be used as a template for learning how to integrate EDA and core knowledge tools. With the use of more bookstore libraries, notably Vibrant in R, the concept can be improved in the future to create more graphical user interfaces. It is necessary to create interactive pages, where the same values as the chart attribute (like plot or histogram) will also change if the attribute value is changed at length. By combining our effects, we can obtain more comprehensive conclusions.

## REFERENCES

1. Kaggle.com. (n.d.). Titanic: Machine Learning for Disaster. Retrieved October 29, 2019, from http://www.kaggle.com/

2. Jain, N., & Srivastava, V. (2013). Data mining techniques: A survey paper. IJRET: International Journal of Research in Engineering and Technology, 2(11), 2319-1163.

ISSN: 2366-1313

3. Zhao, Z., & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. Proceedings of the 24th international conference on Machine learning. ACM.

4. Farag, N., & Hassan, G. (2018). Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. In ICSIE'18 Proceedings of the 7th International Conference on Software and Information Engineering (pp. 1-7). ACM.

5. E. Lam and C. Tang, CS229 Titanic–Machine Learning From Disaster, 2012.

6. Liu, J. (2017). Arkham/Jack-Dies. GitHub. Retrieved August 30, 2017, from https://github.com/Arkham/jack-dies

7. Singh, A., Saraswat, S., & Faujdar, N. (2017). Analyzing Titanic disaster using machine learning algorithms. 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE.

8. Han, J., & Morag, C. (1995). The influence of the sigmoid function parameters on the speed of back propagation learning. In From Natural to Artificial Neural Computation (pp. 195-201). Springer.

9. Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. Retrieved from http://web.arch.usyd.edu.au/wpeng/DecisionTree2.pdf

10. Ekinci, E. O., & Acun, N. (2018). A comparative study on machine learning techniques using Titanic data set. 7th International Conference on Advanced Technologies.

11. Xiao, Y., Wang, T., & Wu, J. (2014). Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection. Knowledge-Based Systems, 59, 75-84.

12. Prasadu Peddi, and Dr. Akash Saxena. "studying data mining tools and techniques for predicting student performance" International Journal Of Advance Research And Innovative Ideas In Education Volume 2 Issue 2 2016 Page 1959-1967