www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Titanic Survival Prediction: An Exploratory Data Analysis Using Machine Learning Techniques

[1] S. Srinivas, [2] B. Meghana,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

In order to prevent such catastrophes in the future, it is crucial to identify the reasons behind past human disasters. Tragically, around 1,500 people—passengers and employees alike—lost their life in an incident that occurred on April 15, 1912. Modern, ongoing research suggests that lesser human destruction may be achievable with the right statistical evaluation. Nowadays, there are a plethora of advanced tools at our disposal that allow us to do precise statistical computations. This study used machine learning techniques to examine the number of people who managed to survive the Titanic disaster. With a total of 891 entities utilized for training and 418 for the test set, this research study is significant since it compares several machine learning techniques.

Keywords-Data Analysis, Pattern Recognition, Machine Learning, and Prediction.

I. INTRODUCTION

Machine learning is an advanced branch of AI that employs statistics, or more specifically mathematics, to find patterns in data sets, which it then uses to assess crucial processes like prediction. Crucially, machine learning generates logic as output, but conventional programming necessitates logical code. [1] Review Figure 1.

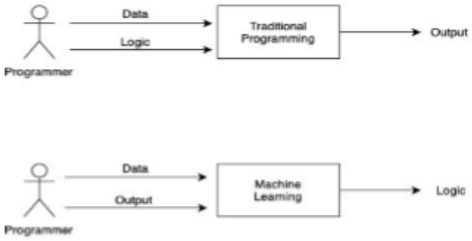

Figure1:Difference between Traditional

Programming and Machine Learning

Figure 1: The traditional method involves the programmer entering data and logic into the program, and the computer then produces an output. However, in machine learning, the model receives input from the programmer in the form of data and output, and then returns logic or code. With the use of statistical analysis and supervised machine learning methods such as Logistic Regression, Random Forest, Decision Tree, K-nearest neighbor, etc., this research endeavor examines the Titanic data set to determine the likelihood of the ship's survival. According to the records, there were many kinds of individuals on board, thus any projections must take that into account in order to maximize the likelihood of survival. It is crucial to preprocess the data set in every way before beginning to train the model [2][3]. This includes checking for outliers, missing values, and similar formats. Figure 2 is a flowchart that illustrates the research process clearly.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Figure 2: Workflow of Data analysis

The data analysis workflow and the author's conclusion-making process are illustrated in Figure 2. For a more genuine understanding of the accuracy, it is best to compute all the statistical methods in a sequential fashion. In order to forecast how many people on the Titanic made it out alive, this article will use machine learning methods to scour the ship's dataset. Logistic regression, decision trees, stochastic gradient descent, random forests, and k-nearest neighbor are some of the supervised learning techniques used to categorize passengers as either survived or did not. A variety of assessment criteria, including accuracy, F1-score, recall, and precision, will be used to compare the algorithms' performance in this study. In order to better understand the dataset, the article also delves into data preparation, feature engineering, and data visualization. Future marine transport systems may benefit from this study's findings in terms of enhanced safety measures and emergency procedures.

II. LITERATURE REVIEW

Many research have made heavy use of the Titanic dataset to investigate different approaches to predictive modeling, feature selection, and machine learning algorithms. Analysis of the dataset has shown the most important variables for survival prediction, and it has been used as a standard for contests and seminars. This literature review shows how the Titanic dataset has been useful for machine learning and predictive modeling research, and how versatile it is. Research in the field of machine learning and predictive modeling has covered a vast array of topics, from a review of data mining techniques to an examination of sigmoid function parameters in artificial 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.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

5	2012	Predictive Modeling	It proposed a predict modelling approach using t Titanic dataset and offers comprehensive review related concepts and methods.
6	2017	Predictive Modeling	This GitHub repository contai a predictive model for the Ja Dies competition, which based on the Titanic dataset a serves as a similar benchmar for machine learning.
7	2017	Predictive Modeling	This study uses vario machine learning algorithms analyze the Titanic disaster a identifies the most import factors for survival prediction.
8	1995	Neural Networks	The paper investigates t impact of sigmoid functi parameters on t backpropagation learnin algorithm in artificial neur networks.
9	2009	Decision Trees	This paper presents implementation of the ID decision tree learning algorith and provides a tutorial on ho to apply it to predict modeling tasks.
10	2018	Predictive Modeling	This study compares t performance of differe machine learning techniques the Titanic dataset a identifies the most accurate method for survival prediction
11	2014	SVM	The paper proposes t methods for selecting Gaussi kernel parameters in one-cla SVM and applies them to fa detection in industrial systems
12	2014	Predictive Modeling	This study uses vario machine learning algorithms classify Titanic passenger da and predict their chances survival in the disaster.
13	2012	Predictive Modeling	This website provides a tutori on predictive modeling usi the Titanic dataset and offers comprehensive review related concepts and methods.
14	2011	Sentiment Analysis	The paper presents a sentime analysis method for Twitt data and shows its effectiveness in predicting the polarity tweets.
15	2020	Predictive Modeling	This study uses a machi learning approach to predict t prognosis of breast cancer patients and identifies the mo important features for outcom prediction.

III. DESCRIPTION OF DATA AND EXPERIMENTAL SETUP

the columns are embarked and the cabin is. Our primary value is survival, with features. Family relationships are organized in this way in the dataset. All blood, adoptive, and step-relative relatives are considered siblings; mistresses and fiancés were not. In this way, the dataset characterizes family relationships. The parents are both biological. The term "child" may mean any of many different genders, including biological children, stepchildren, and stepbrothers. There were no parch value for the youngsters as they were accompanied just by a babysitter. The Author must clean the dataset before applying any data analytics to it. Additionally, the dataset has some missing values that must be addressed. A random pick from the existing age is used to fill in missing values in variables like Embarked, Cabin, and age. Here, the mode value from the Embarked miss value column is substituted for the cabin column. Fill in the blank in the age column using the mean of the column. Analyzing and exploring data Our first step will be to conduct a data-driven exploratory examination of the issue at hand. Exploratory data analysis is used to go through the dataset and find the variables that affect the survival rate. A comprehensive examination of the data is carried out by establishing a link between each feature and survival. The impact of sex on the survival rate is seen in figure 3.

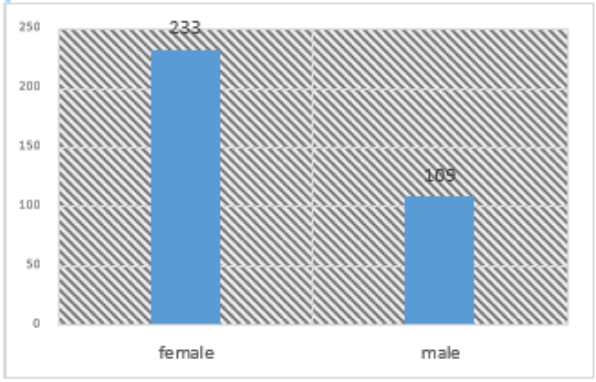

Figure 3: SUM of Survived by Sex

According to Figure 3, females are more likely to survive than men, as seen in Figure 6. The survival rates for males were determined to be 18.89081% and for females to be 74.20382%. Fare, cabin, title, family, P-class, embarked, and surviving are some of the other traits that have this relationship. With the use of the property's name, the title was formed. Parch and Sibps were brought together. Thanks to this method, we can now rank all passenger survival attributes in order of priority.

IV. MACHINE LEARNING MODELS

Validation and survival prediction are accomplished using many machine learning methods. One simple

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

classification model that works well with classes that are easily separable along linear dimensions is Logistic Regression, which is more accurately described as a classification model than a regression model [4][5]. Since our outcome is survival or non-survival, the logistic regression model is appropriate here. When dealing with linear and binary classification issues, logistic regression is the way to go since it is both simple and efficient. One kind of supervised learning is Random Forest (B). Its machine learning applications include solving regression and classification problems. "A Random Forest is a classifier that incorporates numerous decision trees on different subsets of the supplied dataset and takes the average to enhance the predicted accuracy of that dataset" [6] in accordance with its name. The random forest in the Titanic dataset gathers predictions from all the decision trees and uses the most popular ones to predict the final outcome, rather than depending on just one. C. Stochastic Gradient Descent (SGD): After selecting a random weight vector, the author uses a massive dataset to improve gradient descent throughout each search. First, hypotheses are always parameterized using gradient descent. Second, errors are differentiable according to the parameters. This allows the approach to search throughout a large or infinite hypothesis field. Using the provided dataset (the Titanic dataset), the method initializes the weights in SGD and updates the weight vector with a single data point. Upon completion of an error computation, gradient descent incrementally modifies it to improve convergence. While D. Decision Tree is most often used to handle classification issues, it is capable of handling regression difficulties as well, such as those seen in the Titanic dataset [7]. It resembles a tree, with leaves representing outcomes, branches representing decision-making procedures, and core nodes representing dataset attributes. A decision tree consists of two types of nodes: the Decision Node and the Leaf Node. Although these options lead to leaf nodes, which do not have any branches, Several branches are included in decision nodes, which are used for decision-making. Dataset attributes provide the basis of the evaluations or tests. E. K-nearest Neighbors (KNN) is a supervised learning method that encompasses both regression and classification. In order for KNN to attempt to forecast the proper class for the test data, it calculates the distance between the test data and all of the training points in the provided Titanic dataset. The next step is to choose the K locations that closely resemble the test data [8][9]. By determining the likelihood of test data belonging to each of the "K" training data classes, the KNN method finds the class with the greatest probability. In a regression scenario, the value is calculated by averaging the 'K' chosen training points.

V. MODEL PERFORMANCE EVALUATION

To ensure that the model is accurate, a "confusion matrix" is used. One way to measure the efficacy of a model or algorithm is via a confusion matrix, a kind of design table. To determine how accurate the categorization model is, one may utilize a confusion matrix [10][11]. Based on the actual results of the data, it determines the exact amount of correct or incorrect predictions. The order of the matrix is denoted by M*M, where M is the number of values. A lot of people utilize the matrix data to see how well these models worked. Classification models are also evaluated using accuracy as a parameter [12][13]. Accuracy is the proportion of times our model gets the prediction right. Here is the conventional wisdom on what constitutes accuracy: Here is an example of how the positive and negative outcomes may be used to measure the accuracy of binary classification: The acronyms TP, TN, FP, and FN stand for True Positives, False Positives, True Negatives, respectively. If our model correctly predicted 80% of the 100 photos in the X test set, we would get a score of 80/100, 0.8, or 80% accuracy. A machine learning algorithm's accuracy may be expressed as the sum of its parts, which is the sum of all positive and negative results, divided by the total number of tests. (C) Remember Finding all relevant examples in a dataset is a capability of a model. One way to think about recall is as the sum of all the true positives divided by all the true positives plus the number of false negatives[14][15]. True Positives divided by the sum of True Positives and False Negatives is the recall formula used in machine learning. D. Accuracy: The capacity of the categorization system to identify pertinent information. If you add up all of the positive results and divide them by the total number of positive results, you'll get the precision. This is the formula for machine learning precision: True Positives divided by the sum of True Positives and False Positives. A. F1-Score: We may use

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

the F1 score to combine the two criteria when trying to establish the appropriate recall and accuracy ratio. 2*(Precision * Recall) / Precision + Recall) is the formula for the F1 score in machine learning.

VI. RESULT AND CONCLUSION

Data cleansing is the first stage of any data analysis process. Datasets and their interrelationships may be better understood via exploratory data analysis. Make use of a variety of visual approaches. The one mentioned before makes use of ggplot and histograms. Exploratory data analysis allows for the drawing of certain conclusions and the discovery of some facts. The feature engineering process relies on exploratory data analysis to determine the exact parameters needed to construct the training and prediction models. The monetary value of the survivors is predicted using machine learning algorithms. In classification issues, the Random Forest approach is used to provide predictions. According to the confusion matrix, Random Forest is the most accurate model with an accuracy of 0.827261504, recall of 0.813453456, F1-score of 0.8237261504, and precision of.827261504. Based on these results, it seems that Random Forest is rather good at making predictions using these attributes on this dataset. Table 2 provides a comprehensive overview of the statistical study.

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 becomes abundantly clear that the models' accuracy might vary depending on the feature modeling technique used. Given their great accuracy, Random Forest and Decision Tree are the best models to use for categorization tasks. Figure 4 displays the experimental findings, which indicate how well different machine learning algorithms predicted whether or not the Titanic passengers would survive. We used accuracy, F1-score, recall, and precision to measure the algorithms' performance. An F1-score of 0.82, recall of 0.81, and precision of 0.82 were the highest results achieved by the Random Forest algorithm, which achieved an accuracy of 82%. The Decision Tree algorithm achieved 79% accuracy and the Logistic Regression method 78% accuracy. Stochastic Gradient Descent, on the other hand, performed poorly, achieving an accuracy of only 58%. An accuracy of 66% was the best the K-nearest neighbor algorithm could get. Based on these findings, it seems that the Random Forest method is the best option for making machine learning predictions about the fate of the Titanic passengers.

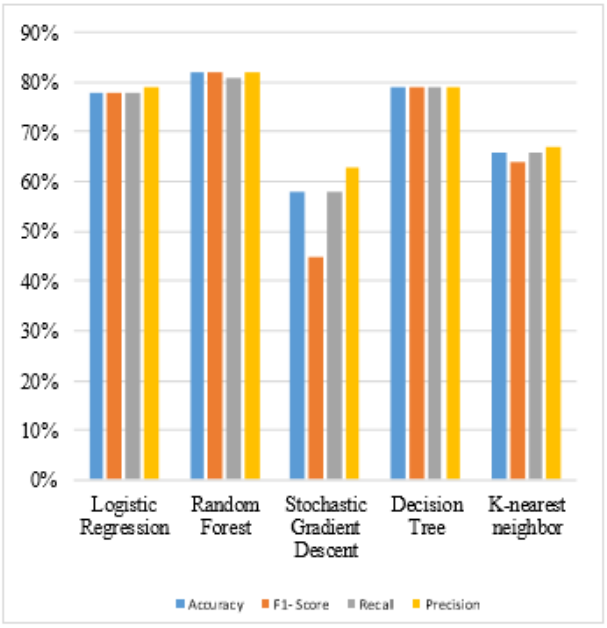

Figure 4: Performance Measures

Show the outcomes of the algorithms in Figure 4. You can see how well the algorithm did in terms of accuracy and other metrics on this graph.

VII. CONCLUSION AND FUTURE SCOPE

Machine learning models estimate how much the passengers who made it out alive are worth. In a classification challenge, the random forest approach is used to produce predictions. Using the confusion matrix to compare the models' accuracy, we find that the Random Forest model is the most effective, with a score of 0.82. What this means is that the Random forest, when trained with these characteristics, has very good prediction power on this dataset. It becomes abundantly clear that the models' accuracy might vary depending on the feature modeling technique used. Random forest models have the highest accuracy for classification challenges. Data analytics and machine learning are being used in this endeavor. Implementing EDA with machine learning from the ground up may be modeled after the work done in this project. In the future, the idea may be developed to construct more complicated graphical user interfaces using more contemporary frameworks, like shiny in R. You may make an interactive website where changing the value of an attribute on the scale also changes the values that correspond to the graph of that attribute (like a ggplot or histogram). If we combine our findings, we can also draw far more exact 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.

[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 dataset. 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] Cicoria, S., Sherlock, J., Muniswamaiah, M., & Clarke, L. (2014). Classification of Titanic Passenger Data and Chances of Surviving the Disaster. Proceedings of Student-Faculty Research Day CSIS, 1-6.

[13] Lam, E., & Tang, C. (2012). Titanic Machine Learning From Disaster. Lampang-Titanic Machine Learning From Disaster.

[14] Xie Agarwal, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. Proceedings of the ACL 2011 Workshop on Languages in Social Media.

[15] Andjelkovic Cirkovic, B. R. (2020). Machine learning approach for breast cancer prognosis prediction. Computational Modeling in Bioengineering and Bioinformatics.