JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TITANIC SURVIVAL PREDICTION

Kavya N C, Mr. Srinivasulu M

University B.D.T College of Engineering, Davanagere, Karnataka, India

Abstract

The sinking of the RMS Titanic is without a doubt one of the most infamous and horrific tragedies in history. Out of 2224 passengers and crew, the Titanic tragically sank on her maiden voyage and in the early hours of April 15, 1912, after colliding with an iceberg, killing almost 1502 of them. This made the ship one of the deadliest commercial ships in history up to that point. The laws governing ship safety have been toughened as a result of the horrific disaster that shocked the globe and caused it to feel profoundly sorry and scared. Thomas Andrews, the architect of the building, died in the disaster. It was a grim realisation following the sinking of the Titanic that certain persons had a higher chance of surviving than others. Child and mother priority had been given top priority. The Titanic was a prime example of the Titanic's time, which was the beginning of the 20th century and marked a severe division in socioeconomic groups. Exploratory data analytics (EDA) is utilised at the beginning and used to find facts that have been hidden or previously unknown in the existing data collection. After choosing several machine learning models, it is required to draw a conclusion about the study of which categories of people have a higher likelihood of surviving. After that, comparisons of the outcomes obtained by using different machine learning models were done based on precision.

Keywords: Feature engineering, Machine learning, Data Science, Exploratory Analysis, Model Evaluation, Exploratory Data Analytics, Data mining, ggplot, Logistic Regression, Random Forest,Feature Engineering, Support Vector Machine, Confusion Matrix.

1. Introduction

Machine learning [1] is the umbrella term for any method that can be implemented on a computer and applied to a set of data to search for patterns. In essence, this refers to all algorithms used in data science, regardless of whether they are supervised or unsupervised, employed for segmentation, classification, or regression. Machine learning is an essential tool in many fields, including autonomous driving, handwriting identification, language translation, speech recognition, and image categorization.

The pixels that make up an image's characteristics are used to create predictions for picture categorization, and machine learning algorithms have the ability to make these discoveries. In autonomous vehicles, information from cameras, range sensors, and GPS is combined with the pitch and volume of sound samples for voice recognition.

How many Titanic survivors there will be is predicted using machine learning techniques. A number of features, such as name, title, age, sex, and class, will be used to produce the forecasts. In order to find meaningful and useful patterns in vast amounts of data, predictive analysis uses computational approaches. The likelihood of survival is estimated using machine learning approaches based on different feature combinations. The objective is to do exploratory data analytics on the currently accessible dataset and to examine the effect of each field on passengers' survival by applying analytics between each dataset field and the "Survival" field. Numerous algorithms are examined for accuracy, and the most accurate algorithm is then recommended for predictions.

The most notorious catastrophe is known as the sinking of "The Titanic," which happened on April 15, 1912, more than a century ago. The Titanic suffered extensive damage as a result of the collision with the iceberg. On that terrible night, a wide variety of people of all ages and genders were present, but it was unfortunate that there weren't enough lifeboats available for rescue. There were many men among the dead, and the numerous women and kids on board took their place. The men taking the second-class flight were already dead. [2]

In order to forecast which people survived when the Titanic sank, machine learning techniques are used. Features such as ticket price, age, sex, and class.

By applying analytics between each dataset field and the "Survival" field, it will be possible to undertake exploratory data analytics to mine the available dataset for diverse information and determine the impact of each field on passenger survival. A machine learning method is used to make predictions about newer data sets. The correctness of the data analysis using the implemented algorithms will be examined. The most accurate algorithm is offered for predictions after being compared to other algorithms' accuracy levels. [3]

Many pieces of the Titanic were torn off during that fatal night in the disaster that occurred more than a century ago. Unfortunately, there were insufficient lifeboats to save every one of the 2224 people. Many guys who were missing in action were.

The purpose of this research article is to accurately forecast, given a collection of demographic data, who would survive the Titanic. Gender, age, ticket type, and socioeconomic status all had a role in whether a passenger would be fortunate enough to survive the Titanic or sadly perish, according to a prediction model developed using passenger data. By combining statistical methods, predictive analysis is a technique for identifying meaningful and practical trends in huge data sets. Machine learning techniques based on diverse feature combinations are used to predict survival [2].

The goal of this study is to utilise exploratory data analytics to glean interesting information from the existing data set and to assess how each field would effect the passengers' survival by applying "Survival" field analytics in between each field of the data set. Data was also examined to determine the efficacy of the used algorithms. The best-performing model is selected [3] after comparing different algorithms based on this information. The Titanic dataset analysis yielded two predictions. The first included utilising machine learning algorithms to identify the traits that the fortunate passengers shared that enabled them to escape the shipwreck, and the second involved figuring out if I would have survived had I been on the tragic ship.

1.1 Introduction to python

A general-purpose, high-level programming language, Python has gained popularity recently. It enables programmers to write code in fewer lines, something that is not achievable in other languages. Python programming is notable for its support for several programming paradigms. Python has a huge collection of comprehensive standard libraries that are expandable. Python's key characteristics include its simplicity and ease of learning, freeware and open source status, high-level programming language, platform independence, portability, dynamically typed, both procedure- and object-oriented design, interpreted, extendable, embedded nature, and sizeable library.

1.2 Introduction to Data Science

Data Science is a multidisciplinary field that employs scientific methods, practises, tools, and systems to glean knowledge from both structured and unstructured data. Big data, data mining, and data analytics are all connected to data science. It is aware of the phenomenon behind the data. It uses methods and theories that are derived from a variety of disciplines in the context of mathematics, statistics, computer science, and information science.

1.3 Introduction to Machine Learning

Automatically identifying meaningful patterns in data is a process known as machine learning. In the recent years, it has transformed into a common tool for almost any task needing information extraction from large data sets. The technology that permeates our lives nowadays includes machine learning. Search engines figure out how to provide us the best results while putting profitable adverts, anti-spam software figures out how to filter our email communications, and fraud-spotting software safeguards credit card transactions. Face recognition is possible with digital cameras, while voice recognition is possible with personal assistant apps on smartphones.

1.4 Python for Data Science

The most important data science libraries to be familiar with are as follows:

• Numpy

• Matplotlib

• Scipy

Numpy: Numpy will greatly improve our ability to manage multi-dimensional arrays. Although doing so directly might be challenging, Numpy is the foundation upon which many other libraries (indeed, virtually all of them) are built. Simply put, using Pandas, Matplotlib, Scipy, or Scikit-Learn is challenging without Numpy.

Matplotlib: The visualisation of data is crucial. Data visualisation enables us to more effectively comprehend the data, locate information that would not be seen in the raw form, and present our discoveries to others. Matplotlib is the top-rated and most well-known Python data visualisation library. Although it is not user-friendly, it often offers a variety of capabilities, such as bar charts, scatterplots, pie charts, and histograms, which are helpful for projecting multidimensional data.

Scipy: Numerous concepts that are very significant but also complicated and time-consuming are covered in mathematics. But Python has a whole scipy library that takes care of this problem for us. We will learn how to use this library in this programme, along with a few functions and illustrations of how they work.

2.Literature survey

Every machine learning algorithm operates most effectively under a certain set of conditions. By making sure your algorithm complies with the requirements, you can guarantee superior performance. There is no situation when any algorithm can be employed. As an alternative, you should think about using algorithms like Logistic Regression, Decision Trees, SVM, Random Forest, etc. in these situations. Logistic regression and decision trees are used as prediction models in this study.

Logistic Regression [4], [5], and [6] [7] is used to predict the effects of a series of variables on a binary response variable and to classify observations by modelling the probability of an event occurring depending on the values of the independent variables, which can be categorical or numerical. It also estimates the probability that an event occurs for a randomly selected observations versus the probability that the event does not occur. Biology and social sciences are where it is most frequently utilised.

The efficiency of the logistic regression [8] model can be evaluated using the AIC (Akaike Information Criteria), Null Deviance and Residual Deviance, Confusion Matrix, and McFadden R2 (also known as pseudo R2). AIC is a metric that is comparable to adjusted R2 in logistic regression. A model is penalised by the AIC, a measure of fit, depending on how many model coefficients there are. As a result, we consistently choose models with low AIC values. The Null Deviance and Residual Deviance metrics reflect the response predicted by a model with simply an intercept. Models with lower values are better. Remaining deviation is the outcome that a model forecasts after taking independent factors into account. Models with lower values are more accurate. A table comparison of Actual and Predicted values is all that the Confusion Matrix is. The model's accuracy can be assessed in this way, and overfitting can be avoided. A logistic regression is used to analyse data because there is no statistic that can be compared to R-squared. Instead, McFadden R2 is employed. However, pseudo R-squareds have been created to evaluate the goodness-of-fit of logistic models.

A hierarchical tree structure known as a decision tree [9] can be used to divide a sizable collection of records into more manageable sets of classes by applying a number of simple decision rules. A decision tree model is a set of rules for grouping a significant amount of heterogeneous individuals into more controllable, homogenous (mutually exclusive) groupings. Class characteristics may be any type of variable, including those with binary, nominal, ordinal, and quantitative values; however, class types must be of the qualitative variety (categorical or binary, or ordinal).

To put it simply, a decision tree creates a set of rules (or a series of questions) that can be used to identify a class given a set of qualities and their classes. A hierarchy of segments within segments is created by applying one rule at a time. The hierarchy is referred to as a tree, and each section is referred to as a node. The members of the generated sets are increasingly similar to one another with each subsequent division.As a result, the method for creating decision trees is known as recursive partitioning. Applications for decision trees include determining whether or not to grant a loan, classifying buyers from non-buyers, classifying tumour cells as benign or malignant, and diagnosing various diseases based on symptoms and characteristics.

3.Methodology

It is highly likely that the data we collected contains errors, missing numbers, and corrupted values because it is still in its raw form. Before making any conclusions from the data, feature engineering and data wrangling, often known as data preparation, are required. In order to make large, complicated data sets easy to access and analyse, data wrangling involves organising and cleaning them up. To increase the predictive power of learning algorithms, a method called feature engineering [10] seeks to generate more pertinent features from the raw features of the data.

The first step in our method for resolving the problem is to collect the raw data needed to do so. Then, we import the dataset into the working environment, perform data preprocessing (data wrangling and feature engineering), explore the data, and create a model that will be used to perform analysis using machine learning algorithms. The model is then assessed, and the process is repeated until the model performs satisfactorily. The findings are then compared within the algorithm, and the model that best matches the problem is selected.

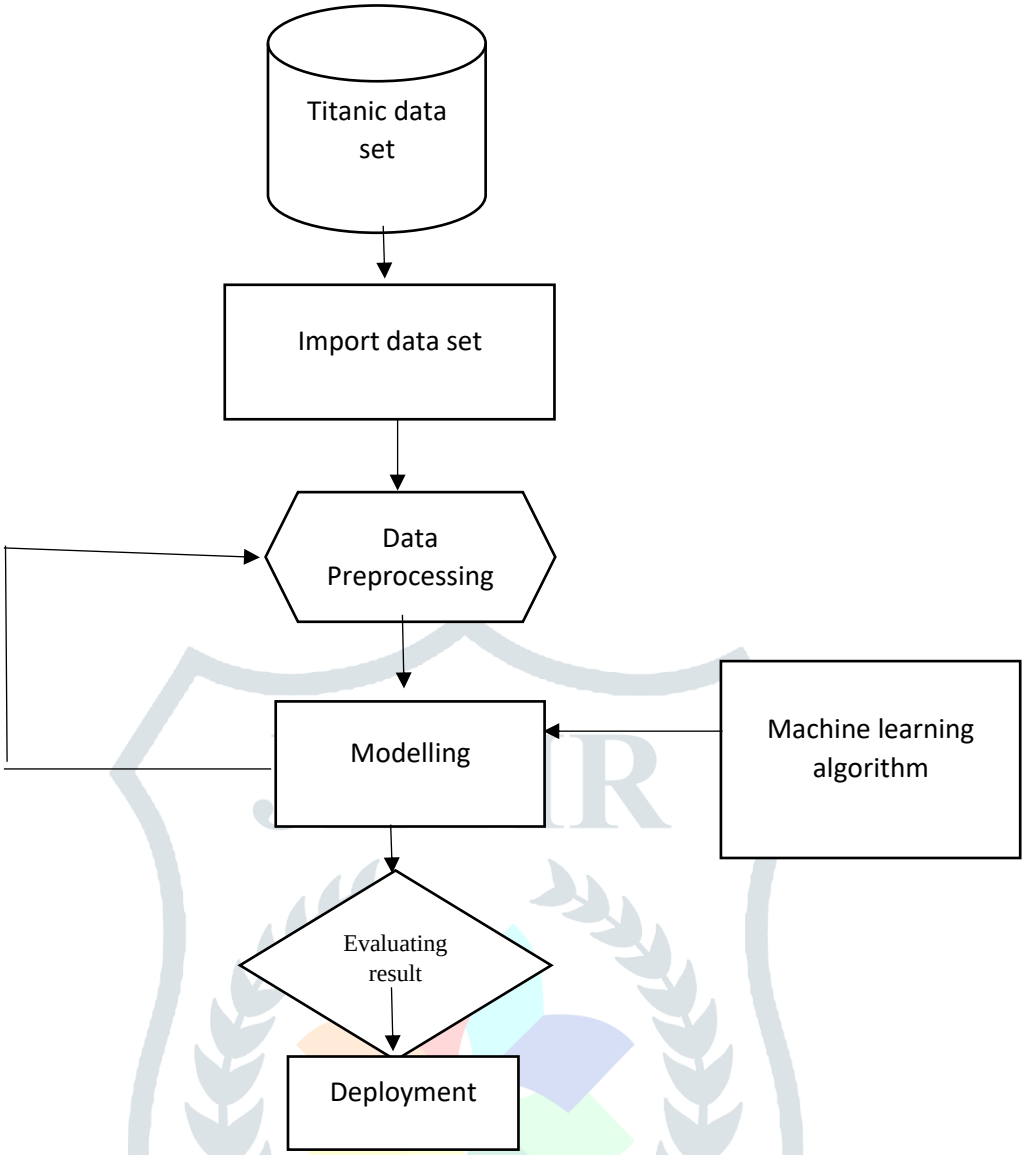

Fig. Operational flow chart

3.1 Feature Engineering

The most critical stage of data analytics is feature engineering. It has to do with creating predictions and choosing the features to employ in training. When developing a machine learning model, features that are useful in the dataset are found using domain expertise in feature engineering. In terms of modelling, it aids in comprehending the dataset. A poor prediction model may result from a poor feature selection. Choosing the appropriate characteristics is crucial for accuracy and predictive power. All the extraneous or pointless features are filtered away.

Age, sex, cabin, title, Placss, family size (parch plus sibsp columns), fare, and embarked are employed based on the exploratory analysis mentioned above. The reaction column is determined by the survival column. These characteristics were chosen because they have values that affect the survival rate. The "x" value in the bar-plots will be these characteristics. Even a smart algorithm may result in inaccurate predictions if the wrong features were chosen. Therefore, feature engineering serves as the foundation for creating an accurate predictive model.

3.2Machine Learning Models

The implementation of several machine learning algorithms validates and forecasts survival.

3.2.1 Logistic Regression:

Logistic regression is the technique that performs the best when the dependent variable is dichotomous (binary or categorical). [11] The data are described and the link between a single binary dependent variable and one or more independent nominal, ordinal, interval, or ratio-level variables are explained using logistic regression. It is employed to solve binary classification issues; some real-world examples include spam detection, which determines whether an email is spam or not, health, which establishes whether a given mass of tissue is benign or malignant, and marketing, which establishes whether a particular user will buy an insurance product or not.

3.2.2 Decision Tree:

The decision tree is an example of supervised learning algorithm. This is usually used in relation to classification-based challenges. It supports both continuous and categorical input and output variables. Each root node represents a split point on a single input variable (x) and the variable itself. The dependent factor is present in the leaf nodes (y). Consider this: Assume that the task of identifying a person's gender based on the given information requires two independent variables, i.e., input variables (x), which are height

in centimetres and weight in kilogrammes. (Hypothetical example used merely as a point of comparison.)

3.2.3 Random Forest:

The supervised classification algorithm known as random forest. The algorithm essentially creates a forest with plenty of trees. Results are more accurate the more trees there are in the forest. Both classification and regression issues can be solved with the random forest approach. To create a model, for instance, 100 observations must be sampled at random, along with 5 initial variables that are also picked at random. After several iterations of the same procedure, the ultimate conclusion is drawn in light of the data collected. Each forecast serves as a function (mean) for the final prediction.

3.2.4 Support Vector Machine:

The supervised machine learning algorithm includes the Support Vector Machine (SVM). Both classification and regression issues are resolved with this technique. Creating hyper planes in a multidimensional space to divide cases with various class labels allows for classification to be done. A dummy variable is made for categorical data variables with values of either 0 or 1. A collection of three dummy variables can therefore be used to represent a categorical dependent variable with three levels, let's say (A, B, and C):

A: {1, 0, 0}; B: {0, 1, 0}; C: {0, 0, 1}

4.Model evaluation

The model's accuracy is assessed using a "confusion matrix." A confusion matrix is a table design that makes it possible to see how well an algorithm performs and is right.

4.1 Confusion Matrix

The accuracy of the categorization model is tested using a confusion matrix. When compared to the data's actual outcome, it provides the precise number of forecasts that were right or wrong. There are N values in the matrix, which has an order of N*N. Utilizing the information in the matrix, performance of such models is frequently assessed.

- **Sensitivity:** In addition to the false negative rate, it describes the proportion of true positives that are correctly detected. True positive/(true negative + false positive) = sensitivity. The lowest and maximum values for sensitivity are "0.0" and "1.0," respectively.
- **Specificity:** It complements the false positive rate by measuring the percentage of negatives that are accurately identified. Specificity is equal to the product of true negatives and false positives. Specificity's minimum value is "0.0," while its maximum value is "1.0."
- **Positive Predictive Value**: It provides the statistical test's performance measurement. It is a ratio of true positives (events that lead to true predictions and subjects' results being true as well) and the total of true positives and false positives (event that makes false prediction and subject result is also false).
- **Negative Predicted Value:** The sum of true negatives and false negatives, as well as the ratio of true negatives (an event that produces a negative prediction and a false result), are what determine the true negatives ratio (event that makes false prediction and subject result is positive).

4.2Accuracy:

It provides a measurement of the model's or algorithm's percentage of correctly predicted outcomes. "1.0" is the best value, while "0.0" is the worst.

5.Description of data

The str () function in R is used to determine the dataset's structure from a csv file.

Table 1. Description of each attribute in our dataset

Attribute	Description	Factors
Survival	Survival of passenger	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	Male/Female
Age	Age of passengers in years	
sibsp	# of siblings / spouses aboard theTitanic	
parch	# of parents /children aboard the Titanic	

ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
Embarked	Port from wherepassenger embarked. C forCherbourg, Q forQueenstown, S for Southampton	C, Q, S

Now let us explore our dataset by knowing the influence of each attribute on survival of passenger. We will create histograms, Bar plots to achieve this.

6.Data cleaning

The data must first be cleaned before any form of data analytics can be applied to it. The dataset has some missing values that need to be addressed. Missing values for variables like Age, Cabin, and Embarked are replaced with a random sample from the current Age. [15]

We discovered that there is one traveller with passenger ID 1044 who has a missing fare in the instance of the column Fare. We first discovered the value of this passenger's Embarked and Pclass to set a relevant value for the fair column. Then the median is determined for all passengers whose embarkation and Pclass were the same as those of passenger ID 1044.

7.Exploratory data analysis

In the initial stage, we'll conduct an exploratory data analysis for our issue. In exploratory data analysis, the dataset is examined to identify the characteristics that might affect the survival rate. By establishing a connection between each attribute and survival, the data is thoroughly studied.

7.1 Age verses Survival

Figure 5 below illustrates how age will impact the survival rate. The odds of survival increase as age decreases, and vice versa.

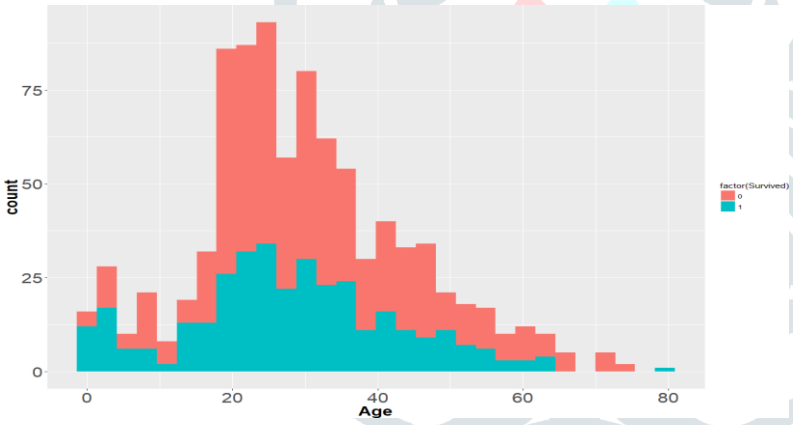

Fig 5: Age v/s Survival

7.2 Sex verses Survival

The survival rate of females is higher than that of men, as seen in Fig. 6. In our calculations, the survival rates for men and women were found to be, respectively, 18.89081% and 74.20382%.

Other characteristics including fare, cabin, title, family, Pclass, and embarked are found to be related to survival in a similar manner. From the attribute "name," the title was taken. Sibsp and Parch were united. The significance of each attribute on the passenger's survival can then be determined in this way.

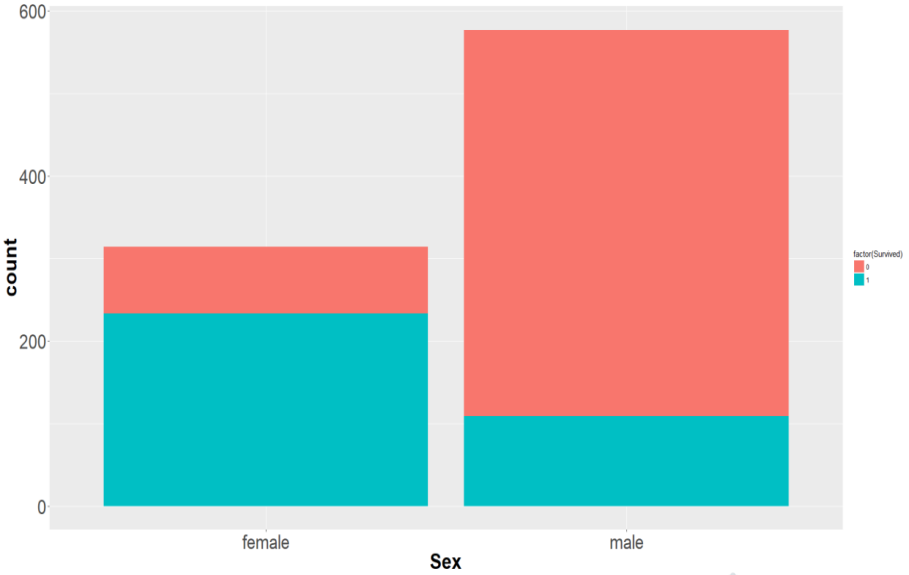

Fig 6: Sex v/s Survival

The table below lists age groups and the survival rates for each age group in the same way that we detected some additional fats.

Table 2. Age Group and Survival Rate

Age Group	Survival Rate (%)
0-10	53.24675
10-20	38.29787
20-30	37.03704
30-40	40.21739
40-50	34.82143
50-60	34.61538

8.Experimental analysis and discussion

8.1 Data set description:

The original data was used to produce the training dataset (70%) and test dataset (30%). With the help of the training set, we build our machine learning models. The training set also contains independent characteristics such as gender, class, fare, and Pclass. Our aim variable, passenger survival status, is one of them. Utilizing the test set, we should assess how well our model performs on unseen data. In the test set, the likelihood of passengers' survival is not made public. We will apply our model to predict the probability of passenger survival. Using the test set will allow you to gauge how well your model works with untested data. For each test passenger, we do not provide the actual situation. To anticipate these outcomes is your obligation.

8.2 Results:

We tested our trained algorithms against a test data set after training them, and we evaluated their performance by comparing their goodness of fit to a confusion matrix. 30% of the data and 70% of the data are used as training data sets. When compared to logistic regression (81.3%) for the given data set, the decision tree approach has a high accuracy in prediction of the survival rate (83.7%).

8.3 Enhancements and Reasoning:

It may be possible to increase the precision of the forecast for the given data set by predicting the survival rate using other machine learning algorithms[8] such as Random Forests and different types of Support Vector Machines.

Conclusion

Interesting trends were found via study of feature individual-level data. The likelihood of survival appears to be influenced by elements like socioeconomic level, social standards, and family structure. However, the information in the available data set was used to draw these conclusions. An extended hyperparameter should be modified on several machine learning models in order to further enhance the end result. By using ensemble learning, it can be further enhanced.This study work started with data exploration, which led to a check for missing data and a discovery of the crucial features. When the data pre-processing component arrived, missing values were computed and turned into numeric characteristics. In the future, some extra features were created. Cross validation was utilised to train and apply the chosen Random Forest model to 8 different machine learning models simultaneously. Examining and evaluating the computer model included looking at its confusion matrix, f-score, precision, and recall. Any data analysis process must start with and focus on purifying the data. The dataset's characteristics and the relationship between them can be determined via exploratory data analytics. EDA is applied to analyse the relationship between the features of the dataset. To accomplish this, many graphical techniques are applied. The truth is discovered by using EDA to draw some inferences. Women are more likely to survive than men, with a survival rate of around 74% compared to a male survival rate of about 12%.It is possible to confirm this by extracting titles from the name column. About 16% of Mr.'s patients survive, while 79% of Mrs.'s patients do. In order to determine a given passenger's family size, we combined the parch and sibsp columns. We found that the survival rate rises when the family size is between 0 and 3. The survival rate, on the other hand, tends to decline as family size increases above 3. Utilizing the exploratory data analytics method, feature engineering identifies the precise parameters that must be employed while designing the prediction and training model. Machine learning methods assess the values of the passengers who made it out alive. The technique of logistic regression is typically utilised to create predictions in classification problems.

References

1.Michalski R S,et al.Machine Learning:Challenges of the eighties.Machine Learning,1986,99-102.

2.Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

3.Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-Titanic Machine Learning From Disaster, 2012.

4.Bircan H., Logistic Regression Analysis: Practice in Medical Data, Kocaeli University Social Sciences Institute Journal, 2004 / 2: 185-208

5.J C. Bezdek Introduction of statistical model 1973.

6.Vapnik V.N. The Nature of Statistical Learning Theory[M]. New YorkSpringer-Verlag.1995

7.V. Vapnik, “Statistical learning theory,” Wiley, New York, 1998.

8.Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara

9.M Jamel Selim S Z The construction of decision tree vol. 61 pp. 177-188 1994.

10.https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

11.Unwin A, Hofmann H (1999). \GUI and Command-line{ Conict or Synergy?" In K Berk,M Pourahmadi (eds.), Computing Science and Statistics.

12.Galit Shmueli and Otto R. Koppius MIS Quarterly, Predictive Analytics in Information System Research, , Vol. 35, No. 3(September 2011), pp. 553-572.

13.Michalski R S,et al.Machine Learning:Challenges of the eighties.Machine Learning,1986,99-102