Home Assignment Report on

Building a Predictive Logistic Regression Model to Determine Passenger Survival Based on Given Parameters.

Submitted in partial fulfillment of the requirements
of the Subject/Course

Data Science

By

Sr. No.	Class & Div.	Name of Student	G.R. No.	Roll No.
1	CS-D	Balivada Priyanka	12220002	75
2	CS-D	Sontakke Vedika	12220206	82
3	CS-D	Tambe Ajinkya	12220014	85
4	CS-D	Thakare Prathamesh	12220016	86

Under the Guidance of
Prof.

Kiran Bhojraj Ingale

(Department of Multidisciplinary Engineering)

Bansilal Ramnath Agarwal Charitable Trust's

Vishwakarma Institute of Technology, Pune – 37

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Academic Year: 2023 – 2024 (Sem-IV)

Introduction

The problem we are solving is a classic example of binary classification, where the task is to predict the survival status (i.e., whether a passenger survived or not) based on a set of input variables or features (i.e., Pclass, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked).

To solve this problem, we are using logistic regression, which is a popular statistical method for binary classification. The logistic regression model estimates the probability of a binary outcome (i.e., survival or not) based on the input features. In other words, it calculates the likelihood of a passenger surviving based on their characteristics such as age, gender, ticket class, etc. The logistic regression model uses a logistic function or Sigmoid function to convert the output into a probability value between 0 and 1. If the predicted probability is greater than or equal to a threshold value (usually 0.5), the model classifies the passenger as a survivor (1); otherwise, it classifies the passenger as a non-survivor (0).

The logistic regression model is trained on the given train dataset, which contains both input features and target variable (Survived). The training data can be used to learn the coefficients or weights of the logistic regression model, which determine the importance of each input feature in predicting the outcome. The Titanic dataset presents interesting challenges for building a predictive model. The dataset contains missing values and features that are not useful for predicting survival, such as passenger ID and name. Therefore, data preprocessing and feature engineering techniques may be necessary to prepare the data for modeling.

Once the model is trained, it can be used to predict the survival status of a new passenger by inputting their characteristics into the model. The model will output a probability value, which can be compared to the threshold value to make a binary prediction of survival or not. Overall, the logistic regression model is a powerful tool for predicting the survival status of passengers based on their characteristics, and it can be applied to other similar classification problems as well.

Literature

A. Singh, S. Saraswat and N. Faujdar performed predictive analysis on a titanic dataset using four algorithms namely Navie Bayes, Logistic Regression, Decision Tree, Random Forest. Among all the algorithms Logistic Regression outperformed with the accuracy of 94.26 %. They also performed feature engineering (creating new features from the existing data that may be more informative for the model) in which they introduced some extra columns such as Mother, Children, Family and Respectable. These new columns show that these people are influential and respectable which might increase their survival rates [1].

The Paper "A Comparative Study on Machine Learning Techniques (2018)" using the Titanic Dataset proposed models for predicting whether a person survived the Titanic disaster or not. First, a detailed data analysis is conducted to investigate features that have correlation or are non-informative. And as preprocessing some new features are added to the dataset such as family_size and some of them are excluded such as name, ticket and cabin. Secondly, in classification step fourteen different machine learning algorithms are used for classifying the dataset formed in preprocessing steps like Logistic Regression, k-Nearest Neighbors, and Random Forest to predict survival . As a conclusion, logistic regression gives the accuracy of 80.2% and Voting (GB, ANN, kNN) gives the highest accuracy of 86.9%. This paper presents a comparative study on machine learning techniques to analyze Titanic dataset to learn what features affect the classification results and which techniques are robust. [2]

In Reference [3] Singh, Karman, Renuka Nagpal, and Rajni Sehgal mainly focused on data visualization of the dataset. Various plots such as bar graph, pie chart, correlation matrices were plotted to analyze the features. Authors also showed the importance of all the features using the graph which shows that Sex has the highest importance of all. Algorithms such as Decision Tree, KNN, Logistic Regression and SVM were used for classification purposes in which Decision Tree secured the highest accuracy of 93.60% and Logistic Regression with the accuracy of 83.63%.

Kakde Yogesh and Shefali Agrawal, along with exploratory data analysis, model training created a GUI using shiny library in R. Multiple stats were visualized in the GUI environment such as age v/s survival, sex v/s survival, cabin v/s survival and many more. They also derived the survival rates depending on the age group, passenger class. ML models like Logistic Regression, Decision Tree, Random Forest and SVM were used. Measures such as sensitivity, specificity, positive predictive value and negative predictive value were taken under care to evaluate the performance of the models. With the help of these logistic regression was proven to be the best among all with the accuracy of 83.72% [4].

The research paper "Prediction of Titanic Data Analysis Using Logistic Regression Compared with Naive Bayes for Better Accuracy" aims to predict the survival of passengers on the Titanic using Logistic Regression (LR) and Naive Bayes (NB) machine learning algorithms. The study used a dataset for Novel Logistic Regression and Naive Bayes, and found that the accuracy of Logistic Regression was 92.94% while that of Naive Bayes was 88.95%. The results showed that the accuracy of Logistic Regression in predicting the survival of both male and female passengers was better than Naive Bayes. Overall, the study suggests that the Logistic Regression algorithm provides better information about the Titanic data analysis than Naive Bayes algorithm [5].

● Dataset description

The dataset we are using is the famous "Titanic: Machine Learning from Disaster" dataset. It contains data on passengers who were aboard the Titanic during its maiden voyage which sank after colliding with an iceberg in 1912. The dataset contains the following features:

1. PassengerId: a unique identifier for each passenger
2. Survived: indicates whether a passenger survived or not (0 = No, 1 = Yes)
3. Pclass: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: passenger's name
5. Sex: passenger's gender (Male or Female)
6. Age: passenger's age
7. SibSp: the number of siblings/spouses aboard the Titanic
8. Parch: the number of parents/children aboard the Titanic
9. Ticket: ticket number
10. Fare: passenger fare (price)
11. Cabin: cabin number
12. Embarked: port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

The goal of the problem is to build a logistic regression model using all the variables (except PassengerId and Name) to predict whether a passenger will survive or not based on the given parameters.

- ## Methodology

Logistic regression is a statistical modeling technique used for binary classification tasks, where the outcome variable is binary, that is, it has only two possible values, such as yes or no, true or false, or 0 or 1. The goal of logistic regression is to model the probability of a particular binary outcome given a set of input variables.

The logistic regression process typically involves five steps: data preparation, model specification, model training, model evaluation, and model deployment.

1. Data preparation: The first step in logistic regression is to prepare the data for modeling. This involves cleaning the data, dealing with missing values and outliers, and selecting relevant features.
2. Model specification: The logistic regression model is specified by defining the relationship between the input variables and the binary output variable. The logistic function is used to transform the linear combination of input variables into a probability value.
3. Model training: The logistic regression model is trained using a dataset with known binary outcomes. The model is optimized to minimize the difference between the predicted probabilities and the true outcomes.
4. Model evaluation: The trained logistic regression model is evaluated on a separate dataset to assess its performance in predicting binary outcomes. Common evaluation metrics for binary classification tasks include accuracy, precision, recall, and F1 score.
5. Model deployment: The final step is to deploy the trained logistic regression model to make predictions on new data.

- ## Algorithm

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability of an instance belonging to a given class. It is used for classification algorithms; its name is logistic regression. It's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Logistic Regression Equation: The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

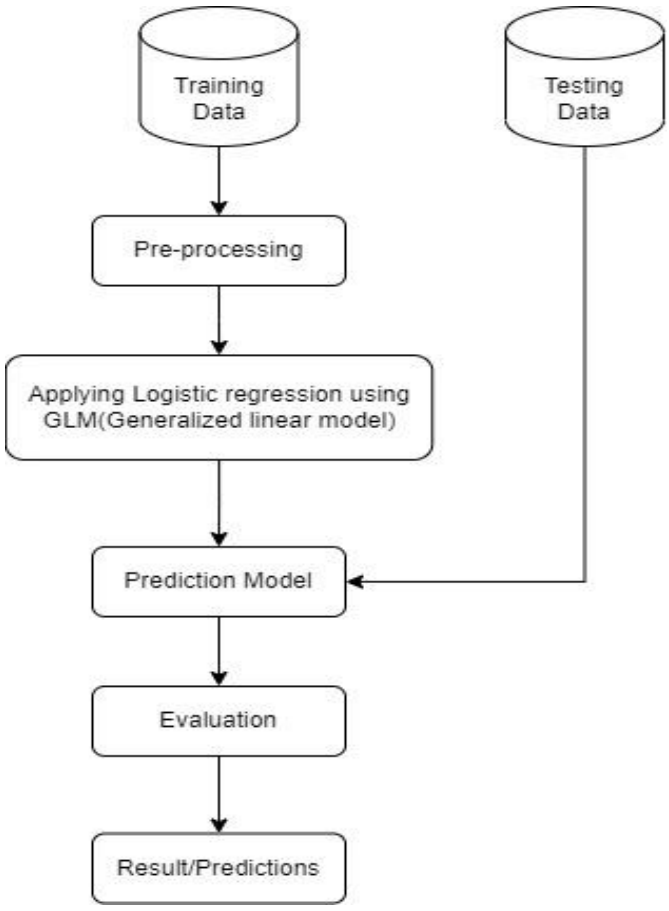
$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Type of Logistic Regression:
On the basis of the categories, Logistic Regression can be classified into three types:

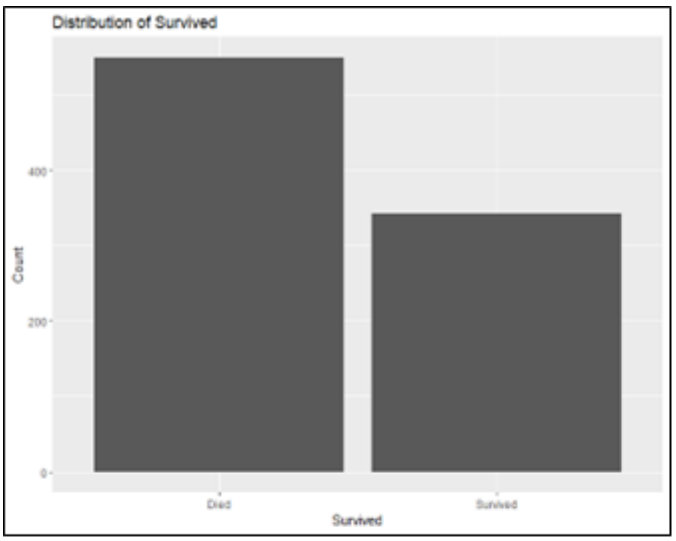
- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

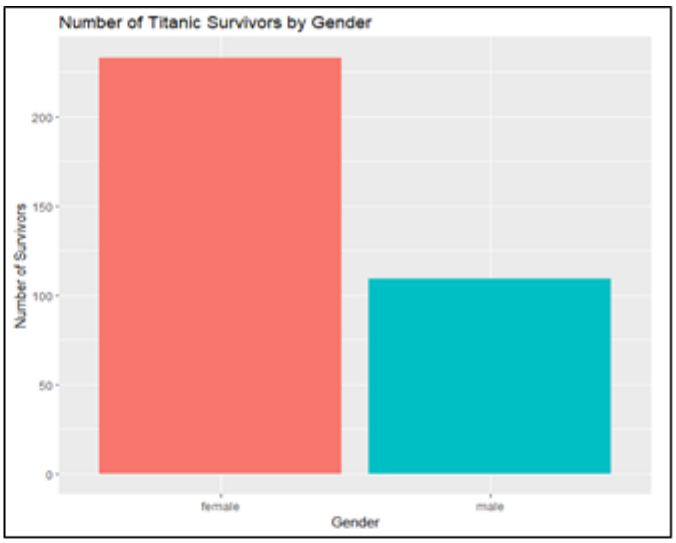
Flowchart

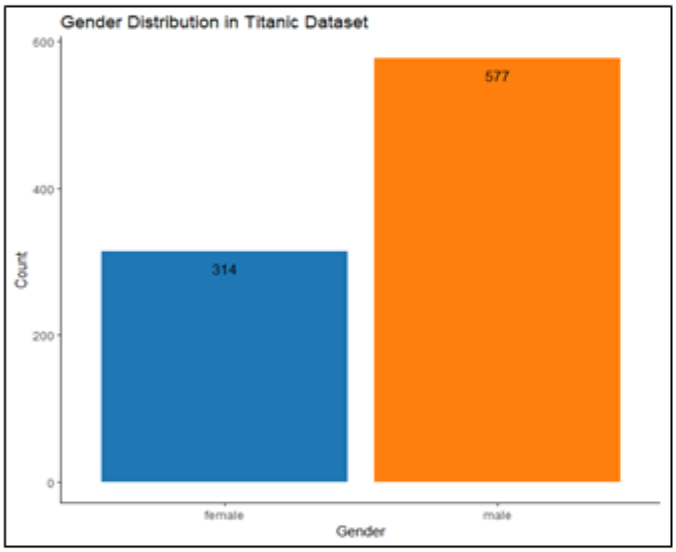


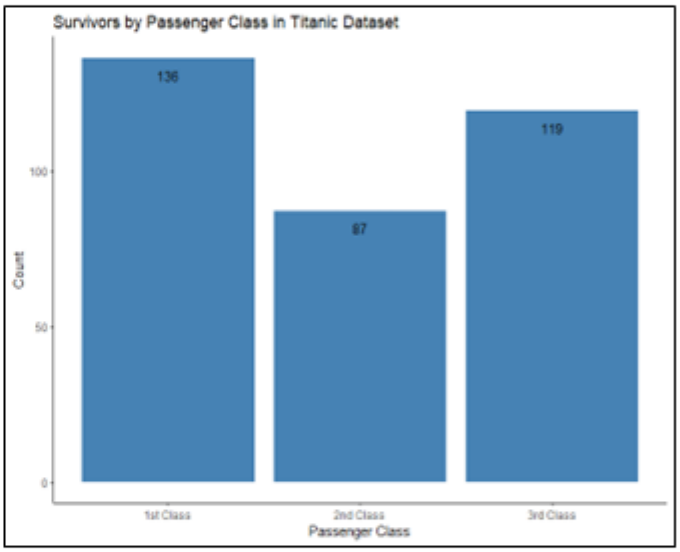
- **Results and its Interpretation**

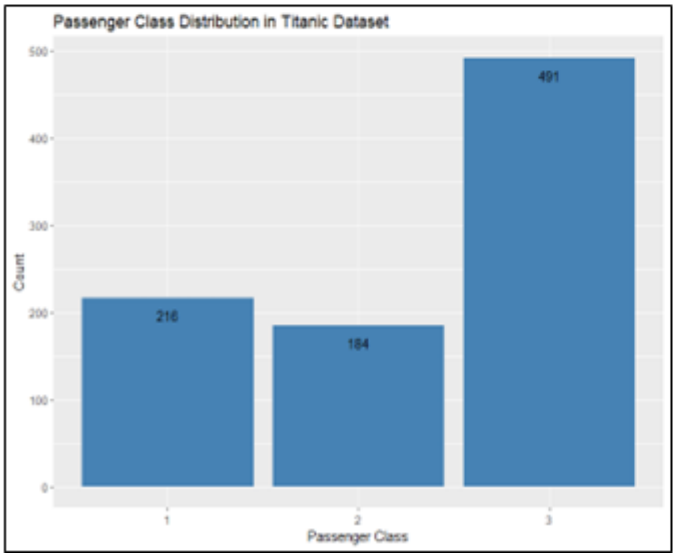
  Following are the plot we derived from the given dataset to have brief idea about how the attributes or features are related to each other

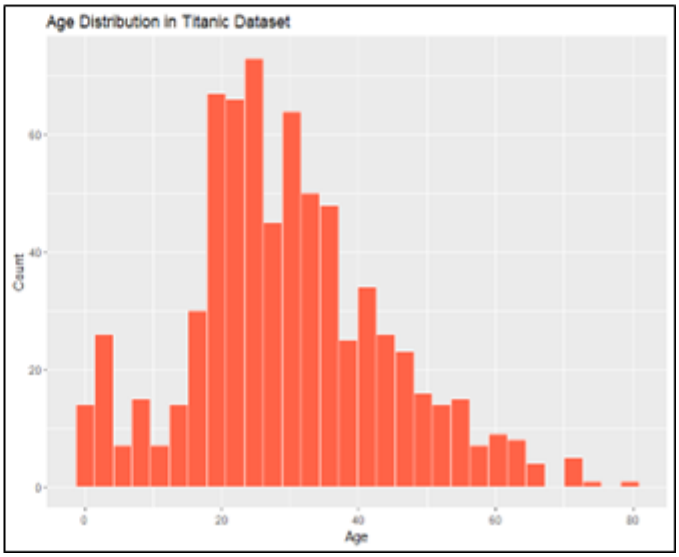




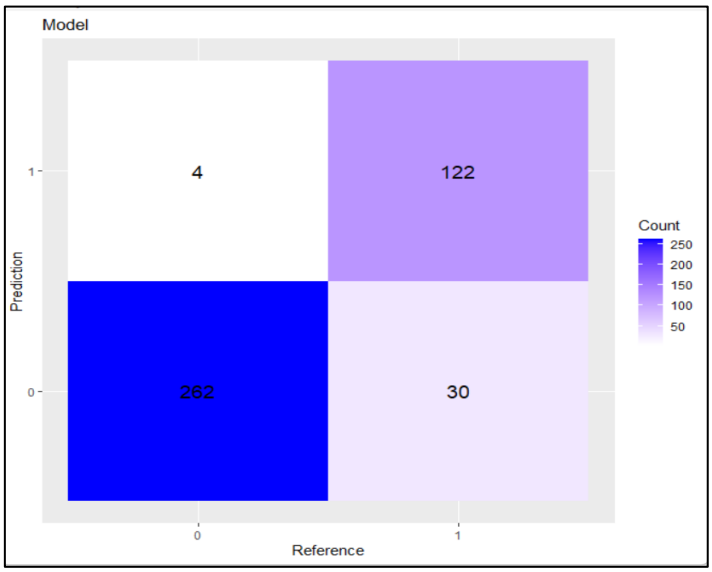








The confusion matrix for our model evaluation results is as follows:



From the confusion matrix, we can calculate various evaluation metrics of our model.

1. The accuracy of the model is the proportion of correct predictions over the total number of predictions, and can be calculated using the formula:

   Accuracy = (TP + TN) / (TP + TN + FP + FN), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

   For the given confusion matrix, the accuracy is:

   Accuracy = (122 + 262) / (122 + 262 + 30 + 4) = 0.918

2. The precision of the model is the proportion of true positives over the total number of positive predictions, and can be calculated using the formula:

   Precision = TP / (TP + FP)

   For the given confusion matrix, the precision is:

   Precision = 122 / (122 + 30) = 0.802

3. The recall of the model is the proportion of true positives over the total number of actual positives, and can be calculated using the formula:

   Recall = TP / (TP + FN)

   For the given confusion matrix, the recall is:

   Recall = 122 / (122 + 4) = 0.968

4. The F1 score of the model is the harmonic mean of precision and recall, and can be calculated using the formula:

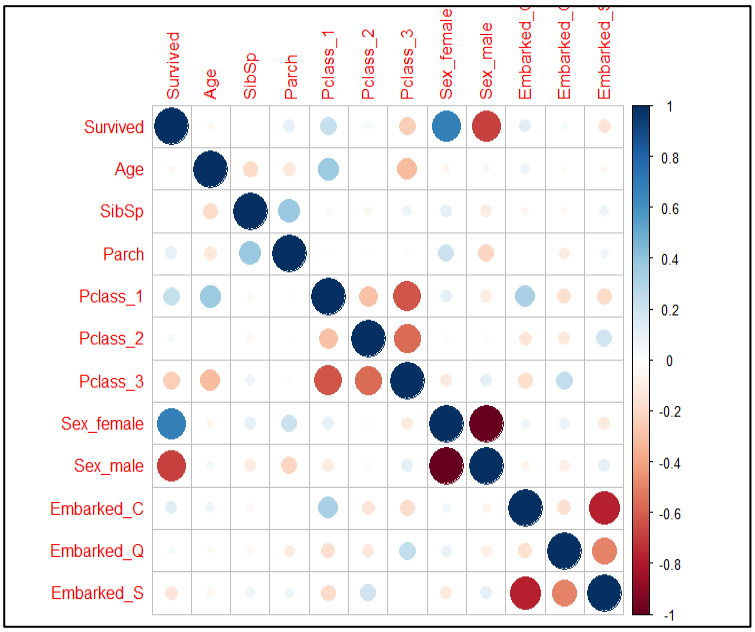
   f1_score = 2 * Precision * Recall / (Precision + Recall)

   For the given confusion matrix, the F1 score is:

   f1_score = 2 * 0.802 * 0.968 / (0.802 + 0.968) = 0.877

By the observation of above calculations we can say our model has achieved a high accuracy of 0.918 (i.e. 91.8%), indicating that it is able to correctly predict the outcome for the majority of the samples. Additionally, the precision of the model is high at 0.802 (i.e. 80.2%), indicating that the model is able to make accurate positive predictions. The recall of the model is also high at 0.968 (i.e. 96.8%), indicating that the model is able to correctly identify the majority of the positive samples. The F1 score of the model is also high at 0.877 (i.e. 87.7%), indicating that the model has a good balance between precision and recall.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 262  30
         1   4 122

               Accuracy : 0.9187
                 95% CI : (0.8882, 0.943)
    No Information Rate : 0.6364
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8176

 Mcnemar's Test P-Value : 1.807e-05

            Sensitivity : 0.9850
            Specificity : 0.8026
         Pos Pred Value : 0.8973
         Neg Pred Value : 0.9683
             Prevalence : 0.6364
         Detection Rate : 0.6268
   Detection Prevalence : 0.6986
      Balanced Accuracy : 0.8938

       'Positive' Class : 0
```



• Conclusion

In conclusion, building a logistic regression model to predict survival on the Titanic dataset is a classic example of binary classification using machine learning. The dataset provides an opportunity to practice data preprocessing and feature engineering techniques, as well as model training and evaluation. The resulting model can have practical applications in predicting survival in similar scenarios.

• References

[1] A. Singh, S. Saraswat and N. Faujdar, "Analyzing Titanic disaster using machine learning algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 406-411, doi: 10.1109/CCAA.2017.8229835.

[2] Ekinci, Ekin, S. İlhan Omurca, and Neytullah Acun. "A comparative study on machine learning techniques using the Titanic dataset." 7th international conference on advanced technologies. 2018.

[3] Singh, Karman, Renuka Nagpal, and Rajni Sehgal. "Exploratory data analysis and machine learning on titanic disaster dataset." *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2020.

[4] Kakde, Yogesh, and Shefali Agrawal. "Predicting survival on titanic by applying exploratory data analytics and machine learning techniques." *International Journal of Computer Applications* 179.44 (2018): 32-38.

[5] Sreenivasulu, L., and V. Chandrasekar. "PREDICTION OF TITANIC DATA ANALYSIS USING LOGISTIC REGRESSION COMPARED WITH NAIVE BAYES FOR BETTER ACCURACY."