SURVIVOR PREDICTION ON TITANIC DATA

Mini Project report submitted
in partial fulfilment of the requirement for the degree of

Bachelor of Technology

By

Nitin Bhattacharyya (180610026031)
Priyankar Maroti (180610026060)

Under the guidance
of
Prof. D.S. Pegu
&
Ankur Jyoti Sarma

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING
ASSAM ENGINEERING COLLEGE
JALUKBARI- 781013, GUWAHATI

August, 2021

ASSAM ENGINEERING COLLEGE, GUWAHATI

CERTIFICATE

This is to certify that the report entitled "SURVIVOR PREDICTION ON TITANIC DATA" submitted by Nitin Bhattacharyya (18/310) & Priyankar Maroti (18/311) of B. Tech 6th semester, Electronics & Telecommunication Engineering, is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the report has not been submitted to any other University/Institute for the award of any Degree or Diploma.

Signature of Supervisor(s)
D.S. Pegu & Ankur Jyoti Sarma
Electronics & TeleCommunication
Assam Engineering College
August, 2021

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty andintegrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)	(Signature)
Priyankar Maroti	Nitin Bhattacharyya
180610026060	180610026031

Date: _____

ACKNOWLEDGMENT

We would like to take this opportunity to thank our Head of the Department, Prof. Dinesh ShankarPegu for giving us the opportunity to do a mini project on 'Survivor Prediction on Titanic Data'. We would also like to thank Ankur Jyoti Sarma along with our guide Dinesh Shankar Pegufor guiding us in my project, for providing valuable suggestions, for his ongoing support during the project, from initial advice, and provision of contacts in the first stages through ongoing advice and encouragement, which led to the final report of this mini project.

We are also grateful to other teachers of the department for contributiong their inputs in this project

A special acknowledgement goes to our colleague who helped us in completing the project by exchanging interesting ideas and sharing their experience.

Nitin Bhatacharyya (180610026031)

Priyankar Maroti (180610026060)

ABSTRACT

The sinking of the Titanic caused the death of thousands of passengers and crew is one of the deadliest maritime disasters in history. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. The interesting observation which comes out from the sinking is that some people were more likely to survivethan others, like women, children were the one who got the priority to rescue. The objective isto first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive.

In this project, we see how we can use machine-learning techniques to predict survivors of theTitanic. With a dataset of 891 individuals containing features like sex, age, and class, we attempt to predict the survivors of a small test group. In particular, compare different machinelearning techniques like Naive Bayes, SVM, and decision tree analysis.

Of the differents models used, it was observed that the Logistic Regression has the highest accuracy rate of around 80.5% of all the models.

LIST OF FIGURES

CONTENTS

CHAPTER 1
INTRODUCTION

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of “The Titanic”. The collision with the iceberg ripped off many parts of theTitanic. Many classes of people of all ages and gender where present on that fateful night, butthe bad luck was that there were only few life boats to rescue. The dead included a large numberof men whose place was given to the many women and children on board. The men travellingin second class were dead on the vine.

Machine learning algorithms are applied to make a prediction which passengers survived at thetime of sinking of the Titanic. Features like age, sex, class will be used to make the predictions.Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms survival is predicted on different combinations of features. Machine Learning uses Python programming language. Various libraries of Python were used in this project like numpy, pandas, matplotlib, seaborns, etc. Skitlearn library is also used in this project

The goal of the project was to predict the survival of passengers based off a set of data. We used Kaggle competition "Titanic: Machine Learning from Disaster" to retrieve necessary dataand evaluate accuracy of our predictions. After getting the dataset, we did the Exploratory DataAnalysis of the dataset. We also saw different plots which gave us the overview of the dependency of survival on different features like Embarkment, Sex, Pclass, etc. Then the datahas been split into two groups, a 'training set' and a 'test set'. For the training set, we are providedwith the outcome (whether or not a passenger survived).

We used this set to build our model to generate predictions for the test set. For each passengerin the test set, we had to predict whether or not they survived the sinking. We shall get the result in the form of percentage or generally we shall get the result that if the passenger survived or not.

CHAPTER 2
LITERATURE REVIEW

Every machine learning algorithm works best under a given set of conditions. Making sure our algorithm fits the assumptions / requirements ensures superior performance. We can't use any algorithm in any condition. Instead, in such situations, we should try using algorithms such as Logistic Regression, Decision Trees, SVM, Random Forest etc. Logistic regression, decision trees, etc. are the models used in this paper for prediction.

Logistic Regression is used to model the probability of an event occurring depending on the values of the independent variables which can be categorical and numerical and to estimate the probability that an event occurs for a randomly selected observations versus the probability that the event does not occur and it is used to predict the effects of series of varibles on a binary response variable and it is used to classify observations by estimating theprobability that an observation is in a particular category. It is most commonly used in social and biological sciences. The performance of Logistic regression model can be measured using AIC (AkaikeInformation Criteria), Null Deviance and Residual Deviance, Confusion Matrix and McFadden R2 is called as pseudo R2. AIC is an analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value. Null Deviance and Residual Deviance measure indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model. Confusion Matrix is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. McFadden R2 isused to analyze data with a logistic regression, an equivalent statistic to Rsquared does not exist. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squared'shave been developed. To find the accuracy of model in confusion matrix the formula is

$$accuracy = (true\ positives + true\ negatives)/\ (true\ positives + true\ negatives + false\ positives + false\ negatives)$$

Decision tree is a hierarchical tree structure that can be used to divide up a large collection ofrecords into smaller sets of classes by applying a sequence of simple decision rules. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutually exclusive) classes. The attributes of the classes can be any type of variables from binary, nominal, ordinal, and quantitative values, while theclasses must be qualitative type (categorical or binary, or ordinal). In short, given a data

of attributes together with its classes, a decision tree produces a sequence of rules (or series of questions) that can be used to recognize the class. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segmentis called a node. With each successive division, the members of the resulting sets become more and more similar to each other. Hence, the algorithm used to construct decision tree is referred to as recursive partitioning.

CHAPTER 3
METHODOLOGY

The data we collected is still raw-data which is very likely to contains mistakes, missing valuesand corrupt values. Before drawing any conclusions from the data, we need to do some data preprocessing which involves data wrangling and feature engineering. Data wrangling is the process of cleaning and unify the messy and complex data sets for easy access and analysis. Feature engineering process attempts to create additional relevant features from existing raw features in the data and to increase the predictive power of learing algorithms.

3.1 *Feature Engineering***-**

Feature engineering is the most important part of data analytics process. It deals with, selectingthe features that are used in training and making predictions. In feature engineering the domainknowledge is used to find features in the dataset which are helpful in building machine learningmodel. It helps in understanding the dataset in terms of modeling. A bad feature selection maylead to less accurate or poor predictive model. The accuracy and the predictive power depend on the choice of correct features. It filters out all the unused or redundant features. Based on the exploratory analysis above, following features are used age, sex, Pclass, family size (parchplus sibsp columns), embarked. Survival column is chosen as response column. These featuresare selected because their values have an impact on the rate of survival. These features will bethe value of "x" in the bar-plots. If wrong features were selected then even the good algorithmmay produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

3.2 Machine Learning Models-

Various machine learning models are implemented to validate and predict survival.

3.2.1 Logistic Regression-

Logistic regression is the technique which works best when dependent variable is dichotomous(binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independentvariables is done with the help of logistic regression. It is used to solve binary classification problem, some of the reallife examples are spam detection- predicting

if an email is spam or not, health-Predicting if a given mass of tissue is benign or malignant, marketing- predicting ifa given user will buy an insurance product or not.

3.2.2Decision Tree-

Decision tree is a supervised learning algorithm. This is generally used in problems based onclassification. It is suitable for both categorical and continuous input and output variables. Eachroot node represents a single input variable (x) and a split point on that variable. The dependent variable (y) is present at leaf nodes. For example: Suppose there are two independent variables,i.e., input variables (x) which are height in centimeter and weight in kilograms and the task tofind gender of person based on the given data.

3.2.3 Support Vector Machine-

Support Vector Machine (SVM) falls in supervised machine learning algorithm. This algorithmis used to solve both classification and regression problems. The classification is performed byconstructing hyper planes in a multidimensional space that separates cases of different class labels. For categorical data variables a dummy variable is created with values as either 0 or 1.So, a categorical dependent variable consisting three levels, say (A, B, C) can be represented by a set of three dummy variables: A: {1, 0, 0}; B: {0, 1, 0}; C: {0, 0, 1}

3.2.4 Gaussian Naive Bayes-

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier. Naive Bayes Classifiers are based on the Bayes Theorem. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature.

3.3 Confusion Matrix-

The atccuracy of the model is evaluated using “confusion matrix”. A confusion matrix is a tablelayout that allows to visualize the correctness and the performance of an algorithm. A confusionmatrix is a method to verify how accurately the classification model works. It

gives the actualnumber of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N*N, here N is the number of values. Performance of suchmodels is commonly evaluated using the data in the matrix.

Some parameters can be identified using the confusion matrix like Sensitivity, Specificity, Accuracy, etc.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	a/(a+b)
	Negative	c	d	Negative Predictive Value	d/(c+d)
		Sensitivity	Specificity	**Accuracy=**	
		a/(a+c)	d/(b+d)	**(a+d)/(a+b+c+d)**	

Figure 3.1: Design of confusion matrix

Where a=True Positive (TP) b=False Positive (FP)
c=False Negative (FN)d=True Negative (TN)

3.3.1 *Precision/Sensitivity*-

It defines the percentage of actual positive which are correctly identified, and is complementaryto the false negative rate.

Sensitivity= TP/ (TP + FN).

The ideal value for sensitivity is “1.0” and minimum value is “0.0”.

3.3.2 *Specificity*- It measures the proportion of negatives which are correctly identified, and iscomplementary to the false positive rate.

Specificity= TN/ (TN + FP).

The ideal value for specificity is “1.0” and least value is “0.0”.

3.3.3 *Accuracy*- It gives the measure of percentage of correct prediction done by the model/algorithm. The best value is “1.0” and the worst value is “0.0

Accuracy=(TP+TN) / (TP+FP+FN+TN)

CHAPTER 4
RESULT ANALYSIS

We started the project firstly by collecting the dataset of the passengers from kaggle. As the data we collected was raw, we first cleaned the data and did the exploratory data analysis. Basically, we examined that all the values in our dataset are filled and the data we have should be well designed which can be suitable for fitting in the model. We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the survival rate. The data is deeply analysed by finding a relationship between each attribute and survival. Then we tested our model by providing test data or data for any individual and we were able to know that if the person will survive or not.

4.1 Correlation Heatmap-

A **Heatmap** contains values representing various shades of the same colour for each value to be plotted. Usually, the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used.
The heatmap below gives the correlation between various features of the dataset.

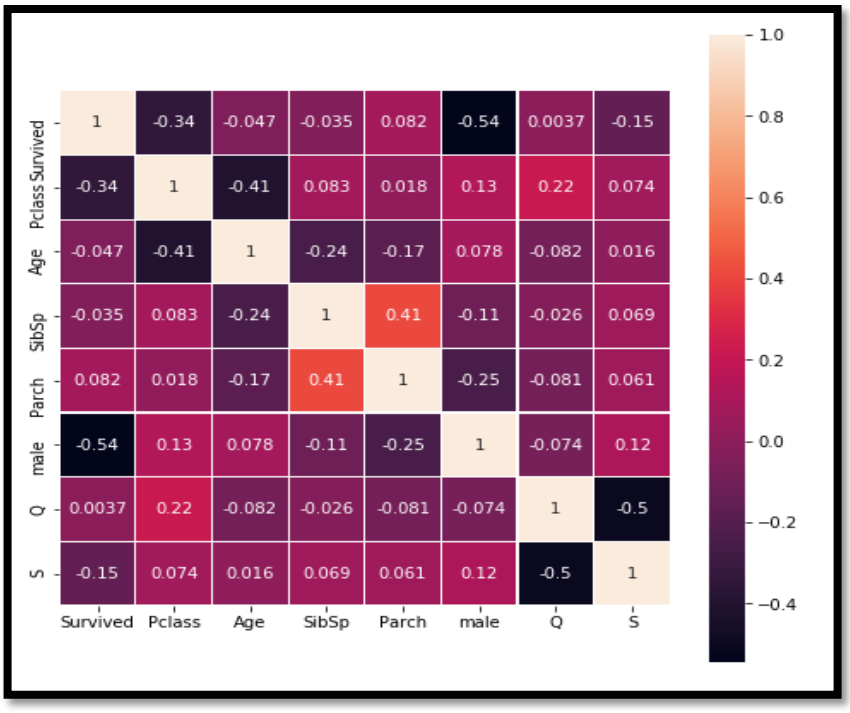

Figure 4.1: HeatMap obtained from the dataset

Explanation of the significance of the correlation coefficients between two features corresponding to each block given by the heatmap

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables. Positive correlation is a relationship between two variables in which both variables move in the same direction. This is when one variable increases while the other increases and visa versa. For example, positive correlation may be that the more you exercise, the more calories you will burn. Whilst negative correlation is a relationship where one variable increases as the other decreases, and vice versa.

The correlation coefficient does not hold any other significance other than to show how strong or weak the relationship is between two features i.e. if one feature increases or decreases how fast or how slow the other feature increases or decreases compared to the first feature.

The following part will illustrate all the major conclusions we found during our EDA and relate those findings to the heatmap.

4.2 Result between different features-

Age & Survival

The above histogram shows how survival will be affected by age. We can see that as age increases, survival of people is going down. So we can say that Age and Survival are negatively correlated as shown by our correlation heatmap.

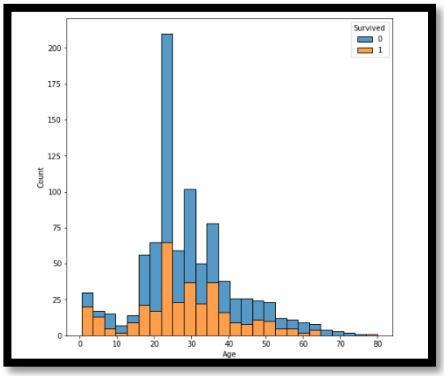

Figure 4.2: Variation of Age & Survival

	Age Group	Total Count	Total Survived	Male Count	Female Count	Male Survived Count	Female Survived Count	Total Survived(%)	Male Survived(%)	Female Survived(%)
0	0-20	179	82	102	77	29	53	45.81	28.43	68.83
1	20-40	577	208	386	191	65	143	36.05	16.84	74.87
2	40-60	141	56	90	51	17	39	39.72	18.89	76.47
3	>60	22	5	19	3	2	3	22.73	10.53	100.00

Figure 4.3: Table for count of passengers on basis of different age group

The data from the above plot is also verified by this table that clearly shows that %survival goes down as the age increases, furthermore the table also shows the %survival for men and women for the age groups.

Sex & Survival

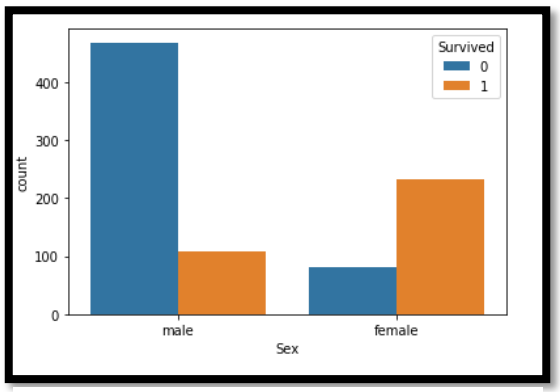

Figure 4.4: Variation of Sex & Survival

	Label	Values
0	Male Total	597.00
1	Male Survived	113.00
2	Male Survived %	18.93
3	Female Total	322.00
4	Female Survived	238.00
5	Female Survived %	73.91

Figure 4.5: Count of passengers on the basis of survivality & sex

The plot shows that more female survived compared to male. This means that females had a high chance of survival compared to male. We found that 73.91% females survived compared to 18.93% of males. This shows if a passenger was a male atop titanic his survival rate would decrease, which shows the negative correlation between Male and Survival as shown by the correlation heatmap. The exact values are given by the table.

Pclass & Survived

As Pclass increases, survivality goes down which signifies the negative correlation between Pclass & Survival.

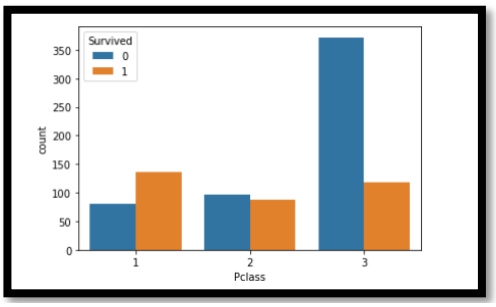

Figure 4.6: Variation of Pclass & Survival

Pclass & Age

As Pclass increases, the no. of people in the lower age group increases and the no. of people in the higher age group decreases which shows the negative correlation between the two.

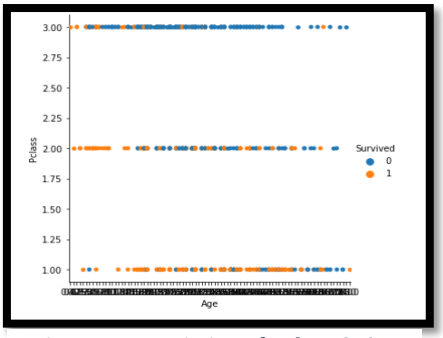

Figure 4.7: Variation of Pclass & Age

Pclass & SibSp

1) A passenger with more siblings/spouse chose a lower passenger class (for Pclass higher the no. lower is the class) i.e. as Pclass increases SibSp value increases which shows the positive correlation

2) If a person has more siblings/spouse their survival chances will decreases which shows the negative correlation

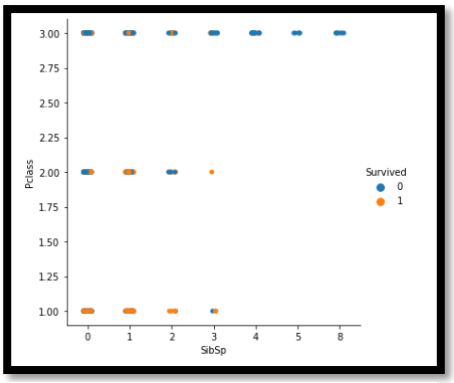

Figure 4.8: Variation of Pclass & SibSp

Pclass & Parch

A passenger with more parents/children chose a lower passenger class(for Pclass higher the no. lower is the class) i.e. as Pclass increases Parch value increases which shows the positive correlation.

If a person has more parents/children their survival chances will increases.

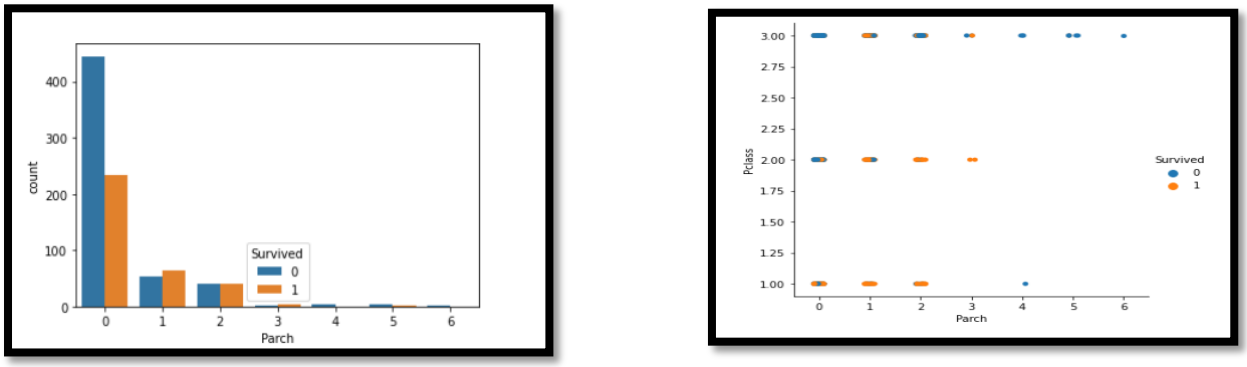

Figure 4.9: Variation of Pclass & Parch

Pclass & male

This plot shows the positive correlation between the pclass and male features (From the plot both no. of females and no. of males increases as Pclass increases but we can see the no. of males increases much more than the no. of females. Thus it can be said that as Pclass increases the binary value (female=0 male=1) increases from 0 to 1 which shows the positive correlation. Although the survival rate of female is high but a female belonging to the lowest passenger class will have higher chance of not surviving.

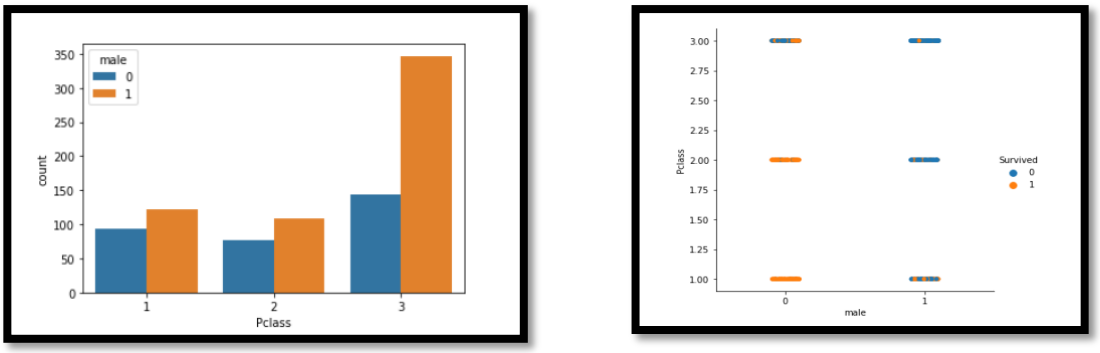

Figure 4.10: Variation of Pclass & male

4.3 Confusion Matrix and Accuracy Of Various Prediction Models

After we made these conclusions, we made our prediction models based on various Machine Learning Algorithms and fit our data into those models to get our prediction. We then plot the confusion matrix for all these models and compared the accuracy of prediction among the various models. The machine learning algorithms applied and confusion matrix have already been discussed in the **METHODOLOGY** section of this report.

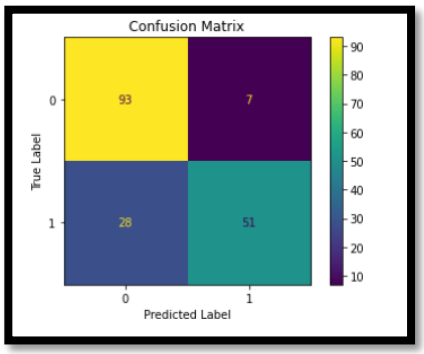

Figure 4.11(a): Confusion matrix for Logistic Regression

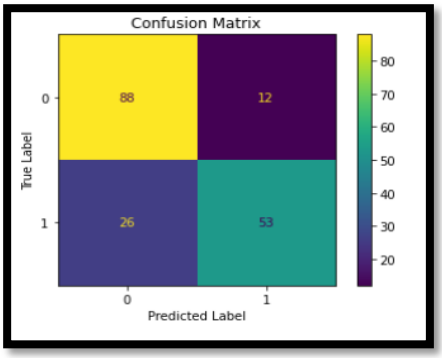

Figure 4.11(b): Confusion matrix for Gaussian Naive Bayes

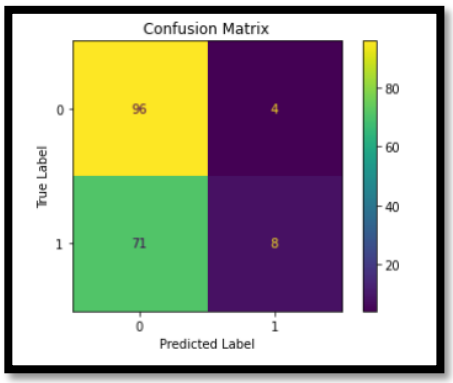

Figure 4.11(c): Confusion matrix for Support Vector Machine

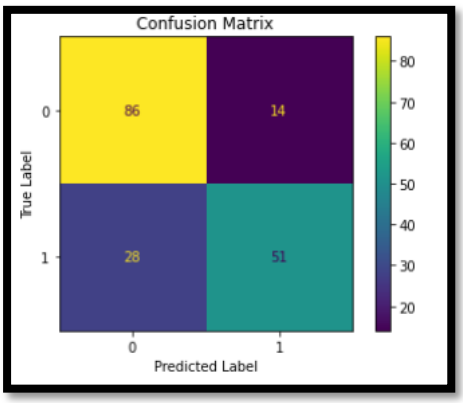

Figure 4.11(d): Confusion matrix for Decision Tree

4.4 Accuracies of each model

	Model	Score
0	Logistic Regression	80.45
3	Gaussian Naive Bayes	78.77
2	Decision Tree	76.54
1	Support Vector Machines	58.10

Figure 4.12: Accuracy Score of different models

From the above score (Accuracy) table, we can see that Logistic Regression model has the highest accuracy with a score of 80.45. This means that applying a logistic regression model for this prediction will predict an outcome that is the most likely to be true.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE OF WORK

5.1 *Conclusion*-

After analyzing the data and making the data suitable for fittting into different modelsfor training the model. We splitted our data into training dataset and testing dataset. Wefitted our training dataset to different models like Logistic Regression Model, DecisionTree, Support Vector Machine and Gaussian Naïve Bayes. Then we fed the testing dataset to the models to check how good our model is to predict the survivality of apassenger.

From the figure 4.12 we can say that logistic regression model will give more accurate prediction on survivality as compared to other models. So, we will chose the Logistic Regression model in our case for the prediction of survivality of the passengerin the incident.

As we got the accuracy of the moddel by providing the testing dataset and comparing the output from the model with the actual output we had, we also tried to manually enterthe details of a passenger and check the status of his survival.

The result obtained for the particular passenger is shown in the figure.

It was seen that the chances for the survival of the passenger is low (10.4% only). Theremany factors on which the survival depends which are mentioned in the Result Analysis section.
So, we can conclude that the result obtained by our model is stisfactory as from the analysis we can get some idea on the survivality of the passenger.

Analysis

In [64]:
```
pclass=int(input('Enter the PClass: '))
age=int(input('Enter the Age of the Passenger: '))
sibsp=int(input('Enter the no. of Sibling or Spouse: '))
parch=int(input('Enter the no. of Parch: '))
```

```
Enter the PClass: 3
Enter the Age of the Passenger: 20
Enter the no. of Sibling or Spouse: 1
Enter the no. of Parch: 3
```

In [65]:
```
gender=input('Enter male or female: ')
gender=gender.lower()
if gender=='male':
    sex=1
else:
    sex=0
```

```
Enter male or female: male
```

In [66]:
```
emb=input('Enter the port of embarkmaent ("C","Q","S"): ')
emb=emb.upper()
if emb=='C':
        Q=0
        S=0
elif emb=='Q':
        Q=1
        S=0
else:
        Q=0
        S=1
```

```
Enter the port of embarkmaent ("C","Q","S"): Q
```

In [96]:
```
demo = logreg.predict([[pclass,age,sibsp,parch,sex,Q,S]])
n=int(demo)
y_pred1=logreg.predict_proba([[pclass,age,sibsp,parch,sex,Q,S]])

for idx,i in enumerate([[item] for sub in y_pred1 for item in sub]):
    if idx==0:
        a=round(i.pop()*100,2)
    else:
        b=round(i.pop()*100,2)
```

In [97]:
```
print(f'The chances of the survival of the passenger is : {b}%')
```

```
The chances of the survival of the passenger is : 10.4%
```

Figure 5.1: Program to check the surviavality of any passenger

5.1 Future Scope-

This project involves implementation of data analytics and machine learning. This project workcan be used as reference to learn implementation of EDA and machine learning from very basic. In future the idea can be extended by making more advanced graphical user interface with the help of newer libraries like shiny in R. An interactive page can be made, i.e., if the value of a attribute is changed on the scale the values corresponding to its graph (ggplot or histogram) will also change. We can also draw much focused conclusions by combining resultswe obtained.

REFERENCES

[1] 2021. [Online]. Available: https://www.researchgate.net/profile/Yogesh-Kakde/publication/325228831_Predicting_Survival_on_Titanic_by_Applying_Exploratory_Data_Analytics_and_Machine_Learning_Techniques/links/5c068f63a6fdcc315f9c0bb9/Predicting-Survival-on-Titanic-by-Applying-Exploratory-Data-Analytics-and-Machine-Learning-Techniques.pdf.

[2] K. Fessel, *Youtube.com*, 2021. [Online]. Available: https://www.youtube.com/channel/UCirb0k3PnuQnRjh8tTJHJuA.

[3] "scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation", *Scikit-learn.org*, 2021. [Online]. Available: https://scikit-learn.org/stable/.