Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset

Bachelor of Technology

In

Information Technology

By

CHATAPARTHI SIVA SHANKAR (22PD1A5403)

From

West Godavari Institute of Science and Engineering

Jul, 2024

Preface

The Machine Learning has seen an enormous growth in the recent past. With the systems getting smarter and automated, the days are not far when we will become completely dependent on them. The most important feature for success of a system is the amount of data that went into training it and the amount of data to which it gives correct results. Thus, data plays the most important role in Machine Learning models, as it is the result of the data available that will decide the future of the system at hand.

The data is present in enormous forms and formats. Thus, it becomes important to study the kind of data and select an algorithm that works the best with that kind of data.

INTRODUCTION

Machine learning is a subset of artificial intelligence (AI) that enables computers to learn from data and make predictions or decisions without being explicitly programmed. It involves using algorithms and statistical models to analyze and draw inferences from patterns in data. Machine learning has applications across various domains, including finance, healthcare, marketing, and even disaster management.

One intriguing project that demonstrates the power of machine learning is the "Titanic - Machine Learning from Disaster" project. This project uses data from the infamous Titanic disaster to build predictive models that determine the likelihood of passengers' survival based on various features.

Project Overview: Titanic - Machine Learning from Disaster

Objective: The main goal of this project is to predict whether a passenger survived the Titanic disaster based on a set of features such as age, sex, class, and other relevant attributes.

Dataset: The project utilizes the Titanic dataset, which includes three files:

- train.csv: Contains the training data, with features like passenger name, age, sex, ticket class, and a label indicating survival.
- test.csv: Contains the test data without the survival label.
- gender_submission.csv: A sample submission file that provides a template for how to format the predictions for the test set.

Key Features:

1. **Name and Title:** Passengers' names and titles, which can be used to infer social status and possibly age.
2. **Age:** Age of the passengers, which is crucial since survival rates varied significantly by age group.
3. **Sex:** Gender of the passengers, with women and children having higher survival rates.
4. **Pclass:** Passenger class, with higher classes generally having better survival chances.
5. **SibSp and Parch:** Number of siblings/spouses and parents/children aboard, which can indicate family size and social connections.

Process:

1. **Data Cleaning:** Handling missing values, correcting data types, and removing outliers.
2. **Feature Engineering:** Creating new features or modifying existing ones to improve model performance. This can involve extracting titles from names, categorizing age groups, or encoding categorical variables.
3. **Exploratory Data Analysis (EDA):** Visualizing and analyzing the dataset to uncover patterns and relationships between features and survival rates.
4. **Model Selection:** Evaluating various machine learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosted Trees.
5. **Model Tuning:** Optimizing model hyperparameters to enhance predictive accuracy.
6. **Model Evaluation:** Assessing model performance using metrics like accuracy, precision, recall, and the confusion matrix.

Outcome: By following these steps, one can develop a robust predictive model that estimates the likelihood of survival for Titanic passengers. This exercise not only showcases the practical applications of machine learning but also provides insights into data preprocessing, feature engineering, and model evaluation techniques.

Conclusion

The Titanic - Machine Learning from Disaster project serves as an excellent introduction to the fundamentals of machine learning. It emphasizes the importance of data preprocessing, feature engineering, and the iterative process of model selection and tuning. By leveraging historical data and machine learning techniques, we can gain valuable insights and make informed predictions, demonstrating the potential of machine learning to solve real-world problems.

Abstract

The Royal Mail Steamer, TITANIC, was the largest cruise ship ever made. The British cruise ship collided during it's only journey, with a huge iceberg. The collision happened in the Pacific Ocean, when the Cruise was moving from Southampton to the New York city. There were approximately 2400 passengers on the Cruise when the accident happened and more than half of them could not survive. This unfortunate yet one of the biggest incident forces the researchers and data analysts to analyse and go deep in the data set. The aim of the various studies going on is to explore the data available and find a pattern and impact of various features on the survival of a person if he/she was on the ship.

The survival of the passengers has been analysed using various different algorithms and they have been compared. A new algorithm has been proposed that will give more accurate results than all the previously analysed algorithms, using the features of those algorithms only.

Contents

Chapter 1

Introduction

The Machine Learning has acquired an inevitable position in today's world. Everything is getting automated and we rely on some software to predict values for us, analysing which, we make the important decisions. These software need to be trained using enormous amount of data so that they understand the underlying pat- tern and develop a knowledge based on its observations. Then this knowledge is used to analyse any data and observations are made. These observations are corre- lated to the previous events of that type which have been used to train the software. Hence, these values are highly accurate and important future decisions can be made keeping those observations in reference.

With the growth of such systems, a lot of researchers are working on huge amounts of data, trying to gather as much useful data as possible and analysing it and training the models using this data which then will be used to predict more values. The quality of data is an important aspect here. The data available is falling short of the demands of the researchers, maybe because sufficient information is not available or there has been no work in that particular field.

In this project, we have analysed the information of the passengers that were present on the RMS Titanic when it collided with the Iceberg. Only 712 out of the 2456 people present could survive the mishap. We have worked upon:

1. The attributes of the people who survived.
2. Does the gender of the person determined the chance of survival.

3. Does the class of the Compartment determined their chances of survival.
4. We have analysed the data available on various models including Decision Tree, Logistic Regression, etc. and compared their results.
5. A new algorithm has been proposed which gives more accurate predictions than the existing models.

Understanding the data is the key for all the analytical processes. The analysis has been done on Jupyter Notebook in Python language.

1.1 Motivation

This is the era of Machine Learning, with everything getting automated and this has even made the process of making decisions depending upon the previous similar incidents automated. The models need to be trained on a number of previous incidents and they are capable of predicting the future based on the recognised pattern.

With the growth of such systems, a lot of researchers are working on huge amounts of data, trying to gather as much useful data as possible and analysing it and training the models using this data which then will be used to predict more values. The quality of data is an important aspect here. The data available is falling short of the demands of the researchers, maybe because sufficient information is not available or there has been no work in that particular field.

This motivated us to analyse the data available of the people on the titanic ship and make observations. The aim is to find the underlying pattern and then to predict chances of survival of a person had he/she been on the ship.

Chapter 2

Related Work

The data analysis done is used for building of both Predictive as well as descriptive models. The predictive model allows us to get the missing values in the data set and then predict if a person would survive in such a scenario or not. This may be used to predict if such incident could reoccur in future. The descriptive model is useful for the first model, as it will tell which features are of more importance while predicting the survival chance.[1]

Ju Liu has provided the information regarding the need of analysing the data in segmented manner and finding the correlations and establishing relations and then predicting using this.[2]

Lin, Kunal and Zeshi have established the fact that it isn't necessary that the model's accuracy will improve if more features are provided to it. It is also stated that the dimensionality reduction plays an important role too.[3]

The data set of the people present on the ship is provided by the Kaggle team.[4]

Chapter 3

Proposed Work

3.1 Objective

The main objective of our project is to analyse the data available of the people who were present on the Titanic when the ship collided with the Iceberg and drowned. We have done exploratory data analysis on the features such as Gender, Age, Class of Compartment, etc. of the passengers using a number of Predictive models. These models predict if the person would have survived or not if he/she was on the ship. We have then compared all the models and used the results to produce an algorithm that has the best accuracy out of all the existing predicting models. Our work includes the comparison of existing models to determine which gives the best results in such situations and then design an algorithm that can be used to predict the chances of survival to avoid any such loss in future.

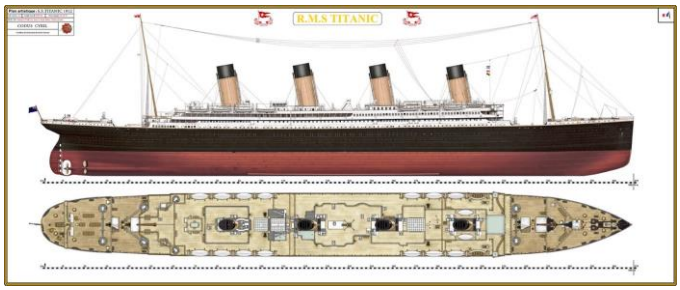


Figure 1: The TITANIC

3.2 Software Requirement Specification

3.2.1 Introduction

- *Purpose*

The aim of this document is to provide a detailed analysis of the data of the pas- sengers of the RMS Titanic. The various models are compared on the basis of their accuracy to predict if a person survived or not using the information available. In the last, using the results, one algorithm has been proposed that suits the best for such kind of data.

- *Scope*

The scope of our work extends to both the predictive and the descriptive models. The data that is being considered in this case is of a particular incident, and it is assumed to be correct. Using the analysis done on this data, new model has been proposed that could be used to predict the chances of survival of a person in any of such unforseen event.

- *Overview of Document*

The next section, the Overall Description section, of this document gives an overview of the functionality of the project. It describes the informal requirements and is used to establish a context for the technical requirements specification in the next section. The third section, Requirements Specification section, of this document is written primarily for the developers and describes in technical terms the details of the functionality of the product. Both sections of the document describe the same software product in its entirety, but are intended for different audiences and thus use different language.

3.2.2 Requirement Specification

- *Functional Requirements*

The analysis has been done on Jupyter Notebook using Python Language. The basic requirement is of sufficient data to train the Machine Learning models. The data should consist of a number of attributes, features so that the analysis can be done on a broader scale and more accurately.

- *Non-Functional Requirements*

The proposed work is more of analysis and hence we will get better results if the data available is vast and true. The quality of data plays a vital role. The accuracy of various models has been compared to improve the performance of the system and a new algorithm has been devised which will further improve the performance and accuracy.

3.3 Dataset

The data has been taken from Kaggle.com, where it is available in the Comma-Separated format. The data set contains 891 rows with attributes including the name of the passenger, the number of siblings, the number of parents or children, the cabin, the ticket number, the fare of the ticket and the place where the person is from.[3]

Pre-processing of the data had to be done because the data had missing values and also some data was present in string format, which had to converted into numeric types so that our model could consider it for analysis. The missing values have been filled with the median of the available values. The data has been split into training and testing to calculate the efficiency of our models. Before the algorithm for the models is built, a number of exploration graphs have been plotted to find out the features which will influence the model the most.

S.No	Attributes
1	PassengerId
2	Survived
3	Pclass
4	Name
5	Sex
6	Age
7	SibSp
8	Ticket
9	Fare
10	Cabin
11	Embarked

Figure 2: Data Attributes

3.4 Building Machine Learning Models

We have trained several Machine Learning models and compared their results. The data set did not provide labels for the training-set, hence we use the predictions on the training set to compare the models. Later, we have compared using cross validation as well.

3.4.1 Logistic Regression

1. This statistical algorithm is used to predict the probability of binary outcomes based on one or more independent variables. It means, this is used to predict an outcome which has 1 or 0 ,yes or no, pass or fail.

2. Probabilities are estimated using sigmoid function. The graph of sigmoid function is as:

Figure 3: Sigmoid Function

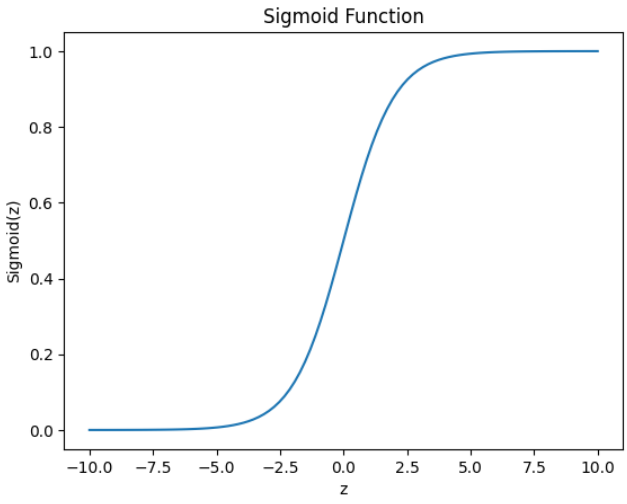
3. Mathematical expression of the sigmoid function is given by:

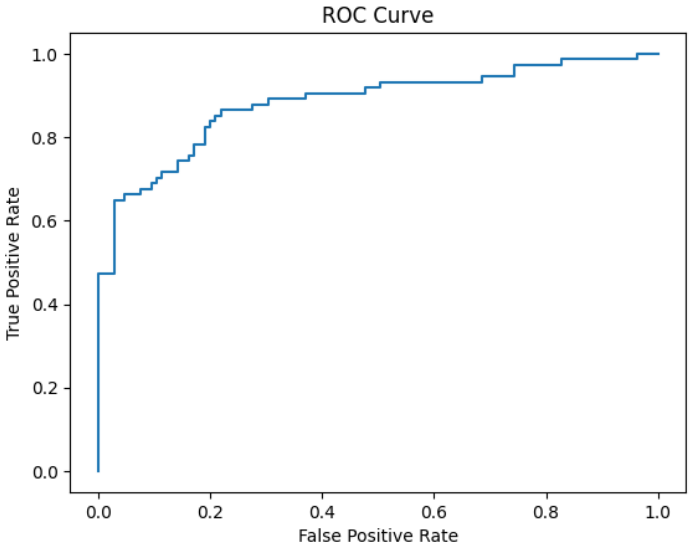
$$F(x) = 1/(1 - e^{-z})$$

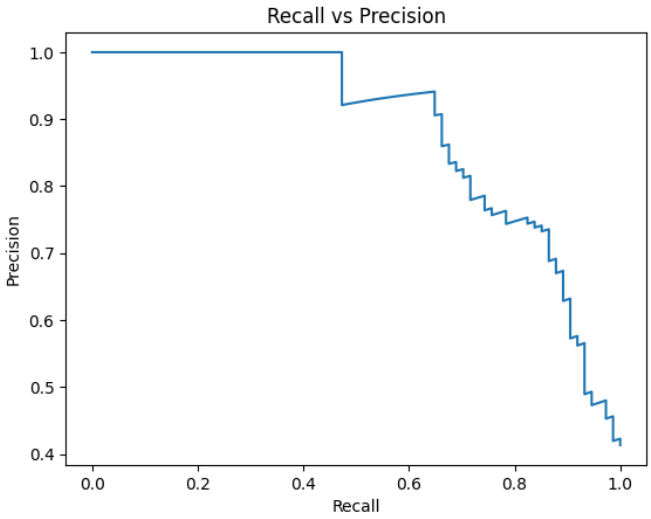
Where z=w0+w1*x1+w2*x2+.....+wn*xn and x1,x2,x3,...,xn are independent variables and F(x) is probability of binary outcome.

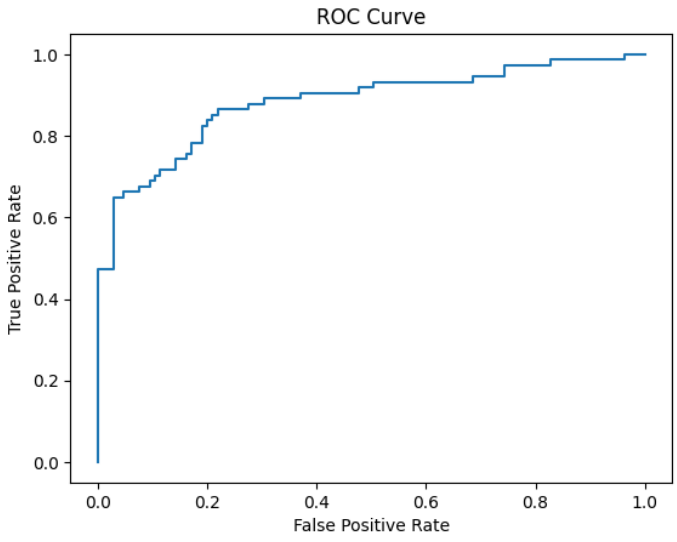
4. Values greater than 0.5 have been classified as survived (1) and others to 0.

5. We get a score of **81% correct predictions.**




ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Recall vs Precision
Precision
Recall
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

3.4.2 K-Nearest Neighbours

1. K nearest neighbours is a classification algorithm highly useful in classifying an unknown data set primarily on the similarity of the neighbouring results.

2. The object is assigned its membership to a class by majority vote of its k nearest neighbours.

3. The degree of 'closeness' is calculated by distance of a said 'object' to its k neighbours.

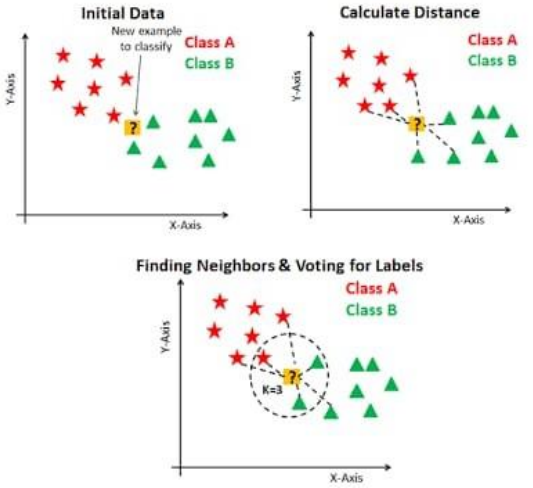
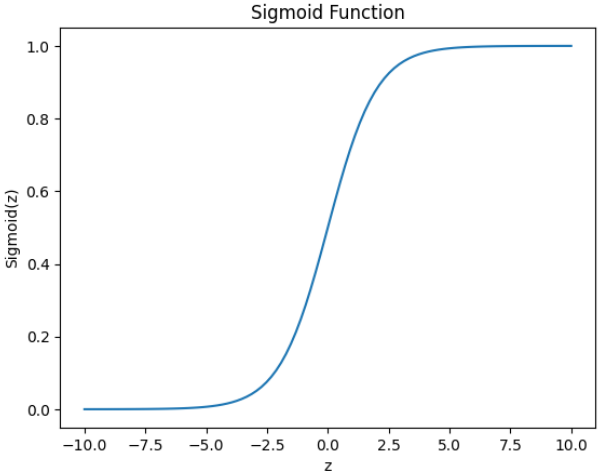
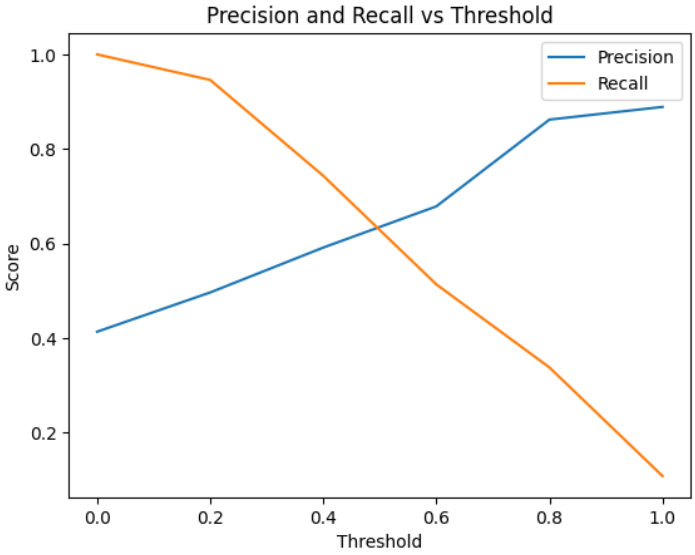
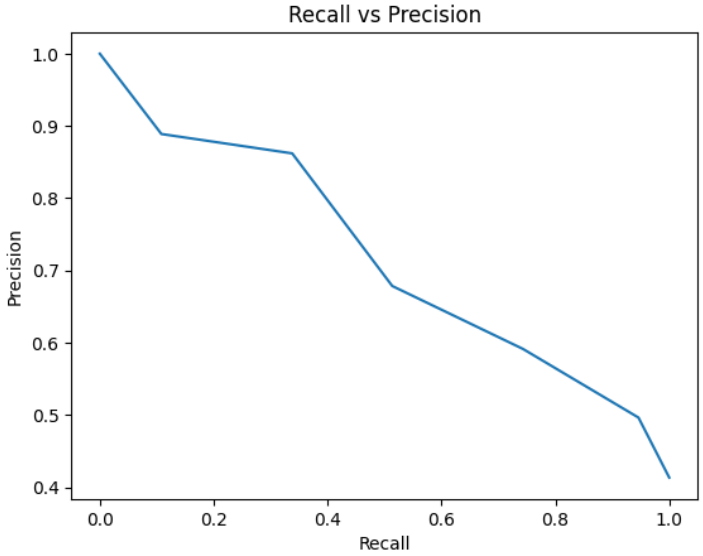


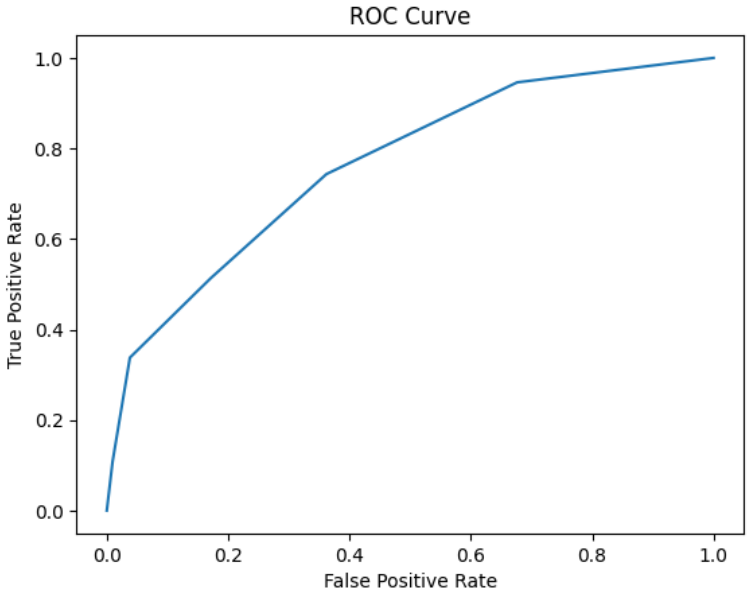
Figure 8: Working of KNN algorithm

4. The model is trained with the value of k equal to 3 which yielded **69.83% correct** predictions.









3.4.3 Decision Tree

1. In decision tree, the results are branched according to a particular condition and this branching continues with further conditions.
2. It is used as an application of supervised machine learning.
3. The algorithm itself evaluates particular decision's importance which is called 'feature classifier'.

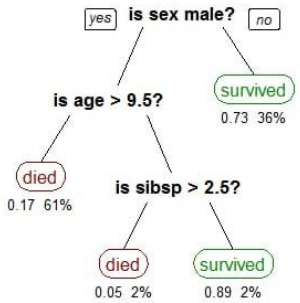
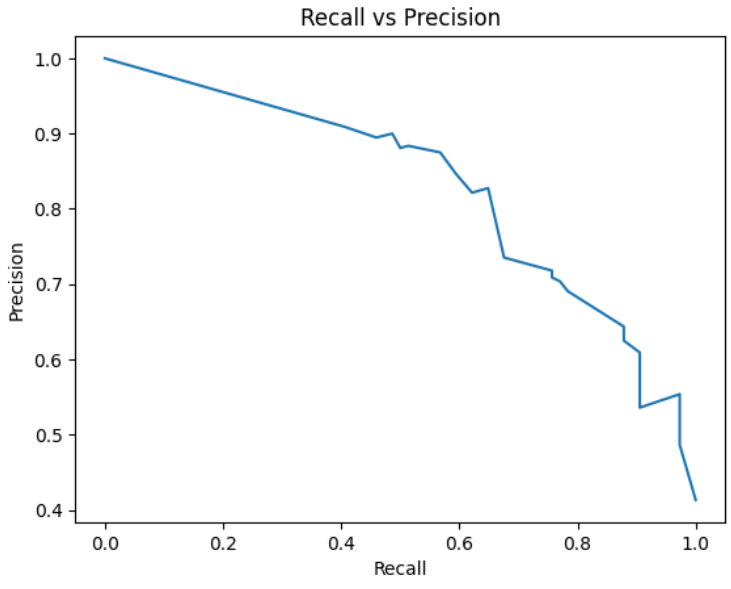
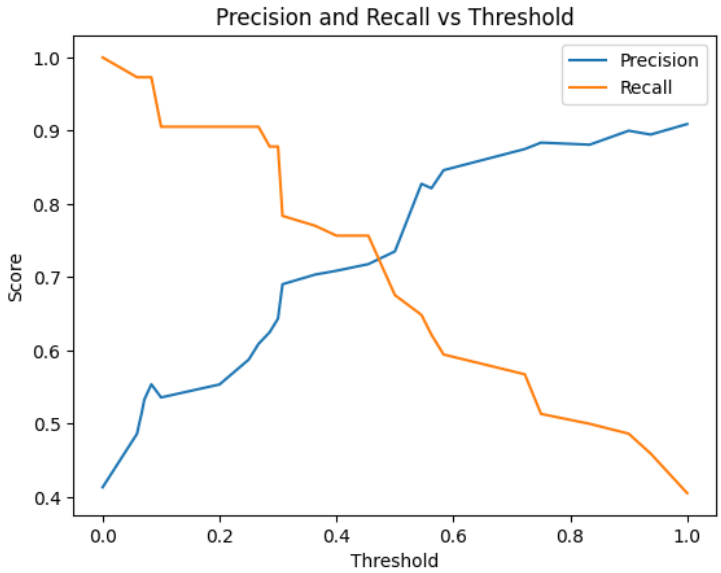
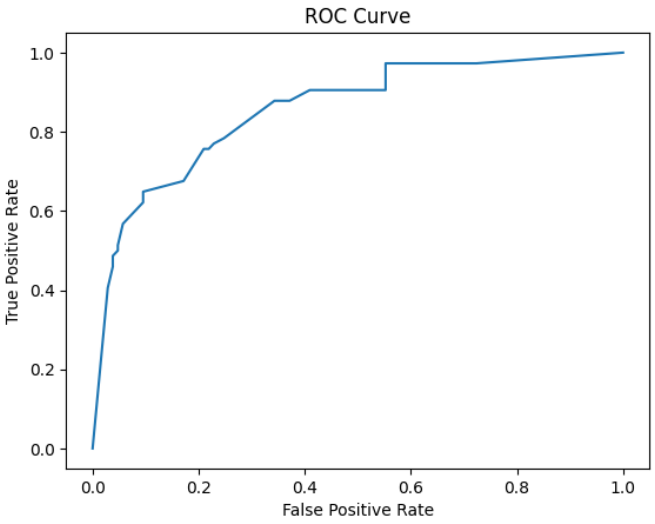


Figure 13: Decision Tree

4. On applying the model with same parameters we get an accuracy of **79.88%** **% correct** predictions.




Precision and Recall vs Threshold
Precision
Recall
Score
Threshold
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

3.4.4 Stochastic Gradient Descent

1. The word ‘Stochastic’ means a system or a process that is linked with random probability.

2. In a typical Batch Gradient Descent, a batch of samples is taken at each iteration to find global minima. However this becomes computationally expensive when we have huge data.

3. This problem is solved by stochastic gradient descent where at each iteration a batch of size 1 is used for applying gradient decent.

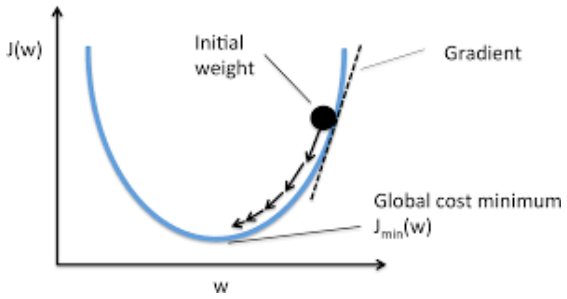
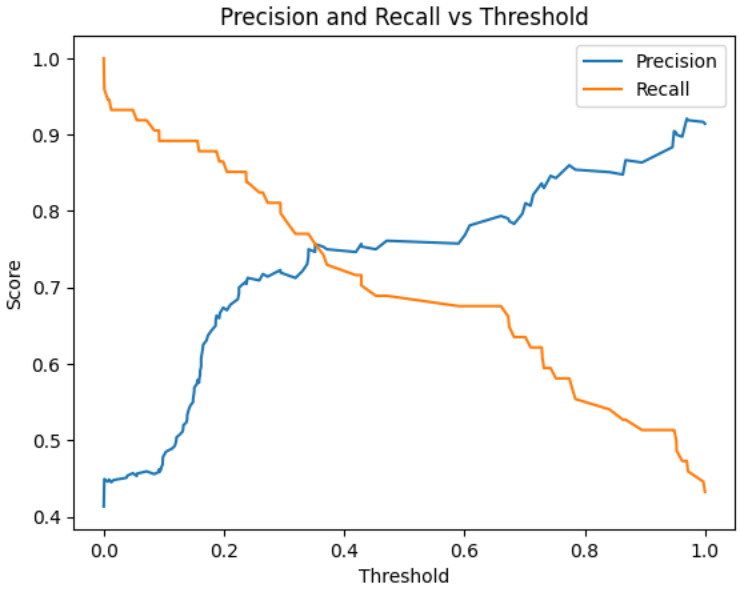
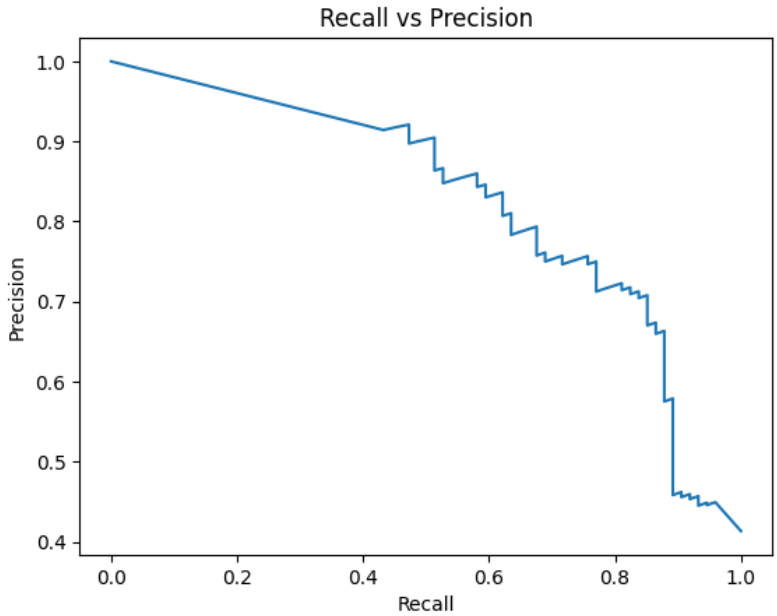
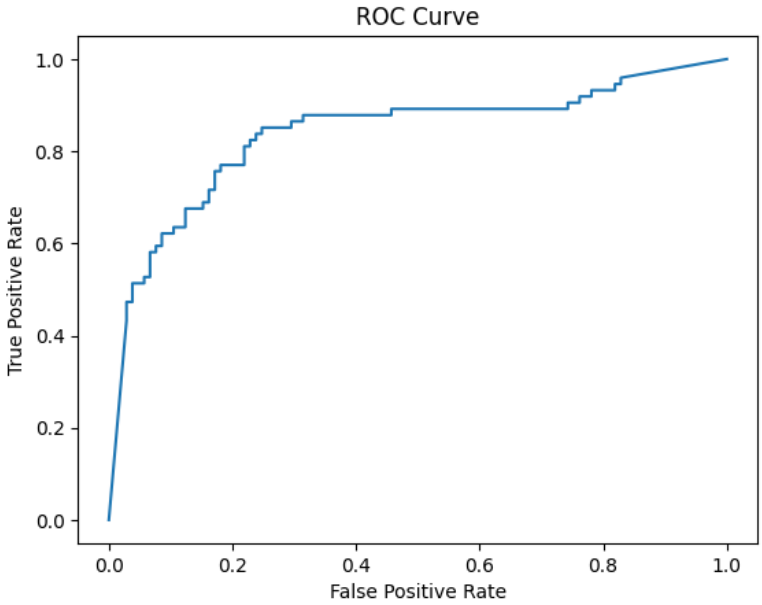


Figure 32: Stochastic Gradient Descent

4. The algorithm yielded us an **accuracy of 77.65%.**


Precision and Recall vs Threshold
Precision
Recall
Score
Threshold
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0


Recall vs Precision
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Chapter 4

Results And Analysis

4.1 Results

Result	Score
Logistic Regression	81.00
K-Nearest Neighbours	69.83%
Decision Tree	79.88%
Stochastic Gradient Descent	77.65%.

Figure 52: The accuracy of various models

1. We see that from all of the existing algorithms , Decision Tree and random Forest give the best performance i.e. 77.65 percent.

2. Naive bayes gives poor performance, reason being the assumption of Naive Bayes that all features contribute equally and are independent to each other which is not in our case as seen in correlation table.

3. Stochastic Gradient Decent inspite of one of the good algorithms, fails here due to under fitting and its more linear nature.

4. The hyper-parameter tuned Decision Tree gives accuracy less than the untuned Decision Tree. Tuned model is over-fitted over training data. Moreover this model is more complex as well takes more time in training as well as in making predictions.

5. Our Stacked model performs extremely well over test data with a accuracy over 94.1675 percent. This is the best mean accuracy achievable so far.

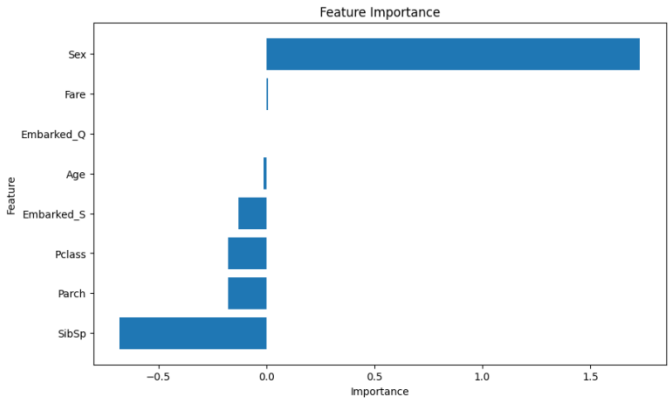


Figure 53: The Importance of various features

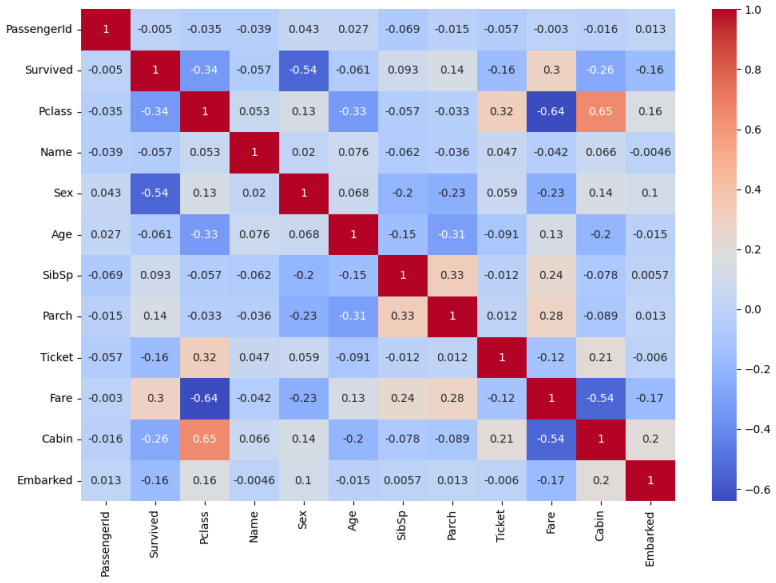


Figure 54: Correlation plot

4.2 Analysis

1. The analysis showed that the **attributes Age, Class of compartment and Gender** are the important features to the model building.

2. Further analysis shows that the majority of passengers who survived occupied the first class.

Figure 55: Survivors vs Class of Compartment

3. The majority passengers surviving were of age group between 20-30. Further, the majority of the men who survived were between 18-30 years of age. While the women who survived ranged between 14-40 majorly.
4. The passengers surviving were majorly female. The gender is found to me the most important feature for classification.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

1. The comprehensive research gives us a result with decision tree having the highest score with 79.88% correct predictions and lowest false discovery rate.
2. The research also aware us of the features that are highly relevant to the prediction of survival of a passenger, with gender being the feature with the highest importance.
3. The correlation between factors first evaluated using a basic formula was also justified in some cases while being defied in others.
4. An algorithm has been proposed which has a better accuracy than all the existing models.
5. The accuracy of various complex models was relatively low for the data. This is because we had only 891 instances of the data, which lead to over-fitting in most of the cases.

5.2 Applications

1. We can use the model both for predictive and classification purposes.
2. We can predict the missing values of attributes and even the events that may happen in future.
3. The predictive model can be used to predict chances of survival and can be used to prevent any mishap in future.
4. The decision tree gave the best accuracy, hence we know which model works the best for such data.

5.3 Future Work

1. Future work includes using other algorithms like K means, gradient boosting, adaboost, further hyper tuning the decision tree algorithm and even using advanced neural networks as well as Reinforcement learning.
2. Validating other techniques like assigning feature importance, introducing a new feature, a more robust pre-processing could improve the accuracies and may yield different results for different algorithms.