**REPORT**

**TITANIC SURVIVAL PREDICTION**

****

**COURSE NAME:**

MACHINE LEARNING FOUNDATION (INT 247)

**SUBMITTED TO:**

DR. Sagar Pande

**SUBMITTED BY:**

NAME : M. Karthik

ROLL NO : B55

REG.NO : 11908893

School : School of Computer Science and Engineering

Name of the University : Lovely Professional University

Date Of Submission : 29th MARCH 2022

**DECLARATION**

I, MAYARA KARTHIK of 6th Semester certify that the project report entitled of **-“TITANIC SURVIVAL PREDICTATION”,** in partial fulfilment of the requirements for the award ofthe Bachelor of Technology in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of Lovely Professional University, Punjab, has been prepared by me and is my personal and authentic work under the guidance of Dr. Sagar Pande.

**ACKNOWLEDGEMENT**

For students, preparation of the project report is great challenge, As it requires in-depth knowledge of a particular field of engineering and the recent advancements. This requires immense guidance and help from experienced person in that field.

I would like to express my gratitude and appreciation to all those who gave me the possibility to complete this report. Special thanks is due to my supervisor Dr. Sagar Pande whose help, stimulating suggestions and encouragement helped me in all time of fabrication process and in writing this report. I also sincerely thanks for the time spent proofreading and correcting my many mistakes.

Many thanks go to the lecturer who have given his full effort in guiding me and achieving the goal as well as his encouragement to maintain our progress in track. My profound thanks go to all classmates, especially to my friends for spending their time in helping and giving support whenever I need it in fabricating my project.

DATE:29TH MARCH 2022

**MAYARA KARTHIK.**

**ABSTRACT**

The Titanic incident has led the scientist and investigators to comprehend what can have prompted the survival of a few travelers and death of the rest. The sinking of the RMS 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 survive than others, like women, children were the one who got the priority to rescue. The objective is to 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. After this the results of applying machine models are compared and analyzed on the basis of accuracy. Many machine learning algorithms contributed in predicting the survival rate of passengers. In addition to the this, a dataset of 891 rows which includes the attributes namely Age, PassengerID, Sex, Name, Embarked, Fare etc. has been used. In this paper, survival of passengers is figured out using various machine learning techniques namely decision tree, logistic regression and linear SVM. The main focus of this work is to differentiate between the three different machine learning algorithms to analyze the survival rate of traveller based on the accuracy.

**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 the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine. Machine learning algorithms are applied to make a prediction for passengers survived after sinking of Titanic.

Different features like name, title, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that includes the utilization of computational methods to find out the important and useful patterns in large data. Using machine learning algorithms, survival is predicted on different combinations of features. The target is to perform exploratory data analytics on the available dataset and to understand the effect of every field on the survival of passengers by applying analytics between every field of the dataset with the “Survival” field. Different algorithms are compared based on their accuracy and therefore the best performing model is recommended for predictions.

The predictions are done for newer data sets by applying machine learning algorithm. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking the Titanic. Features like ticket fare, 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.

**AIM AND SCOPE OF THE PROJECT**

**AIM OF THE PROJECT:**

The goal of the project was to predict the survival of passengers based off a set of data. We retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with 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 passenger in the test set, we had to predict whether or not they survived the sinking. Our score was the percentage of correctly predictions.

**SCOPE OF THE PROJECT:**

The requirements:

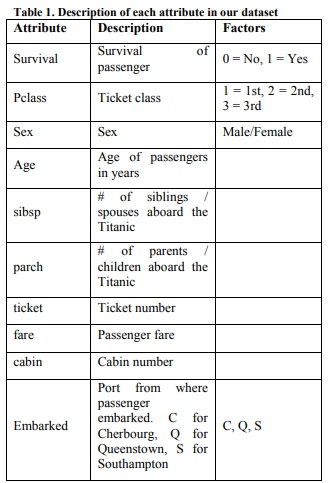
1. Programming language Python and its libraries NumPy (to perform matrix operations) and SciKit-Learn (to apply machine learning algorithms)
2. Several machine learning algorithms (Random Forest classifier, Stochastic gradient descent, Logistic regression)
3. Feature Engineering techniques

Sources used:

1. Google colaboratory
2. Python 3.7 with the libraries NumPy, sklearn, and matplotlib

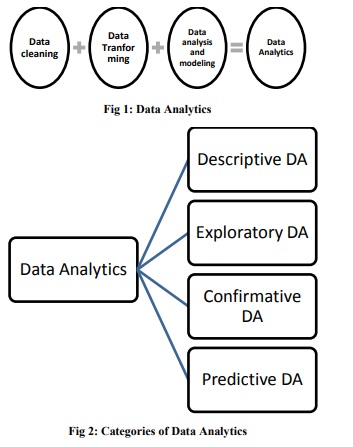
**OBJECTIVE**

The objective of this project is then to build a predictive model to predict which of the passengers survived the ship wreck. In particular, the response variable Survived will be modelled given ten possible predictors. The remainder of this report includes background on the methods used to build the predictive model, specifically classification and regression trees, and random forests. A case study based on the RMS Titanic data implementing the methods will be conducted.



**PURPOSE:**

* 1. The Purpose is to perform exploratory data analytics to mine various information in the dataset available and to know effect of each field on survival of passengers by applying analytics between every field of dataset with “Survival” field.
  2. The predictions are done for newer data sets by applying machine learning algorithm. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions.



**WORK & IMPLEMENTATION**

**DATA DESCRIPTION:**

1. survival: Survival of passenger (0 = No; 1 = Yes)

2. pclass: Passenger Class (1=First, 2=Second,3=Third)

3. name: Name

4. sex: Sex (Male/Female)

5. age: Age of passengers in years

6. sibsp: number of siblings and spouses traveling

7. parch: number of parents and children traveling

8. ticket: Ticket Number

9. fare: Passenger Fare

10. cabin: Cabin number

11. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

**SIBSP:**

This dataset defines family relations

Sibling’s: sister, brother, stepsister, stepbrother.

Spouse: husband, wife.

**AGE:**

Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5.

**PARCH:**

This dataset defines the family relations

Parent’s: father, mother

Children’s: son, daughter, stepson, stepdaughter.

There were also a few children who traveled solely with a

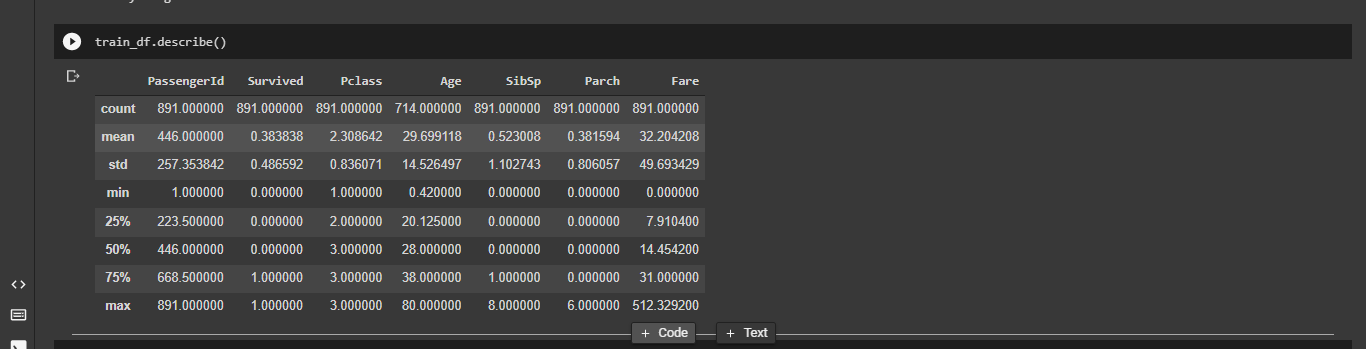
packages and libraries to make writing the program simpler,

the data from the seaborn package is needed to be loaded and

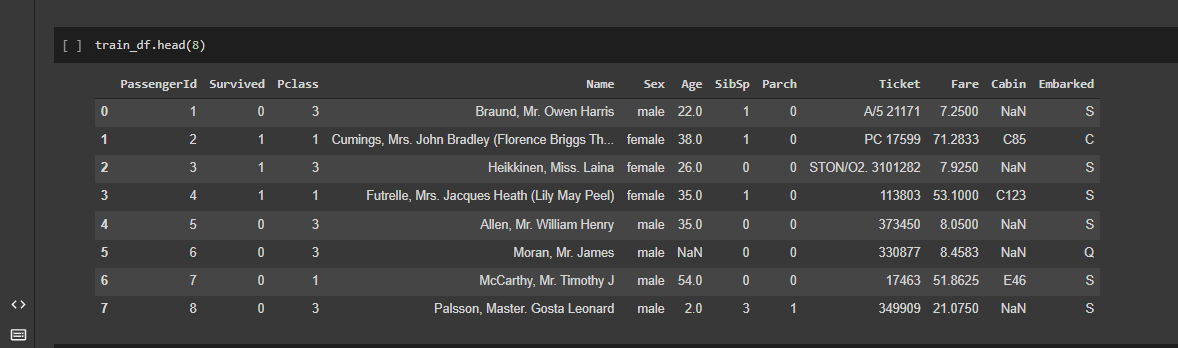
thereby few rows are to be printed.

There were also a few children who travelled solely with a nursemaid, so for them parch=0.

After importing the packages and libraries to make writing the program simpler, the data from the seaborn package in needed to be loaded and thereby few rows are to be printed.

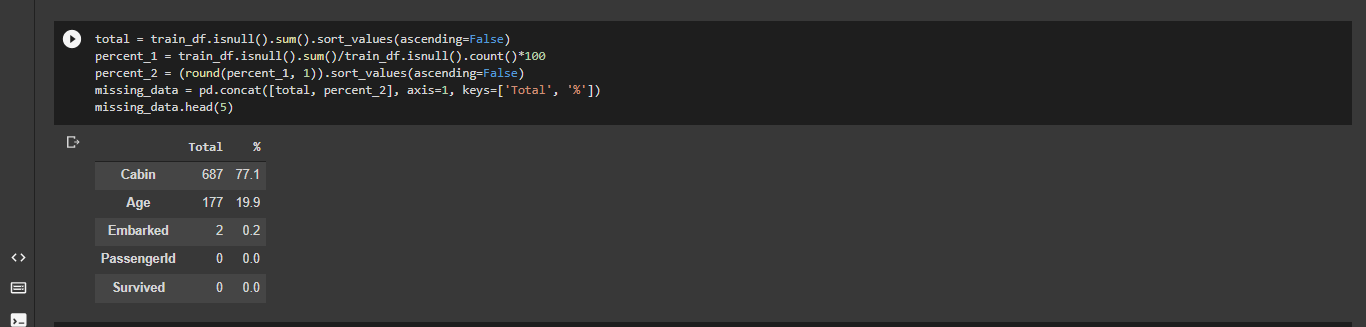


From the above table we can see that 38% out of the training set survived the Titanic. We can also see that the passenger ages range from 0.4 to 80. On top of that we can already detect some features, that contain missing values, like the **AGE** feature.

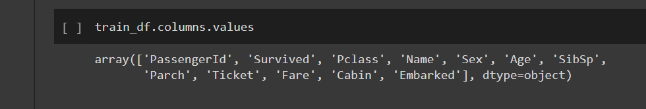


From the above table, first of all that we need to convert a lot of features into numeric ones later on, so that the machine learning algorithms can process them. Furthermore, we can see that the features have widely different ranges, that we will need to convert into roughly the same scale. We can also spot some more features, that contain missing values (NaN = not a number), that we need to deal with.

**Let’s look in more detail what data is missing.**

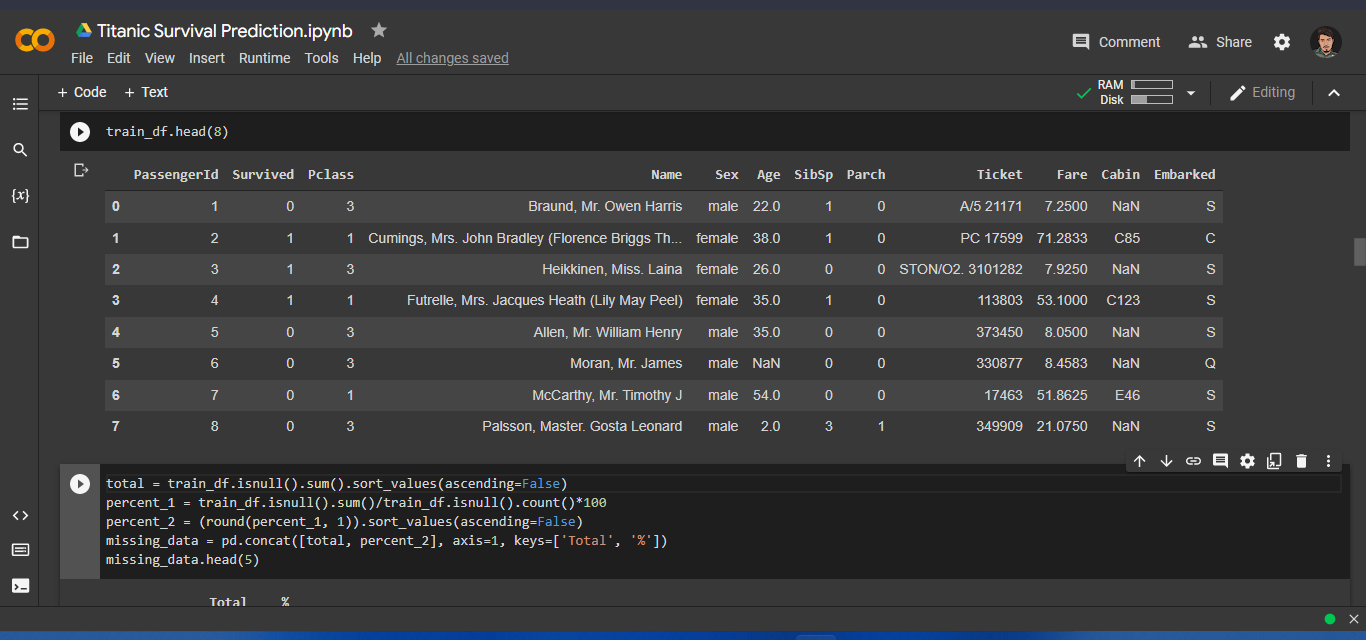


The Embarked feature has only two missing values which can easily be filled. It will be much more tricky to deal with the **AGE** feature, which has 177 missing values. The **CABIN** feature needs further investigation, but it looks like that we might want to drop it from the dataset, since 77 % of it are missing.



Features for high survival rate is:

To me if everything except **PassengerId, Ticket** and **Name** would be correlated with a high survival rate.



The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts

of it. Following the receipt of a count of the dataset’s rows

and column, it is helpful to keep a record of every rows of the

passengers who were aboard the ship. On the other hand, the

columns are regarded as every commuter’s assigned data.

Theoretically, there were 891 passengers/rows and 15 data

points/columns in the data set. Meanwhile, statistics like mean,

count, standard deviation, etc. was received on the dataset.

Ensuing this, the maximum ticket fare that a passenger likely

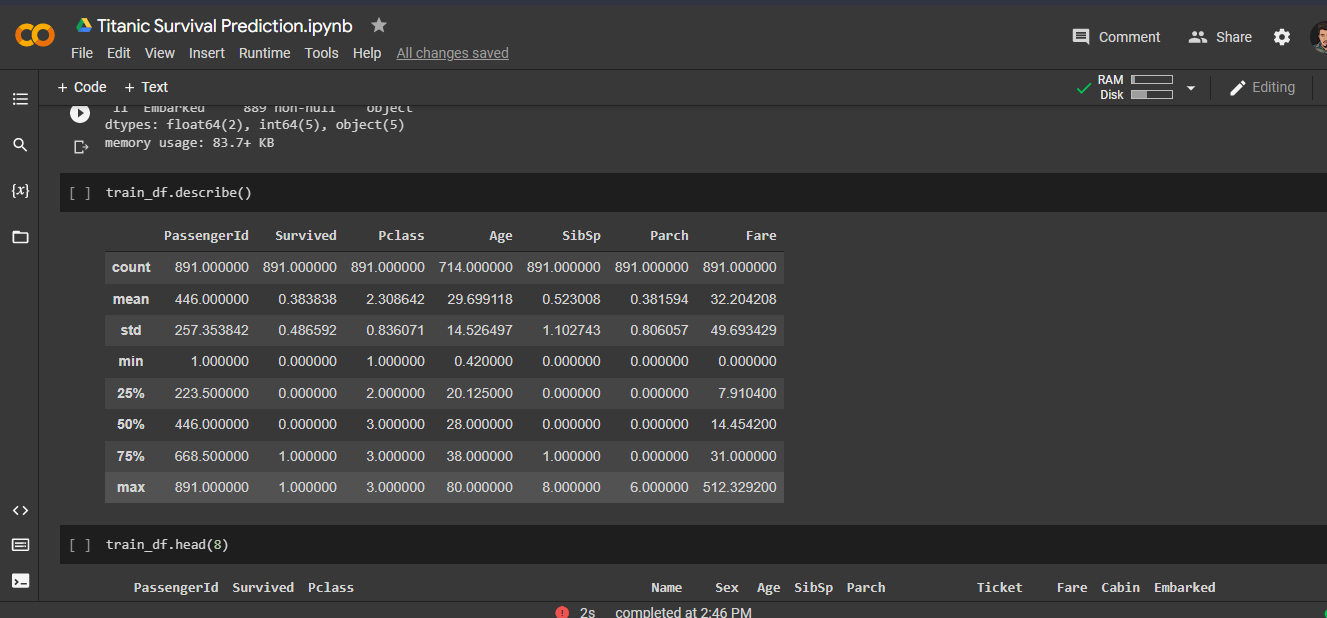
paid for the ticket was around 512.33-pound sterling while the

minimum was as low as 0.

The data was then required to be analyzed by receiving counts of it. Following the receipt of a count of the dataset’s rows and column, it is helpful to keep a record of every row.

**DATA CLEANING**

Before applying any type of data analytics on the dataset, the data is first cleaned. There are some missing values in the dataset which needs to be handled. In attributes like Age, Cabin and Embarked, missing values are replaced with random sample from existing age. The data is first cleaned before implementing some form of data analytics on the dataset. Here, it needs to be analyzed. It is also seen that there are a few absent data values for the column: age as it is smaller than 891. Absent values are restored with arbitrary samples. Present age inside attributes such as Age, Cabin and Embarked. In this data set, the mean age of all the passengers is 29.699 wherein the oldest passenger is 80 years old and the youngest is only 0.42 years.

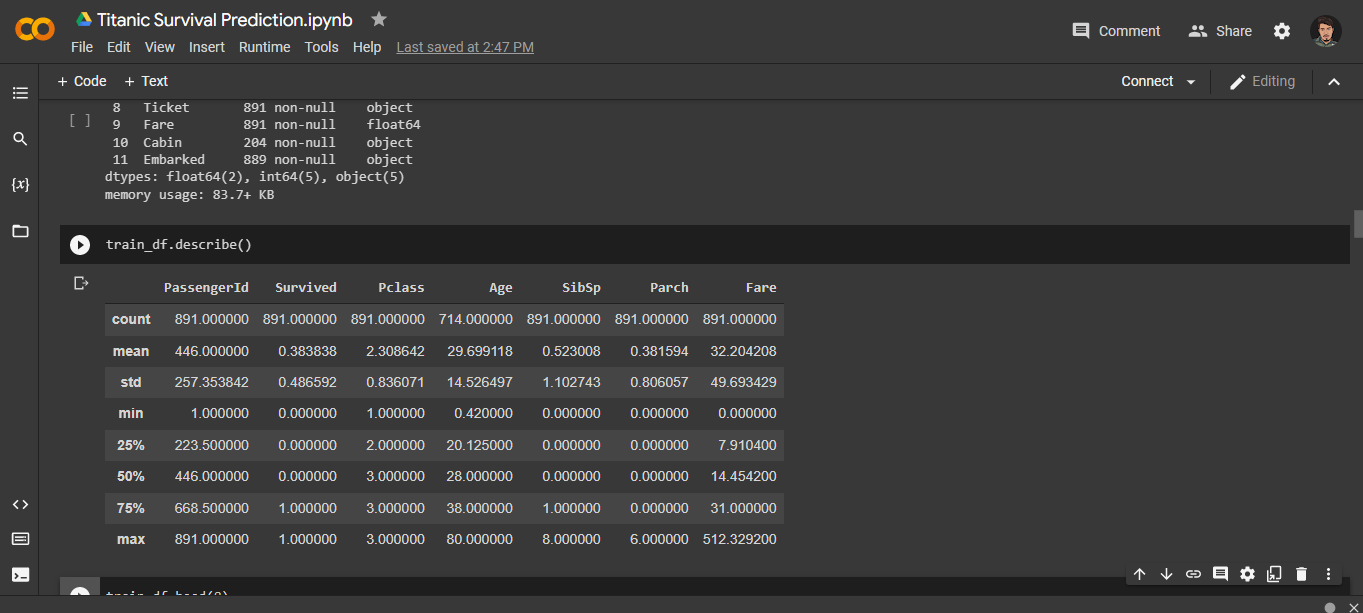


**PROCESS FLOW:**

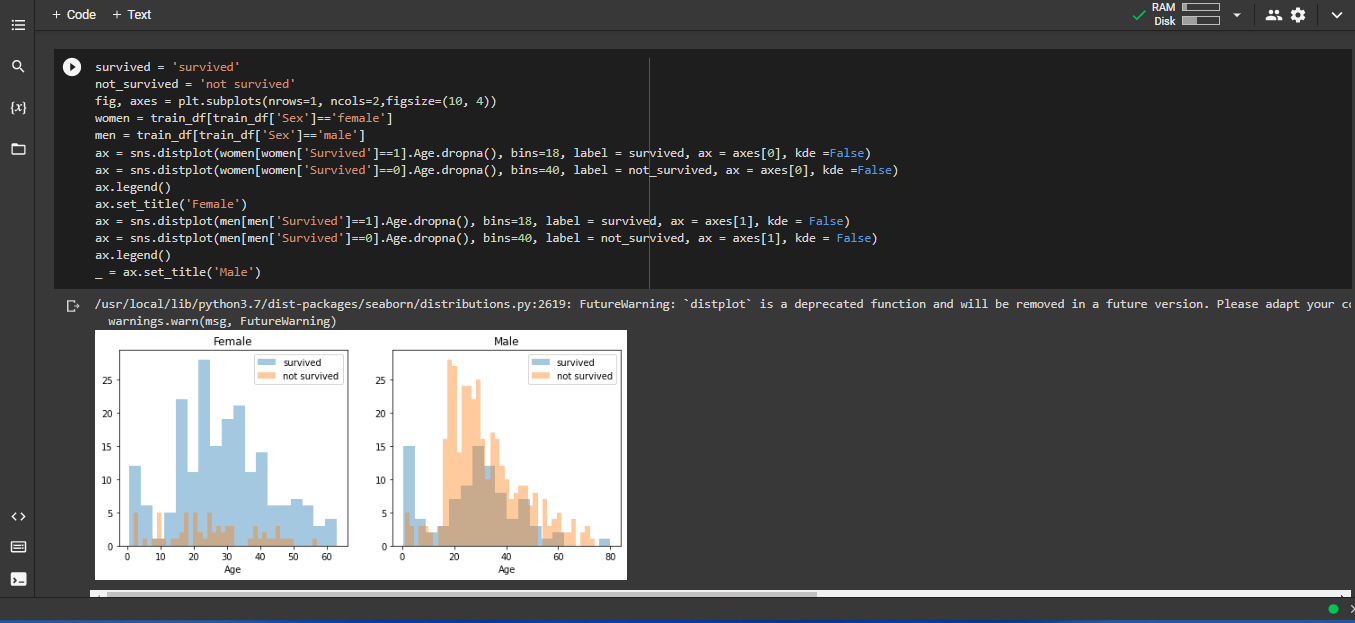
**EXPLORATORY DATA ANALYSIS:**

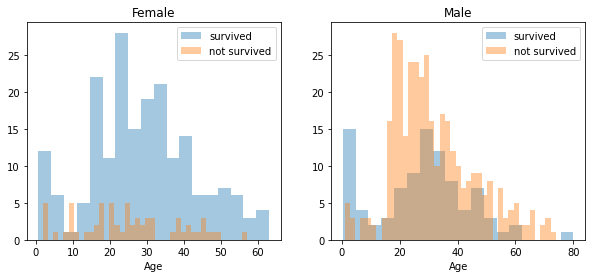
We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the survival rate. The data is deeply analysed by finding a relationship between each attribute and survival.

**Explore and Visualize Dataset:**

****

**Sex and Age:**

****

****

the probability of survival is higher between

ages 18 and 30. Women’s survival percentage is slightly

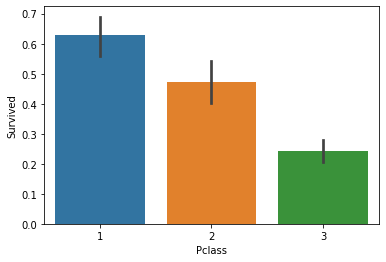
higher in between ages 14 and 40. Between ages 5 to 18, the

likelihood for survival is very low for men but that is not

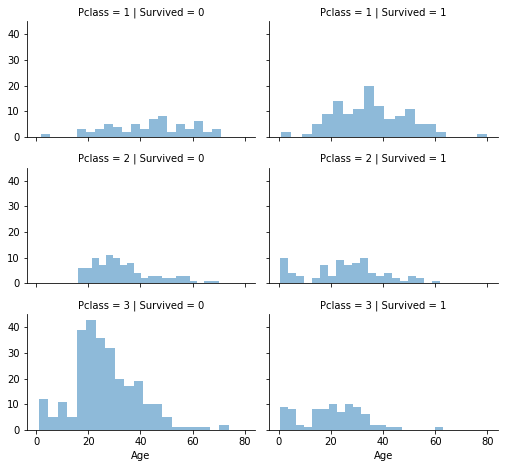
necessarily true for women. Babies and toddlers are however

just a bit more likely to survive.

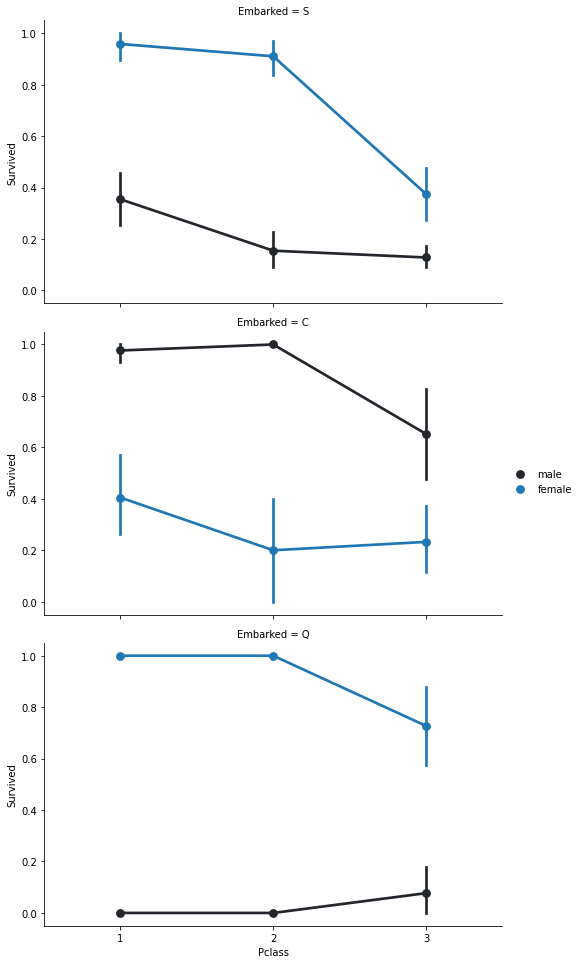
The probability of survival is higher between ages 18 and 30. Women’s survival percentage is slightly higher in between ages 14 and 40. Between ages 5 to 18, the likelihood for survival is very low for men but that is not necessarily true for women. Babies and toddlers are however just a bit more likely to survive. Since there seem to be certain ages, which have increased odds of survival.

****

Here we see clearly, that Pclass is contributing to a person’s chance of survival, especially if this person is in class 1.

****

The plot above confirms our assumption about pclass 1, but we can also spot a high probability that a person in pclass 3 will not survive.

****

Depending on the gender, embarked seems to be associated with survival. Women have a greater chance of survival at ports Q and S. Whereas port C has seen more men surviving but low likelihood at ports Q or S. Even Pclass is correlated to survival rates.

It is observed that Pclass is massively contributing to an

individual’s chance to survival. In class 1, a person’s chance is

remarkably higher than the other two classes. Pclass 3 being

the last, the survival rates are the lowest.

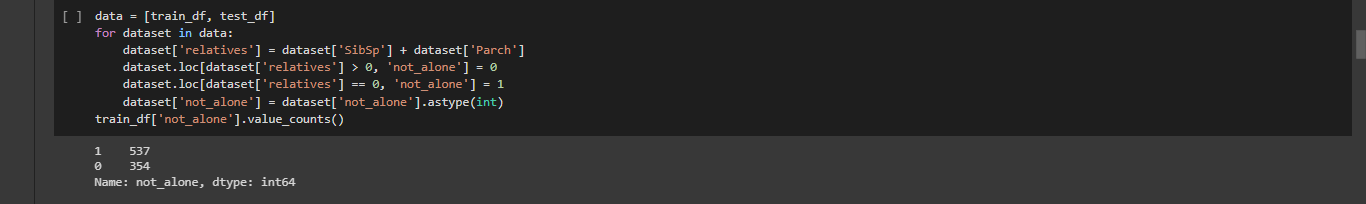
It is observed that Pclass is massively contributing to an

individual’s chance to survival. In class 1, a person’s chance is

remarkably higher than the other two classes. Pclass 3 being

the last, the survival rates are the lowest.

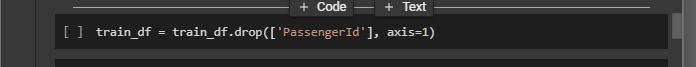
Here embarked seems to be correlated with survival, depending, on the gender. It is observed that Pclass is massively contributing to an individual’s chance to survival. In class 1, a person’s chance is remarkably higher than the other two classes. Pclass 3 being the last, the survival rates are the lowest.



Sibsp and parch would make more sense as a combined feature, that shows the total number of relatives, a person has on the Titanic.

**DATA PREPROCESSING**

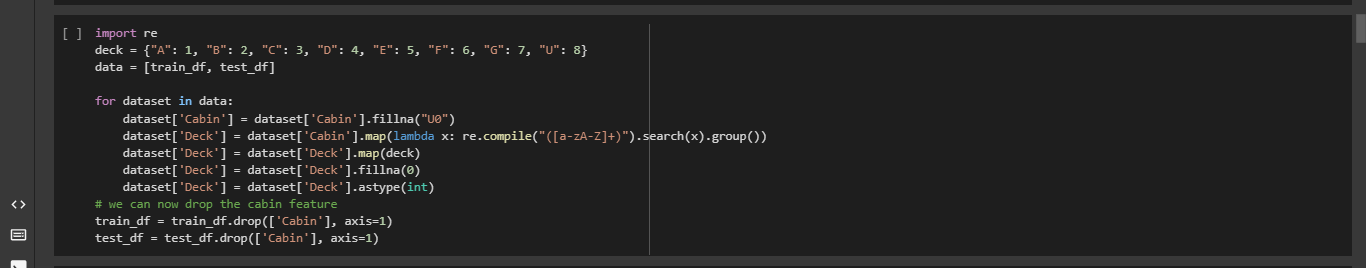
Here I am drop ‘PassengerID’ from the train set, because it does not contribute to a person’s survival probability. I will not drop it from the test set, since it is required there for the submission.



**DATA MISSING:**

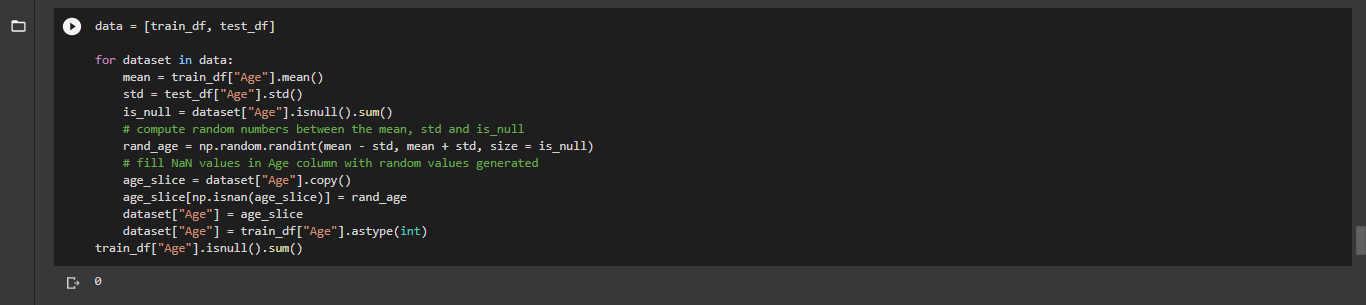
* **Cabin:**

Here we are dealing with missing data, we have to deal with Cabin (687), Embarked (2) and Age (177). we have to delete the ‘Cabin’ variable. A cabin number looks like C123 and the letter refers to the deck. Therefore, we’re going to extract these and create a new feature, that contains a person’s deck. Afterwords we will convert the feature into a numeric variable. The missing values will be converted to zero. In the picture below you can see the actual decks of the titanic, ranging from A to G.



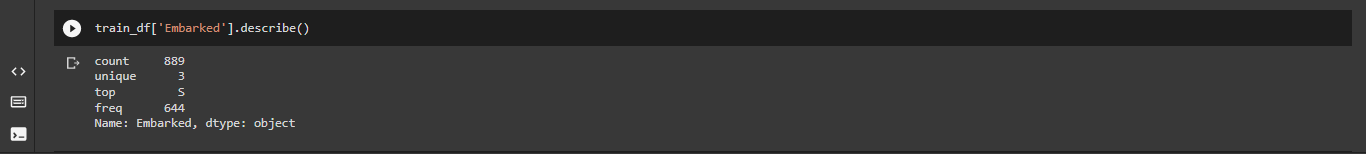
* **AGE:**

Here I am checking the issue with the age features missing values. And I am going to create an array that contains random numbers, which are computed based on the mean age value in regards to the standard deviation and is\_null.

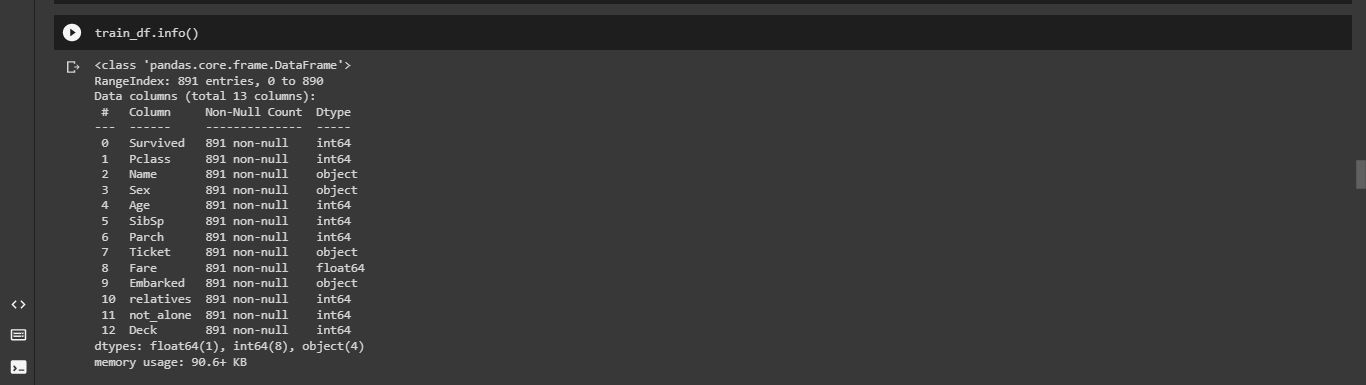


**EMBARKED:**

Since the Embarked feature has only 2 missing values, we will just fill these with the most common one.



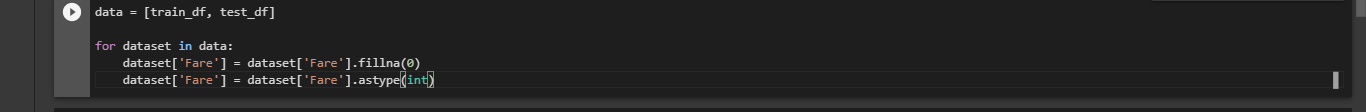
**CONVERTING FEATURES:**



From above fig. I can see that **FARE** is a float and we have to deal with 4 categorical features: Name, Sex, Ticket and Embarked. We should look into it and transform one after another.

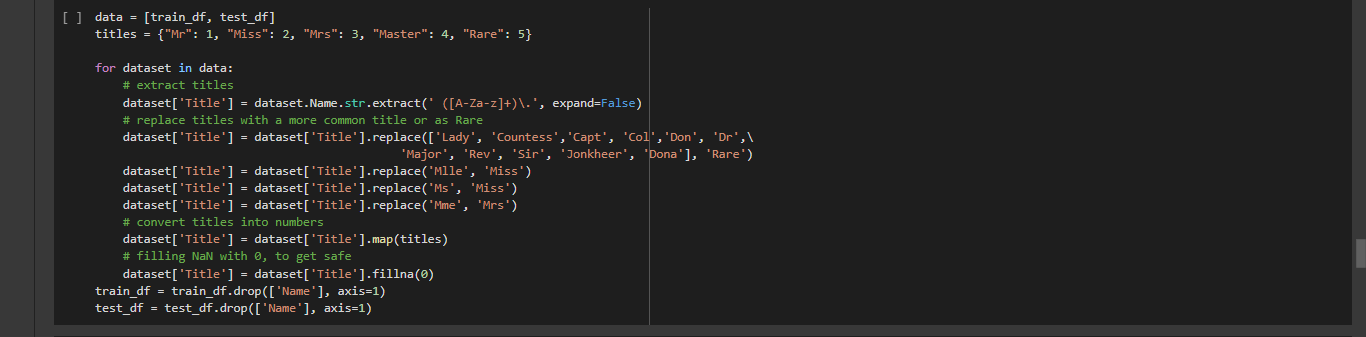
1. **FARE:**

Converting **FARE** from float to int64, using the **ASTYPE()** function pandas provides:



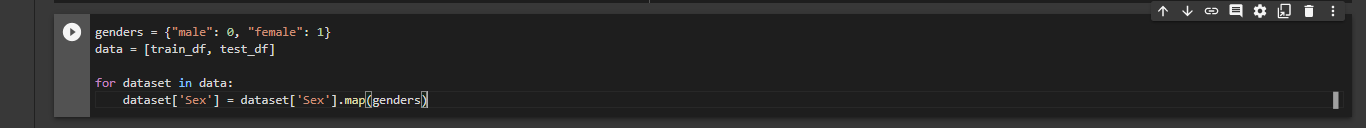
1. **NAME:**

I used the name feature to extract the Titles from the Name, so that I can build a new feature out of that.



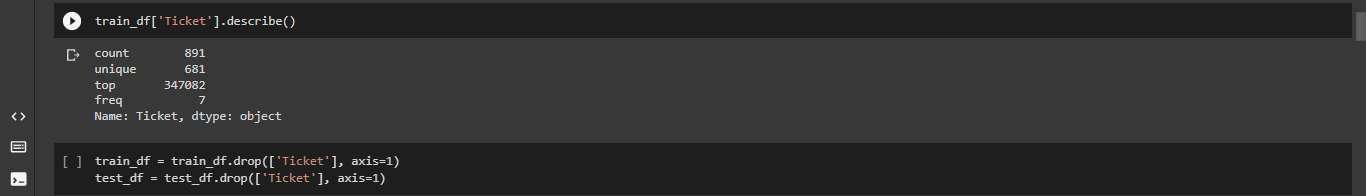
1. **SEX:**

Here I am converting **SEX** feature into numeric.



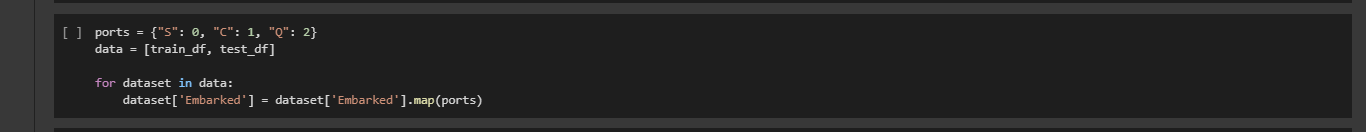
1. **TICKET:**

The Ticket attribute has 681 unique tickets, it will be quite difficult to convert them into meaningful categories. So I drop it from the dataset.



1. **EMBARKED:**

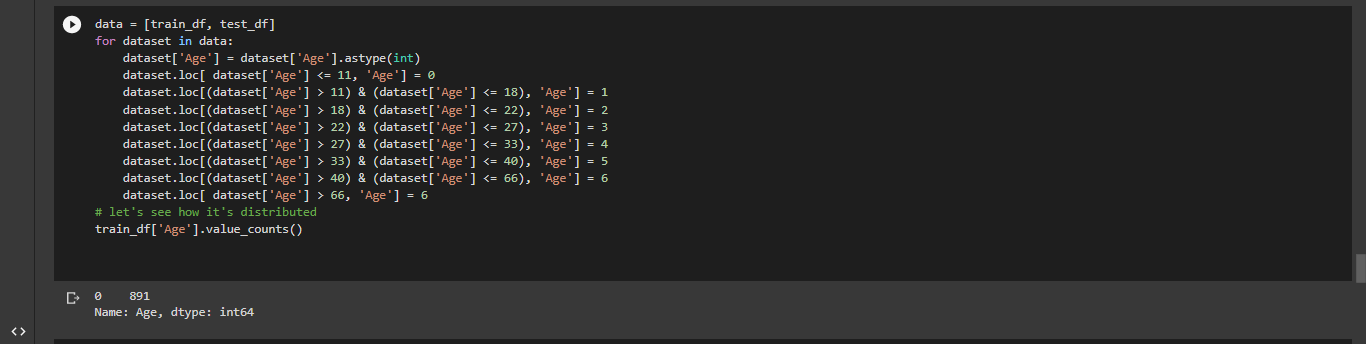
Here I am converting **EMBARKED** into numeric.



**CREATING CATEGORIES:**

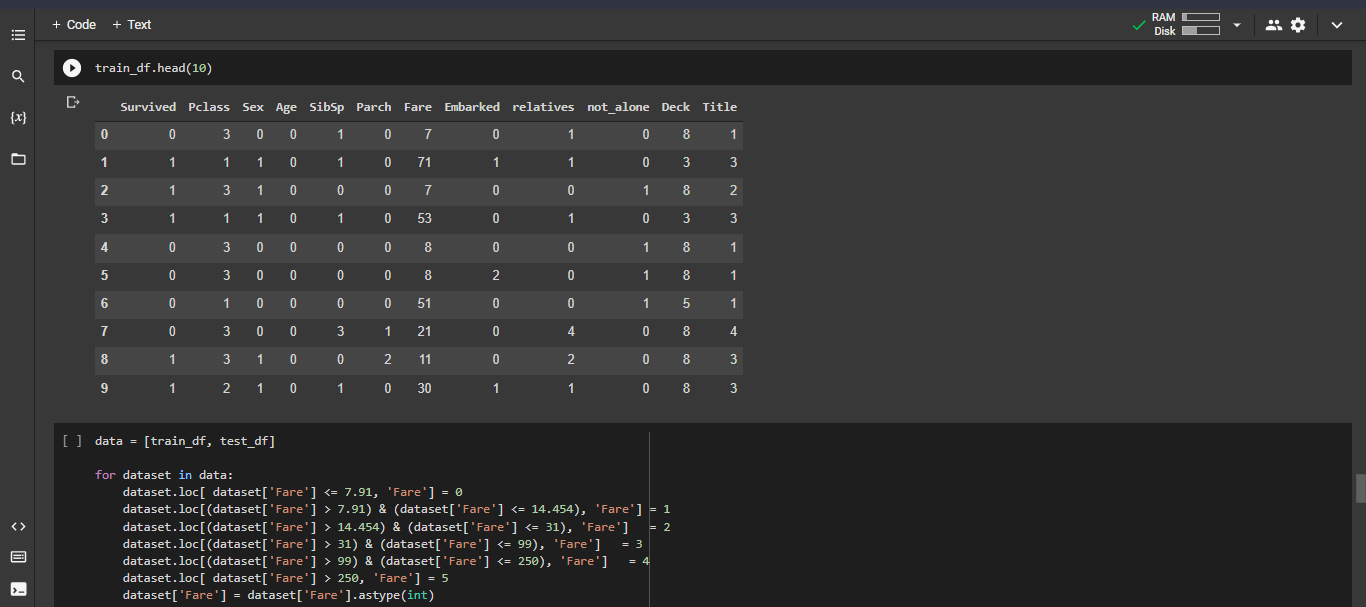
* **AGE:**

First, I will convert the **AGE** feature. And I will convert it from float into integer. Then I will create the new **AGEGROUP** variable, by categorizing every age into a group.



* **FARE:**

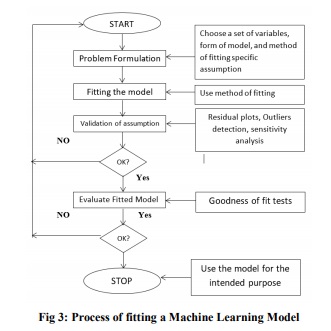
For the **FARE** feature, we need to do the same as with the ‘Age’ feature. This process is quite hard because if we cut the range of the fare values into a few equally big categories, more than 75% of the values would fall into the first category. Luckily, we can use sklearn qcut( )function, that we can use to see, how we can form the categories.



**METHODOLOGY**

**PROCESS FLOW:**

There is a step-by-step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed



**Feature Engineering**

The majorly significant aspect of data analytics is feature

engineering. It acts towards choosing characteristics

implemented during training and therefore during predictions.

Domain awareness is used in feature engineering to identify

aspects in the data set that assist in creating a model for

machine learning. When it comes to modelling, it assists in

interpreting the data collection. A bad selection of features

may result in a less precise and appalling predictive model.

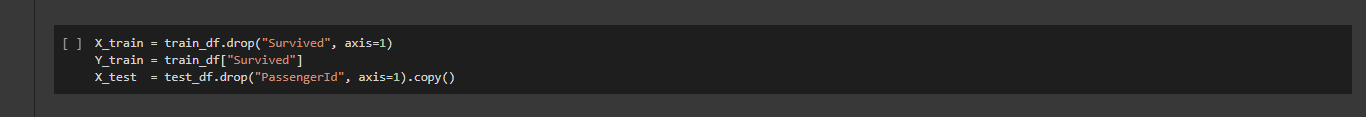
Reliability and predictive capacity are dependent upon

selection of accurate characteristics. It strains out the

unutilized and inessential functions.

The majorly significant aspect of data analytics is feature engineering. It acts towards choosing characteristics implemented during training and therefore during predictions. Domain awareness is used in feature engineering to identify aspects in the data set that assist in creating a model for machine learning. When it comes to modelling, it assists in interpreting the data collection. A bad selection of features may result in a less precise and appalling predictive model. Reliability and predictive capacity are dependent upon selection of accurate characteristics. It strains out the unutilized and inessential functions.

Using the age of a person, their sex, title, cabin number, Pclass, place of embarkation, fare and size of family which includes both parch as well as sibsp, the following features on the exploratory study is done. The survival column is selected as the reaction column. Such characteristics are chosen because their ideals have an effect on the survival rate. If wrong features are chosen, bad predictions may be generated. Thus, in creating an effective predictive model, function engineering works as a backbone. I will train several Machine Learning models and compare their results. Note that because the dataset does not provide labels for their testing-set, we need to use the predictions on the training set to compare the algorithms with each other. The majorly significant aspect of data analytics is feature engineering. It acts towards choosing characteristics implemented during training and therefore during predictions. Domain awareness is used in feature engineering to identify aspects in the data set that assist in creating a model for machine learning. When it comes to modelling, it assists in interpreting the data collection. A bad selection of features may result in a less precise and predictive model. Reliability and predictive capacity are dependent upon selection of accurate characteristics. It strains out the unutilized and inessential functions. Using the age of a person, their sex, title, cabin number, Pclass, place of embarkation, fare and size of family which includes both parch as well as sibsp, the following features on the exploratory study is done. The survival column is selected the reaction column. Such characteristics are chosen because their ideals have an on the survival rate. If wrong features are chosen, bad predictions may be generated. Thus, in creating an effective predictive model, function engineering works as a backbone. NaN, NAN and na are some of the columns that contain empty values. Only four columns, namely age, deck, embarked and embarked\_town have a few absent values. In an attempt to display a few unwanted columns, the next step is to have a look every column’s value name as well as count. The non-redundant rows and columns should be dropped, and their rows and missing values should be removed. It is also advisable to drop the column called deck as it has 688 missing rows of data so an overall of 77.2% of all the data is missing for this column. A look at the data types will tell which all columns are required to be encoded with a digit. Embarked and sex are ideally are two of the columns that must be altered. Later, only distinct values coming from the non-numeric data will be printed while changing them into numeric ones and then printing new values. The data needs to be split twice. First, into independent ‘X’ and dependent ‘Y’ data sets. During the second time, it needs to be split into 80% training and 20% testing data sets. Optionally, it is also possible to scale the data within a specific range. Afterwards there is a need to design a function that within it possesses all the different machine learning models that could potentially come to aid when making decisions about which one to choose. After storing all the models in a variable called model, I found that the Decision Tree Classifier had the most accurate result of 99.29%. The model that came out to reveal the most was Logistic Regression Model with an accuracy only 81.11% on the testing data. However the ultimate chosen model was the Random Forest Classifier. This particular model at position 6 was especially chosen because it had a precision of a whopping 97.53% and 80.41% on the training data and testing data respectively.



**Machine Learning Models**

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

**STOCHASTIC GRADIENT DESCENT:**

Stochastic gradient descent is a method to find the optimal parameter configuration for a machine learning algorithm. It iteratively makes small adjustments to a machine learning network configuration to decrease the error of the network. Error functions are rarely as simple as a typical parabola. Stochastic gradient descent attempts to find the global minimum by adjusting the configuration of the network after each training point.

**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 independent variables is done with the help of logistic regression. It is used to solve binary classification problem, some of the real-life 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 if a given user will buy an insurance product or not.

**Random Forest Classifier:**

This algorithm is specifically used for supervised

classification. Here, the classifier generates forests with a

huge number of trees. With an increase number of trees in the

forest, more reliable can the expected outcomes be[10]. For

both classification and regression problems, the Random

Forest Algorithm may be used for e.g. In order to construct a

model, it will take 5 randomly chosen initial variables from a

random sample of 100 observation. After reiterating the same

over many times, a final decision is generated

Random forest algorithm is supervised classification algorithm. The algorithm basically makes forest with large number of trees. The higher the number of trees in the forest gives the higher accuracy results. Random forest algorithm can be used for both classification and regression problems. For instance, it will take random samples of 100 observation and 5 randomly chosen initial variables to build a model. The same process is repeated a number of times, then the final prediction is made according to the observations. Final prediction is a function (mean) of each prediction.

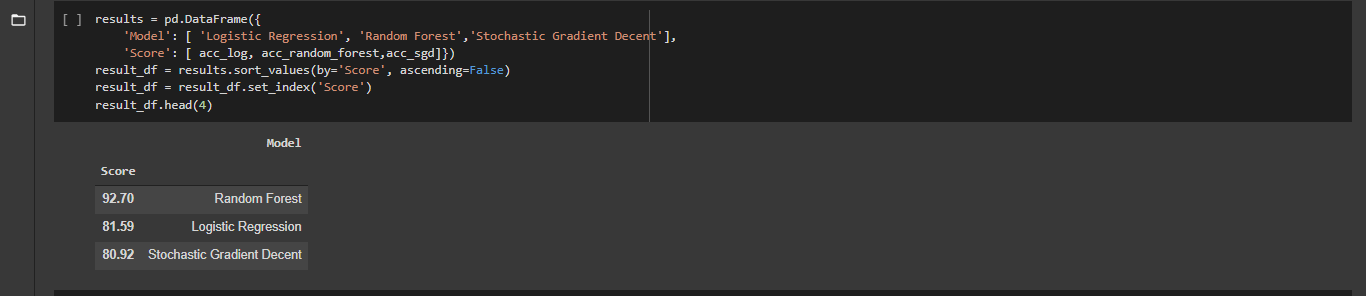
.

****

**MODEL EVALUATION:**

The accuracy of the model is evaluated using “confusion matrix”. A confusion matrix is a ta

Table layout that allows to visualize the correctness and the performance of an algorithm.

****

This is an algorithm for supervised learning. Generally,

classification problems put in use of this. Ideal for input and

output variables, both categorical and continuous. A split

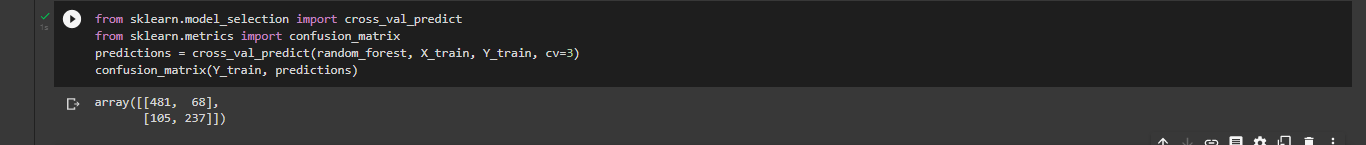
point variable and a single input variable (x) are represented

by any root node. For leaf nodes, the dependent variable (y) is

present

**CONFUSION MATRIX:**

A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number 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 such models is commonly evaluated using the data in the matrix.



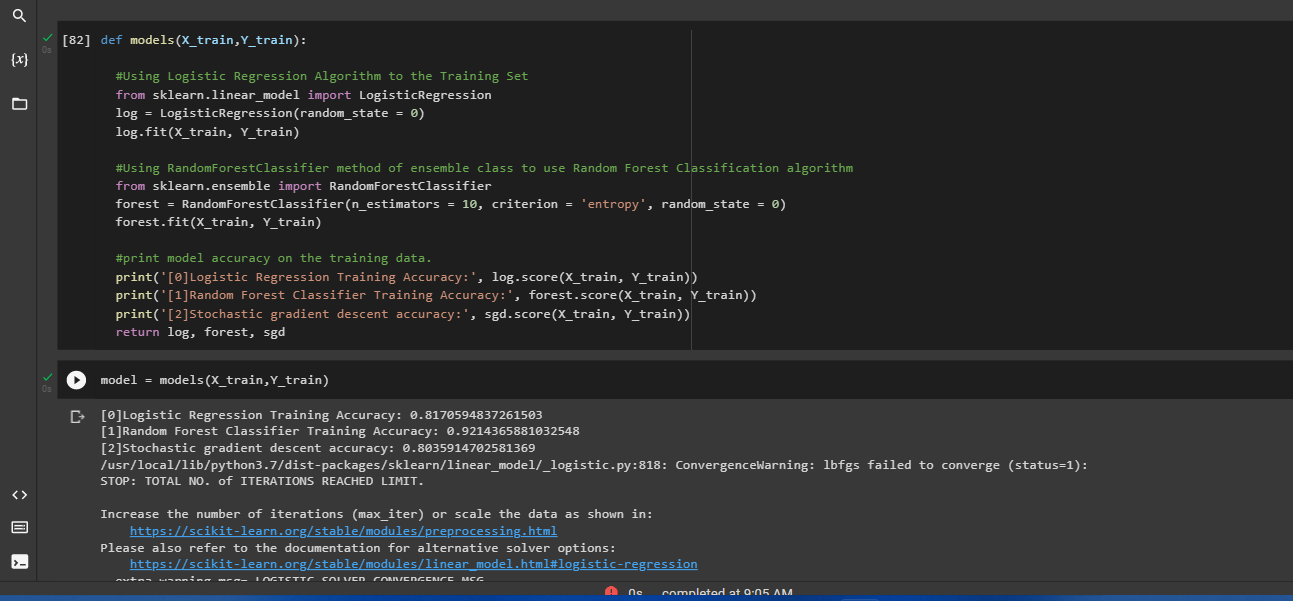
Here the first row is about the not-survived-predictions where 481 passengers were correctly classified as not survived (called true negatives) and 68 where wrongly classified as not survived (false positives).

The second row is about the survived-predictions where 105 passengers were wrongly classified as survived (false negatives) and 239 where correctly classified as survived (true positives).

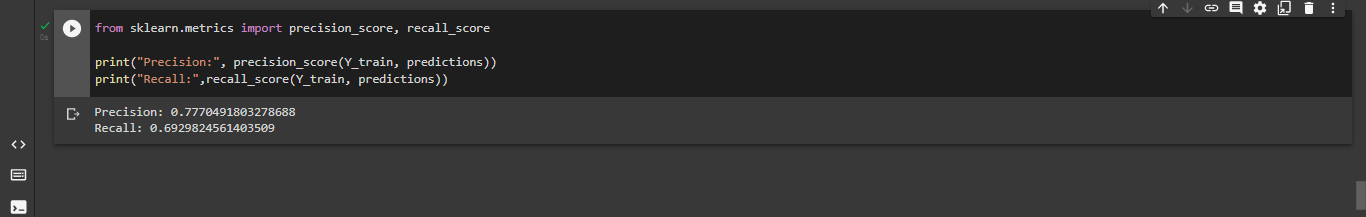
A confusion matrix gives you a lot of information about how well your model does, but there is a way to get even more, like computing the classifiers precision.

**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”



**PRECISION AND RECALL:**

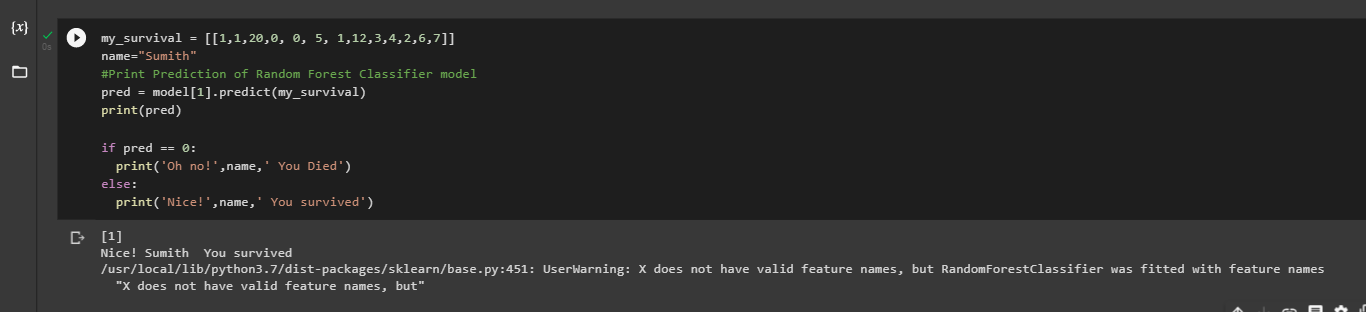
****

The model predicts 78% of the time, a passenger’s survival correctly (precision). The recall tells us that it predicted the survival of 70 % of the people who actually survived.

**PREDICTION:**

Here we can choose any of the models to predict survival of test sample. Since we have evaluated all models by using confusion matrix.

**MY SURVIVAL:**



Here, I am checking whether an individual would have survived in the sinking of the titanic or not.

**WORK:**

I designed a predictive model using python and with help of libraries like pandas, NumPy, seaborn, mat.lib. I used an open-source application “google Collaboratory” to compile the python code. Using seaborn package, titanic data set has been loaded and the whole file and data was analysed by using count plots and graphs which contains data set with many variables like age, gender, embark, p-class, etc... Then we compare the variables for better understanding. my model predicts if we would survive the titanic with actual data provided so that we can enhance the prediction accuracy for individual input which led to accuracy of 90% using model “random forest classifier”. After analysing data, I have created few different models like logistic regression, Stochastic gradient descent, random tree classifier, etc... to check the prediction accuracy of given data. After thorough analysis, Random Forest classifier has optimal prediction accuracy. So, we used Random Forest classifier as our source model.

This is an algorithm for supervised learning. Generally,

classification problems put in use of this. Ideal for input and

output variables, both categorical and continuous. A split

point variable and a single input variable (x) are represented

by any root node. For leaf nodes, the dependent variable (y) is

present.

Using the age of a person, their sex, title, cabin number,

Pclass, place of embarkation, fare and size of family which

includes both parch as well as sibsp, the following features on

the exploratory study is done. The survival column is selected

as the reaction column. Such characteristics are chosen

because their ideals have an effect on the survival rate. If

wrong features are chosen, bad predictions may be generated.

Thus, in creating an effective predictive model, function

engineering works as a backbone

**CONCLUSION:**

In order to further improve the overall result, an extensive

hyperparameter should be tuned on several machine learning

models. It is possible to further improve it by doing ensemble

learning. This research paper began with data exploration, and

subsequently led to checking about absent data and learning

what features are important. At the arrival of the data pre-

processing portion, missing values were computed and

converted features into numeric ones. Later a few more

features were created. 8 different machine learning models

were also simultaneously trained thereby and applied cross

validation onto the chosen Random Forest model. Finally, the

confusion matrix was looked into and evaluated as well as the

f-score, computer model’s precision and recall. When

attempting data analysis, data cleaning is the first and

foremost step. Exploratory data analytics allows one to

recognize the dataset and the dependency linking

characteristics. EDA is implemented in order to find out the

connection between the features of the dataset. This is

achieved with the use of different graphical techniques. Some

conclusions are drawn by applying EDA and the facts are

discovered

In order to further improve the overall result, an extensive hyperparameter should be tuned on several machine learning models. It is possible to further improve it by doing ensemble learning. This research paper began with data exploration, and subsequently led to checking about absent data and learning what features are important. At the arrival of the data pre-processing portion, missing values were computed and converted features into numeric ones. It would be interesting to play more with dataset and introducing more attributes which might lead to better results. Various other machine learning techniques like Naive Bayes, K-NN classification can be used to solve the problem. Later a few more features were created. 8 different machine learning models were also simultaneously trained thereby and applied cross validation onto the chosen Random Forest model.

Data cleaning is the first step while performing data analysis. Exploratory data analytics helps one to understand the dataset and the dependency among the attributes. EDA is used to figure out the relationship between the features of the dataset. This is done by using various graphical techniques. The one used above is ggplot and histograms. By applying EDA some conclusions are drawn and facts are found. There is high influence of age on survival

Finally, the confusion matrix was looked into and evaluated as well as the f-score, computer model’s precision and recall. When attempting data analysis, data cleaning is the first and foremost step. Exploratory data analytics allows one to recognize the dataset and the dependency linking characteristics. EDA is implemented in order to find out the connection between the features of the dataset. This is achieved with the use of different graphical techniques. Some conclusions are drawn by applying EDA and the facts are discovered. Data cleaning is the first step while performing data analysis. Exploratory data analytics helps one to understand the dataset and dependency among the attributes. EDA is used to figure out the relationship between the features of the dataset. This is done by using various graphical techniques. The one used above is histograms. By applying EDA some conclusions are drawn and facts are found. There is high influence of age on survival. We can see from table-2 that as age increases survival decreases. It can be seen that survival rate of female is very high and survival rate of male is very low.

Using the age of a person, their sex, title, cabin number,

Pclass, place of embarkation, fare and size of family which

includes both parch as well as sibsp, the following features on

the exploratory study is done. The survival column is selected

as the reaction column. Such characteristics are chosen

because their ideals have an effect on the survival rate. If

wrong features are chosen, bad predictions may be generated.

Thus, in creating an effective predictive model, function

engineering works as a backbone

**FUTURE ENHANCEMENT**

This project involves implementation of data analytics and machine learning. This project work can be used as reference to learn implementation of EDA and machine learning from very basic level. In future, the idea can be extended by making more advanced graphical user interface with the aid of newer libraries like shiny in R. An interactive page can be created, i.e., if the value of an attribute is varied on the scale, then the values corresponding to its graph (plot or histogram) will also change. It will be helpful to draw much focused conclusions by combining results we obtained

**REFERENCES**

Segal, M. R. (2004). Machine learning benchmarks and random forest

regression.

1. Segal, M. R. (2004). Machine learning benchmarks and random forest regression.
2. Dr. Neeraj Bhargava, Girja Sharma, Decision Tree Analysis on J48 Algorithm for Data Mining. Volume 3, Issue 6, June 2013.
3. MICHAEL AARON WHITLEY, using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015.