**TITANIC SURVIVAL**

**PREDICTION**

****

**PRESENTED BY:**

Vikas sharma

Reshmitha Reddy

Satti vijay naga sri dhananjay reddy

**SUBMITTED TO:**

Skill Vertex, Akash Maurya

**TABLE OF CONTENT:**

->Goals and Objectives

->Software Required

->Introduction

->Implementation

->importing the necessary libraries

->importing the dataset

->Preprocessing

->EDA

->Visualization

->preparing the data

-> Split the data

->scale the data

->Building various model

-> Finding accuracy of each model

->conclusion

**GOALS AND OBJECTIVES:**

->The purpose of this project is to document the process we went through to create our predictions for Titanic survivor prediction

->The objective of this project was to build a classification model that could successfully determine whether a titanic passenger lived or died

**SOFTWARE REQUIRED:**

**TOOLS USED:**

Anaconda Navigator-1.9.6

Jupyter Notebook-5.7.4

**LIBRARY USED:**

Analyzing :Numpy,Pandas,sci-kit learn

Visualization:Matplotlib,Seaborn

**INTRODUCTION:**

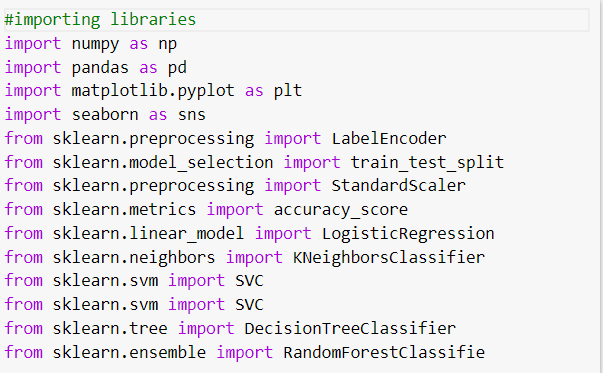
->The sinking of the RMS Titanic is one of the most infamous shipwrecks in history

->in this challenge,we ask you to complete the analysis of what sort of people were likely to survive

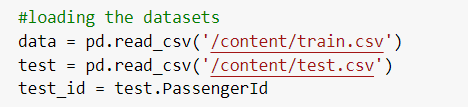
->in particular,we ask you to apply the tools of machine learning to predict which passengers survived the tragedy

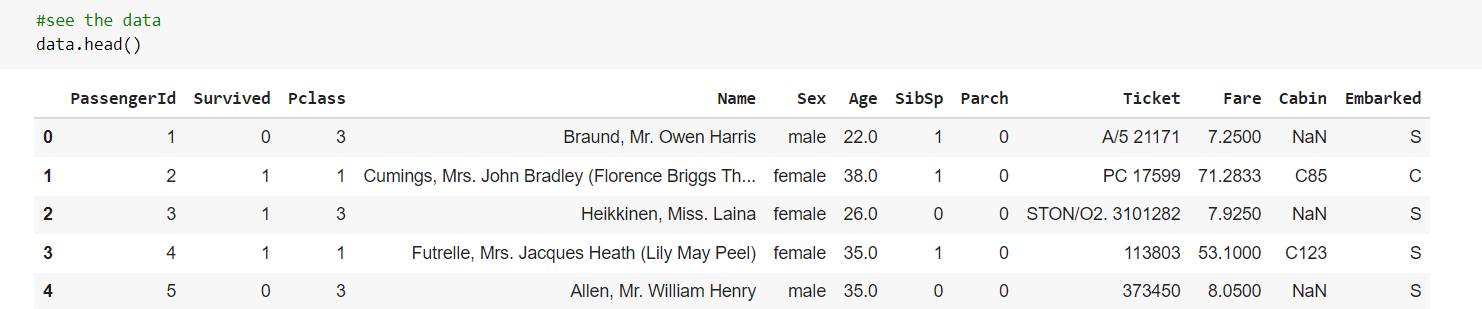
**IMPLEMENTATION:**

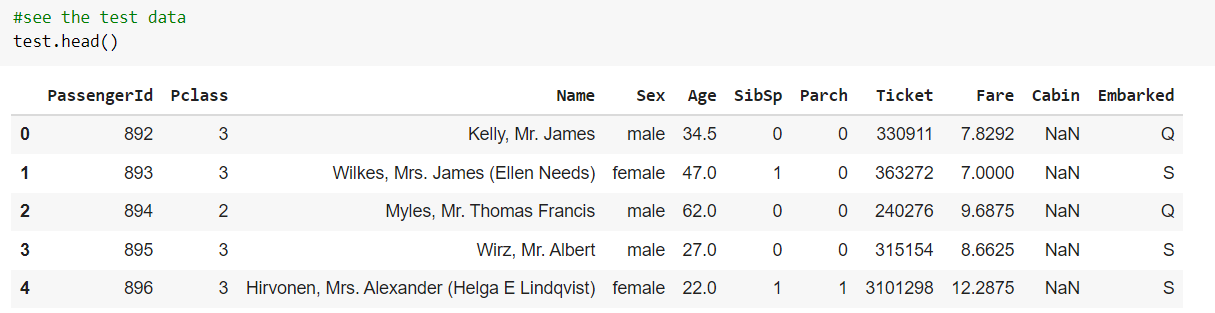
**Importing the necessary libraries**



**Importing the Dataset:**







**Decription about the dataset:**

pclass: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

survived: Survival (0 = No; 1 = Yes)

name: Name

sex: Sex

age: Age

sibsp: Number of siblings/spouses aboard

parch: Number of parents/children aboard

fare: Passenger fare (British pound)

embarked: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

adult\_male: A male 18 or older (0 = No, 1=Yes)

deck: Deck of the ship

who: man (18+), woman (18+), child (<18)

alive: Yes, no

embarked\_town: Port of embarkation ( Cherbourg, Queenstown, Southampton)

class: Passenger class (1st; 2nd; 3rd)

alone: 1= alone, 0= not alone ( you have at least 1 sibling, spouse, parent or child on board)

When we look at the train and test data set we can see that there is no “survived” column in the test dataset because that is the column we need to predict by using our model

**EXPLORATORY DATA ANALYSIS (EDA)**

1. Analysis of the survived and not survived from the dataset by using a plot graph

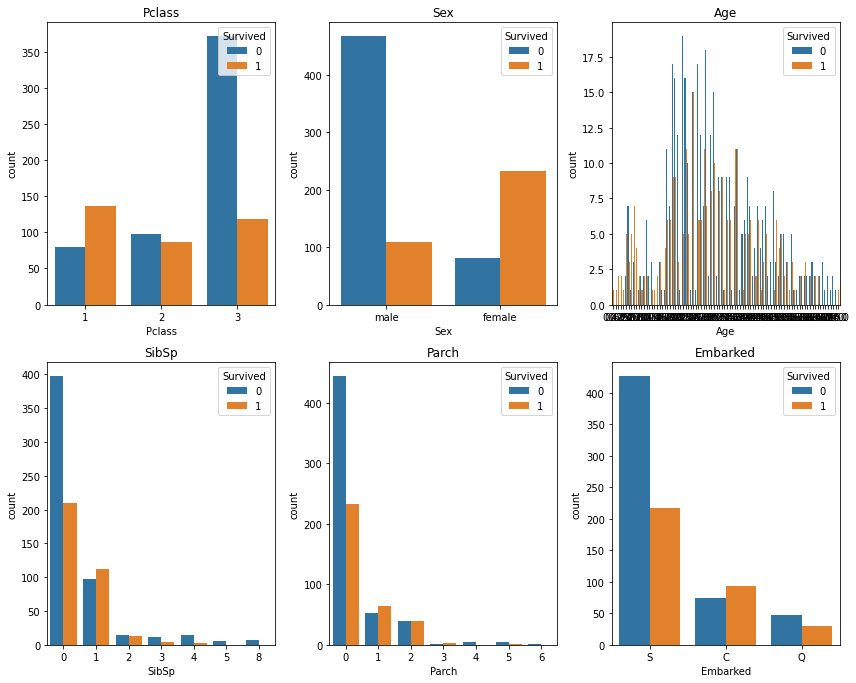


Here 0 indicates the number of passengers who havn’t survived

And 1 indicated the number of passengers who have survived

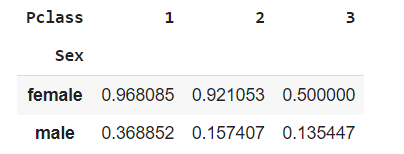
From the graph it is evident that there are more passengers that havn’t survived they are 549 passengers who havn’t survived and 342 passengers who have survived

If we dig deep into the analysis of survived rate against few columns in dataset ie.”Pclass”,”Sex”,”Age”,”SibSp”,”Parch”,”Embarked”



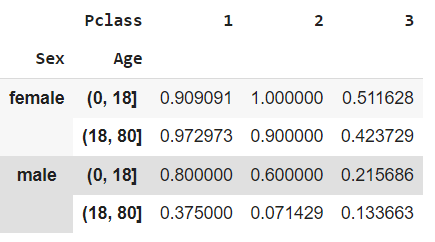
When we look at the survival rate by sex it is evident that more female have survived than male passengers

Now if we dig deep and try to know which Pclass was found more save for the passengers



It seems to be that 1st class is more save for both male and female and also from all the classes it is clear that female from 1st class are more save when compared to other classes for both female and male

The survival rate for a particular age group ie age(0-18yr)and (19-80yr)

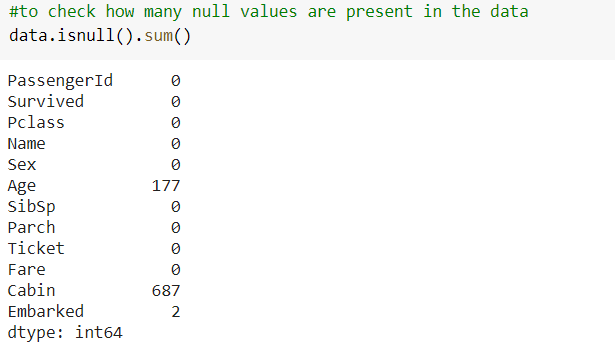


Female of age between 19-80yr of pclass 1 are having high survival rate when compared to the female of all age groups and pclass

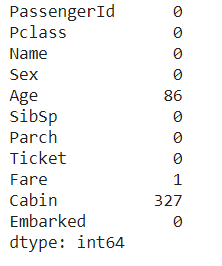
Whereas male of age between 0-18yr of 1st class are having high survival rate

By all EDA done so far we can conclude that passengers of pclass 1 are more save

Let us now check if our data has any null values



When we used isnull().sum() function it returned the number of null values in each columns and the columns Age,cabin and embarked are having 177,687and 2 null values respectively in the data dataset



When we used isnull().sum() function it returned the number of null values in each columns and the columns Age,cabin and fare are having 86,327and 1 null values respectively in the test dataset

We will fill the null values by using the mean value of the respective column

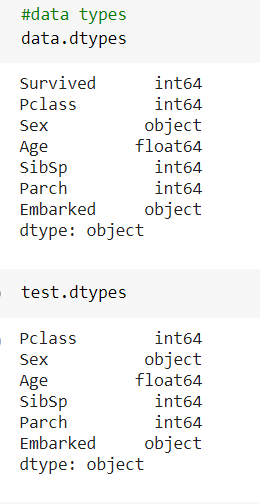
We will fill only age,embarked column

When it comes to the cabin column the error percentage is more and if we fill the cabin column it will be over-fitting

So dropping the cabin column in both datasets is the best approach

And in addition to it we would like to drop even PassengerId ,name,ticket and fare column because they do not contribute in the survival prediction

And cleaning the datasets lets check the data types in both datasets



The data type which is object indicates that the column is not a numeric column and non-numeric data cannot be used in the model ,so we should either drop those columns or else change these column into numeric data

**PREPARING THE DATA**

**SPLIT THE DATA**

For our model we have taken variable X and y where

X contains the columns all the columns other then survived column and

y contains survived column

here we have taken the test size as 0.2 because the data in the dataset is not huge

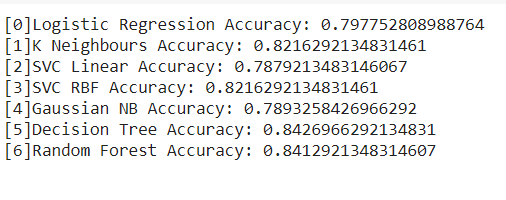
**SCALE THE DATA**

To scale the data we have choosen StandarScaler ove MinMaxScaler because standarscaler follows standard normal distribution therefore,it makes mean=0 and scale the data to unit variable whereas minmaxscaler scales all the data features in the range[0,1] or else in the range[-1,1] if there are negative values in the dataset

**PREPARING VARIOUS MODELS**

The models we developed are logistic regression,KNeighboursclassifier,svc(linear kernel),svc(RBF kernel),gaussianNB,decision tree,randomforestclassifier

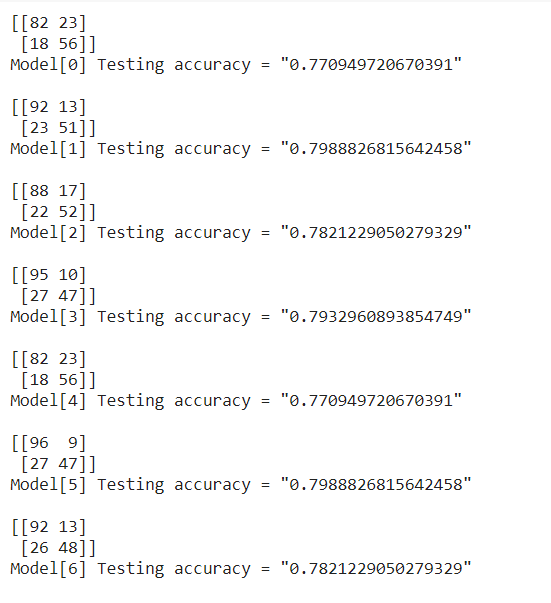
The accuracy we got for each model is



This accuracy is for training data

We can see that the accuracy for Decision tree is more when compared to other models

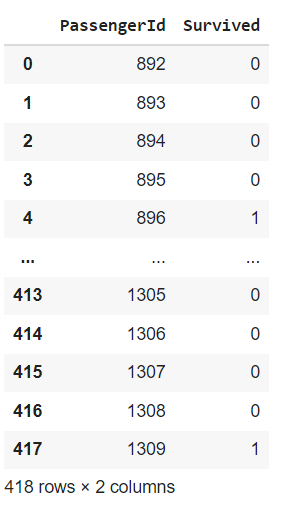
When we check the accuracy of models on our testing data using confusion\_matrix



The KNN and Decision tree are the highest and equal

But when we analyse the accuracy of both train and test dataset then it can be said that decision tree is giving the highest in train set and is more accurate

Now when we predict the values by using decision tree model then



Here this model says that more people have died compared to the survived one

**CONCLUSION**