TITANIC SURVIVED PROJECT-2

SUBMITTED BY:-

NEHA VIBHOR MITTAL

DS2301

PROBLEM IDENTIFICATION -

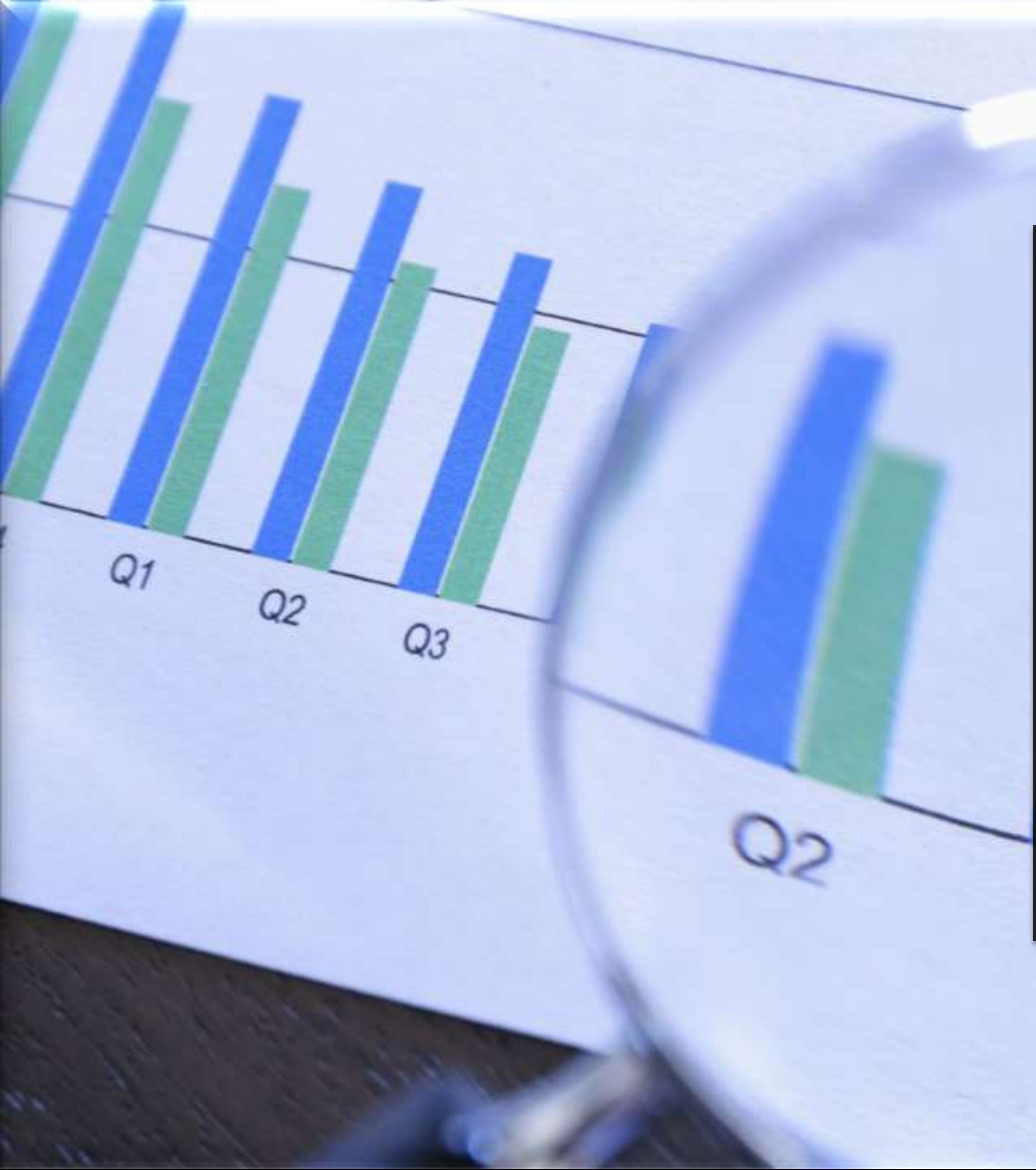
- The Titanic Problem is based on the sinking of the 'Unsinkable' ship Titanic in early 1912.
- It provides information on the fate of the passengers on the Titanic ship, summarized according to economic status(class),sex,age and survival.
- Based on these features, you have to predict if an arbitrary passenger on Titanic would survive the sinking or not.

OBJECTIVE -

The objective of this project is to build a classification model(binary classification) that would successfully determine whether a Titanic passenger got survived or not.

This Dataset includes over 891 records and

12 attributes.



IMPLEMENTATION-

- Importing necessary libraries .
- Importing the dataset.
- Exploring,Cleaning and analysing the data.
- Building the model.
- Using various different algorithms to find out the prediction.
- Performing Cross-validation technique.
- Hyperparameter Tunning for the best model.
- Plotting Roc_auc curve

IMPORTING NECESSARY LIBRARIES-

IMPORTING NECESSARY LIBRARIES-

```python
#importing the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

Importing the dataset.

```
#Loading the dataset

ts = pd.read_csv('C:/Users/nehas/NehaProject/Titanic_Survived Dataset/titanic_train.csv')
ts
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Exploring,Cleaning and analysing the data.

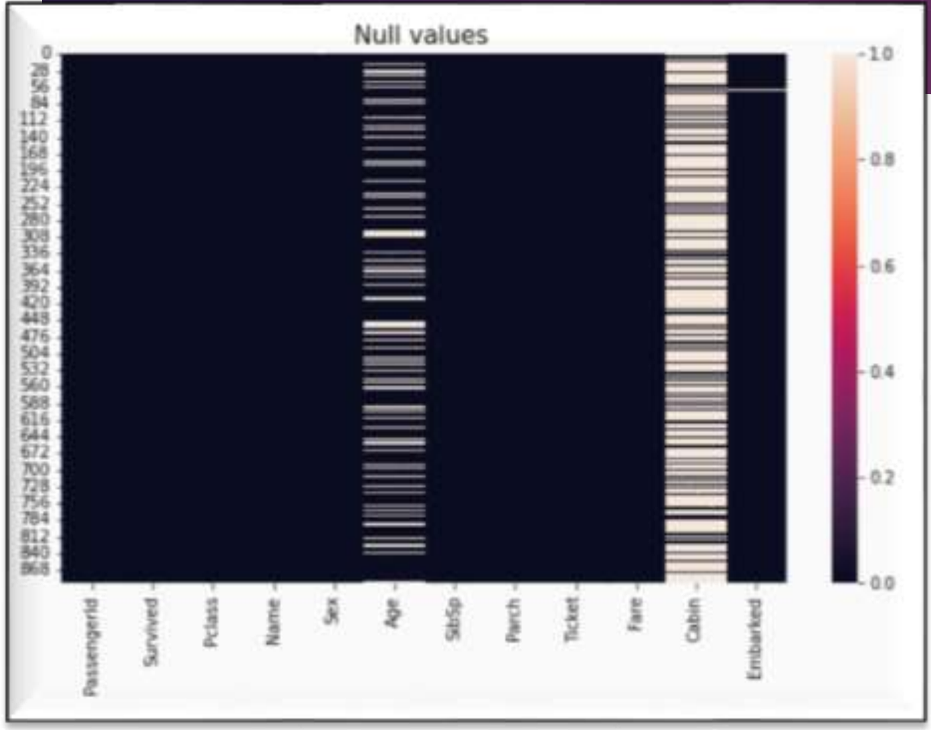
```
# Getting statistical summary

ts.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

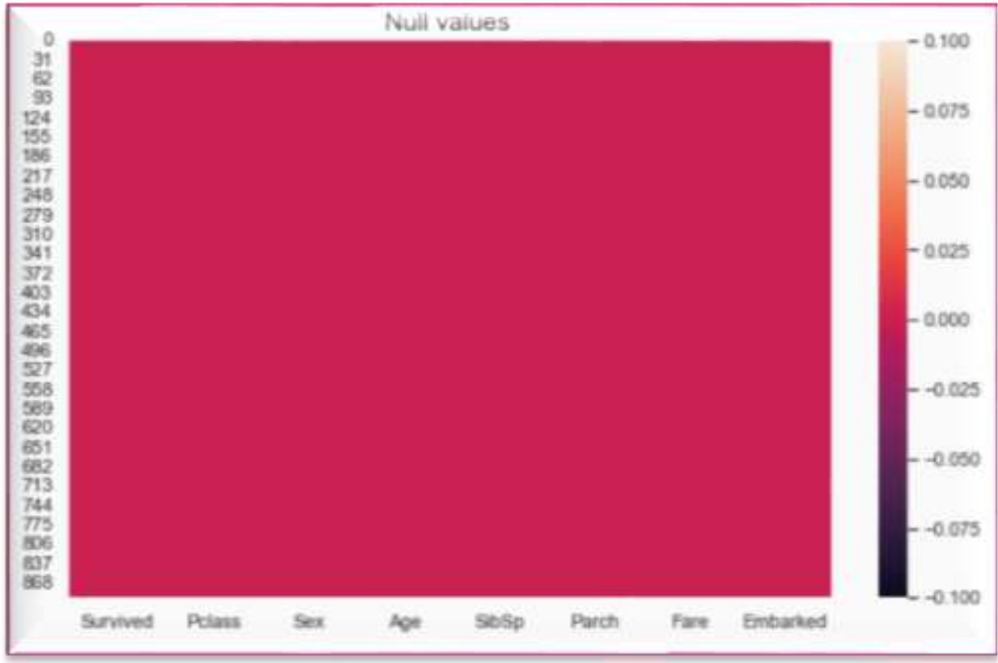
Key Observations -

- With the about data , we can find that Total samples are 891 ,also we can detect some features that contain missing values, like the 'Age' feature (714 out of 891 total).
- Age is normally distributed but 'fare' is right skewed (mean>median>mode).
- As the difference between 75% , standard deviation and max value is very huge , this indicates that the Outliers could also present in 'fare' Attribute .

CLEANING THE NULL VALUES



Removing null values →

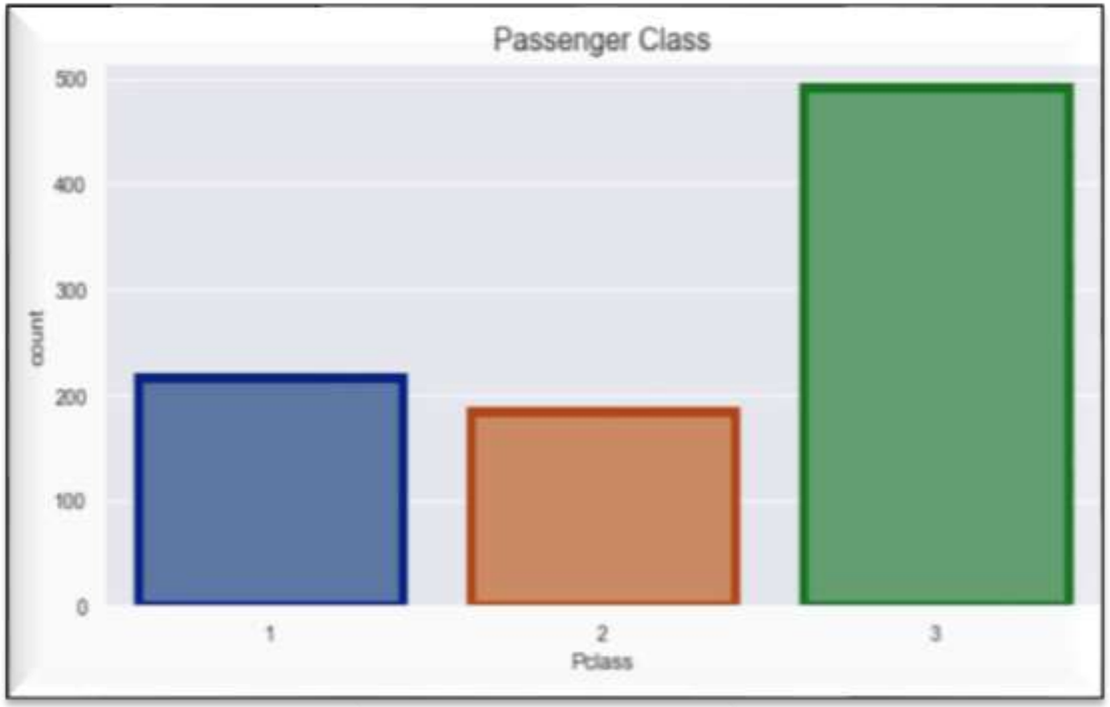


DATA VISUALIZATION-

Analysing the data

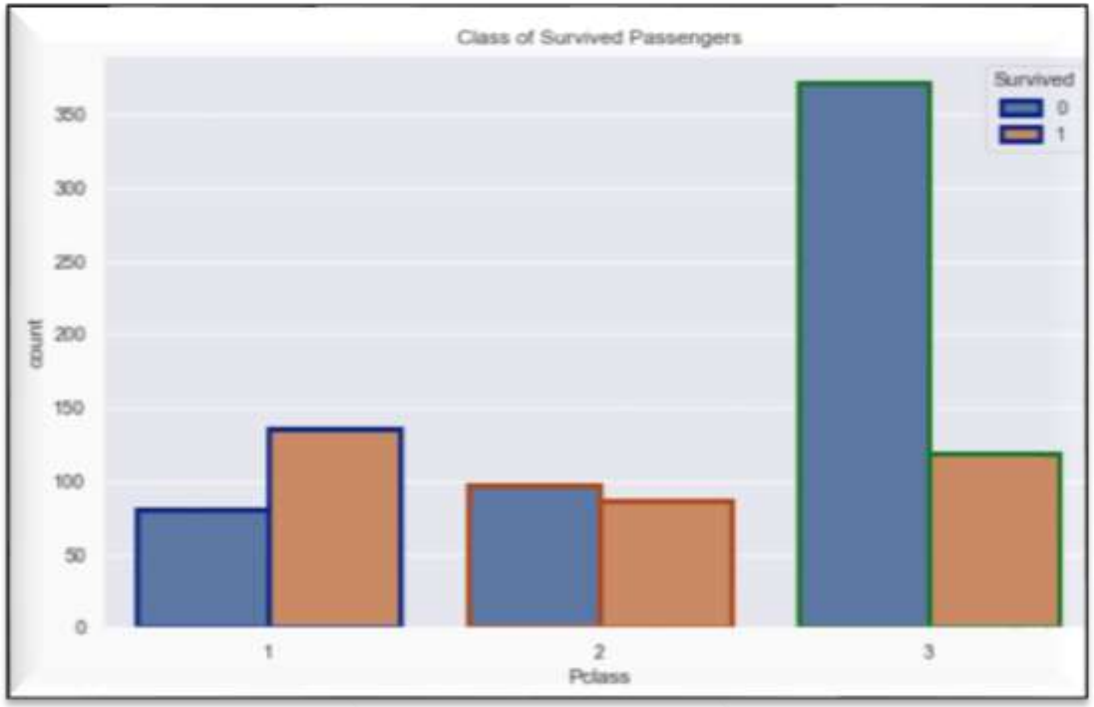
Pclass-

This is a uneven distribution.Passengers in 1st class and 2nd class have almost even distribution while in 3rd class distribution is much higher.



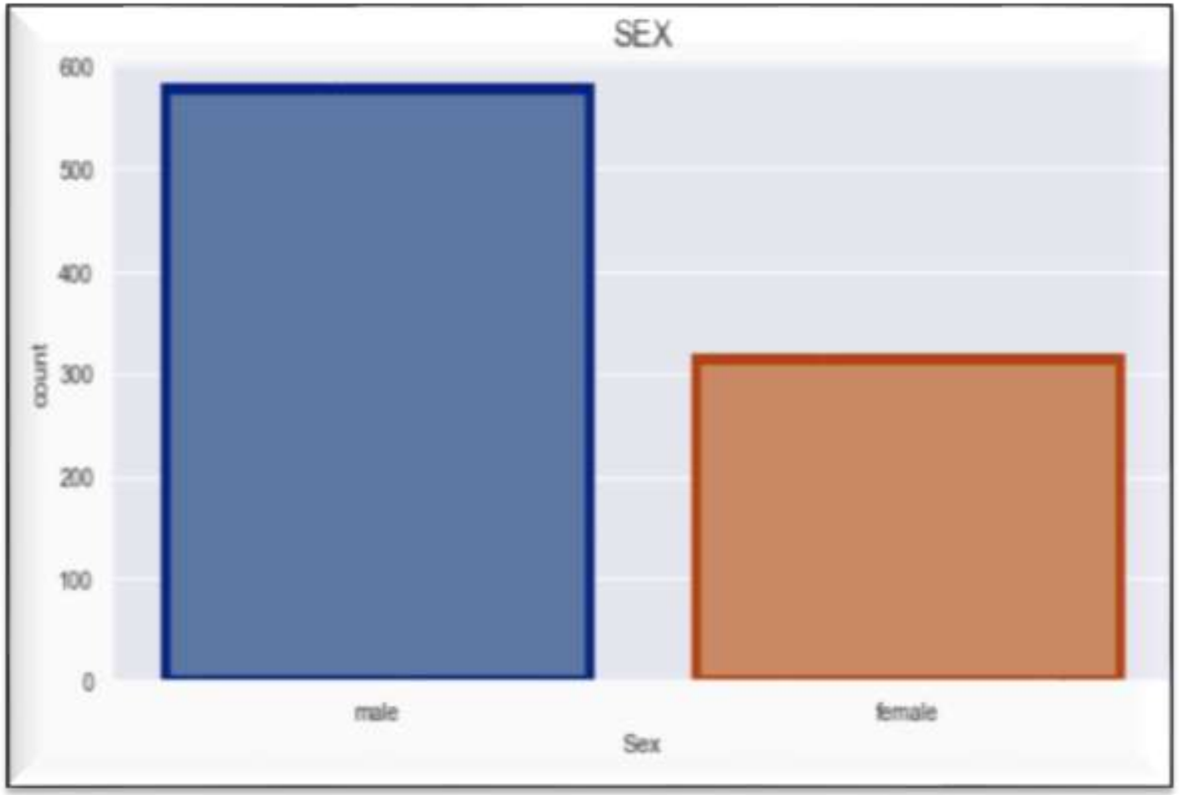
Pclass Survived-

The wealthy people who belongs to 1st class survived mostly whereas , people who bought ticket of 3rd class died mostly.



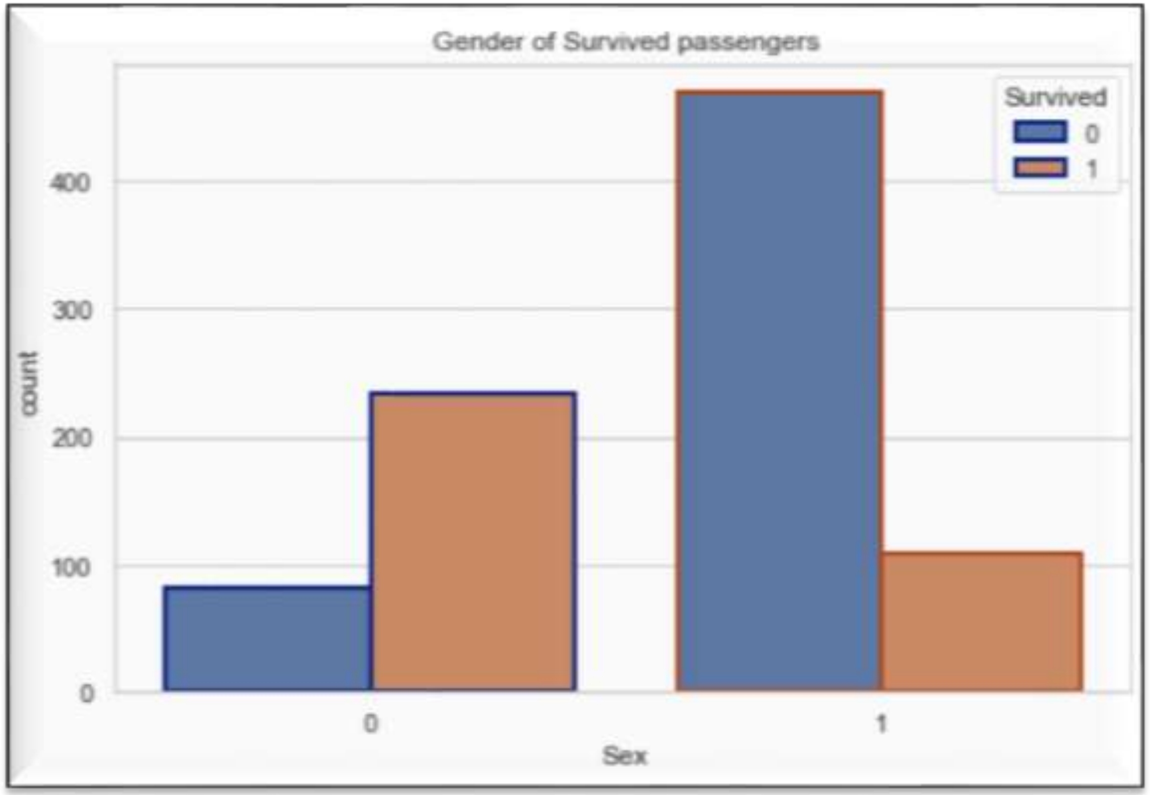
GENDER-

Number of male passengers are higher than the Female passenger



GENDER SURVIVED-

More number of females survived when compared to males.

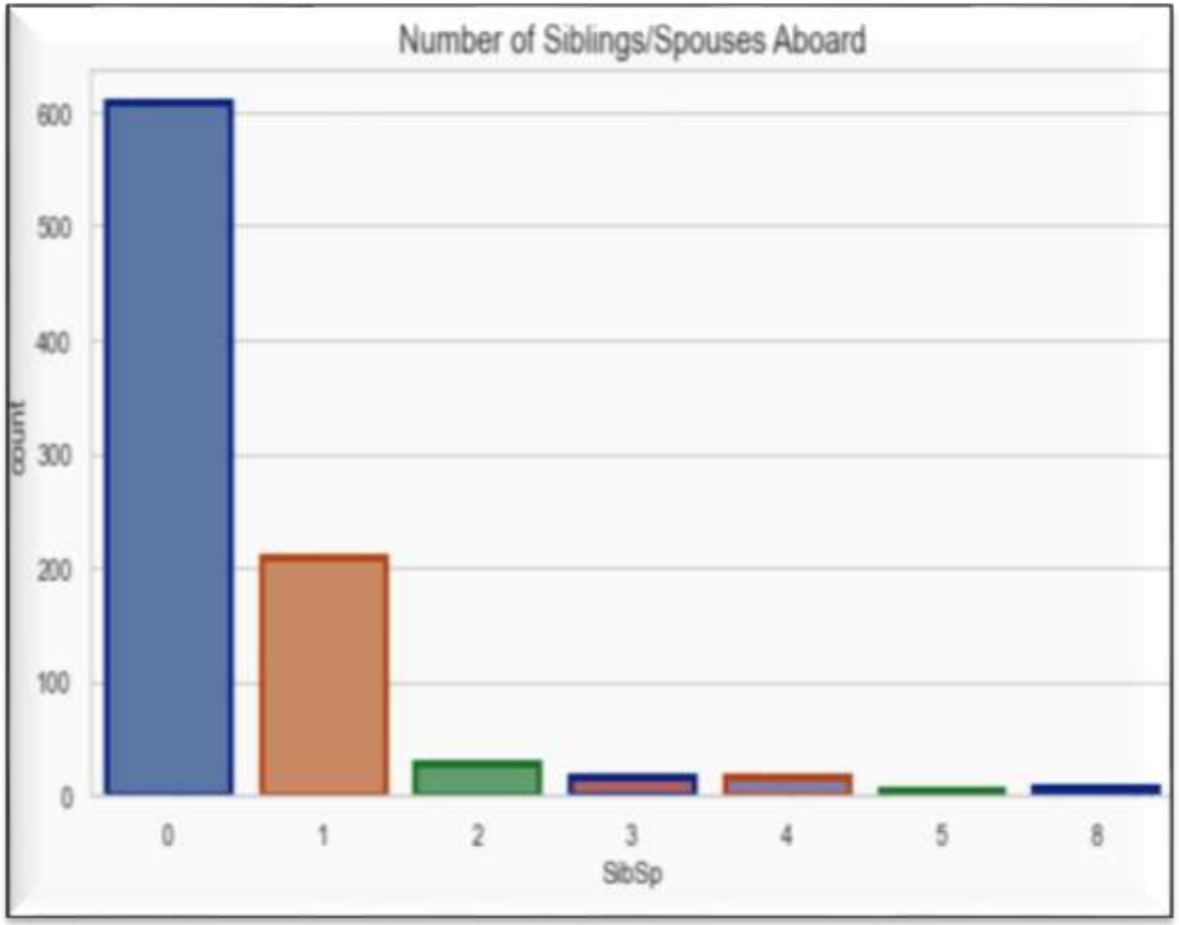


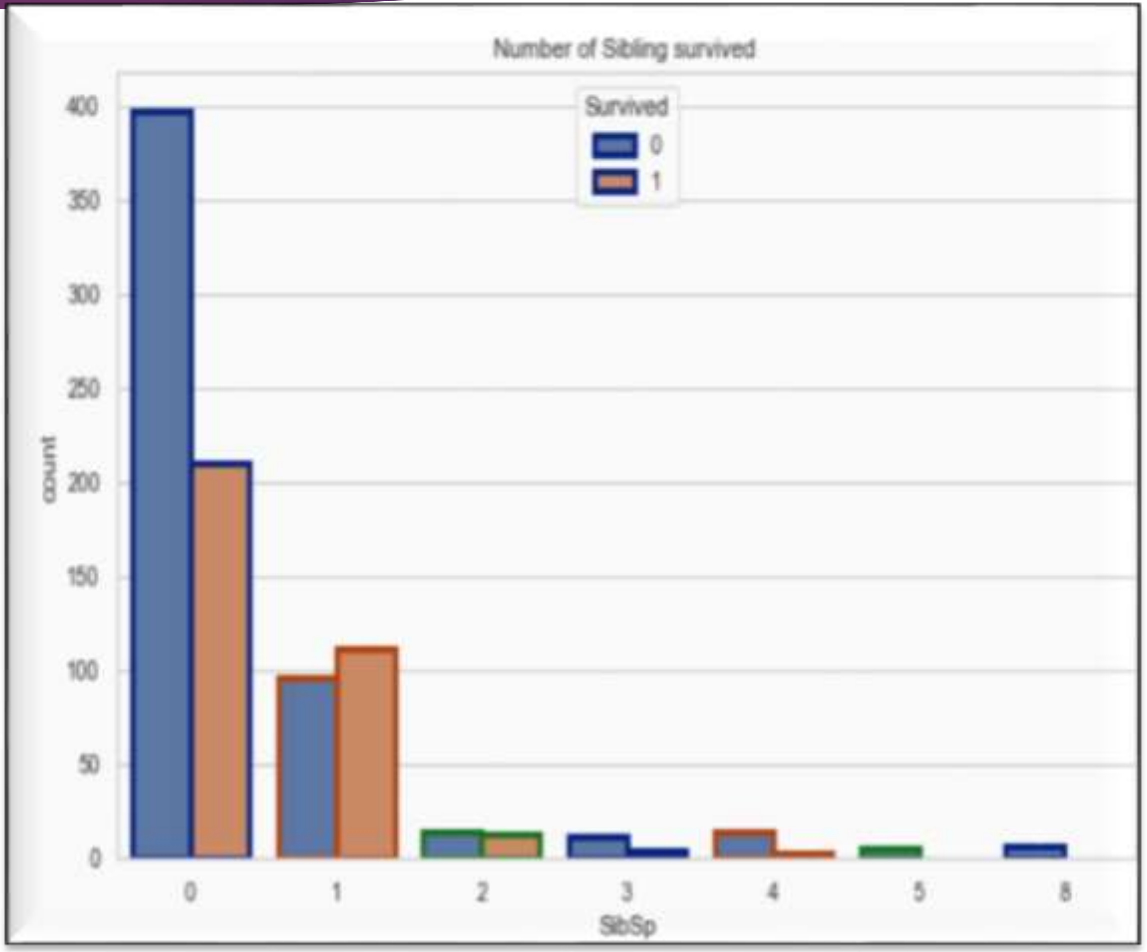
SIBLINGS-

Around 600 people don't have siblings or spouse and around 200 people having 1 sibling or spouse while other people having more that 1 sibling and spouse.

SIBLINGS SURVIVED-

Most familes are with 0 or one sibs who survived morethan those with 2-4 sibilings.





BUILDING MODELS
(using various different algorithms)

MODEL BUILDING

MODEL BUILDING-

```
#for Training-testing data
from sklearn.model_selection import train_test_split

# Models:
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier


# for cross validation
from sklearn.model_selection import cross_val_score,GridSearchCV

# Matrics for Evaluation
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix,roc_auc_score
```

RESULTS-

Key Observations-

The Test accuracy score of all the different models are -

1)Logistic regression - 81.2 %

2)Decision Tree Classifier - 82.7%

3)K-Neighbors Classifier -80.9 %

4)Naive Bayes-82.2 %

5)Support Vector Classifier - 70.9 %

6)Random Forest Classifier - 84.5 %

7)Ada Boost Classifier- 83.6 %

8)Gradient boost Classifier -85.4 %

CROSS-VALIDATION SCORE

```python
#CV Score of ADA boost Classifier -

score = cross_val_score(adb,x,y,cv=5)
print(score)
print(score.mean())

print("Accuracy score :", accuracy_score(y_test,predadb))

print(f"CV Score of ADA:{cross_val_score(adb,x,y,cv = 5).mean()*100:.2f}%")
print('\n')
print('The difference between accuracy score and Cross Validation score is:',accuracy_score(y_test,predadb)-score.mean())
```

```
[0.73636364 0.83181818 0.76363636 0.91324201 0.87671233]
0.8243545039435449
Accuracy score : 0.8363636363636363
CV Score of ADA:82.44%


The difference between accuracy score and Cross Validation score is: 0.012009132420091384
```

HYPERPARAMETER TUNING

Hyperparameter Tuning of ADA boost Classifier ¶

```python
adb=AdaBoostClassifier()
```

```python
param={'algorithm' : ['SAMME.R','SAMME'],
    'n_estimators':[10,25,50,100],
    'learning_rate':[0.1,0.5,1.0]}
```

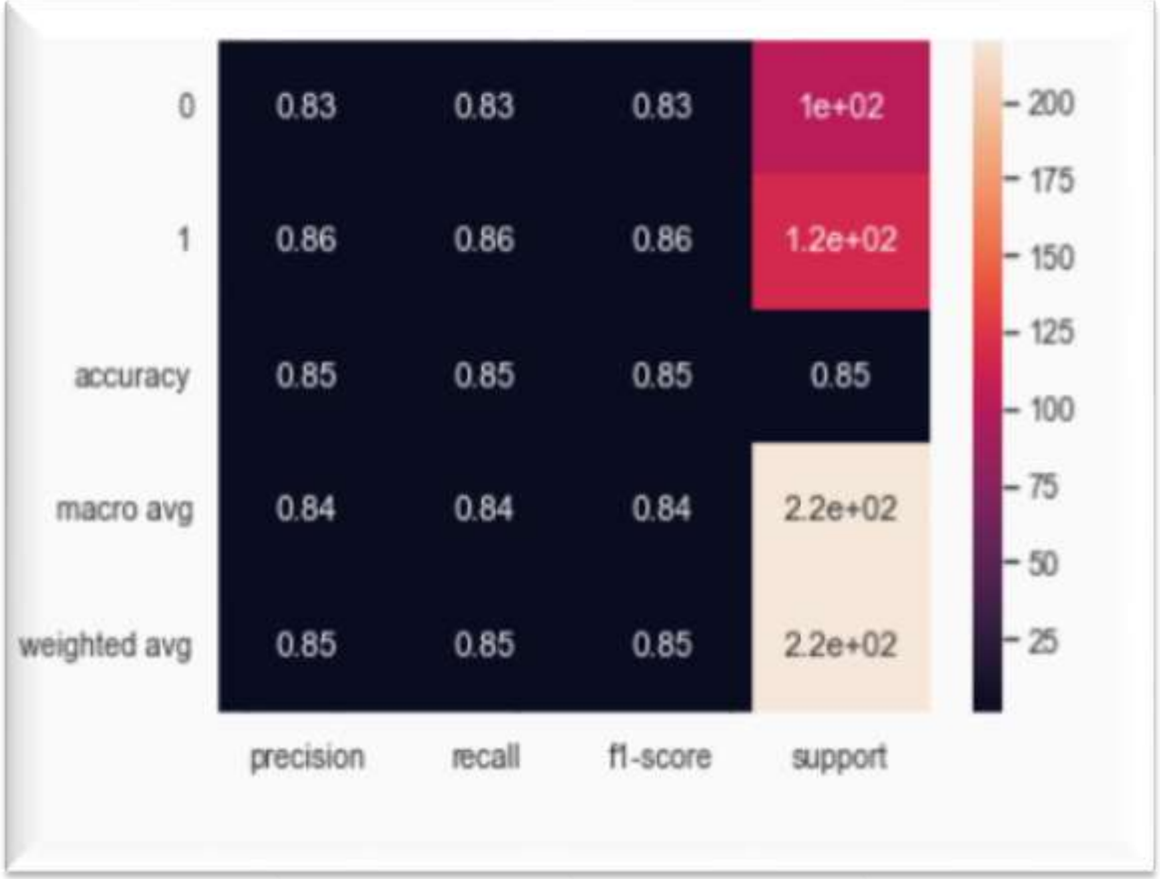
```python
adb_grid=GridSearchCV(AdaBoostClassifier(),param,cv=5,scoring='accuracy')
adb_grid.fit(x_train,y_train)
adb_pred=adb_grid.best_estimator_.predict(x_test)
print("Accuracy after parameter tuning::",accuracy_score(y_test,adb_pred))
adb_grid.best_params_
```

```
Accuracy after parameter tuning:: 0.8454545454545455

{'algorithm': 'SAMME.R', 'learning_rate': 1.0, 'n_estimators': 100}
```

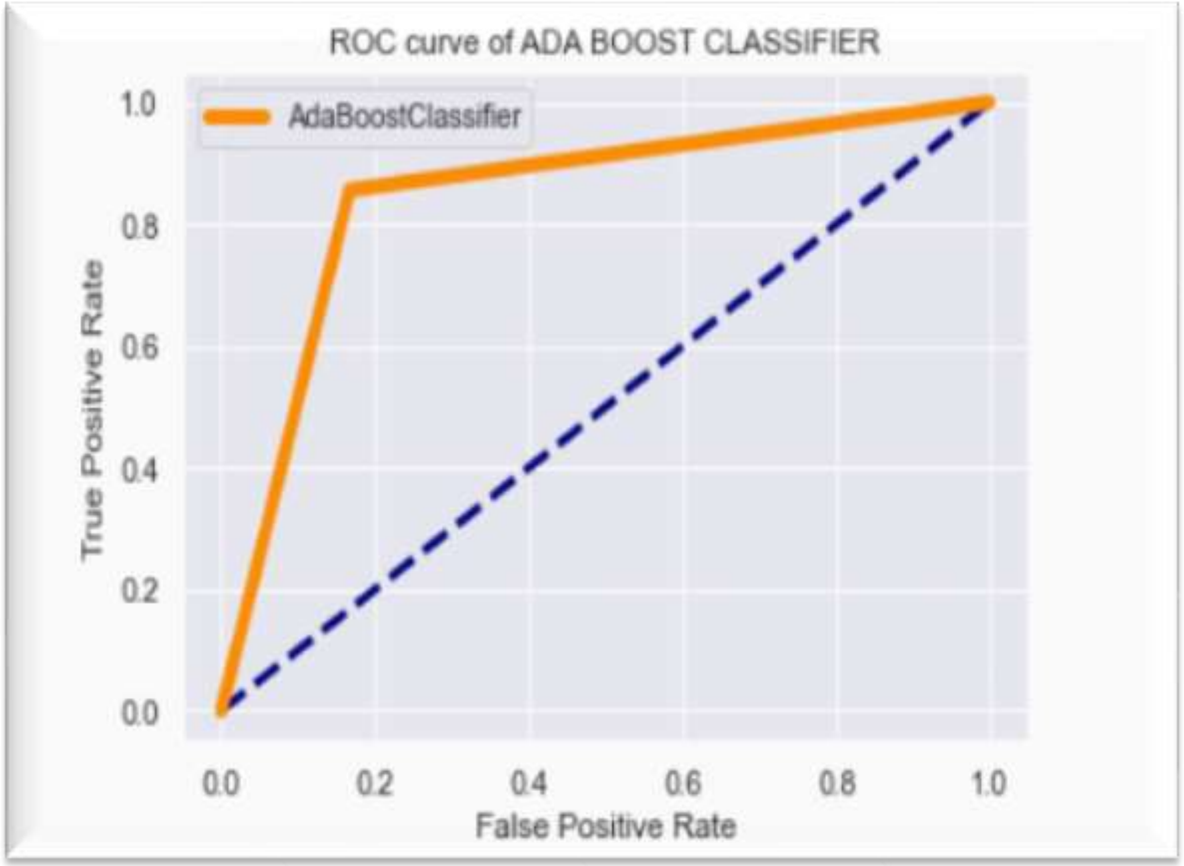
```python
Final_model = AdaBoostClassifier(algorithm ='SAMME.R',learning_rate=1.0,n_estimators= 100)
Final_model.fit(x_train,y_train)
pred=Final_model.predict(x_test)

print('\n')
print('Accuracy Score',accuracy_score(y_test,pred)*100)
print('\n')
print('Confusion Matrix')
print(confusion_matrix(y_test,pred))
print('\n')
print('Classification Report')
print(classification_report(y_test,pred))
print('\n')
print('Roc_auc Score',roc_auc_score(y_test,pred))
```

MODEL EVALUATION-

CLASSIFICATION REPORT



ROC CURVE



CONCLUSION

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.

During the data exploration where we checked about missing data and learned which features are important. During this process, we used seaborn and matplotlib to do the visualizations. The data preprocessing part, we computed missing values, converted features into numeric ones, grouped values into categories and created a few new features.

Afterwards we started training 8 different machine learning models, picked one of and applied cross validation on it. Then we discussed how the selected model works and tuned it's performance through optimizing it's hyperparameter values.

Lastly, we looked at it's confusion matrix and computed the models precision, recall and f-score.

As a result of our work, we gained valuable experience of building prediction systems and achieved our best score for the model.

REFERENCES:-

Learning repository:-
' 'https://github.com/dsrscientist/dataset1/blob/master/titanic_train.csv''

Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-Titanic Machine Learning From Disaster, 2012.