TITANIC DATA SET ANALYSIS USING MACHINE LEARNING

PROJECT MEMBERS

OLIVE DEY

SAGNIK MONDAL

PUSPITA DAS

PRITHA ROY

SUBHROJIT GHOSH

INTRODUCTION

- **RMS *TITANIC*** WAS A LUXURIOUS PASSENGER SHIP OPERATED BY THE “WHITE STAR LINE” IN ENGLAND. IT WAS SANK IN THE NORTH ATLANTIC OCEAN IN THE EARLY MORNING HOURS OF 15 APRIL 1912, AFTER STRIKING AN ICEBERG DURING HER MAIDE VOYAGE MAIDEN FROM SOUTHAMPTON TO NEW YORK CITY . RMS *TITANIC* WAS BELIEVED NOT TO SINK ANYTIME BECAUSE OF ITS TWO RUDDER WHICH WAS NEW AT THAT TIME. OF THE ESTIMATED 2224 PASSENGERS AND CREW ESTIMATED ABOARD, MORE THAN 1,500 DIED, MAKING THE SINKING ONE OF MODERN HISTORY'S DEADLIEST PEACETIME COMMERCIAL MARINE DISASTERS.
- NOW WE HAVE ANALYZED DEPENDENCIES THE SURVIVAL RATE WITH DIFFERENT FEATURES OF THE DATASET.

GOAL OF THIS DATA ANALYSIS

- TO MAKE PREDICTION OF SURVIVAL OF PASSENGERS TRAVELING IN THE SHIP ON THE BASIS OF THE FEATURES GIVEN IN THE TITANIC DATASET.
- ANALYZE THE MOST IMPORTANT FEATURES THOSE ARE RESPONSIBLE FOR DIFFERING THE SURVIVAL RATE OF THE CREW OF THE SHIP USING K BEST FEATURE ANALYSIS.
- COMPARISON THE DIFFERENT PREDICTION ACCURACY SCORE AFTER APPLICATION OF DIFFERENT MACHINE LEARNING MODELS.

LIBRARY USED

FOR ANALYZING:

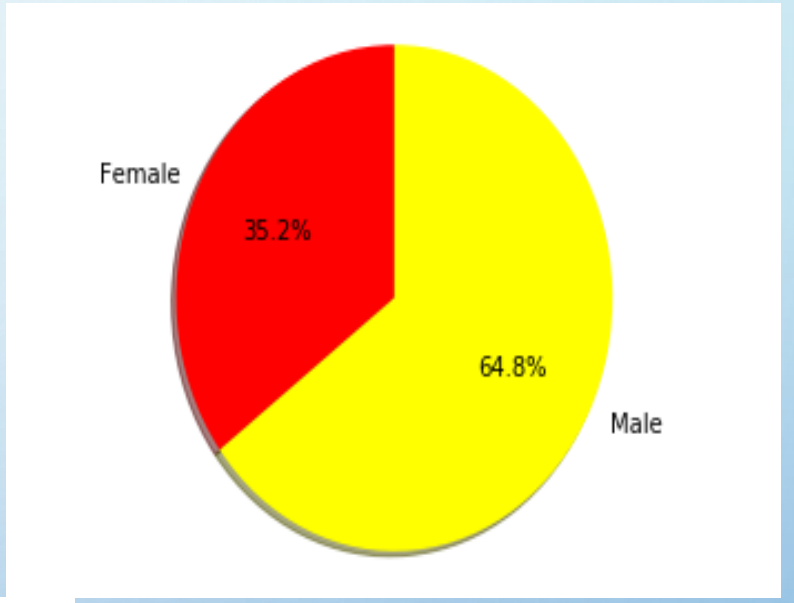
- NUMPY
- PANDAS
- SKLEARN
- MPL_TOOLKITS

FOR VISUALIZING:

- MATPLOTLIB.PYPLOT
- SEABORN
- GRAPHVIZ

PERCENTAGE OF MALE AND FEMALE

- HERE WE HAVE THE DATASET OF 891 PEOPLE CONSISTS OF BOTH MALE AND FEMALE (OUT OF WHICH 577 ARE MALE AND 314 ARE FEMALE). HENCE WE APPLY MACHINE LEARNING ALGORITHM TO FIND RATE OF PERCENTAGE OF MALE AND FEMALE THROUGH A PIE DIAGRAM.

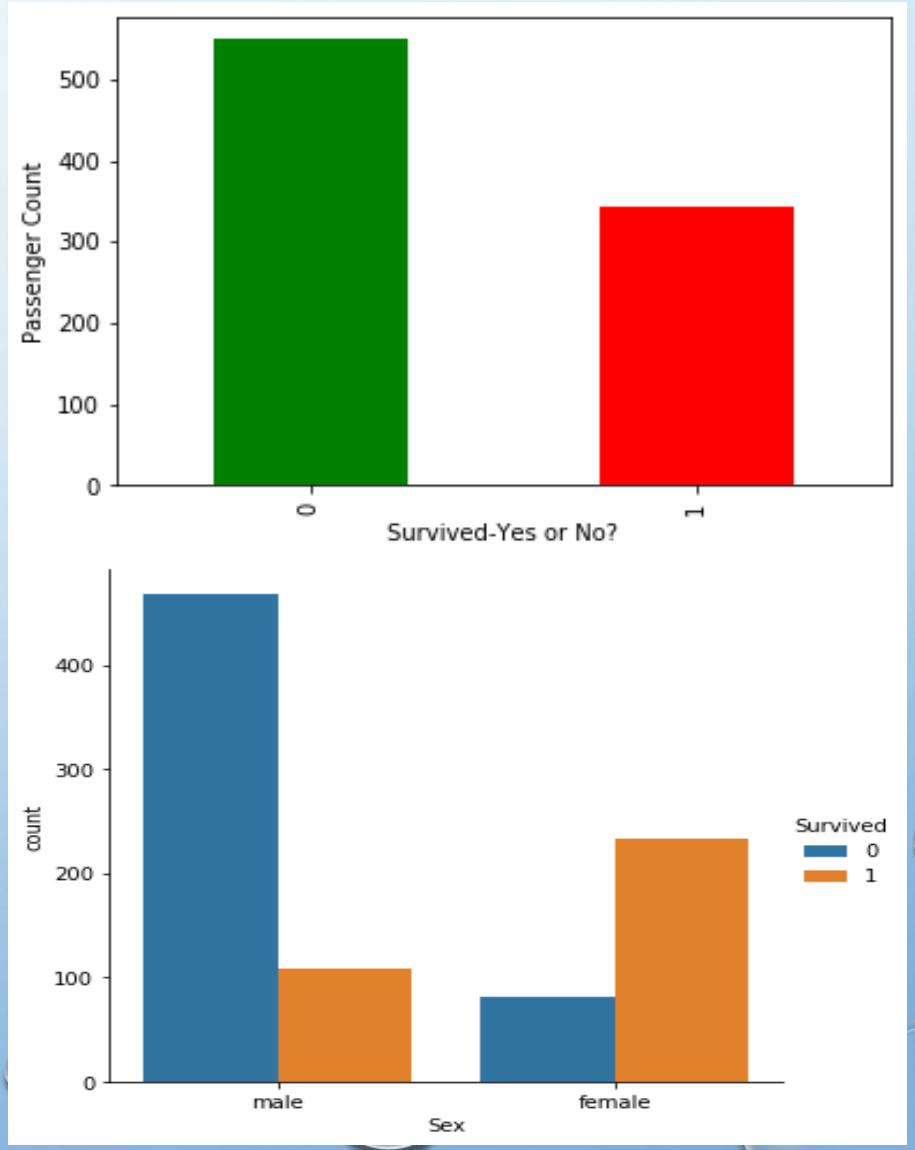



```
189
190 df.Sex.value_counts().sort_index()
191 labels = 'Female', 'Male'
192 sizes = [df.Sex.value_counts().sort_index().values]
193 colors = ['red', 'yellow']
194 plt.pie(sizes,labels= labels, colors=colors,autopct='%1.1f%%', shadow=True, startangle=90)
195 plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
196
```

NUMBER OF DEATH OF MALE AND FEMALE

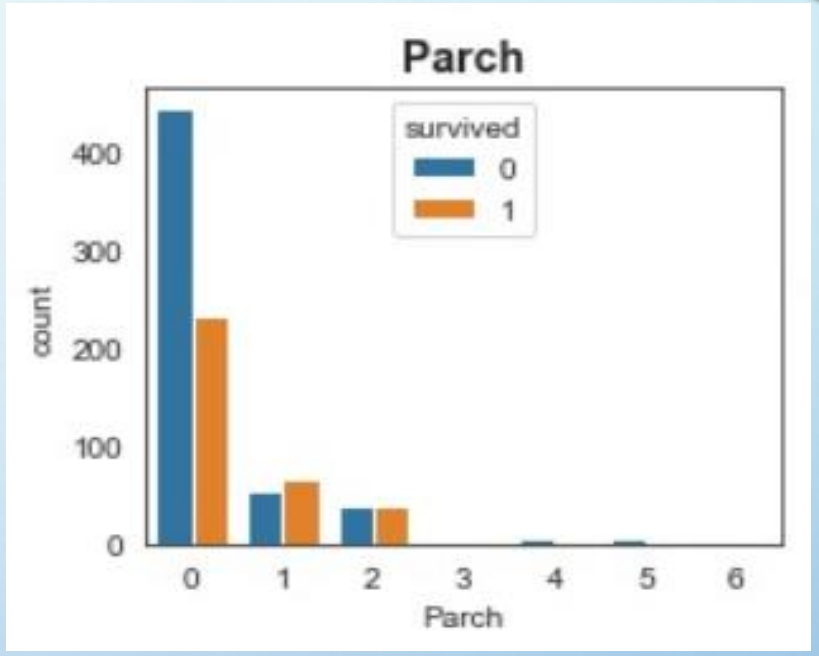
- HERE WE HAVE SHOWED A GROUPED BAR PLOT DENOTING THE DEATH AND SURVIVED OF THE MALE AND FEMALE CREW. FROM THIS WE GET A CONCLUSION THAT THE SURVIVAL OF FEMALE ARE MUCH GREATER THAN THAT OF MALE.

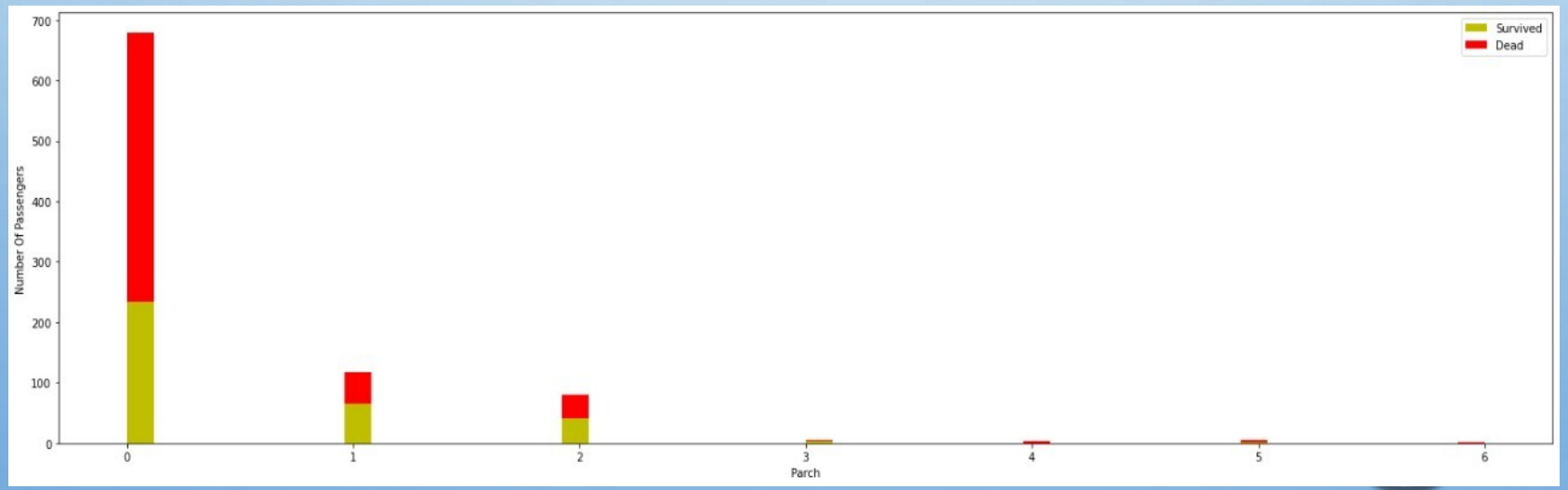
```
28
29 print(df.Survived.value_counts())
30 print(df[["Sex", "Survived"]].groupby("Sex", as_index = False).mean())
```



RELATION OF SERVIVAL WITH PARCH

- IN THE DATASET, THE PARCH COLUMN DENOTES THE PASSENGERS WITH THEIR FAMILY. HENCE WE PLOT THE SURVIVAL OF THE CREW ACCORDING TO PARCH.
- X AXIS : PARCH
- Y AXIS : PASSENGER COUNT
- IT SHOWS THAT THE PASSENGERS WITHOUT FAMILY SURVIVED LESS THAN THE PASSENGERS WITH THEIR FAMILY.



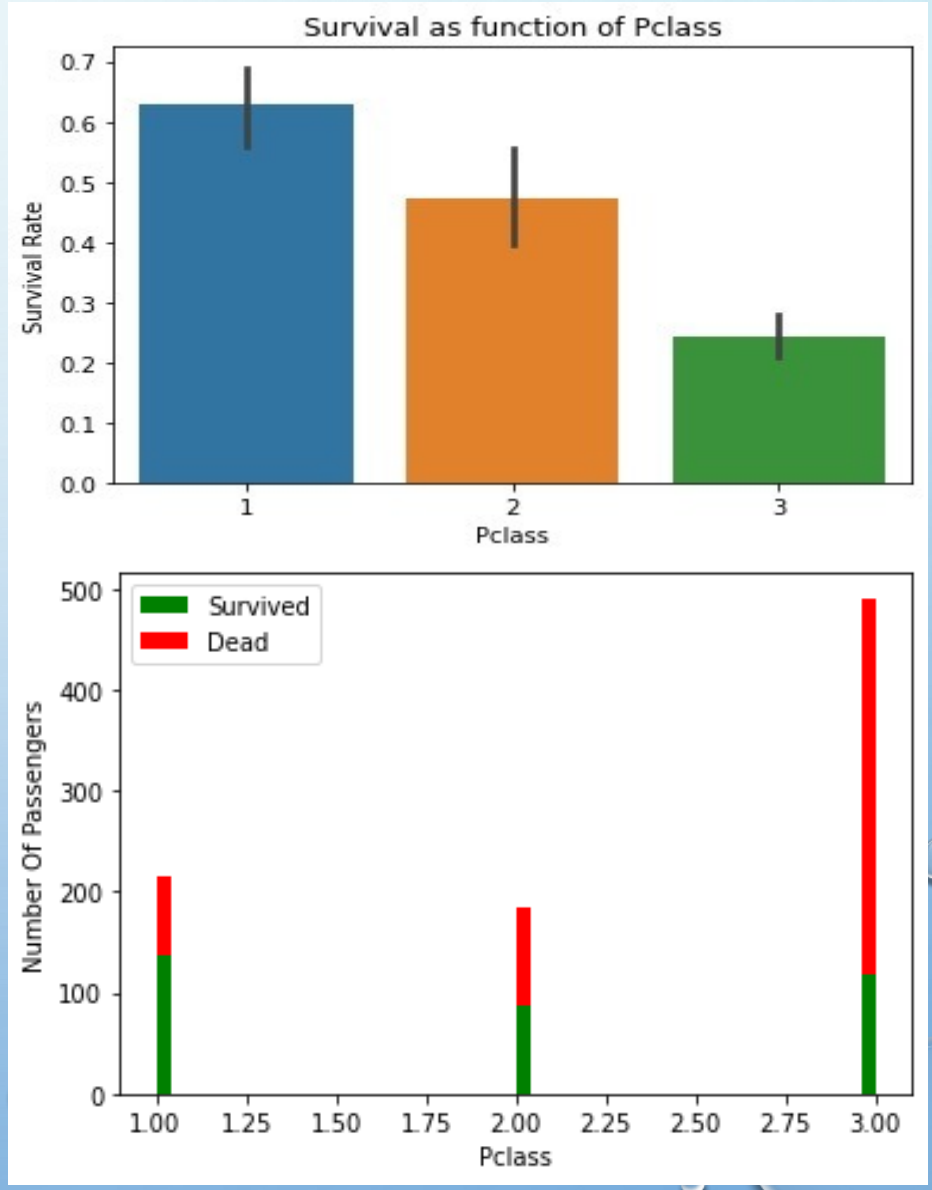


RELATION OF SERVIVAL WITH PCLASS

IN THE DATASET, WE HAVE 3 PCLASSES(1,2,3).HERE WE COMPARE THE SURVIVAL RATE WITH THE PCLASS. THE SURVIVAL RATE OF PCLASS1 IS MUCH HIGHER THAN THE OTHER CLASSES.

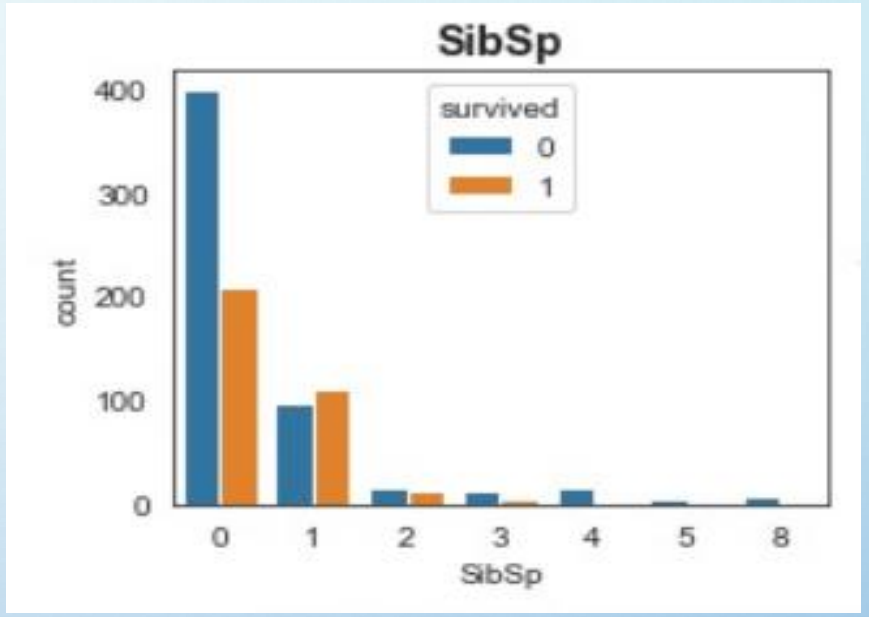
Pclass	Population	Survived	Death
1	216	136	80
2	184	87	97
3	491	119	372

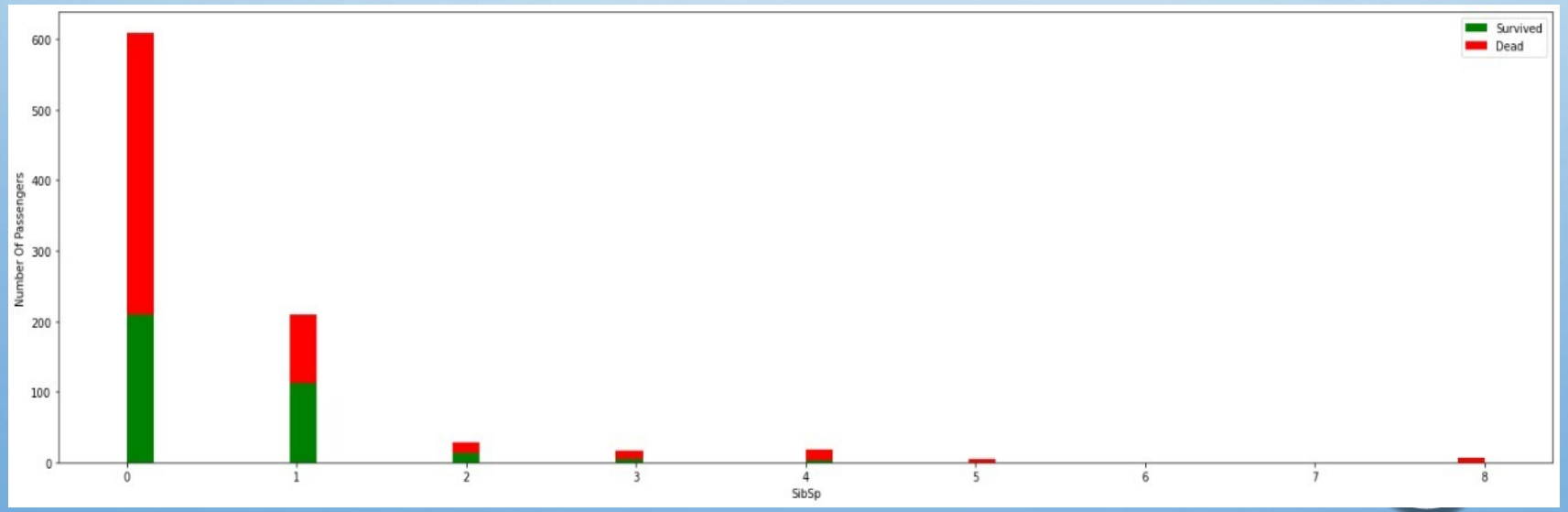
```
73
74 print(df[["Pclass", "Survived"]].groupby("Pclass", as_index = False).mean())
75 print("Passenger number according to Pclass")
76 print(df.Pclass.value_counts())
77 print("Passenger survival number according to Pclass = 1")
78 print(df[df.Pclass == 1].Survived.value_counts())
79 print("Passenger survival number according to Pclass = 2")
80 print(df[df.Pclass == 2].Survived.value_counts())
81 print("Passenger survival number according to Pclass = 3")
82 print(df[df.Pclass == 3].Survived.value_counts())
83
```



RELATION OF SURVIVAL WITH THE SIBLINGS

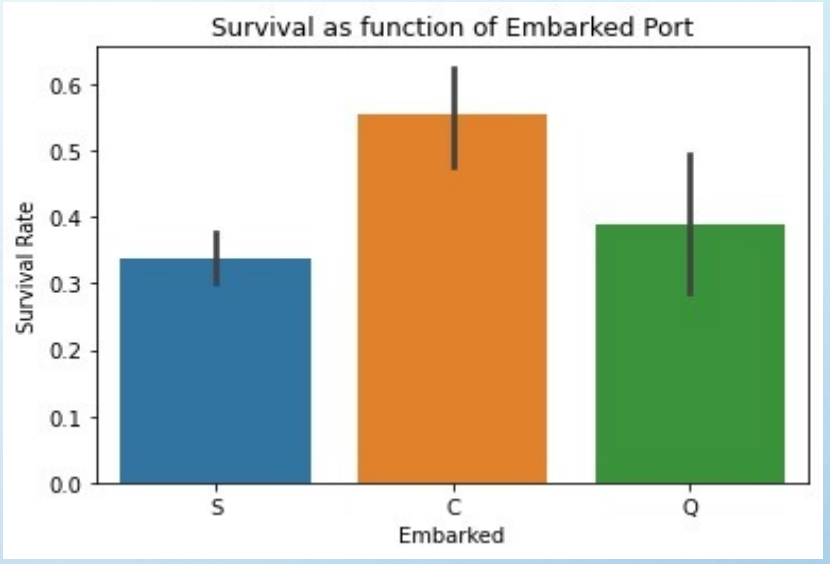
- WE SAW THAT SOME OF THE CREW WERE PRESENT WITH THEIR SIBLINGS.
- X-AXIS : SIBSP, Y-AXIS: NO OF PASSENGERS
- IT IS SEEN THAT PASSENGERS WITH NO SIBLINGS OR SPOUSE DIED IN MAXIMUM AMOUNT.
- SIBSP IS WHEN 1, SURVIVAL RATE GETS HIGHER A LITTLE.
- SIBSP IS WHEN 2 AND 3, SURVIVAL AND DEATH RATE ARE 50-50.
- AFTER SIBSP = 3, SURVIVAL RATE IS 0.

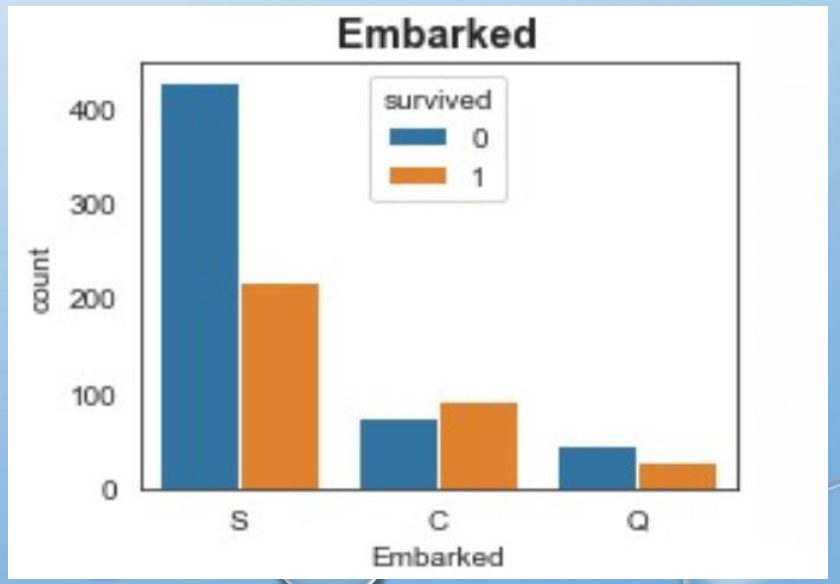




RELATION OF SURVIVAL WITH THE EMBARKED

- BASICALLY EMBARKED CLASS DENOTES THE PLACES WHERE THE SHIP STATIONED AND THE PASSENGERS OF THAT PLACE ENTERED INTO THE SHIP. THERE ARE 3 EMBARKED CLASSES(C, Q, S) IN THE DATASET.
- WE PLOT THE SURVIVAL PROBABILITY OF THOSE EMBARKED CLASS PEOPLE.(FOR FIGURE 1)
- X : EMBARKED CLASS
- Y : SURVIVAL PROBABILITY
- WE PLOT THE SURVIVAL PROBABILITY OF THOSE EMBARKED CLASS PEOPLE.(FOR FIGURE 2)
- X : EMBARKED CLASS
- Y : POPULATION COUNT
- THE SURVIVAL PROBABILITY OF C CLASS CREW IS HIGHER THAN Q AND S.

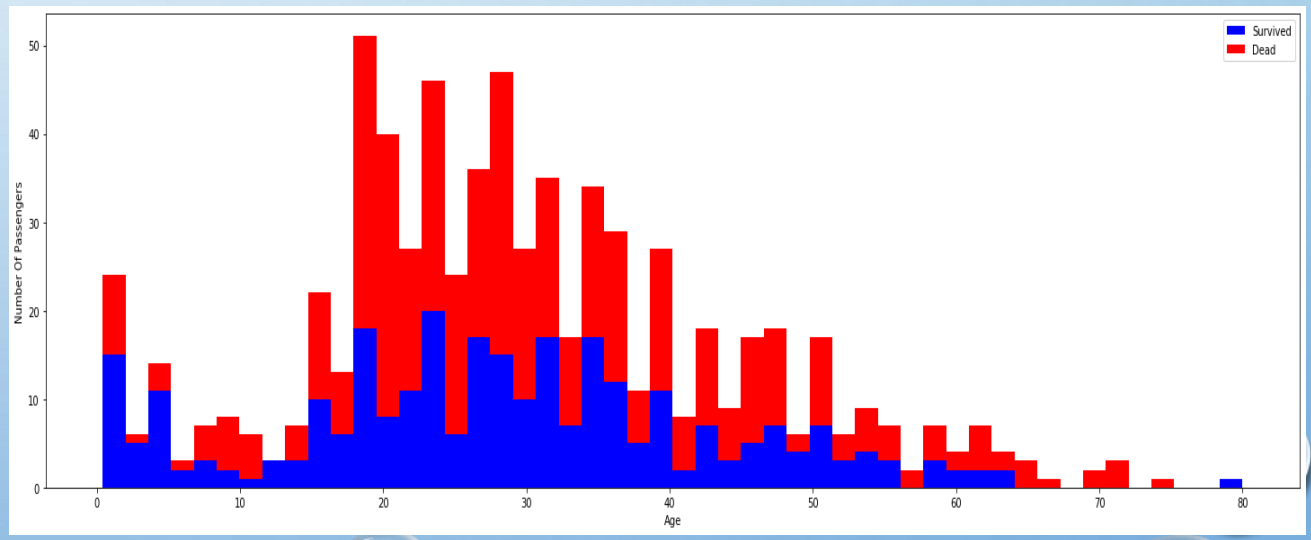




RELATION OF SURVIVAL RATIO WITH AGE

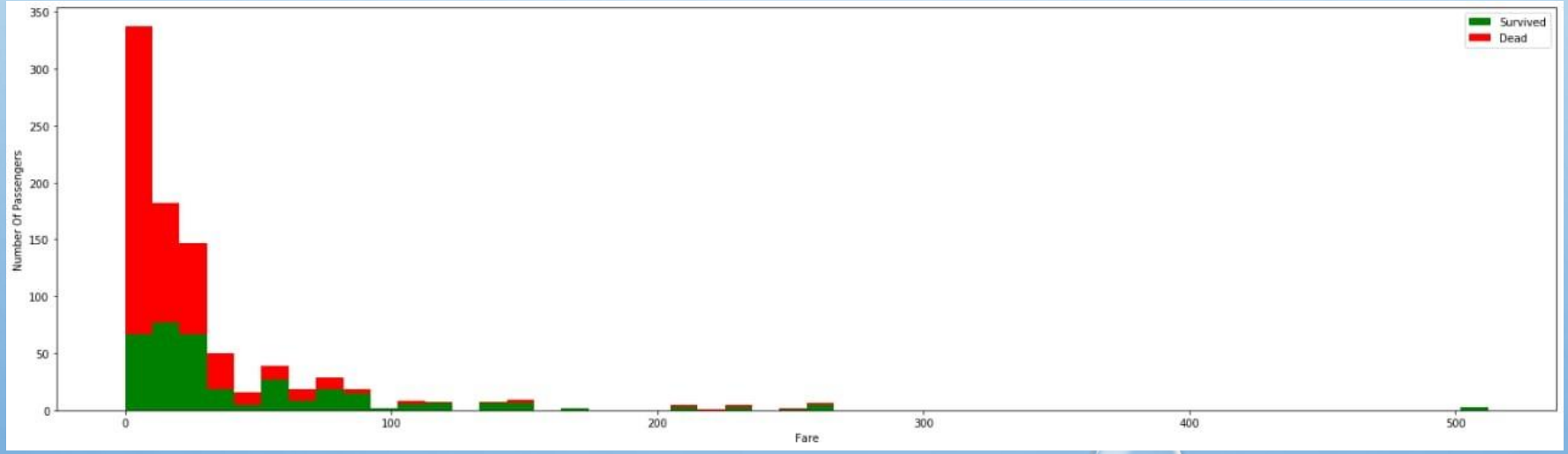
- AT THE TIME OF SINKING OF THE SHIP THE CHILDREN AND THE SENIOR PERSON ARE PRIORITIZED TO USE THE LIFEBOATS TO SAVE THEIR LIFE.
- X : AGE OF THE CREW
- Y : POPULATION
- BETWEEN AGE “0-8” SURVIVAL RATE IS HIGHER THAN DEATH RATE.
- BETWEEN AGE “20-50” DEATH RATE IS MAXIMUM.
- AFTERWARDS DEATH RATE DECREASES BUT STILL ITS HIGHER THAN SURVIVAL RATE.

```
204 df = pd.read_csv("train.csv")
205 figure = plt.figure(figsize=(25,7))
206 plt.hist([df[df['Survived']==1]['Age'],df[df['Survived']==0]['Age']],
207          stacked=True,color=['b','r'], bins=50,label=['Survived' , 'Dead'])
208 plt.xlabel('Age')
209 plt.ylabel('Number Of Passengers')
210 plt.legend()
211
```



RELATION OF SURVIVAL RATIO WITH FARE

- HERE WE COMPARE THE SURVIVED RATE WITH THE FARE VALUE .
- THE PEOPLE WITH HIGHER FARE BELONGS TO THE HIGHER CLASS, HAD GIVEN HIGHER PRIORITIES.
- X-AXIS : FARE
- Y-AXIS: NO OF PASSENGERS.
- THE PEOPLE WITH CHEAPER OR ZERO FARE HAS VERY LOW SURVIVAL RATIO.
- THE PEOPLE WITH HIGHER FARE WERE ABLE TO SURVIVE ALMOST .



FINDING K BEST CLASSIFIER

- THE DATA FEATURES THAT ONE USE TO TRAIN THE MACHINE LEARNING MODELS HAVE A HUGE INFLUENCE ON THE PERFORMANCE ONE CAN ACHIEVE. IRRELEVANT OR PARTIALLY RELEVANT FEATURES CAN NEGATIVELY IMPACT MODEL PERFORMANCE.
- NOW WE HAVE TRIED TO FIND THE BEST K (=4) NUMBERS OF FEATURES , THE VALUE OF THOSE ARE MOSTLY DEPEND ON THE SURVIVAL RATE. USING THIS WE CAN APPLY VARIOUS ML ALGORITHMS.

```
41
42 from sklearn.feature_selection import SelectKBest
43 from sklearn.feature_selection import chi2
44 X = df.drop("Survived",axis=1)
45 y = df["Survived"]
46 fr = SelectKBest(chi2, k=4)
47 fr.fit(X,y)
48 i = fr.get_support()
49 X_new = pd.DataFrame(fr.transform(X), columns = X.columns.values[i])
50 print(X_new.head())
```

	Pclass	Sex	Age	Fare
0	3	1	22	7.25
1	1	0	38	71.2833
2	3	0	26	7.925
3	1	0	35	53.1
4	3	1	35	8.05

MACHINE LEARNING MODELS

- AFTER EXTRACTING THE 4 BEST FEATURES FROM THE TITANIC DATASET WE APPLY DIFFERENT TYPES OF MACHINE LEARNING MODELS ON THE DIVIDED TRAIN (80%) AND TEST(20%) DATASET TO MAKE THE PREDICTIONS AND COMPARE THE ACCURACY WITH THE TRAIN DATASET.
- **MODELS USED :**

- LOGISTIC REGRESSION MODEL
- RANDOM FOREST CLASSIFIER MODEL
- DECISION TREE CLASSIFIER MODEL
- LINEAR SUPPORT VECTOR CLASSIFIER MODEL
- SUPPORT VECTOR CLASSIFIER
- PERCEPTRON MODEL
- GAUSSIAN NAÏVE BIAS.
- K NEIGHBORS CLASSIFIER MODEL

LOGISTIC REGRESSION MODEL

- **LOGISTIC REGRESSION** IS A MACHINE LEARNING CLASSIFICATION ALGORITHM THAT IS USED TO PREDICT THE PROBABILITY OF A CATEGORICAL DEPENDENT VARIABLE.
- HERE WE CHANGE THE HYPER-PARAMETER **'C'**(INVERSE OF THE REGULARIZATION STRENGTH) **= 5** AND **'PENALTY'(**USED TO SPECIFY THE NORM USED IN THE PENALIZATION(L1/L2))**='L2'** TO GET THE BETTER ACCURACY SCORE.

Logistic Regression accuracy Score: 0. 0.8100558659217877
MSE Score: 0.18994413407821228
R2 Score: 0.2134918583613334 4
The confusion matrix in case of Logistic Regression:
[[90 16]
[18 55]]

RANDOM FOREST CLASSIFIER MODEL

- A RANDOM FOREST IS A META ESTIMATOR THAT FITS A NUMBER OF DECISION TREE CLASSIFIERS ON VARIOUS SUB-SAMPLES OF THE DATASET AND USES AVERAGING TO IMPROVE THE PREDICTIVE ACCURACY AND CONTROL OVER-FITTING.
- HERE WE CHANGE THE HYPER-PARAMETER '**MAX_DEPTH'=7, 'MIN_SAMPLES_SPLIT'=10, 'N_ESTIMATORS'=600, 'RANDOM_STATE'=0, 'CRITERION' = 'GINI'** , TO GET THE BETTER ACCURACY SCORE

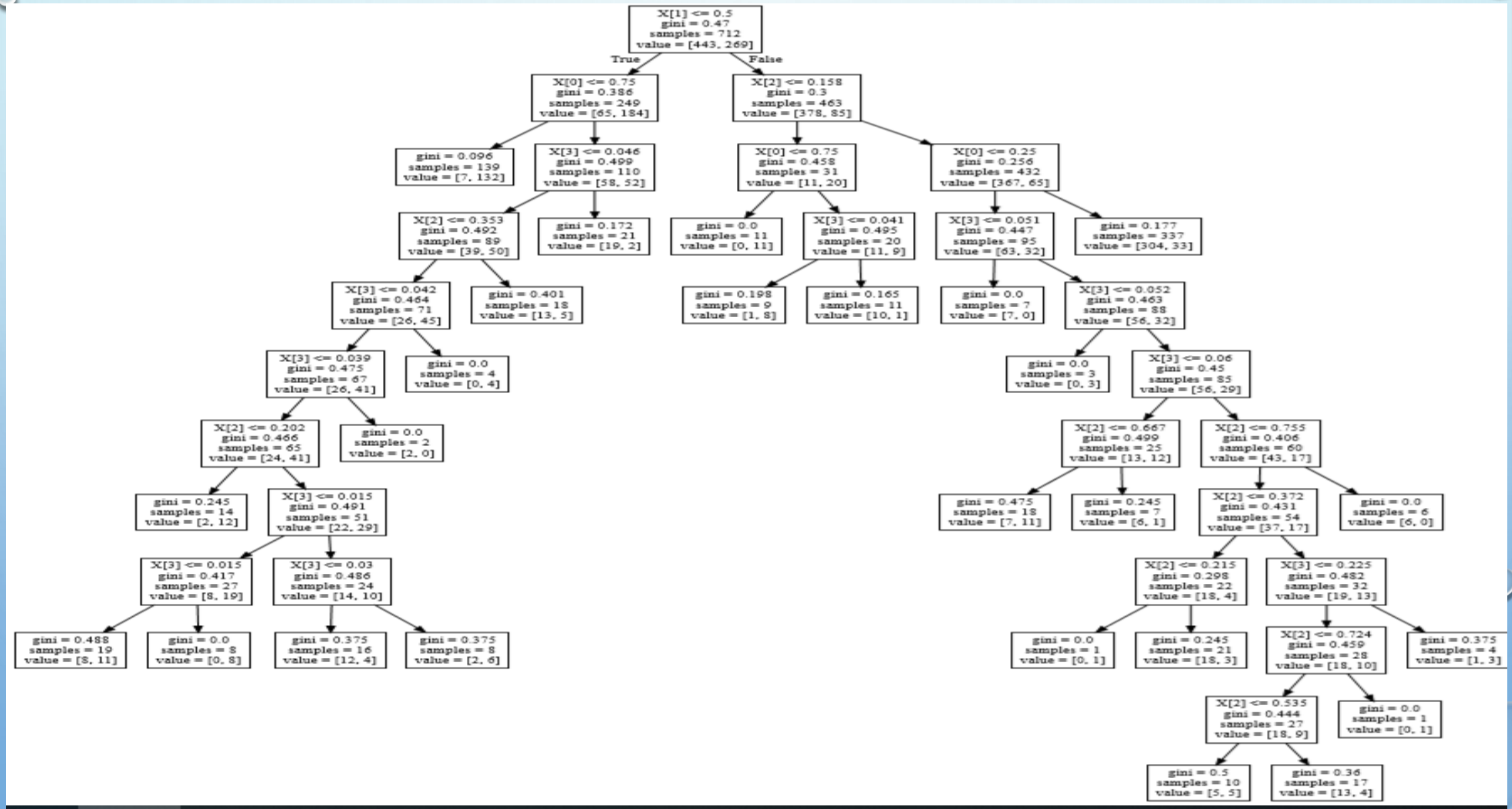
Random forest accuracy score: 0.821229050279329 6
MSE Score: 0.178770949720670 4
R2 Score: 0.259757043163608 04
The confusion matrix in case of Random Forest:
[[102 4]
[28 45]]

DECISION TREE CLASSIFIER MODEL

- **THE DECISION TREE** CLASSIFIER CREATES THE CLASSIFICATION MODEL BY BUILDING A DECISION TREE. EACH NODE IN THE TREE SPECIFIES A TEST ON AN ATTRIBUTE, EACH BRANCH DESCENDING FROM THAT NODE CORRESPONDS TO ONE OF THE POSSIBLE VALUES FOR THAT ATTRIBUTE.
- HERE WE CHANGE THE HYPER-PARAMETER, **'RANDOM_STATE'=0,'CRITERION' = 'GINI', 'MIN_SAMPLES_SPLIT' = 20,** TO GET THE BETTER ACCURACY SCORE.
- HERE WE HAVE SHOWN THE DECISION TREE AT NEXT PAGE WHERE WE CHOOSE 'SEX' AS A ROOT NODE WHICH IS DENOTED BY 'X[1]' AND DIVIDE THE BRANCHES OF THE TREE AS PER THE VALUES OF THE OTHER FEATURES.

Decision tree accuracy score: 0.821229050279329 6
MSE Score: 0.17877094972067039
R2 Score: 0.2597570431636080 4
The confusion matrix in case of Decision tree:
[[99 7]
[25 48]]

DECISION TREE FORMATION USING DATASET



LINEAR SUPPORT VECTOR CLASSIFIER MODEL

- SIMILAR TO SVC WITH PARAMETER KERNEL='LINEAR', BUT IMPLEMENTED IN TERMS OF LIBLINEAR RATHER THAN LIBSVM, SO IT HAS MORE FLEXIBILITY IN THE CHOICE OF PENALTIES AND LOSS FUNCTIONS AND SHOULD SCALE BETTER TO LARGE NUMBERS OF SAMPLES.
- THIS CLASS SUPPORTS BOTH DENSE AND SPARSE INPUT AND THE MULTICLASS SUPPORT IS HANDLED ACCORDING TO A ONE-VS-THE-REST SCHEME.
- HERE WE CHANGE THE HYPER-PARAMETER **'PENALTY' = 'L2', 'RANDOM_STATE' = 42 , 'MAX_ITER'= 5000** ,TO GET THE BETTER ACCURACY SCORE.

Linear SVC accuracy score: 0.776536312849162
MSE Score: 0.223463687150838
R2 Score: 0.0746963039545099
The confusion matrix in case of Linear SVC:
[[90 16]
 [24 49]]

SUPPORT VECTOR CLASSIFIER

- A **SUPPORT VECTOR MACHINE (SVM)** IS A DISCRIMINATIVE CLASSIFIER FORMALLY DEFINED BY A SEPARATING HYPERPLANE . IN OTHER WORDS, GIVEN LABELED TRAINING DATA (SUPERVISED LEARNING), THE ALGORITHM OUTPUTS AN OPTIMAL HYPERPLANE WHICH CATEGORIZES NEW EXAMPLES.
- HERE WE CHANGE THE HYPER-PARAMETER **'C' = 0.1, 'KERNEL' = 'RBF' , 'RANDOM_STATE'=0** ,TO GET THE BETTER ACCURACY SCORE.

Linear SVC accuracy score: 0.776536312849162
MSE Score: 0.22346368715083798
R2 Score: 0.0746963039545099
The confusion matrix in case of Linear SVC:
[[90 16]
 [24 49]]

PERCEPTRON MODEL

- **THE PERCEPTRON ALGORITHM** IS THE SIMPLEST TYPE OF ARTIFICIAL NEURAL NETWORK.
- IT IS A MODEL OF A SINGLE NEURON THAT CAN BE USED FOR TWO-CLASS CLASSIFICATION PROBLEMS AND PROVIDES THE FOUNDATION FOR LATER DEVELOPING MUCH LARGER NETWORKS.
- HERE WE CHANGE THE HYPER-PARAMETER **'N_ITER_NO_CHANGE' = 5000 , 'RANDOM_STATE' = 1,**TO GET THE BETTER ACCURACY SCORE.

Perceptron accuracy score: 0.7932960893854749
MSE Score: 0.2067039106145514
R2 Score: 0.14409408115792177
The confusion matrix in case of Perceptron:
[[84 22]
 [15 58]]

GAUSSIAN NAÏVE BIAS

- **NAIVE BAYES** METHODS ARE A SET OF SUPERVISED LEARNING ALGORITHMS BASED ON APPLYING BAYES' THEOREM WITH THE "NAIVE" ASSUMPTION OF CONDITIONAL INDEPENDENCE BETWEEN EVERY PAIR OF FEATURES GIVEN THE VALUE OF THE CLASS VARIABLE.
- **GAUSSIANNB** IMPLEMENTS THE GAUSSIAN NAIVE BAYES ALGORITHM FOR CLASSIFICATION. THE LIKELIHOOD OF THE FEATURES IS ASSUMED TO BE GAUSSIAN.

Gaussian accuracy score: 0.765363128491620 1
MSE Score: 0.2346368715083799
R2 Score: 0.028431119152235507
The confusion matrix in case of Gaussian:
[[86 20]
 [22 51]]

K NEAREST NEIGHBORS CLASSIFIER MODEL

- THE **K-NEAREST NEIGHBORS (KNN) ALGORITHM** IS A TYPE OF SUPERVISED MACHINE LEARNING ALGORITHMS.
- IT SIMPLY CALCULATES THE DISTANCE OF A NEW DATA POINT TO ALL OTHER TRAINING DATA POINTS. THE DISTANCE CAN BE OF ANY TYPE E.G EUCLIDEAN OR MANHATTAN ETC. IT THEN SELECTS THE K-NEAREST DATA POINTS, WHERE K CAN BE ANY INTEGER. FINALLY IT ASSIGNS THE DATA POINT TO THE CLASS TO WHICH THE MAJORITY OF THE K DATA POINTS BELONG.
- HERE WE CHANGE THE HYPER-PARAMETER '**N_NEIGHBORS'** (NUMBER OF NEIGHBORS TO USE BY DEFAULT FOR :METH:`KNEIGHBORS` QUERIES.) **= 3** ,TO GET THE BETTER ACCURACY SCORE.

KNN accuracy score: 0.7932960893854749
MSE Score: 0.2067039106145514
R2 Score: 0.14409408115792177
The confusion matrix in case of KNN:
[[95 11]
 [26 47]]

COMPARISON BETWEEN THE FITTING OF DIFFERENT MODELS

- HERE WE ARRANGE THE ACCURACY SCORE IN DESCENDING ORDER OF THE TITANIC DATASET AFTER APPLYING DIFFERENT MACHINE LEARNING MODEL. HERE WE CAN SEE THAT THE ACCURACY SCORE OF RANDOM FOREST MODEL IS GREATER THAN THE OTHER MODELS. THE TABLE IS SHOWN BELOW:

Index	Model	Score	MSE	R-square
1	Random Forest	0.821229	0.178771	0.259757
2	Decision Tree	0.821229	0.178771	0.259757
3	Logistic Regression	0.810056	0.189944	0.213492
4	KNN	0.793296	0.206704	0.144094
5	Perceptron	0.793296	0.206704	0.144094
6	Support Vector Machines	0.776536	0.223464	0.074696
7	Linear SVC	0.776536	0.223464	0.074696
8	Naive Bayes	0.765363	0.234637	0.028431

CONCLUSION

- HERE WE HAVE REMOVED THE FEATURES WITH THE COLUMN NAMES - 'NAME', 'PASSENGERID', 'CABIN', 'TICKET'. AS THESE HAVE NOT MUCH DEPENDENCIES ON PREDICTING THE SURVIVAL OF THE CREW.
- AFTER ANALYZING ALL THE ABOVE THINGS, WE GOT THAT THE PERSONS WITH SIBLINGS AND THE PERSONS WHO PAID MUCH FARE AND THE FEMALE CATEGORY HAVE THE SURVIVAL RATE MUCH GREATER THAN OTHERS.
- ULTIMATELY WE GET THE HIGHEST PREDICTION ACCURACY 82%

THANK YOU