PATTERN RECOGNITION AND MACHINE LEARNING

LAB 04
LAB REPORT

QUESTION 1.

1)

- The data was first stored in ‘df’ through pandas and was preprocessed. The preprocessing includes filling NAN values with appropriate values,, dropping certain columns with little or no significance and splitting it into training and testing sets.

2)

- It was identified that the Gaussian Naive Classifier would be the best classifier in this case since it supports continuous values.
- The dataset was also observed to have continuous values in features - Age and Fare.

3)

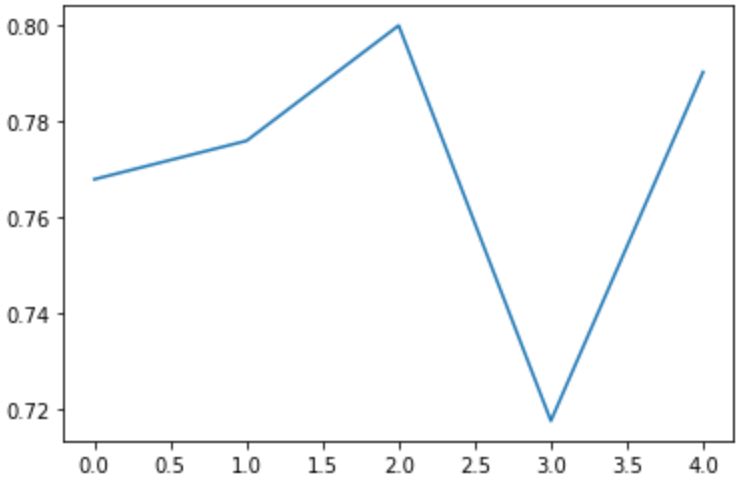
- Gaussian Naive Classifier was implemented from scratch.
- Likelihood probabilities concerning features with discrete values were calculated by treating them independent of each other - and thus by simply multiplying, while those concerning features with continuous values were estimated by considering normal distribution and iid.
- The accuracy of this Classifier was found to be `80.22388059701493 %.`

4)

- 5 fold cross validation was performed with the same from-scratch Classifier and the accuracy was then found out to be `[0.798507462686567, 0.7873134328358209, 0.8022388059701493, 0.8022388059701493, 0.8059701492537313]` for each fold with average accuracy of `0.7704129032258065.`

5)

- The accuracies for each fold were plotted.



- Top class probabilities were also calculated and printed.

6)

- In-built Gaussian model was implemented and its accuracy was found to be `78.35820895522389 %.`
- The Classifier made from scratch has accuracy `80.59701492537313 %.`

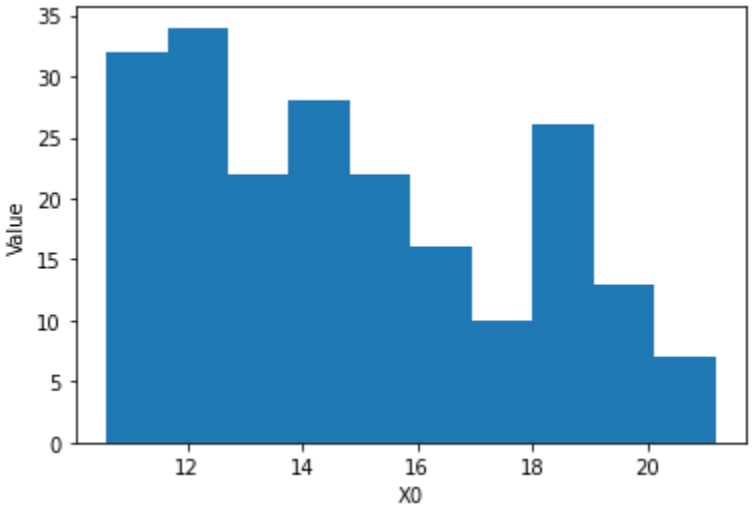
7)

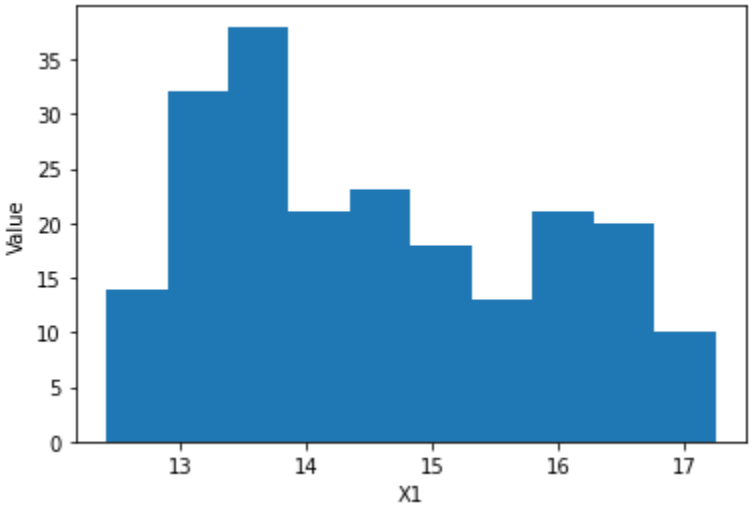
- In build model for Decision Tree Classifier was implemented under KFold and its accuracy turned out to be `77.04129032258065`

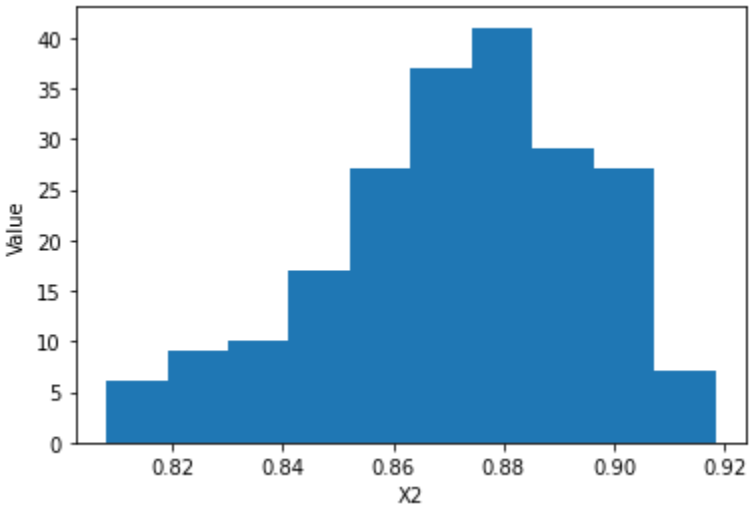
QUESTION 2.

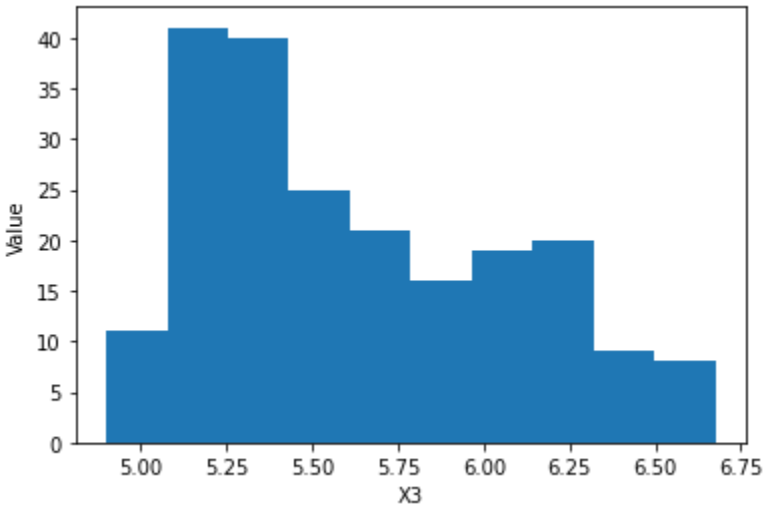
1)

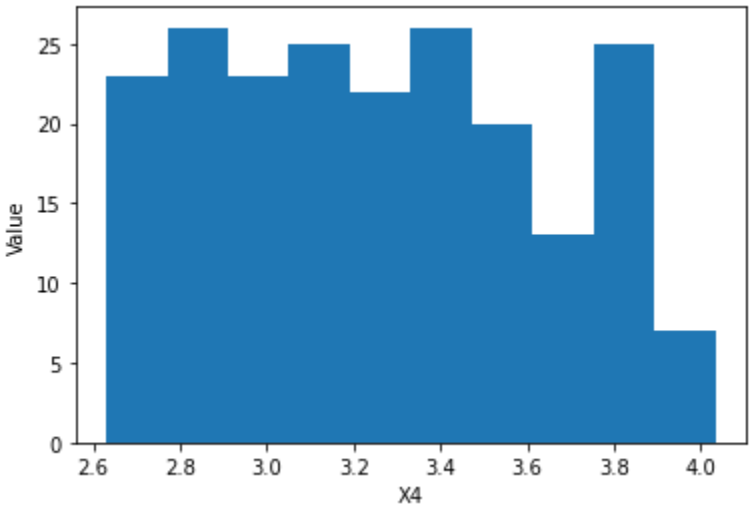
- Histograms were plotted.

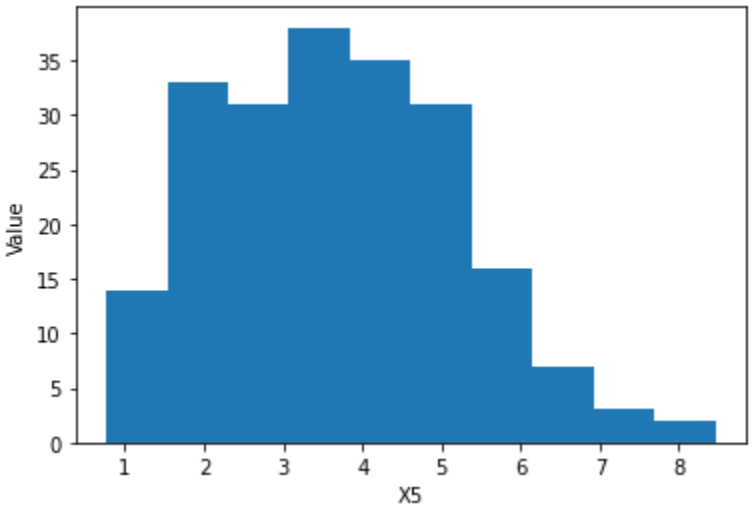


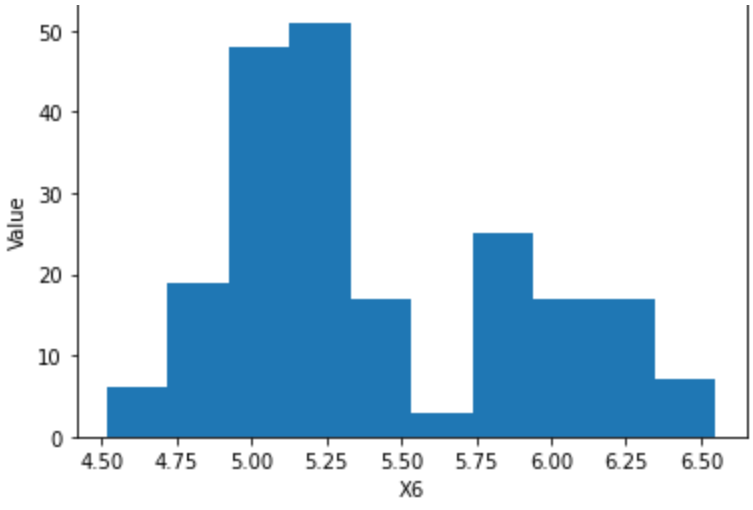



Value
X2
0
5
10
15
20
25
30
35
40
0.82
0.84
0.86
0.88
0.90
0.92


Value
X3
0
5
10
15
20
25
30
35
40
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75


Value
X4
0
5
10
15
20
25
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0


Value
X5
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8


Value
X6
0
10
20
30
40
50
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50

2)

- `Prior Probablility of Case 1: 0.3333333333333333`
- `Prior Probablility of Case 2: 0.3333333333333333`
- `Prior Probablility of Case 3: 0.3333333333333333`
- The entire dataset is equally divided into 3 cases.

3)

- Bins were created off the dataset from scratch.
- The number of bins to be created was set to 10.
- This was decided as the histograms had 10 bars and it gave the best results.

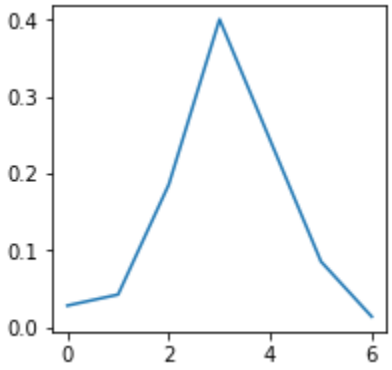
4)

- Likelihood was calculated for each feature and class value wise and was stored in distinctive dictionaries.
- The dictionaries have keys as unique values present in the feature and values as count of respective unique values.

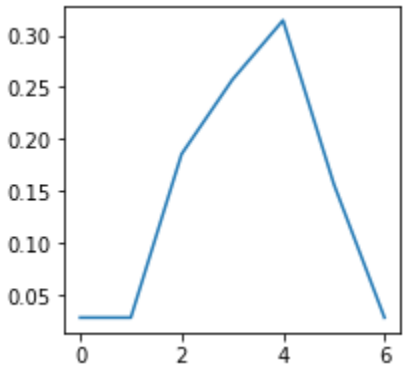
5)

Class 1

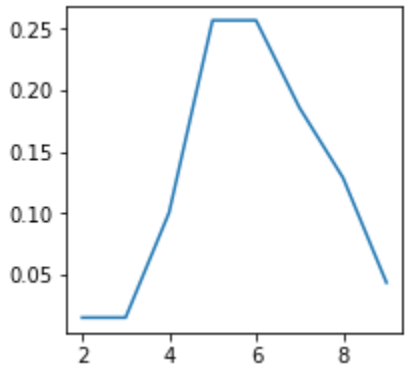
Column X0 Class 1



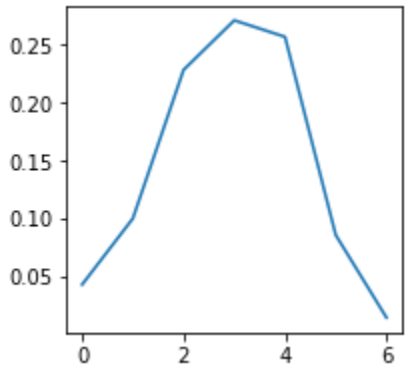
Column X1 Class 1



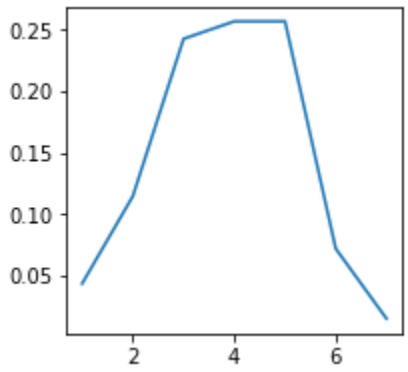
Column X2 Class 1



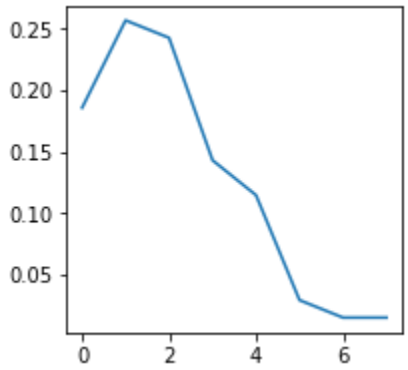
Column X3 Class 1



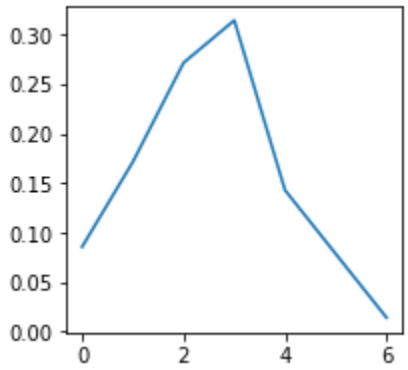
Column X4 Class 1



Column X5 Class 1

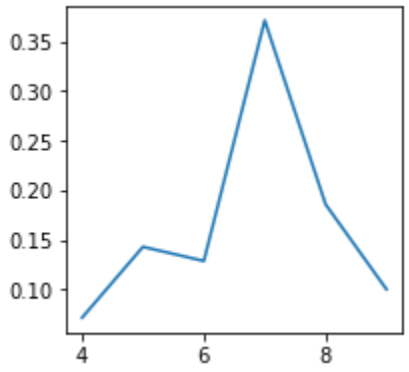


Column X6 Class 1

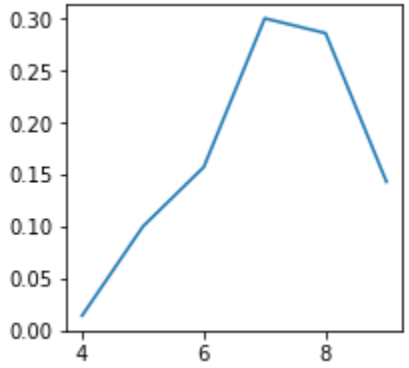


Class 2

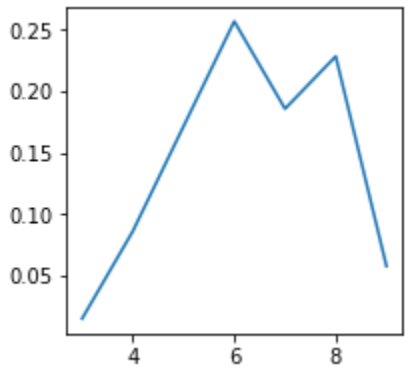
Column X0 Class 2



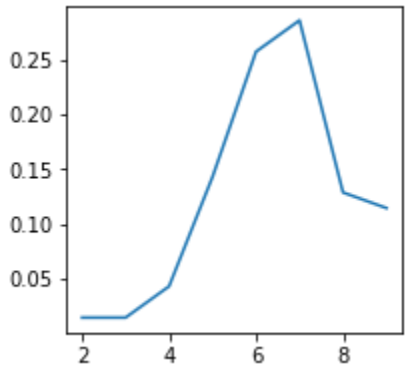
Column X1 Class 2



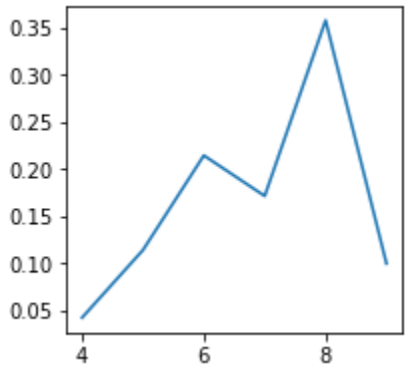
Column X2 Class 2



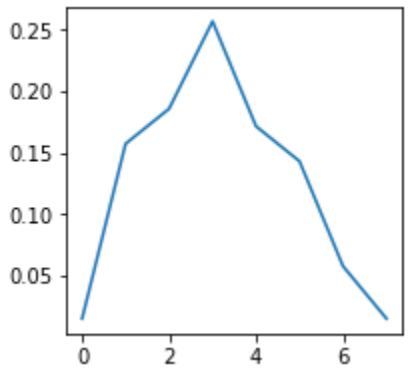
Column X3 Class 2



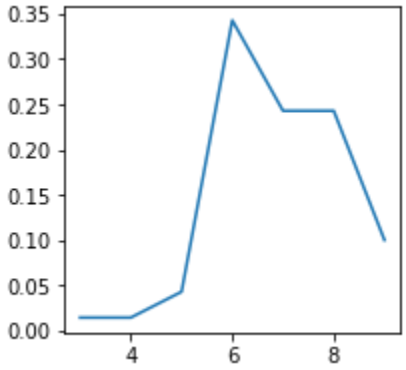
Column X4 Class 2



Column X5 Class 2

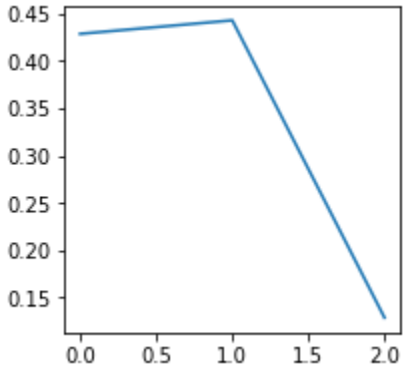


Column X6 Class 2

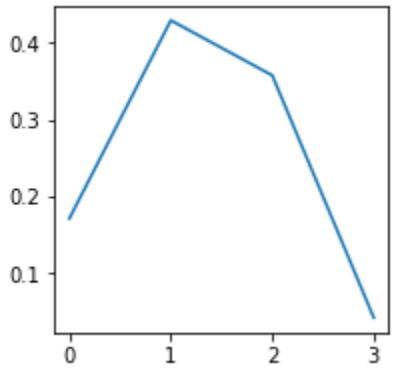


Class 3

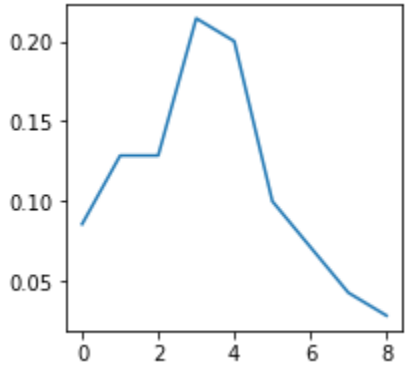
Column X0 Class 3



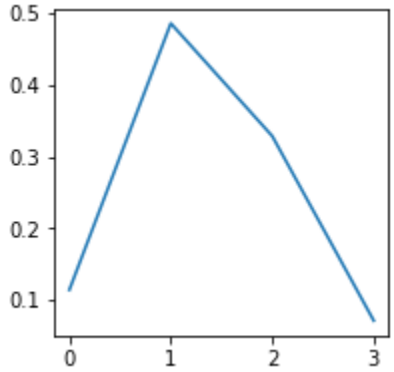
Column X1 Class 3



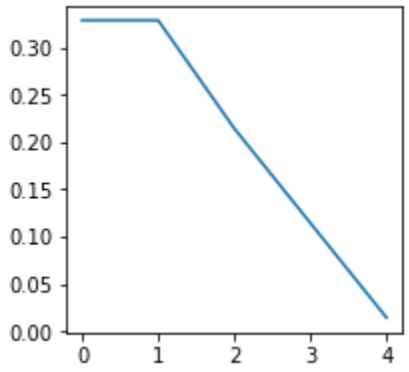
Column X2 Class 3



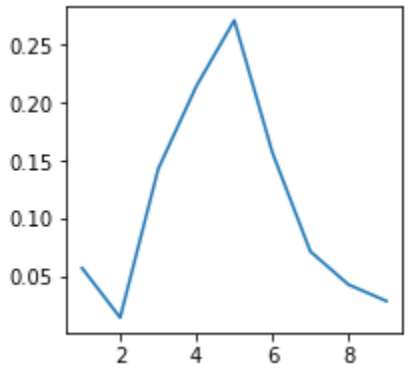
Column X3 Class 3



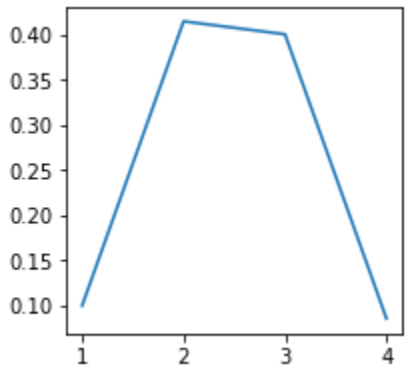
Column X4 Class 3



Column X5 Class 3



Column X6 Class 3



6)

- Posterior Probabilities were calculated for all the 63 testing data entries and were plotted.