PRML LAB 4

Harsh Sharma (B20CS017)

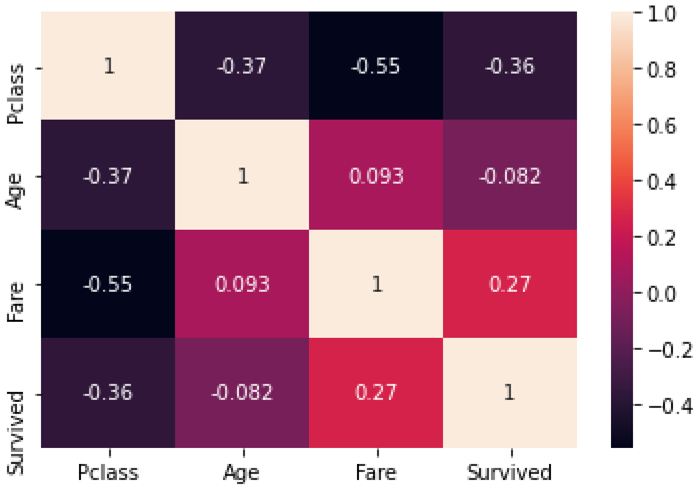
Question 1

Preprocessing and data splitting

- In preprocessing , all the rows containing nan values have been removed .
- For categorical encoding features having dtype equal to object are encoded using label encoder .
- For normalization of data MinMaxScaler is used .
- Data is splitted into 80:20 ratio in which 80 is training and 20 is testing .

The useful columns in the given data are ‘Pclass’,’Sex’,’Age’,’Fare’,’Embarked’, ’Survived’ . Remaining columns are not useful .The dropped columns have almost no effects on the target variable .

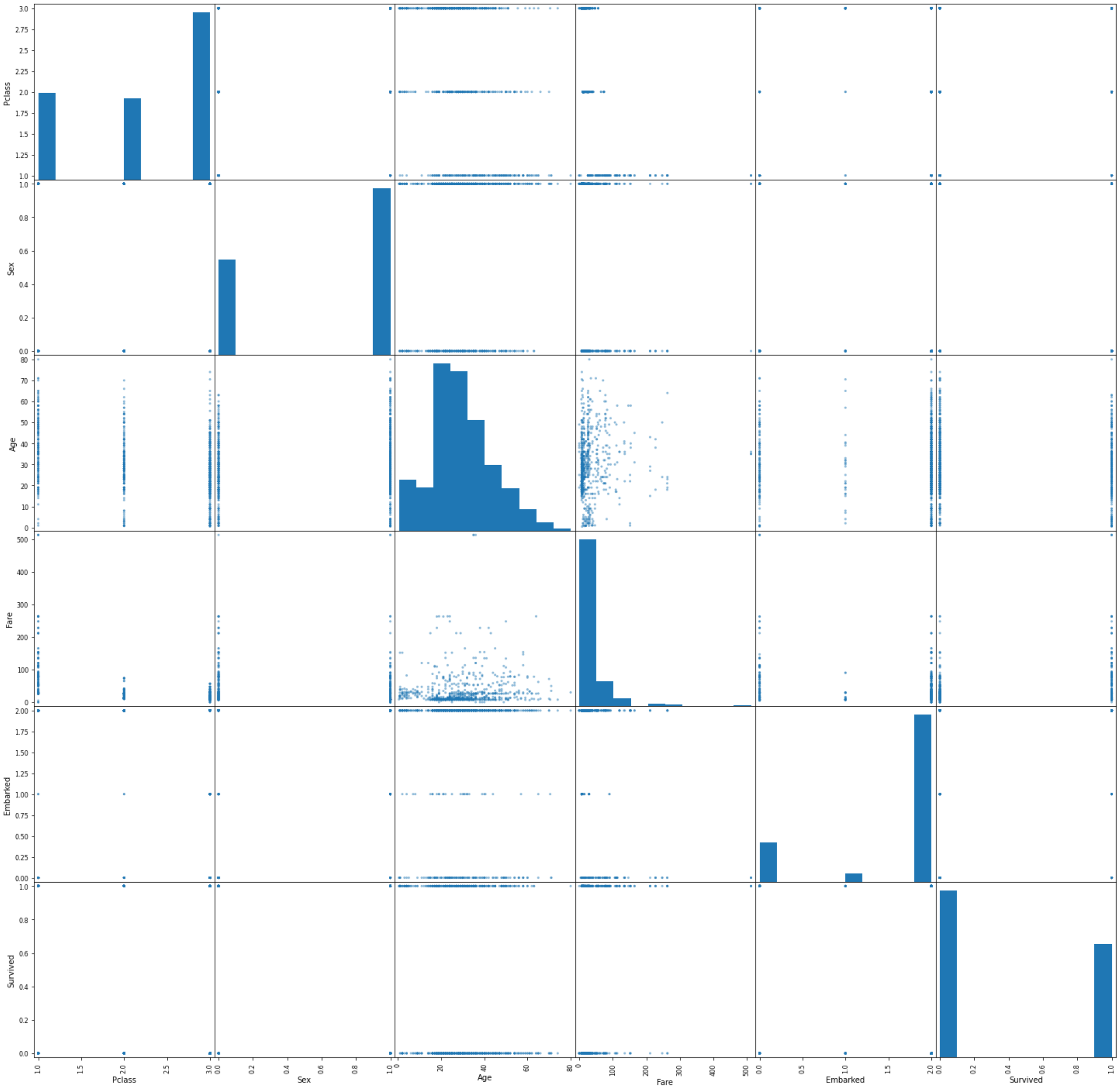
The correlation matrix is shown below

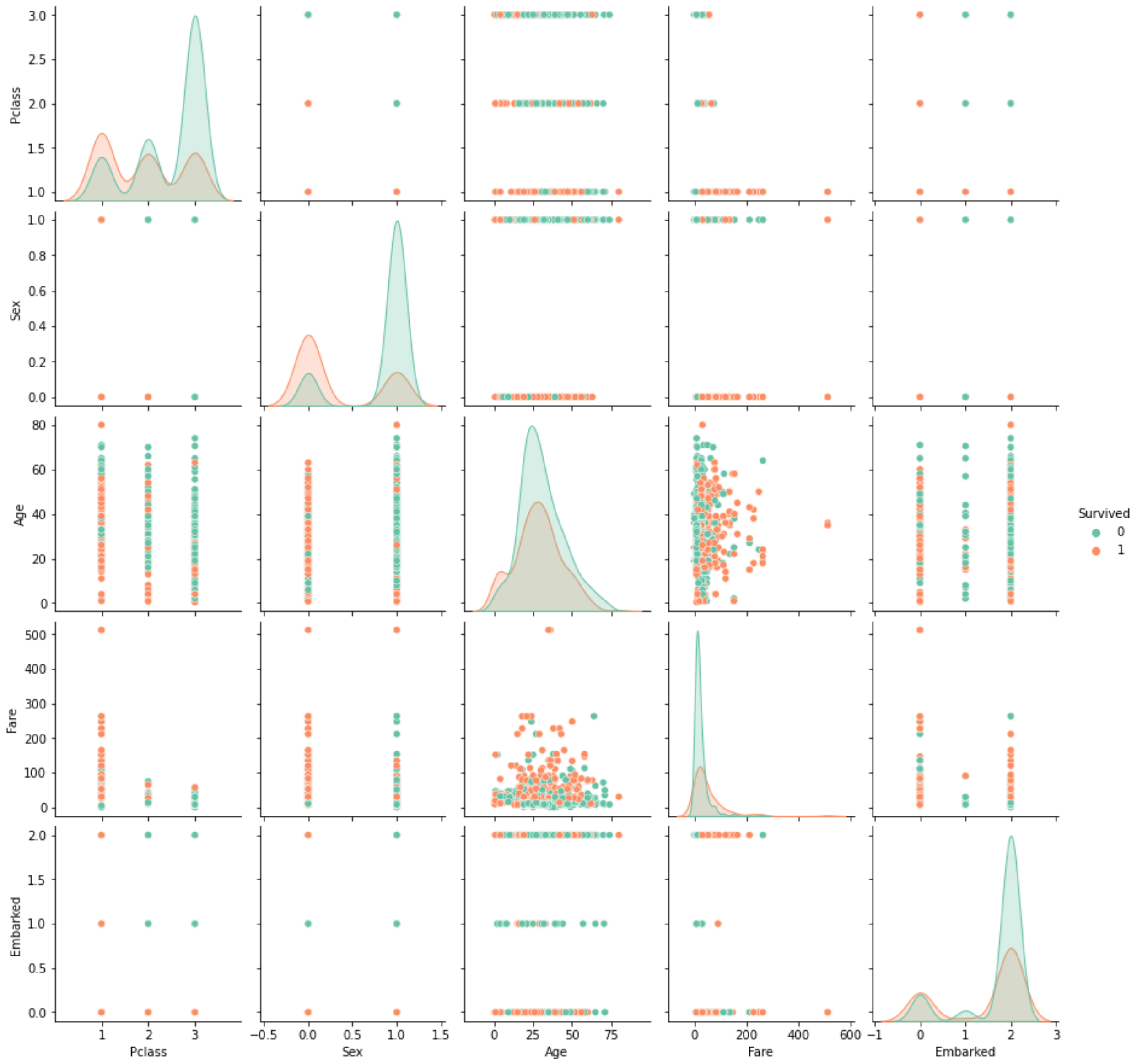


	Pclass	Age	Fare	Survived
Pclass	1	-0.37	-0.55	-0.36
Age	-0.37	1	0.093	-0.082
Fare	-0.55	0.093	1	0.27
Survived	-0.36	-0.082	0.27	1

Data Visualization

For data visualization I have used a scatter_matrix from pandas as well as seaborn’s pair_plot function which plots the graph between all the combinations of two features .



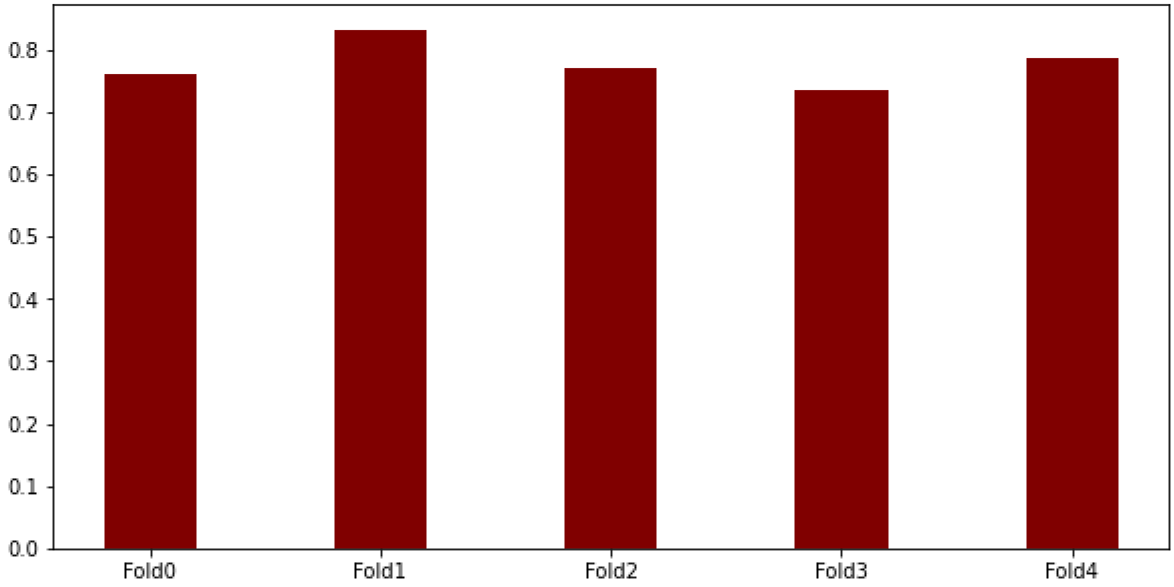


Naive Bayes Classifier

- Implemented Gaussian Distribution variant of naive bayes because from visualization of the data gaussian distribution will fit the most .
- In the implementation first I have calculated prior probabilities of each class
- For every datapoint in the testing data ,I have calculated the likelihood .
- Then calculated posterior probability and on the basis maximum posterior probability predicted the output labels.
- The accuracy of the implemented model is 0.8391608391608392 .

Fivefold Cross Validation

- The training data is divided into five distinct datasets, each of which is exclusive and exhaustive.
- The Naive Bayes Classifier is trained on four partitions and then tested on the fifth.
- This is done five times more until each partition has served as testing data at least once.
- The average accuracy of the five folds is 0.7769911504424779 .
- The accuracy of five fold is shown in the graph below



For top class probability I have calculated probability density for each class in testing data then printed the maximum among the both .

Comparison from the Scikit library

Accuracy of the model from scratch is 0.8391608391608392 .

Accuracy of the model from Scikit library is 0.8321678321678322

We can say that the model from scratch has almost the same accuracy as compared to the library one .

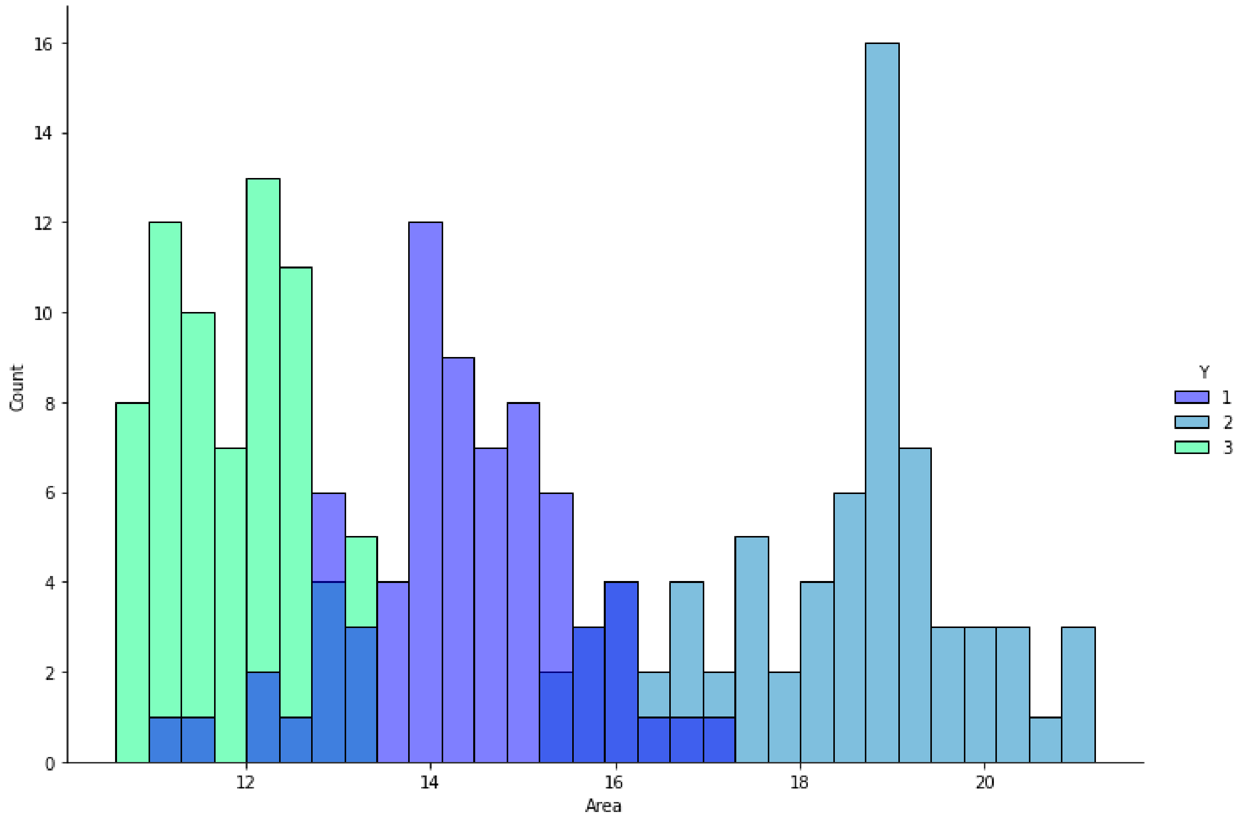
Comparison from the Decision Tree

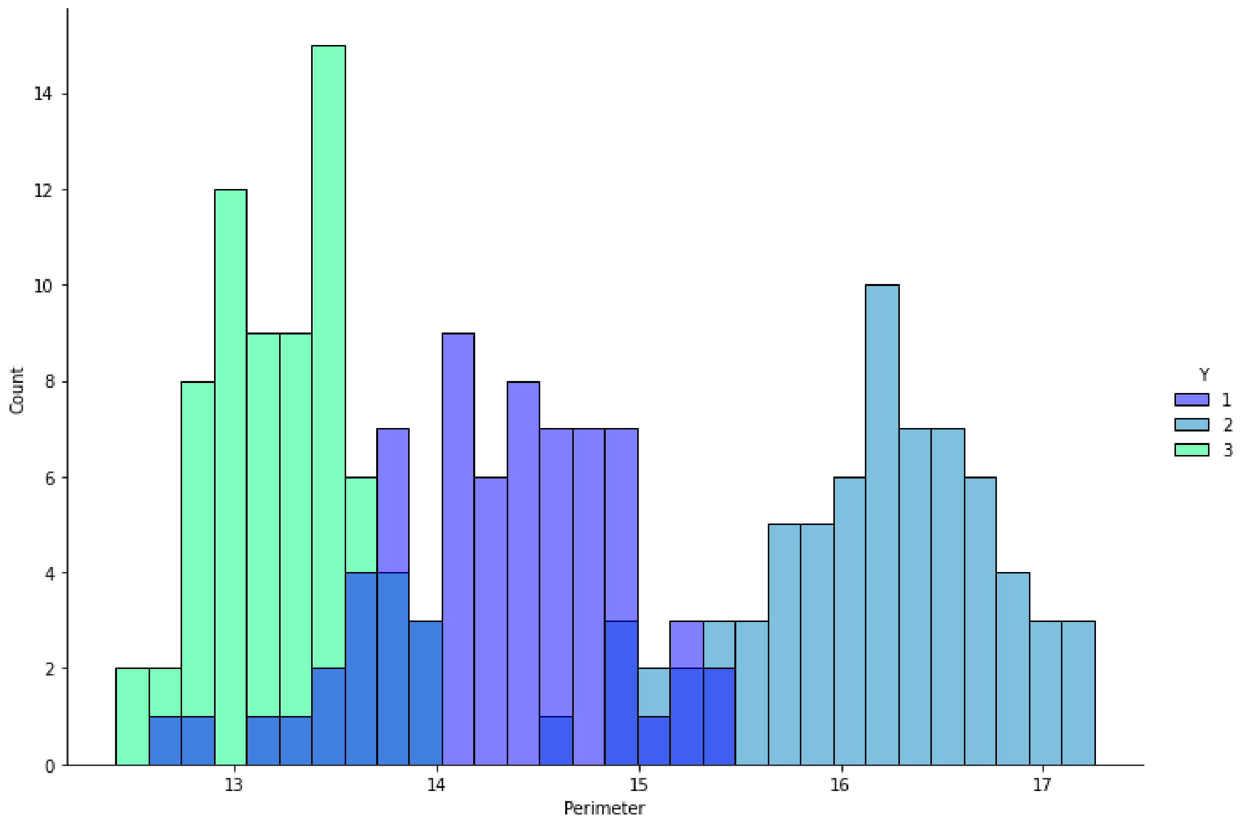
The average accuracy after applying 5 five fold cross validation is 0.7901477832512315.

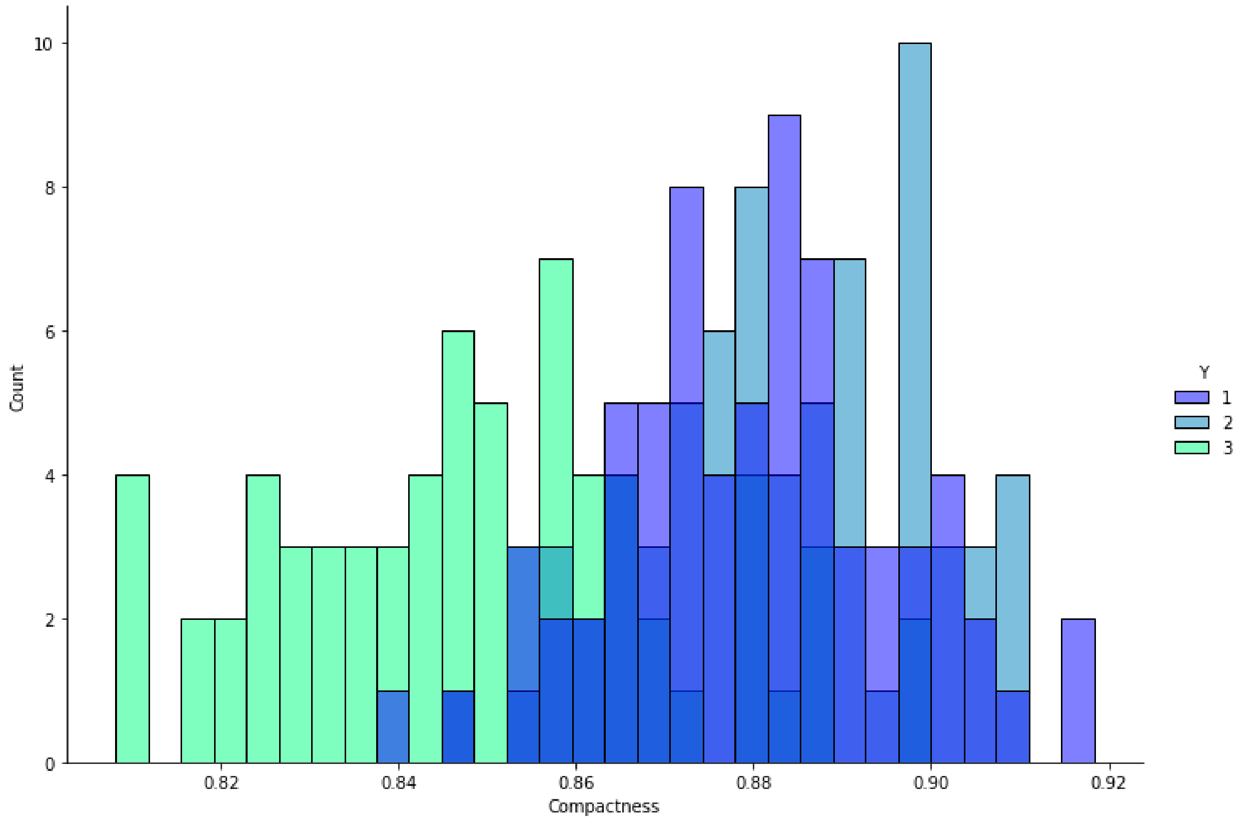
The Naive Bayes model is more accurate than the decision tree .

Question 2

Histogram to plot the distribution of samples




Count
0
2
4
6
8
10
12
14
13
14
15
16
17
Perimeter
Y
1
2
3


Count
0
2
4
6
8
10
0.82
0.84
0.86
0.88
0.90
0.92
Compactness
Y
1
2
3

Prior probability

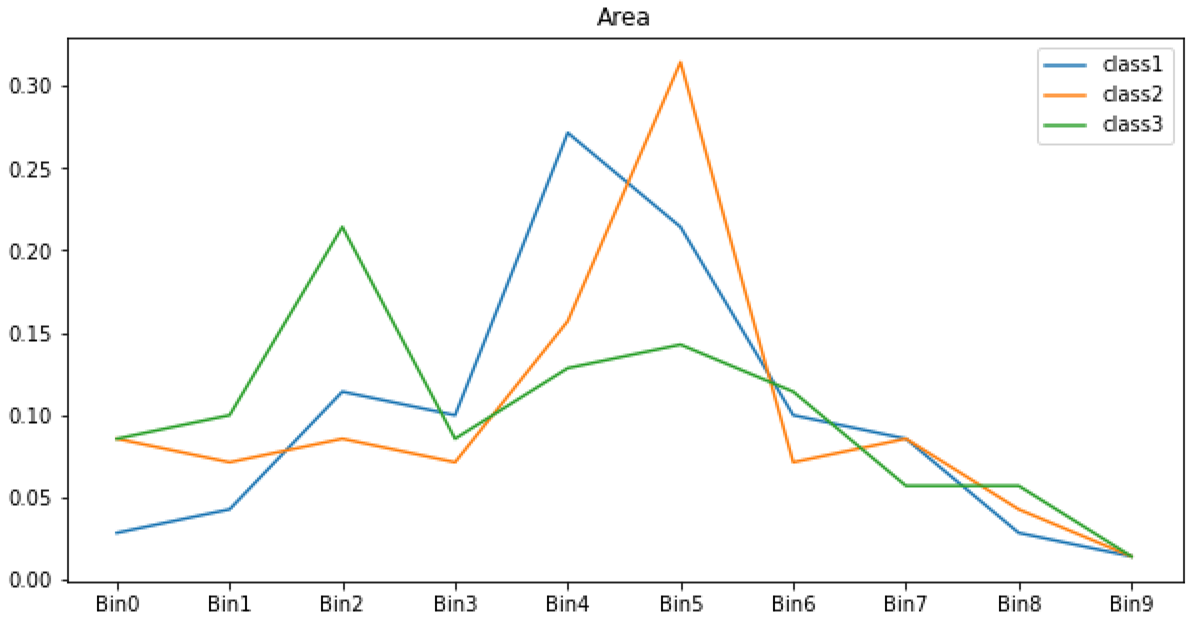
The prior probability is calculated for each class . The prior probability for each class is 0.33333 .

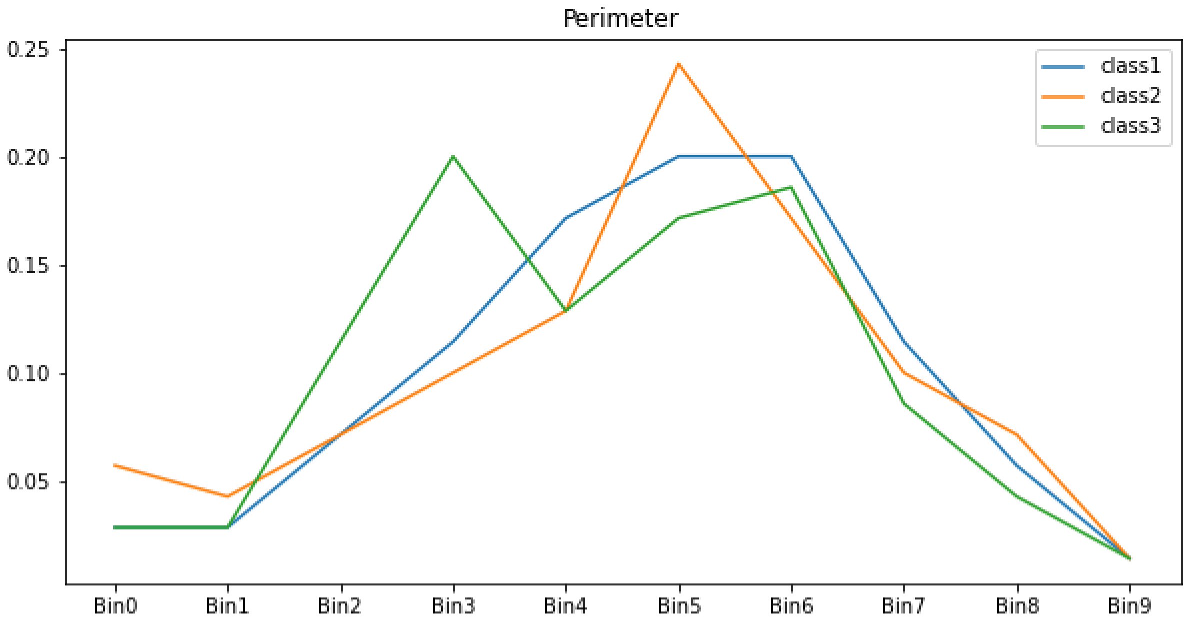
Discretize the features into bins

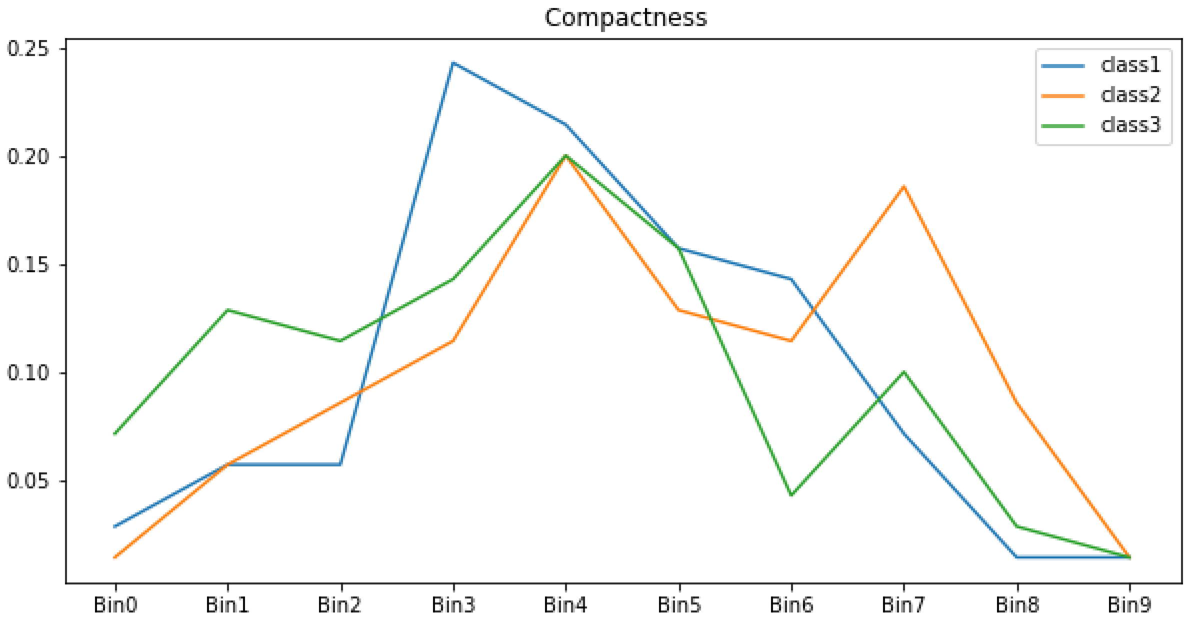
- Three datasets are created for each class .
- For each dataset, bins are created by dividing each feature into 10 parts of equal separation . The count of the datapoint in each feature is calculated on the basis of the partition created .
- Three lists containing 10 bins for each feature were created .

Likelihood/class conditional probabilities

Likelihood conditional probabilities are calculated using the bins created . For each datapoint ,the probability is calculated by dividing its count by the total number of counts of each datapoint in each feature.

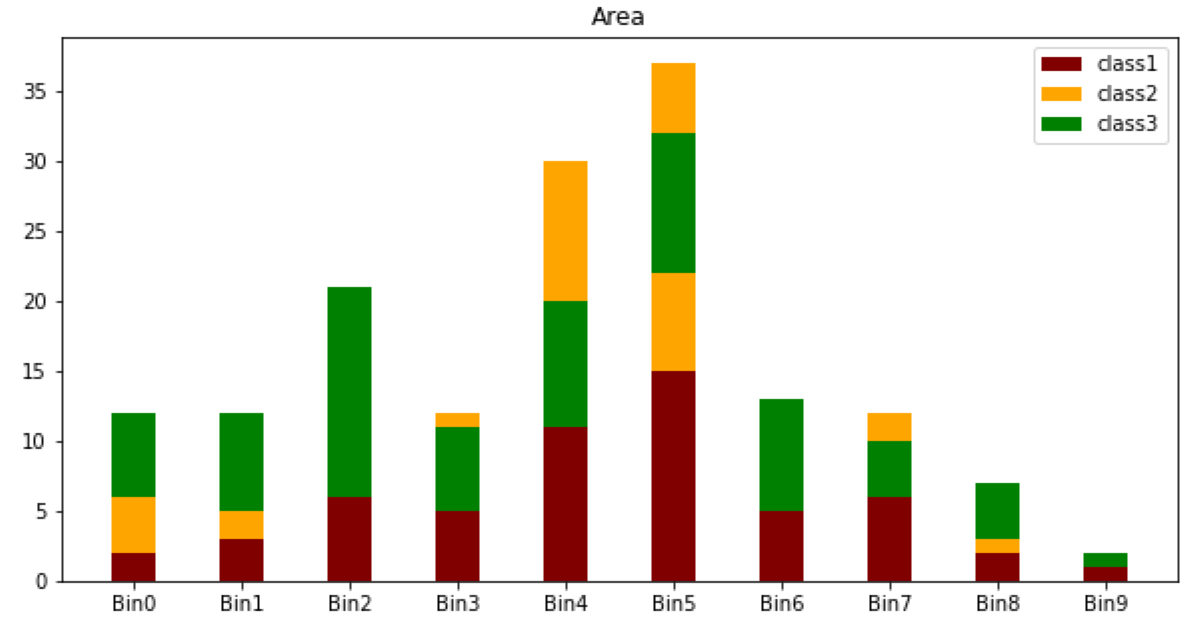


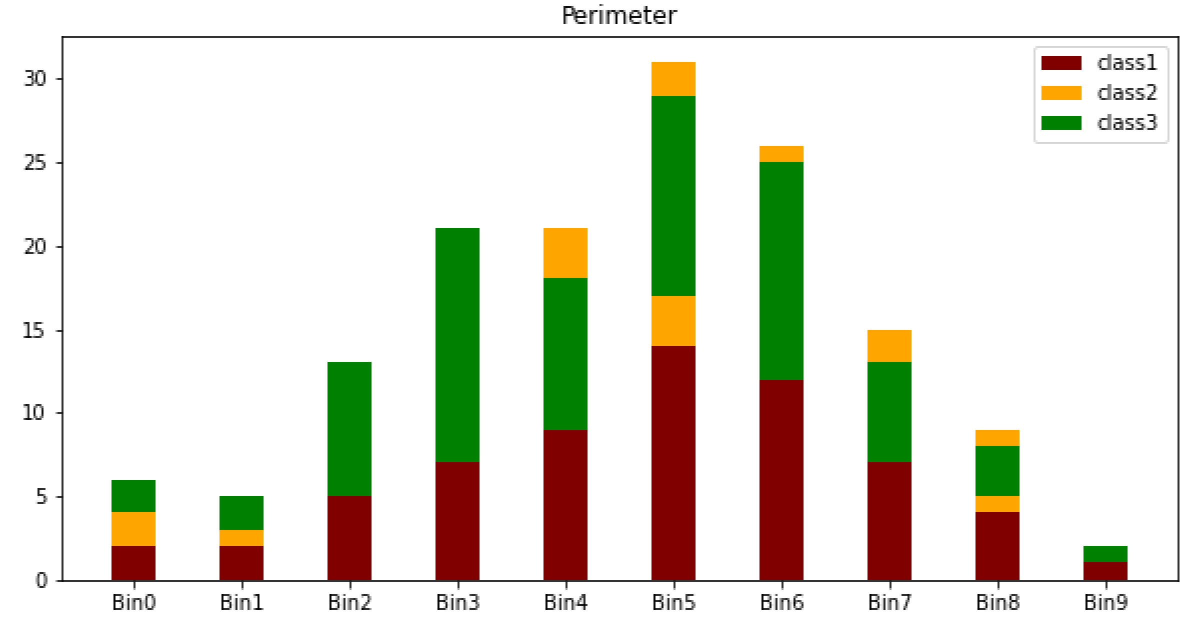

Perimeter
0.25
0.20
0.15
0.10
0.05
Bin0
Bin1
Bin2
Bin3
Bin4
Bin5
Bin6
Bin7
Bin8
Bin9
class1
class2
class3

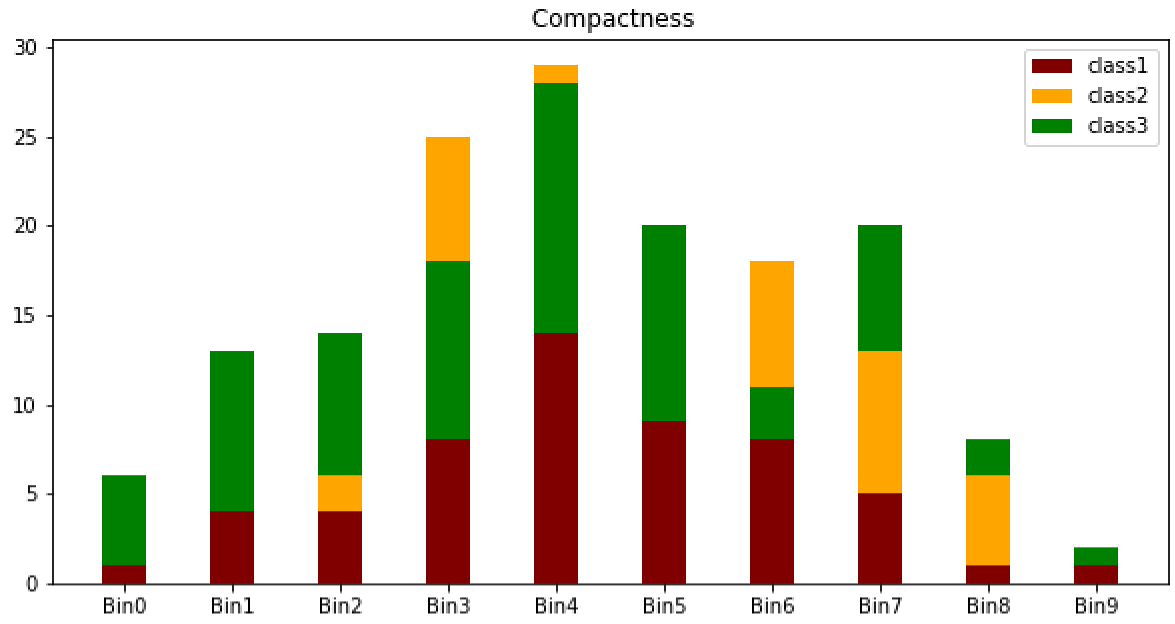

Compactness
0.25
0.20
0.15
0.10
0.05
Bin0
Bin1
Bin2
Bin3
Bin4
Bin5
Bin6
Bin7
Bin8
Bin9
class1
class2
class3

Plot the count of each unique element for each class

For every bin the count in each bin is plotted against count . Few graphs are shown below.

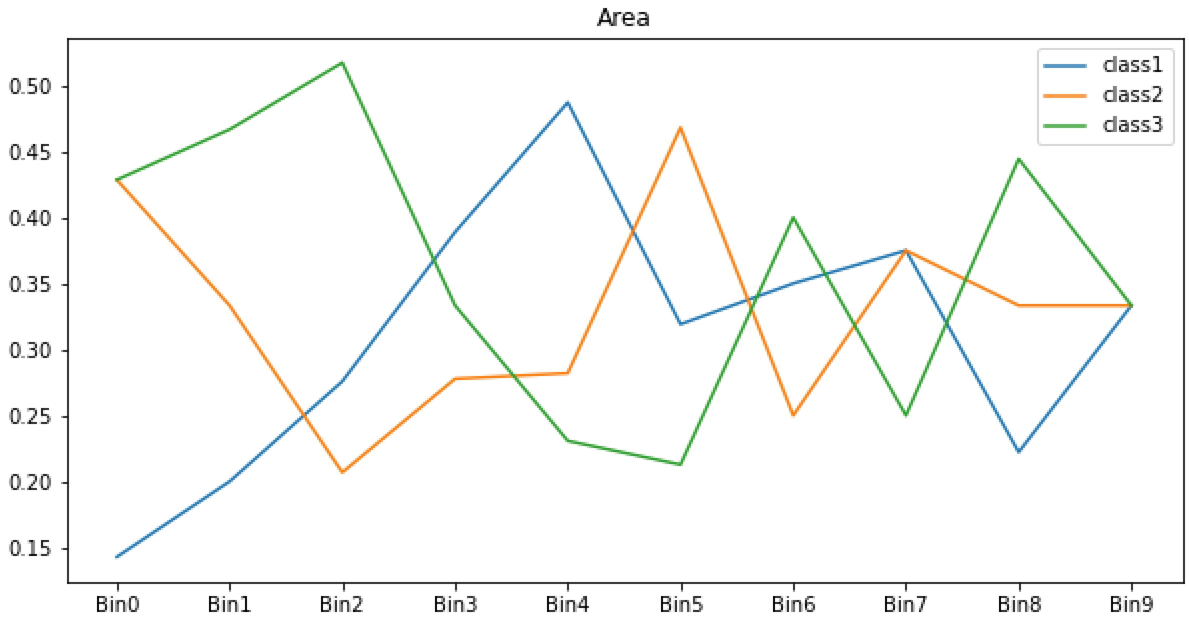


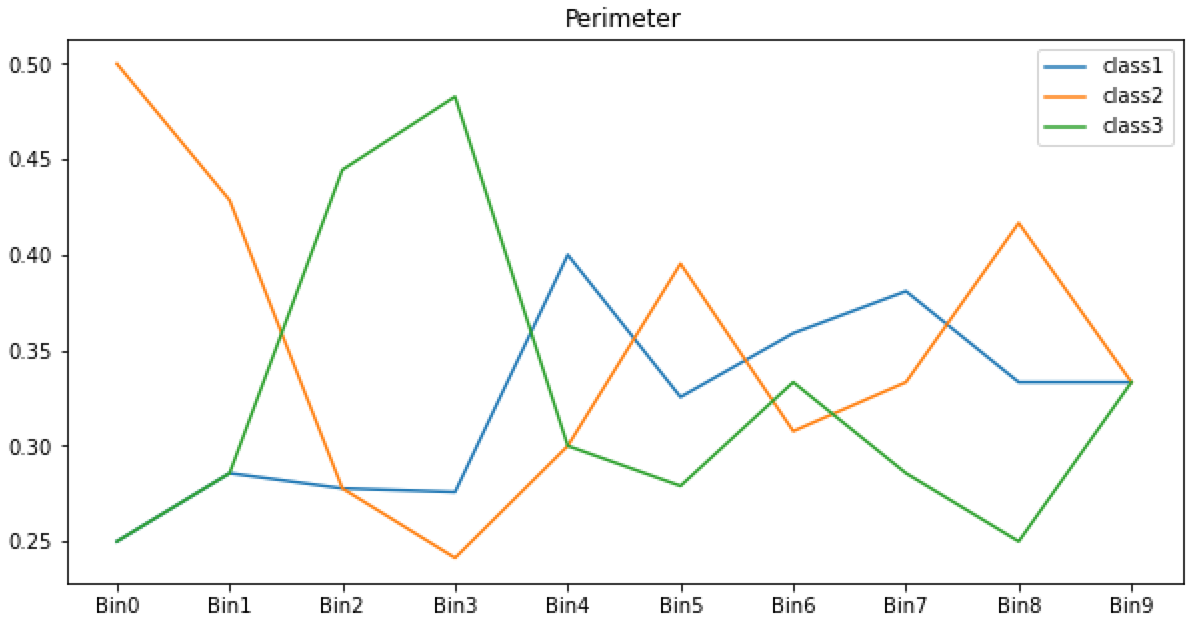




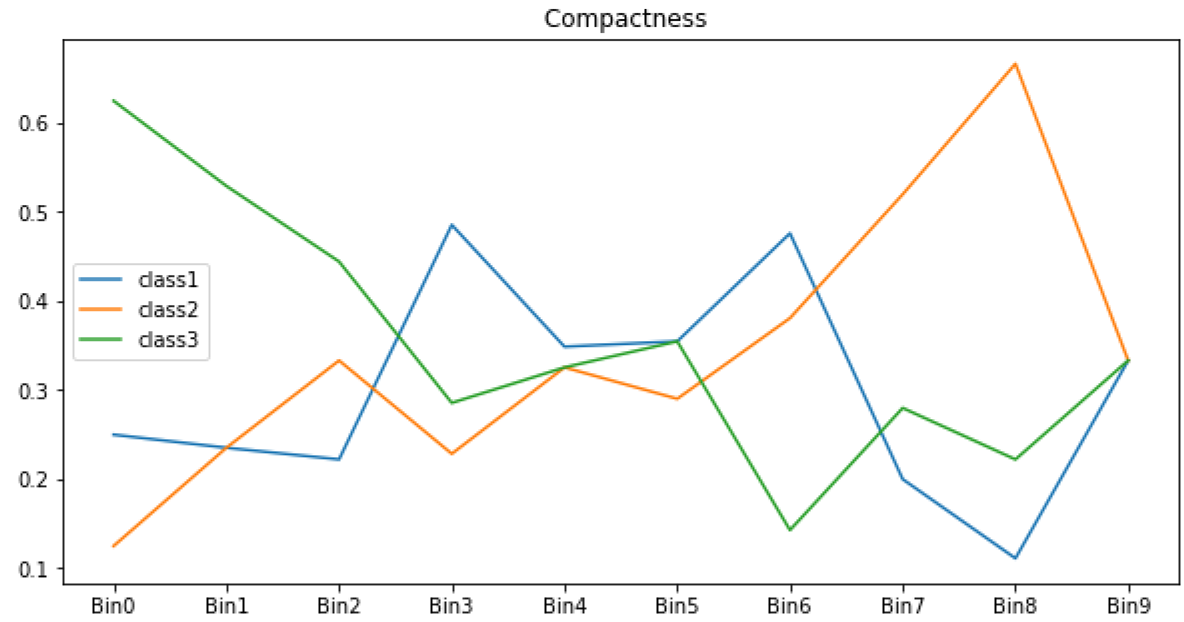
Posterior probabilities and their plots

Posterior probabilities are calculated by multiplying likelihood by prior then dividing it by evidence . The plots are shown below





Perimeter



Compactness