Pattern Recognition and Machine Learning

Lab - 3 Assignment Report

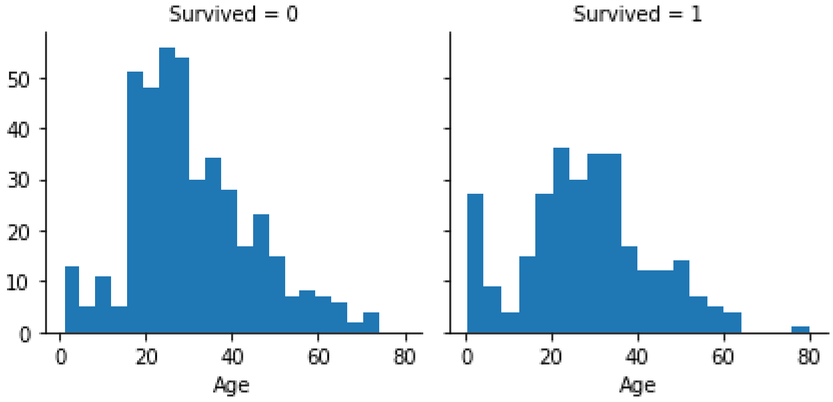
Aryan Himmatlal Prajapati (B21EE012)

Question 1.

Part 1

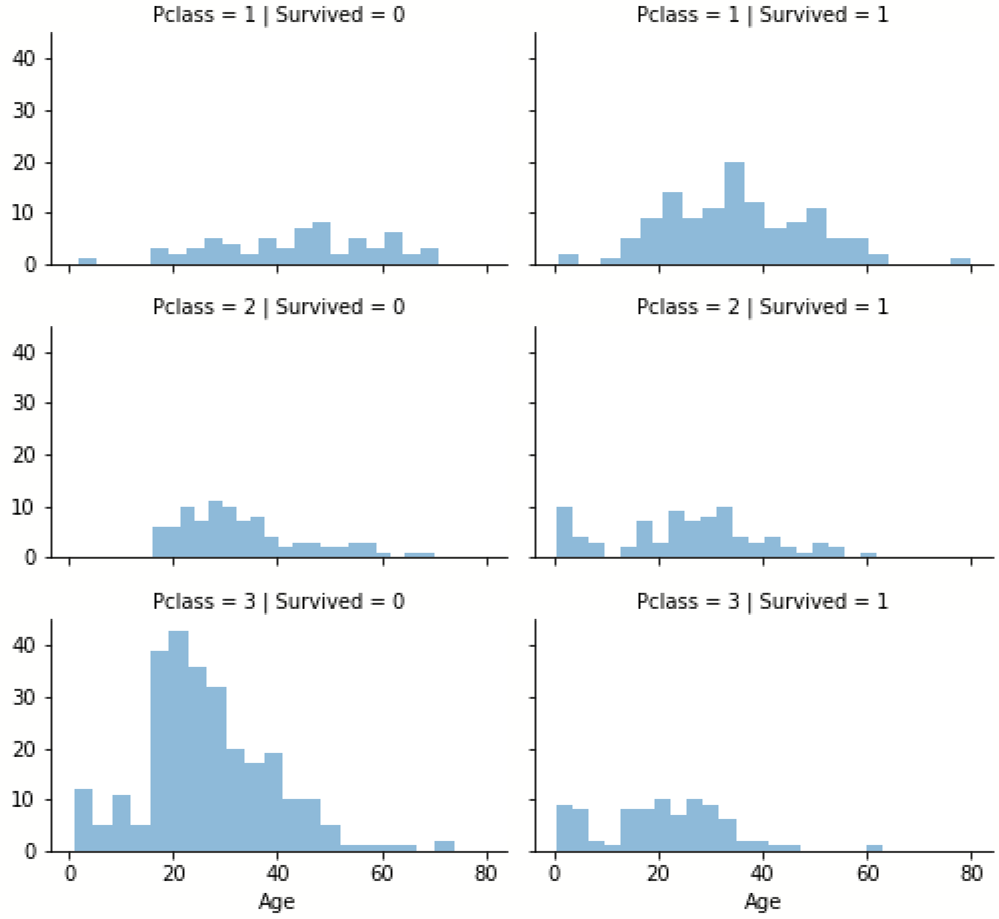
Pre-Processing of Dataset:

1. Data-Type of Features:
    a. Continuous: Age, Fare
    b. Categorical: Survived, Sex and Embarked
    c. Ordinal: Pclass
    d. Alphanumeric: Cabin
2. Analyzing by pivoting features:
    a. In Pclass, We observe significant correlation (>0.5) among Pclass=1 and Survived. We decided to include this feature in our model.
    b. In Sex We confirm the observation during problem definition that Sex=female had a very high survival rate at 74%.
3. Visualization:
    a. Age vs Survived
        i. Observation:
            1. Age <=4 had high survival rate.
            2. Oldest passengers (Age = 80) survived.
            3. Large number of 15-25 year olds did not survive.
            4. Most passengers are in 15-35 age range.



ii. Decision:
We should consider Age in our model training.

b. Pclass | Survived vs Age



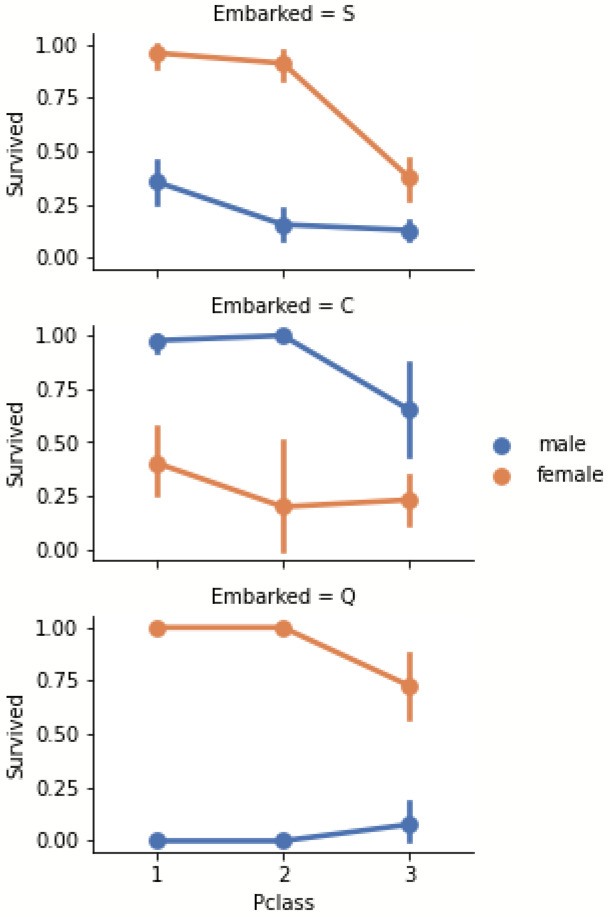
i. Observations:

1. Pclass=3 had most passengers, however most did not survive.
2. Infant passengers in Pclass=2 and Pclass=3 mostly survived.
3. Most passengers in Pclass=1 survived.
4. Pclass varies in terms of Age distribution of passengers.

ii. Decisions:

Consider Pclass for model training.

c. Sex,Embarked, Survived



i. Observations:

1.Female passengers had much better survival rate than males.

2. Males had better survival rate in Pclass=3 when compared with Pclass=2 for C and Q ports.

ii. Decisions:

Add Sex feature and Embarked feature to model training.

d. Embarked, Survived, Fare, Sex

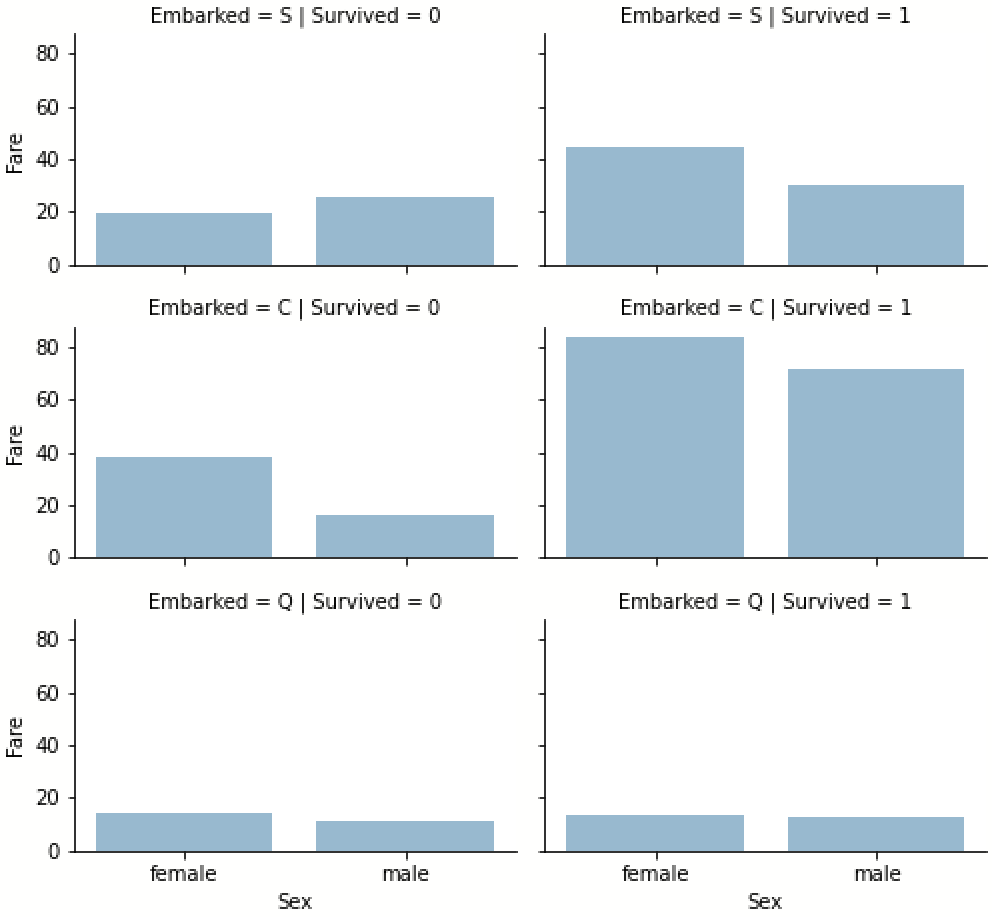
i. Observations:

1.Higher fare paying passengers had better survival.

2. Port of embarkation correlates with survival rates.

ii. Decisions:

Consider banding Fare feature.



4. Removed Unnecessary columns like 'PassengerId','Name','Ticket' and 'Cabin' from dataframe.
5. Replaced Null Values according to features with appropriate estimator:

    Embarked : Mode

    Age: Mean

6. Encoded Categorical data such as Sex and Embarked.
7. Separated Features and Labels.

8.Splitted the data into train and test sets in ratio 70:30.

Part 2

Out of the 5 different Naive Bayes classifiers under sklearn. naive_bayes, the 3 most widely used ones are **Gaussian, Multinomial,** and **Bernoulli.**

Bernoulli's naive bayes is used when features are binary. Since some features in the dataset are not of binary type , we should not use the Bernaoulli Naive Bayes Classifier.

The multinomial naïve Bayes is widely used for assigning documents to classes based on the statistical analysis of their contents. The dataset is not a discretized set , we do not care about the count of fair and age , as they are not discrete , They are Continuous. Hence , We should not use Multinomial Naive Bayes Classifier .

Gaussian Naïve Bayes is used when we assume all the continuous variables associated with each feature to be distributed according to Gaussian Distribution.

As Identified that the dataset contains continuous data points and for obtaining probabilities of such continuous data , Guassian distribution for probability should be used , Hence , the variant of Naive Bayes Classifier best suited for this case is **Gaussian Naive Bayes Classifier.**

Part 3

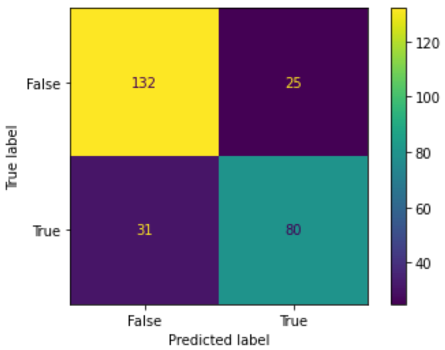
Metrics to Measure Classification Performance:

1. Accuracy:

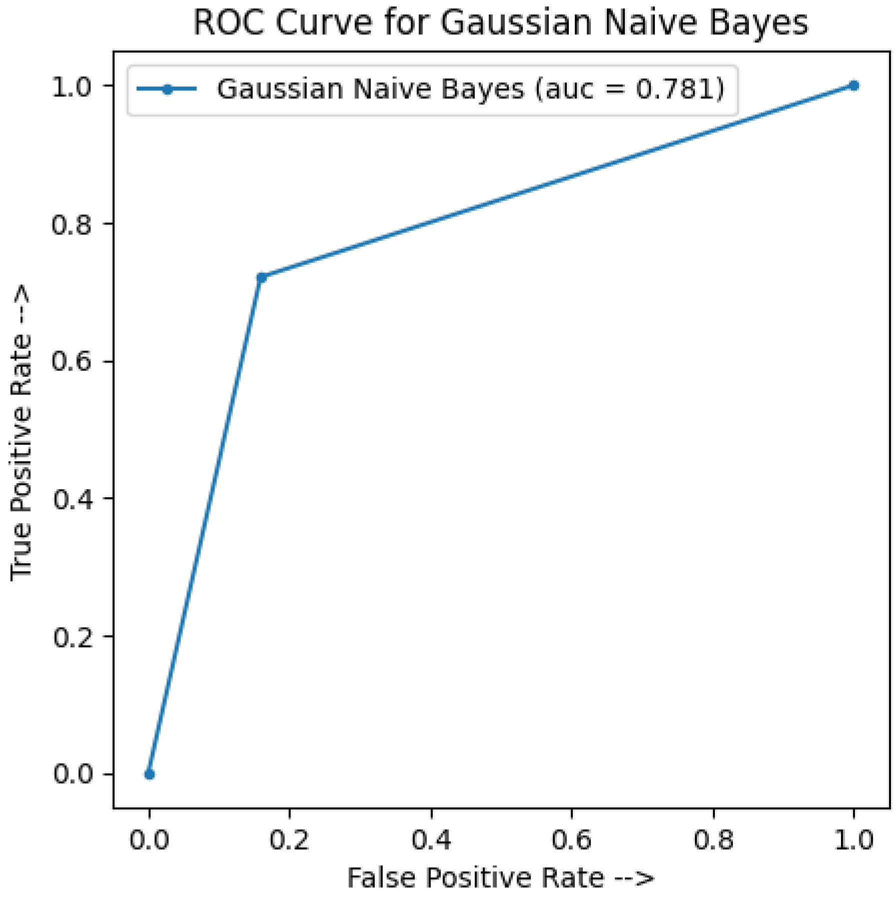
```
Accuracy Score of model on testing dataset is: 79.1044776119403 %
```

2. Confusion matrix:

```
[[132  25]
 [ 31  80]]
```



3. AUC/ROC:



4. Precision: `76.19047619047619 %`

5. Recall: `72.07207207207207 %`

6. F1 Score: `74.07407407407408 %`

7. Kappa:

   The **kappa** statistic compares the observed accuracy to an expected accuracy or the accuracy expected from random chance. One of the flaws of pure accuracy is that if a class is imbalanced then making predictions at random could give a high accuracy score. Kappa accounts for this by comparing the model accuracy to the expected accuracy based on the number of instances in each class.

Essentially it tells us how the model is performing compared to a model that classifies observations at random according to the frequency of each class.

$$kappa = (ObservedAccuracy - ExpectedAccuracy)/(1 - ExpectedAccuracy)$$

```
Kappa: 0.5659667996992307
```

8. MCC (Matthews Correlation Coefficient):

MCC (Matthews Correlation Coefficient) is generally considered one of the best measurements of performance for a classification model. This is largely because, unlike any of the previously mentioned metrics, it takes all possible prediction outcomes into account. If there are imbalances in the classes this will therefore be accounted for.

The MCC is essentially a correlation coefficient between the observed and predicted classifications. As with any correlation coefficient, its value will lie between -1.0 and +1.0. A value of +1 would indicate a perfect model.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

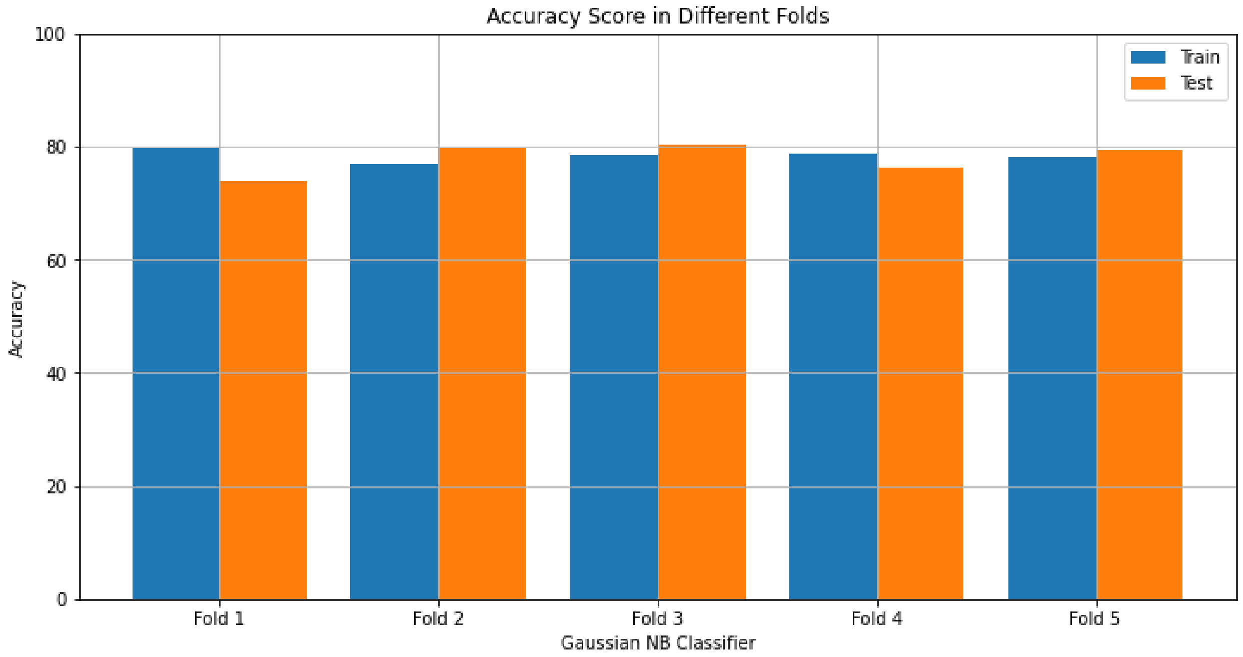
```
MCC (Matthews Correlation Coefficient): 0.5665797685503138
```

Part 4

5 Fold Cross Validation:

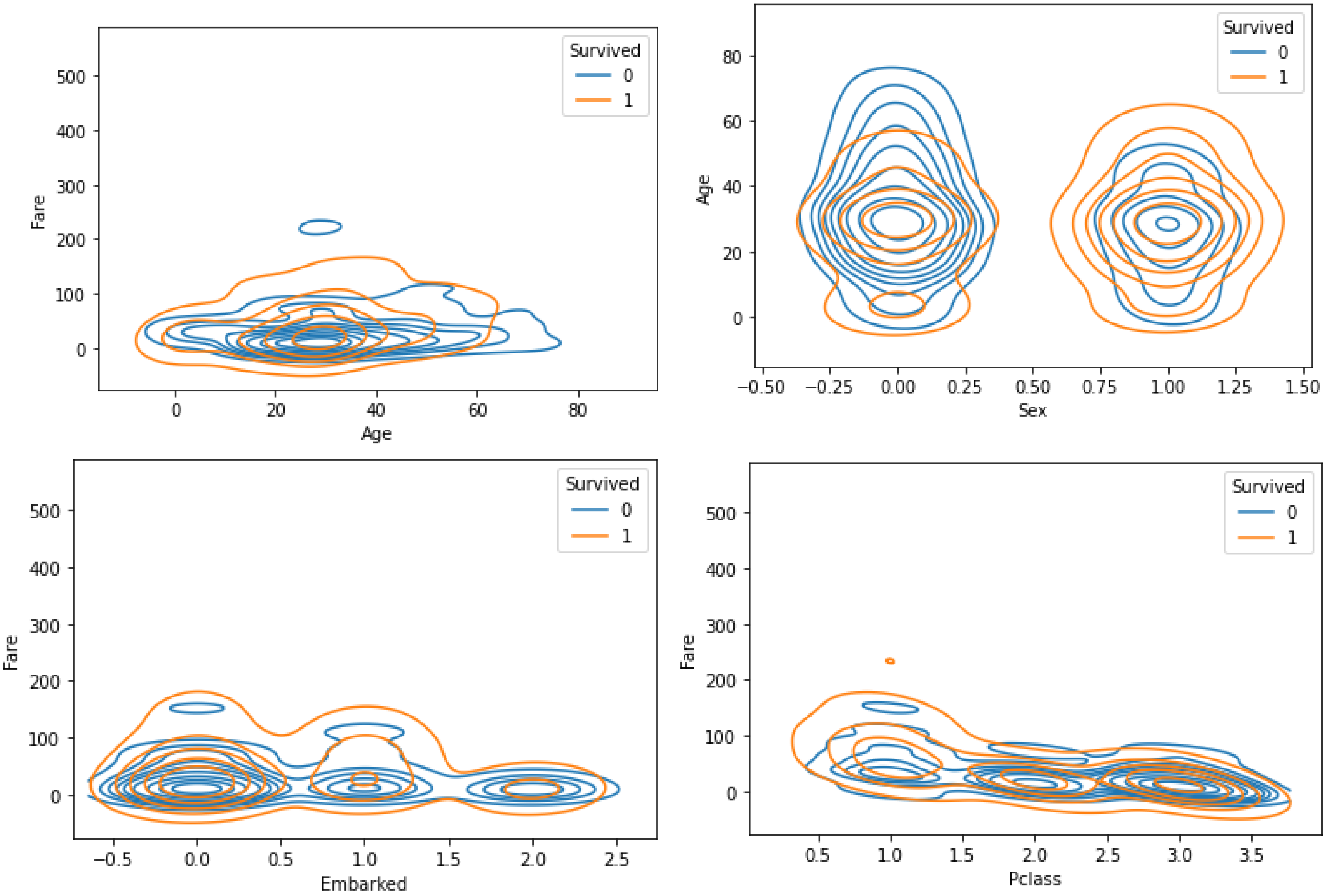
```
Train Accuracy Scores:

Fold 1 : 79.7752808988764 %
Fold 2 : 76.99859747545582 %
Fold 3 : 78.40112201963534 %
Fold 4 : 78.68162692847125 %
Fold 5 : 78.26086956521739 %

Average Train Accuracy score: 78.42349937753124 %

Test Accuracy Scores:

Fold 1 : 73.74301675977654 %
Fold 2 : 79.7752808988764 %
Fold 3 : 80.33707865168539 %
Fold 4 : 76.40449438202246 %
Fold 5 : 79.21348314606742 %

Average Test Accuracy score: 77.89467076768565 %
```

Visualization:



Part 5

Contour Plots:



a Naive Bayes classifier assumes that **the presence of a particular feature in a class is unrelated to the presence of any other feature**.

From Observed Contour Plots, We observe that data is interdependent on each other amongst features. Females have Higher chance of survival than males. Similarly in other features.

So from the observations, We can conclude that assumptions of naive bayes on this particular dataset are not strong enough.

Part 6

Decision Tree Classifier Vs Gaussian Naive Bayes Classification-

1. Accuracy:

```
Accuracy Score for Decision Tree Classifier: 76.86567164179104 %
Accuracy Score for Gaussian Naive Bayes: 79.1044776119403 %
```

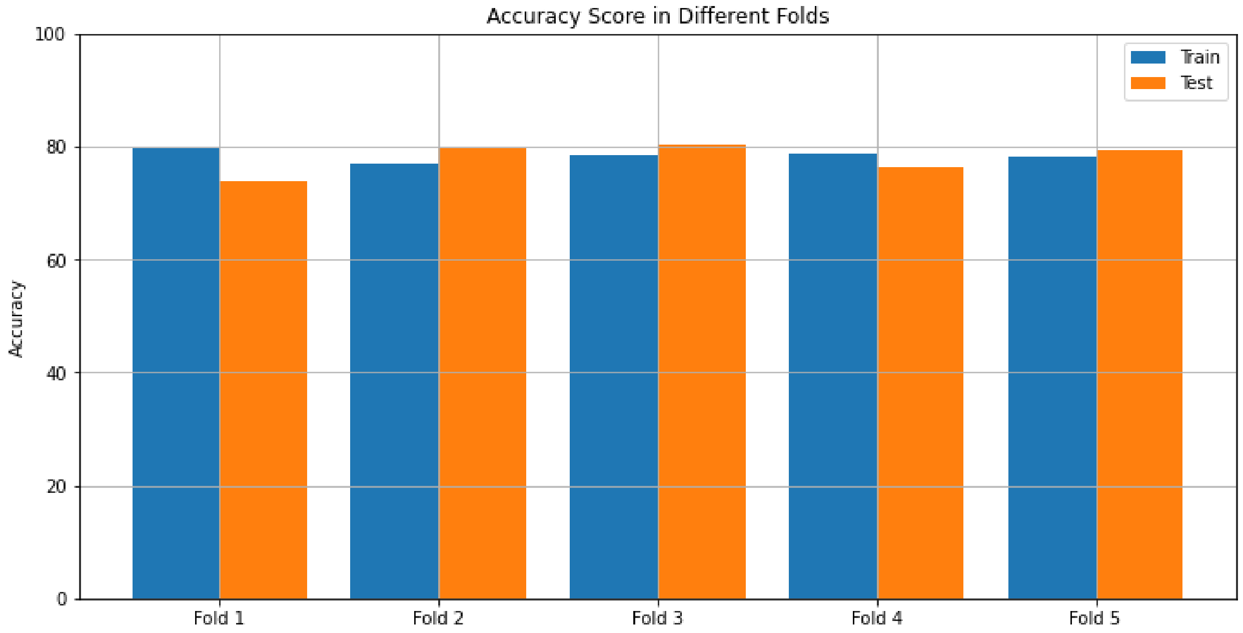
2. 5 fold cross validation:

```
Decision Tree Classifier:

Train Accuracy Scores:
Fold 1 : 98.17415730337079 %
Fold 2 : 98.45722300140253 %
Fold 3 : 98.17671809256662 %
Fold 4 : 98.45722300140253 %
Fold 5 : 98.03646563814866 %

Average Train Accuracy score: 98.26035740737822 %

Test Accuracy Scores:
Fold 1 : 72.06703910614524 %
Fold 2 : 79.21348314606742 %
Fold 3 : 80.89887640449437 %
Fold 4 : 76.40449438202246 %
Fold 5 : 81.46067415730337 %

Average Test Accuracy score: 78.00891343920657 %

--------------------------------------------------------------------
Gaussian Naive Bayes:

Train Accuracy Scores:
Fold 1 : 79.7752808988764 %
Fold 2 : 76.99859747545582 %
Fold 3 : 78.40112201963534 %
Fold 4 : 78.68162692847125 %
Fold 5 : 78.26086956521739 %

Average Train Accuracy score: 78.42349937753124 %
```

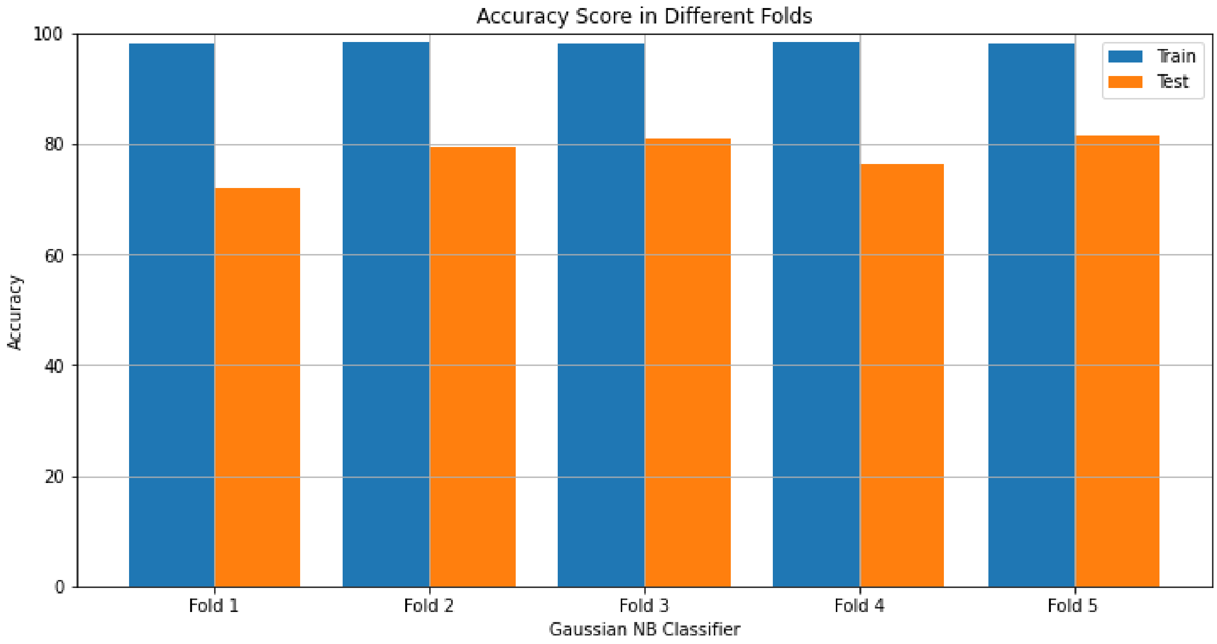
```
Test Accuracy Scores:
Fold 1 : 73.74301675977654 %
Fold 2 : 79.7752808988764 %
Fold 3 : 80.33707865168539 %
Fold 4 : 76.40449438202246 %
Fold 5 : 79.21348314606742 %

Average Test Accuracy score: 77.89467076768565 %
```

Visualization:

Decision Tree Classifier:



Gaussian Naive Bayes:



Accuracy Score of Gaussian Naive Bayes was better than that of the Decision Tree Classifier.

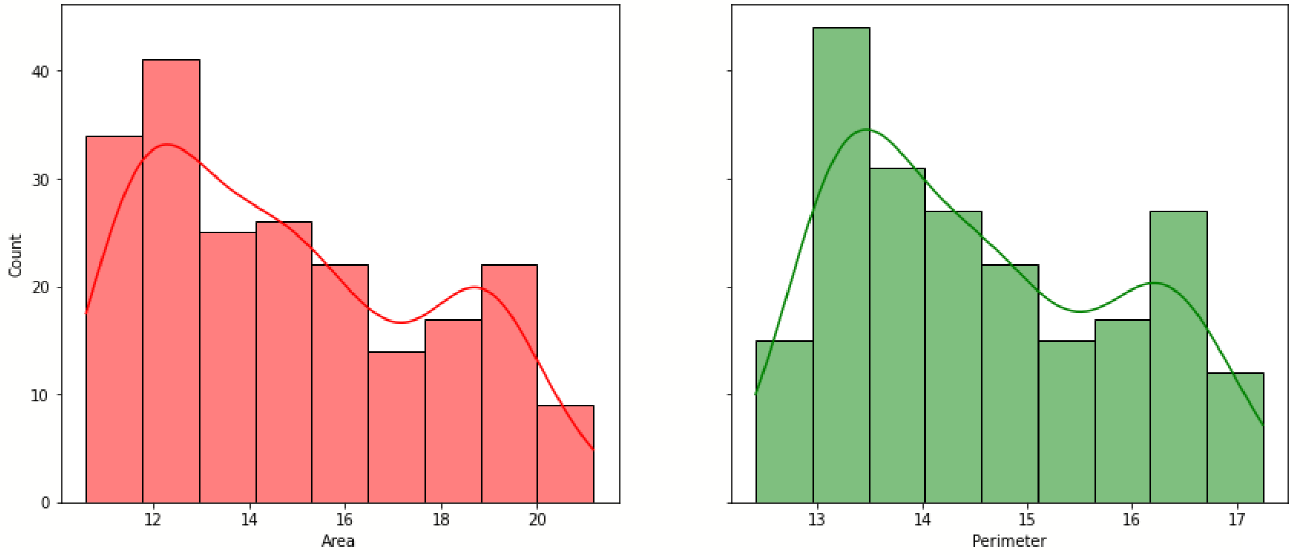
After performing the 5 fold cross validation the Accuracy scores of the decision tree classifier was better than gaussian naive bayes. But this is due to overfitting of the model as we can see the difference between before and after accuracy score for decision tree classifier is high. Also Training data accuracy score is much higher than the testing data accuracy score.

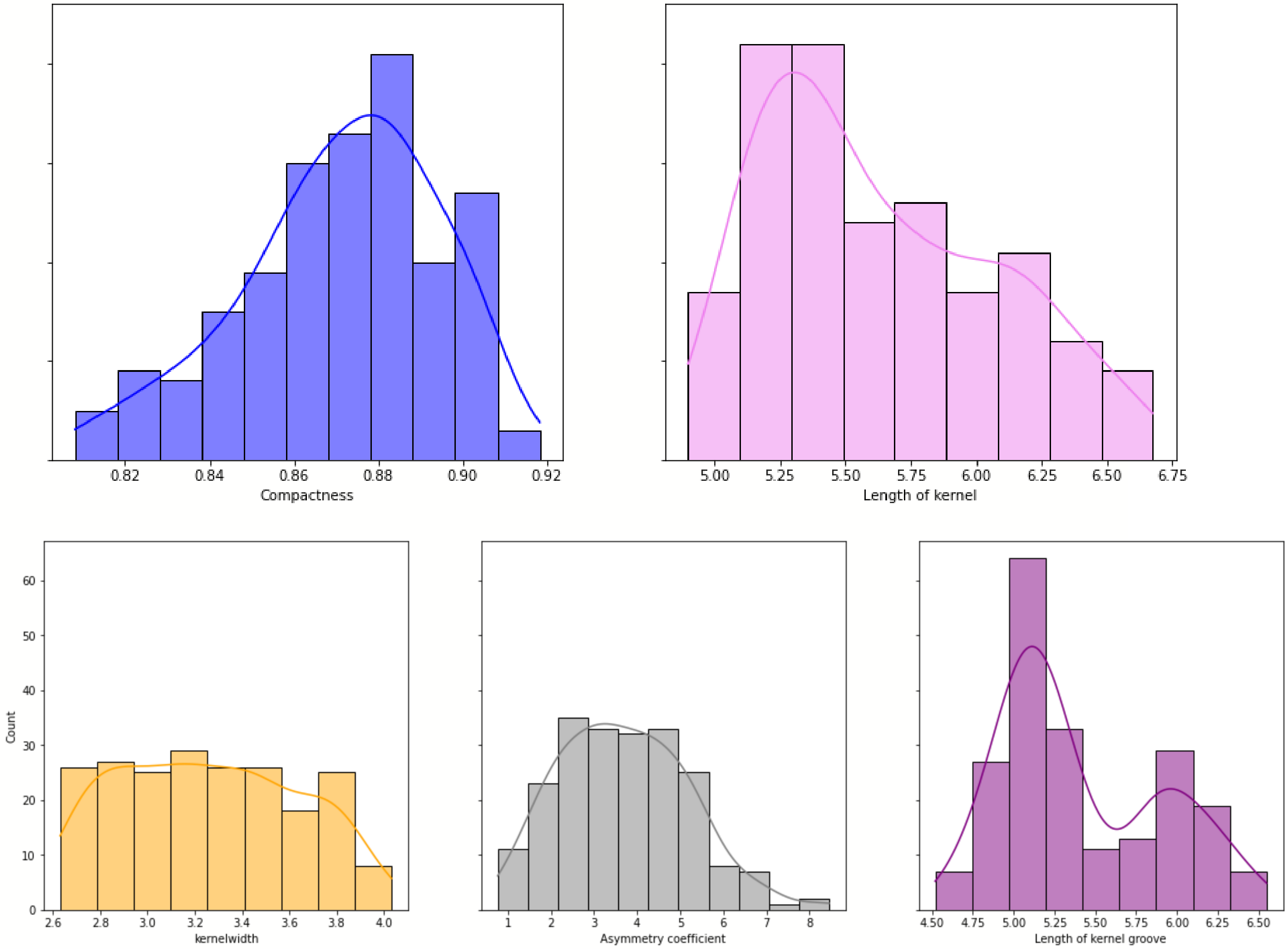
So after these observations we can say that Gaussian Naive Bayes is working better than decision tree classification on this dataset. It is not overfitted because there is not much difference in accuracy scores in different folds.

Question 2.

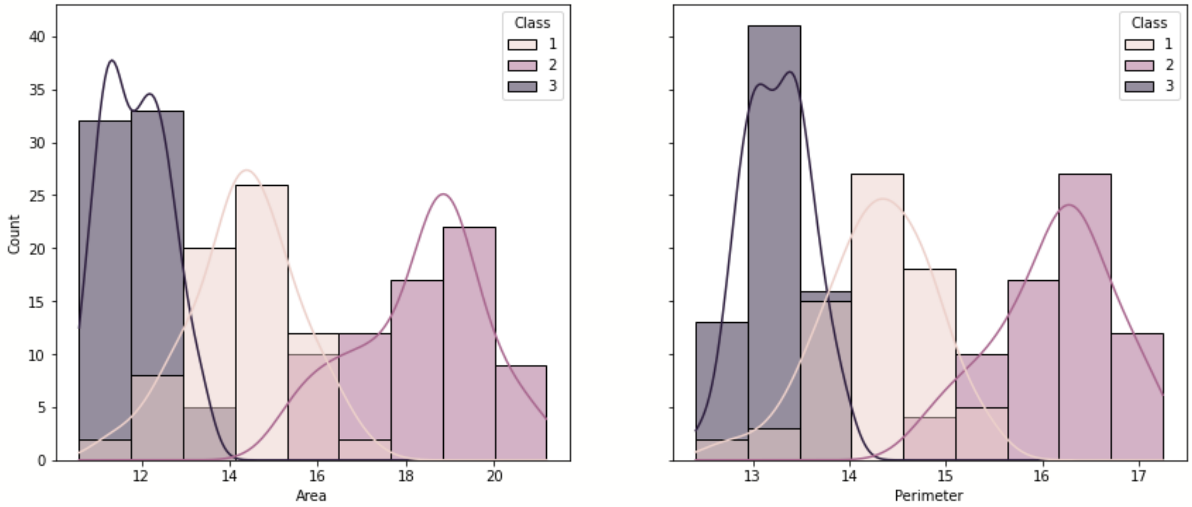
Part 1

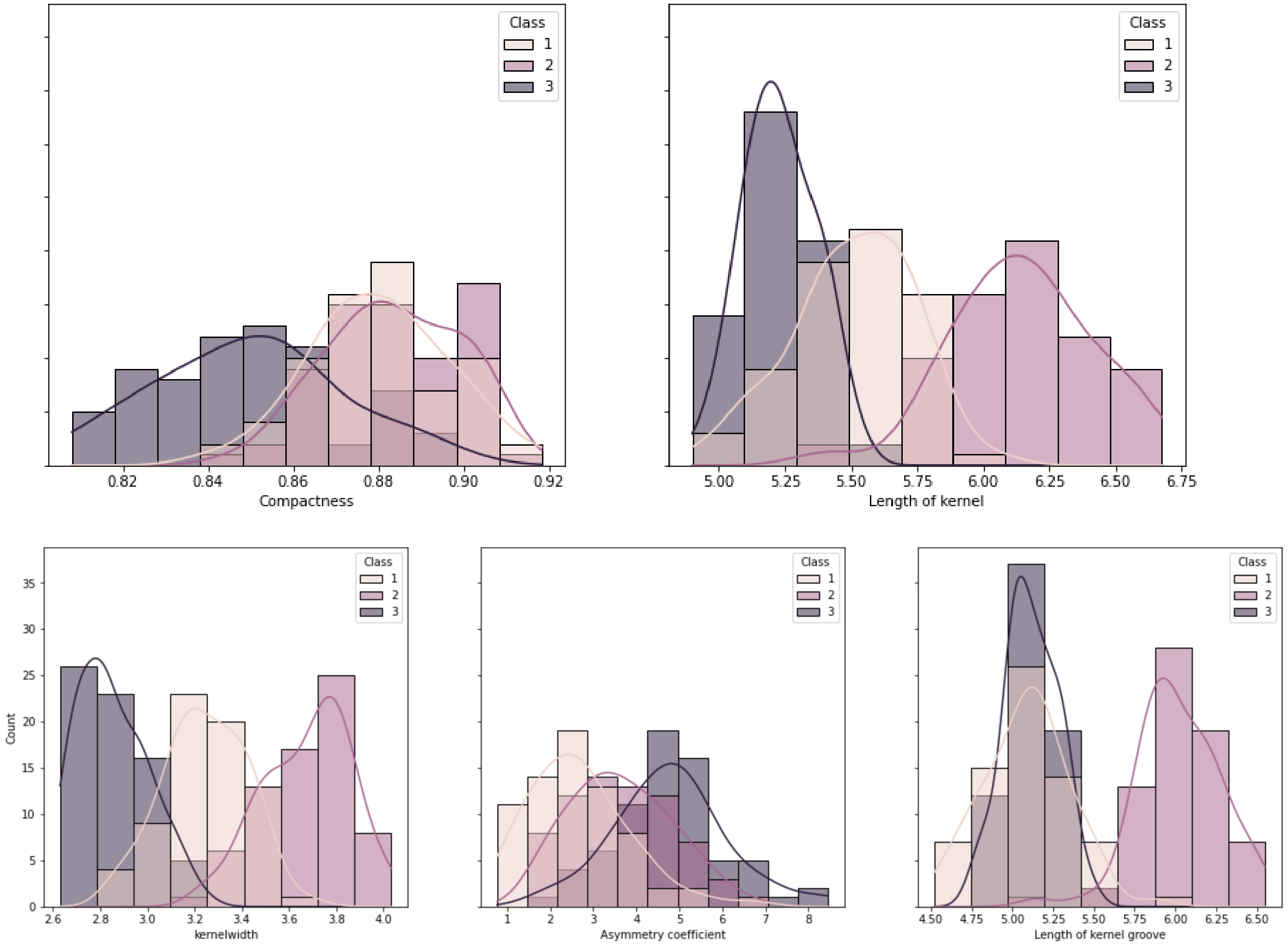
Distribution of Features:





Distribution of Class:





Part 2

```
Prior Probability of Class 1 is: 0.3333333333333333
Prior Probability of Class 2 is: 0.3333333333333333
Prior Probability of Class 3 is: 0.3333333333333333
```

Part 3

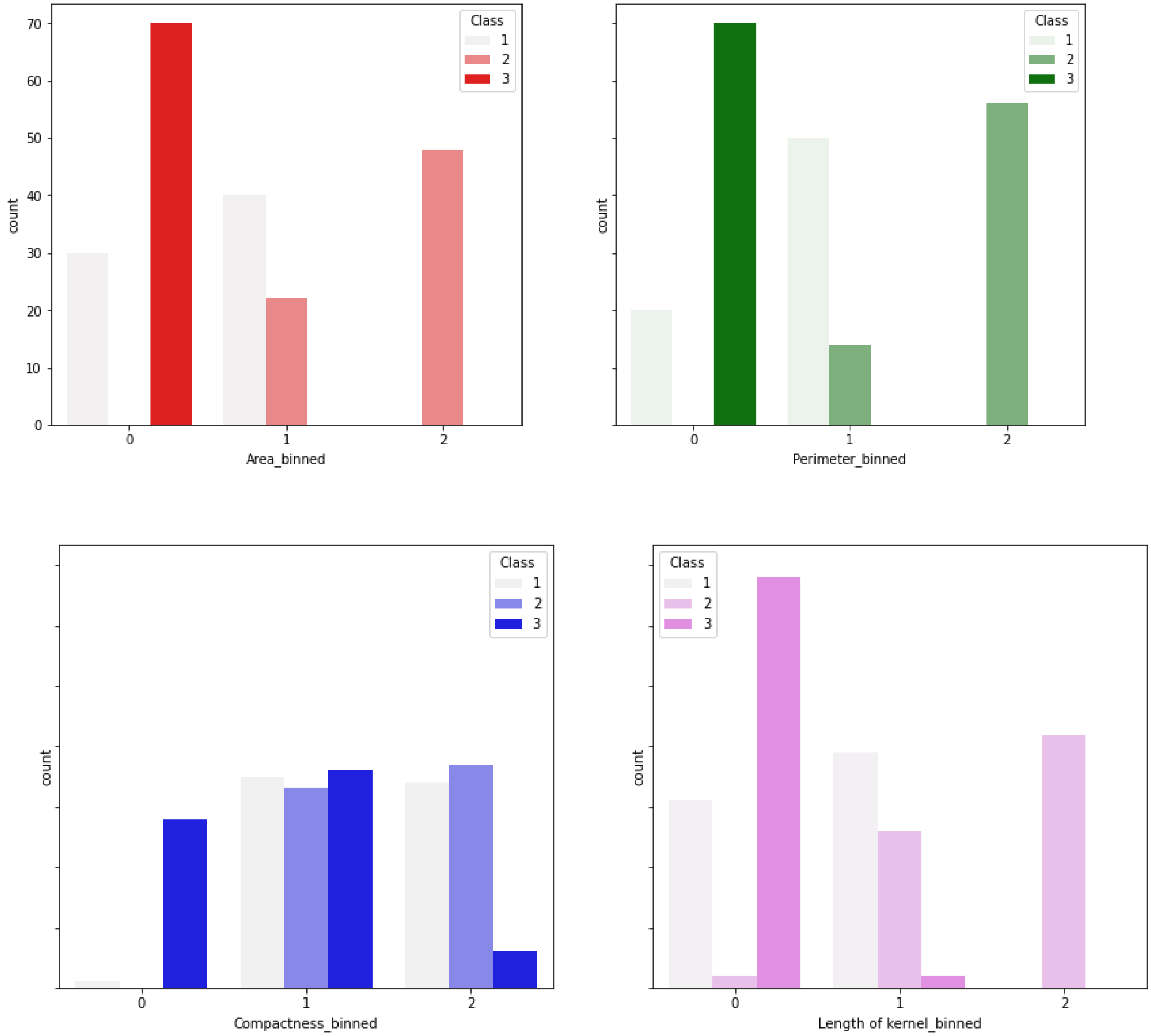
Discretized features dataset into 3 bins.(more details in .ipynb)

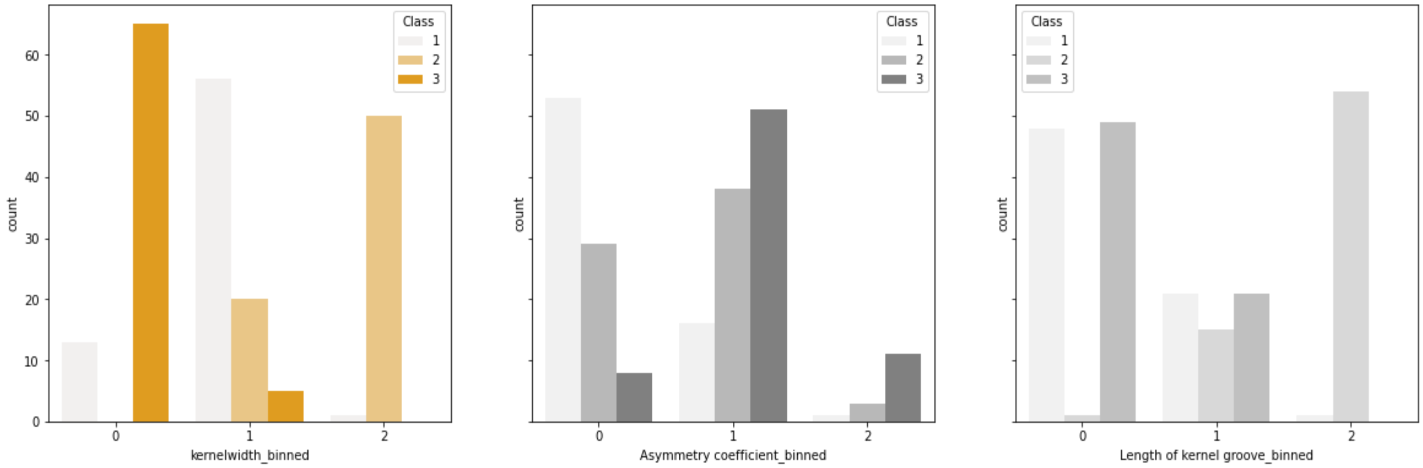
Part 4

Calculated Likelihood Probability without using any inbuilt libraries or functions. Implemented from Scratch. More details in .ipynb file.

Part 5

Count Plot for Each Feature:





Comparing these plots with plots of distribution:

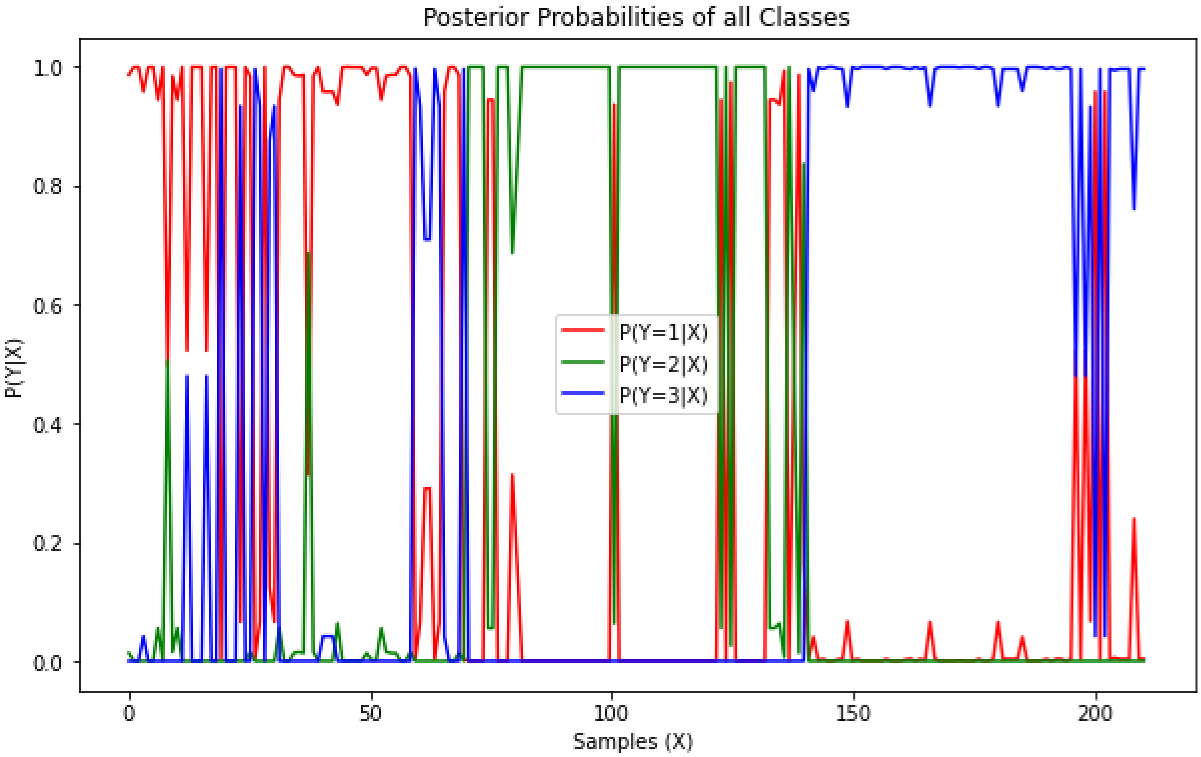
These Plots represent the data and distribution of unique elements of each class in a satisfactory manner , while preserving the original distribution but in a more discrete manner, which is indeed due to the bins that have been created.

For example, We observe that in the count plot of “Area”, there are more low values of “Area” which lie in Class 3, which are also obtained here in the count plot. Similarly, for other features as well.

Part 5

Calculated posterior probabilities and plotted them into Single Graph

Graph:



It is easily seen that the sum of probabilities of each of the three classes is equal to 1. Which we know is obvious. Also we can see that there is a big difference between the probabilities of different classes at the same point. Hence, the probability to classify an Element into some class is greater with a large enough margin than the other 2 classes.

Thank You.