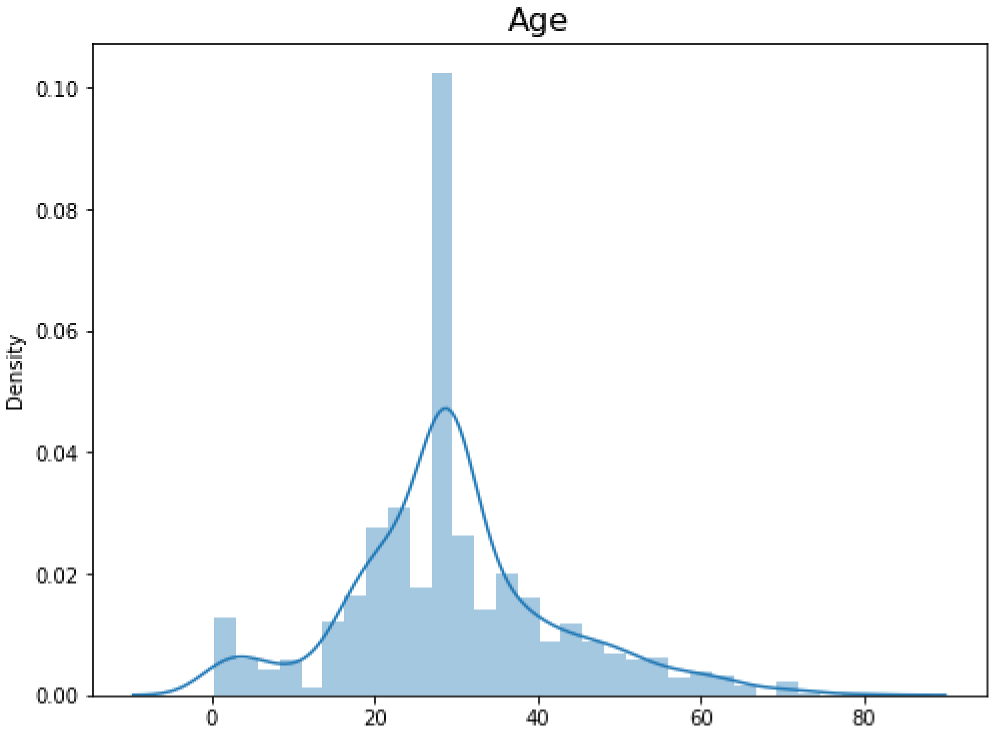
Maniya Yash Rajeshbhai
B20CS033

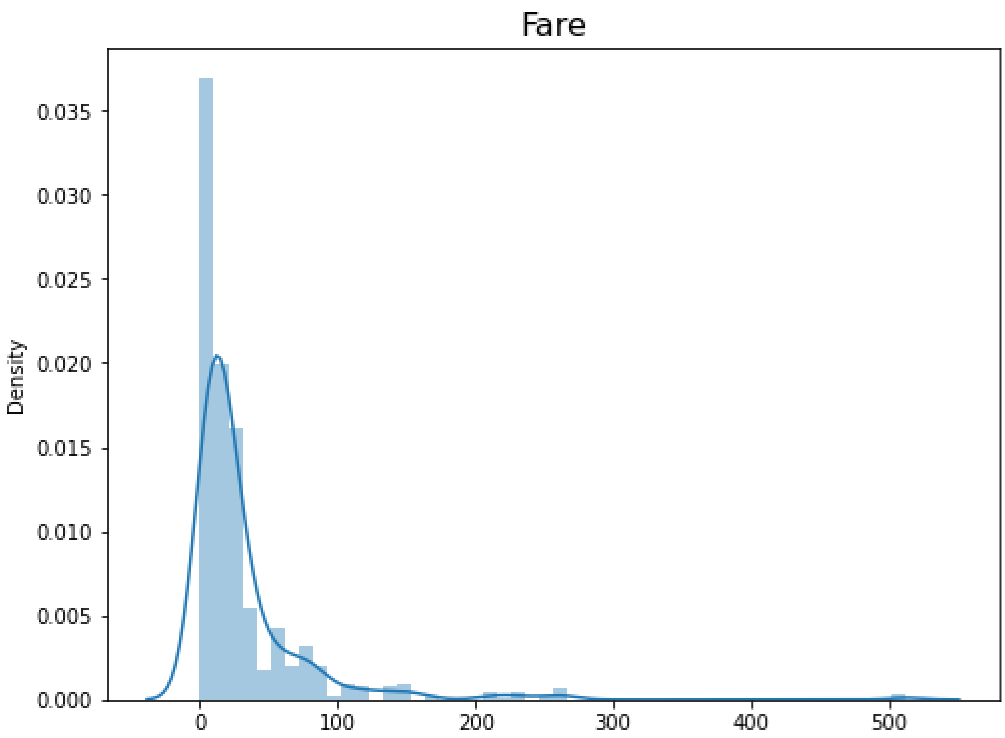
ML Lab 4 Report

Question 1 :

1. Preprocessing and Visualisation :

- Dropped unuseful features like : passengerID, name, ticket. embarked, Cabin.
- Age had many NA values so replaced them with mean age.
- Then plotted distributions of all the features for analysing them. Here are Age, and Fare plots which have continuous variables.





2. Variant Selection :

- Here Age and Fare features have continuous values, and on analysing their distribution plots, we see that they are almost similar to normal distributions, if we ignore some of the initial points.
- We know that for the data following Normal Distributions, we can use Gaussian Naive Bayes Classifier. So, we will use it for classification purpose.

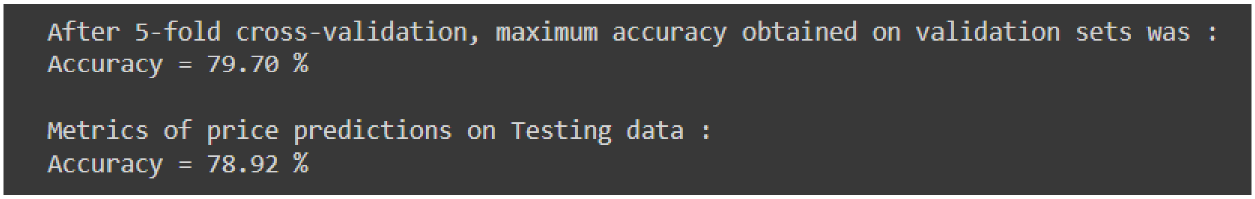
3. Naive Bayes Classifier from Scratch :

- Made a NaiveBayesClassifier() class for the given purpose. It has a fit method which helps to create a model after passing in the training data.
- The model compares Posterior Probability of all the classes and then predicts the class with maximum probability.

4. 5-Fold Cross Validation :

- Applied 5-fold cross validation and selected the optimum model which does not overfit or underfit.

- Also calculated Accuracy scores on validation and testing datasets.
- Below are the metrics for the same :

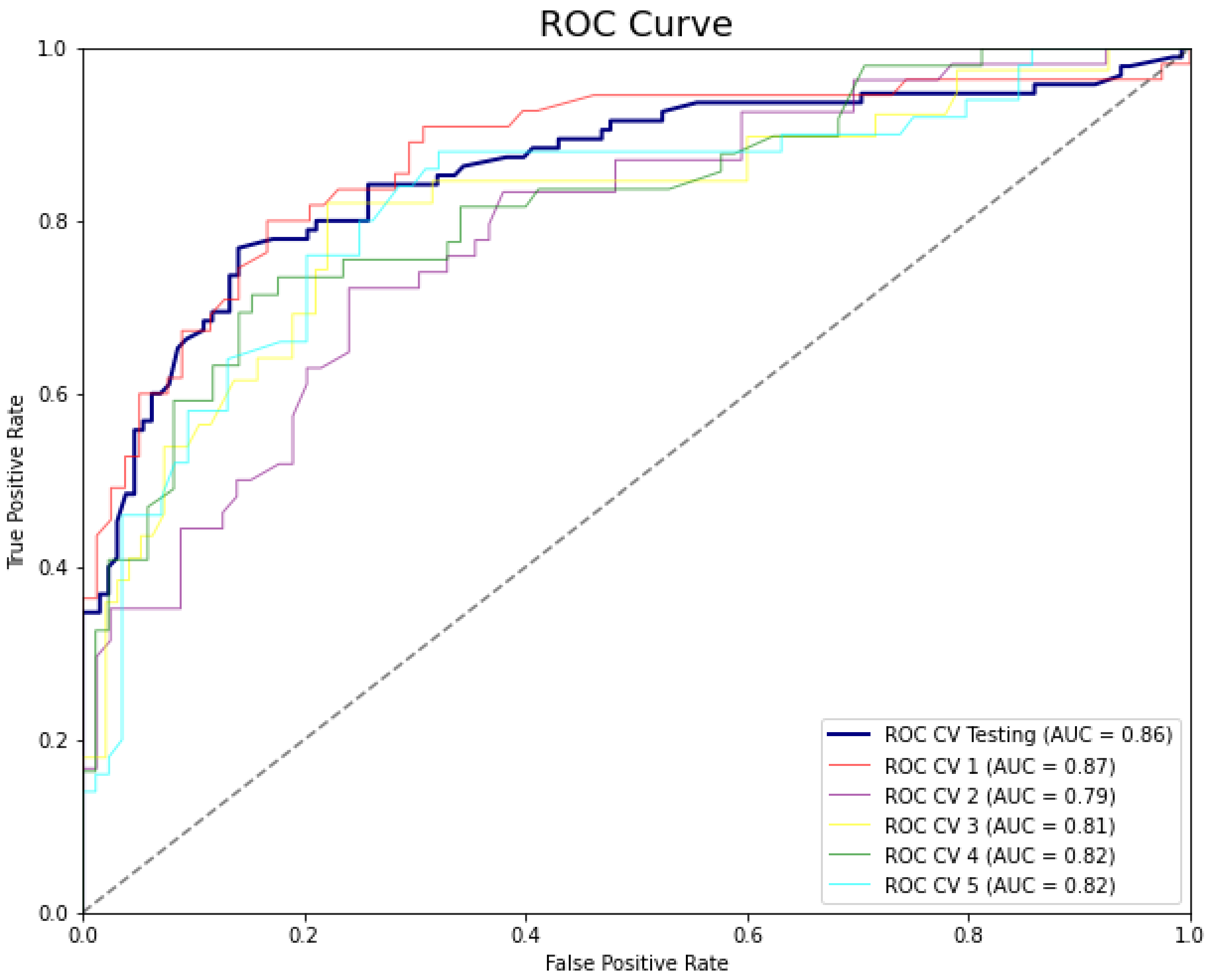


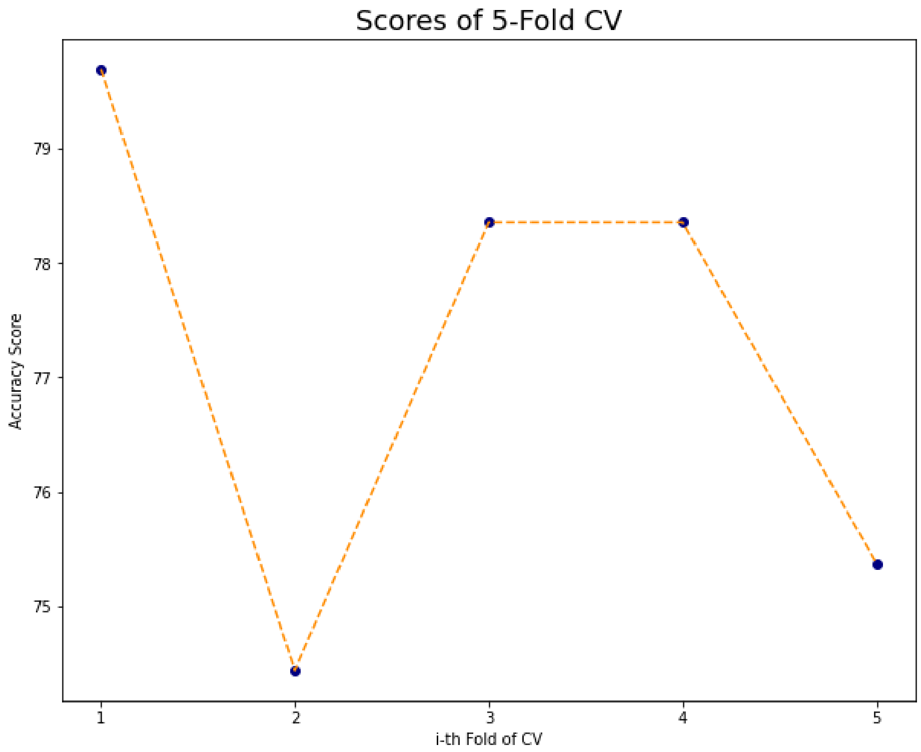
```
After 5-fold cross-validation, maximum accuracy obtained on validation sets was :
Accuracy = 79.70 %

Metrics of price predictions on Testing data :
Accuracy = 78.92 %
```

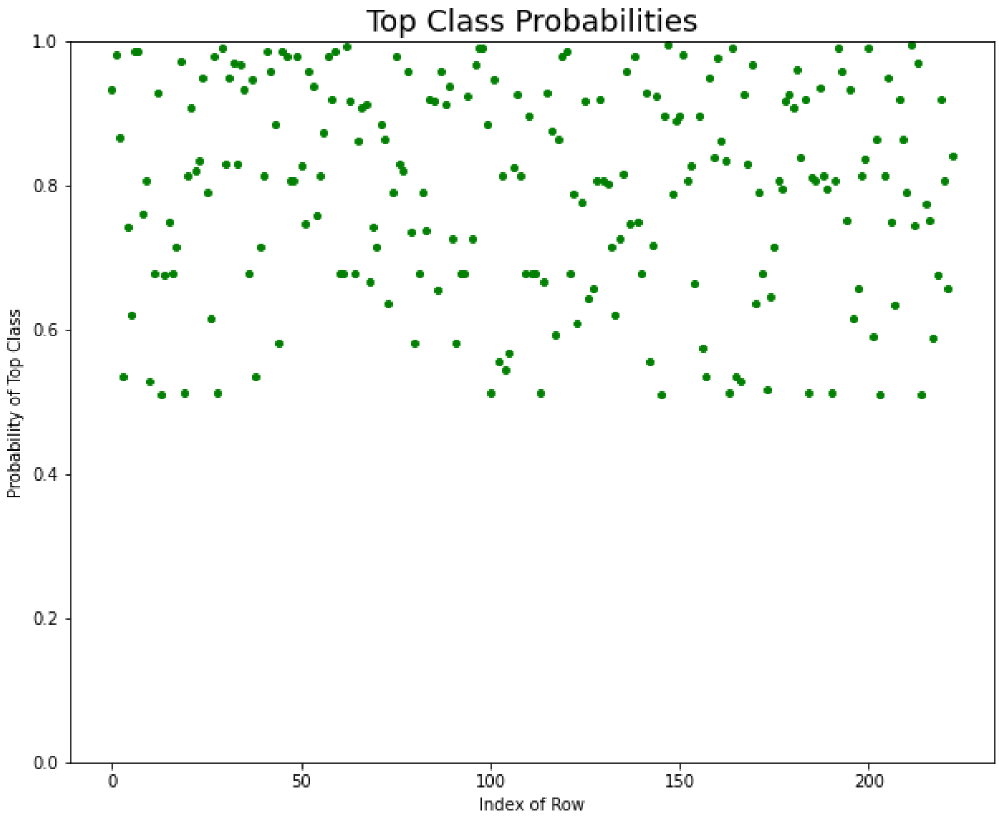
5. Visualising the Cross Validation Results :

- Plotted ROC curves for all 5 cross validation sets, as well as the testing dataset. Plotted Accuracy scores of validation sets too.



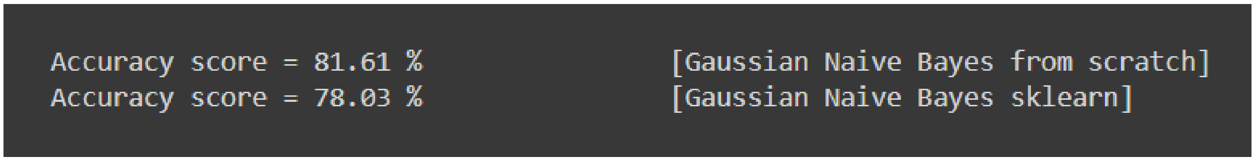


- Under the second sub task, plotted all the top class probabilities of all the rows. We see that **0.5 <= P <= 1**. (as it is winning probability)



6. Comparing Classifiers (scratch built and sklearn) :

- Compared the accuracies of our model and the inbuilt sklearn model.

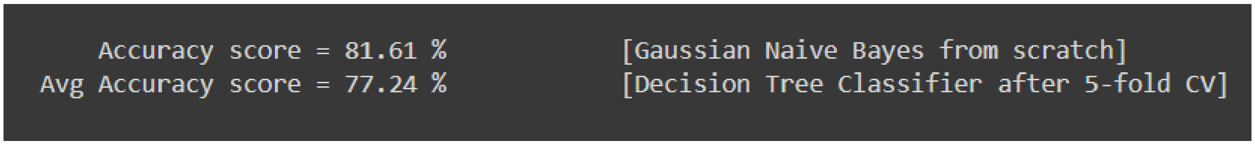


```
Accuracy score = 81.61 %                [Gaussian Naive Bayes from scratch]
Accuracy score = 78.03 %                [Gaussian Naive Bayes sklearn]
```

We see that the model we built gives a better score.

7. Classifying on Decision Tree:

- Compared the accuracies of our model and the inbuilt sklearn Decision Tree model after 5-fold cross validation.



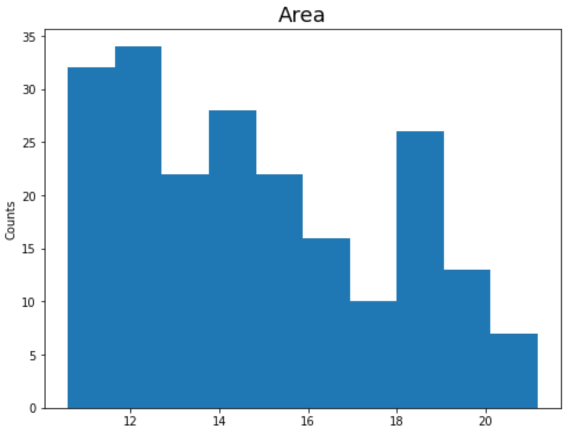
```
    Accuracy score = 81.61 %            [Gaussian Naive Bayes from scratch]
Avg Accuracy score = 77.24 %            [Decision Tree Classifier after 5-fold CV]
```

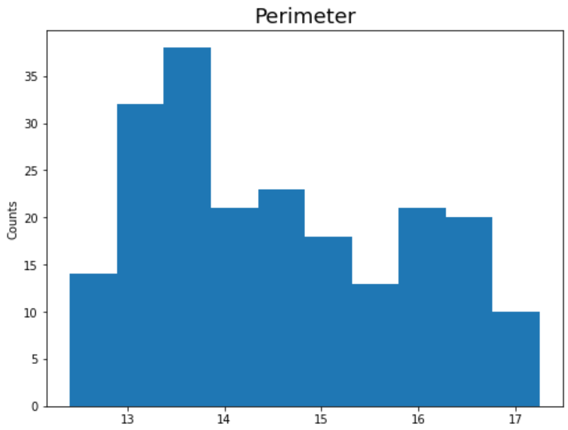
We see that the model we built gives a better score.

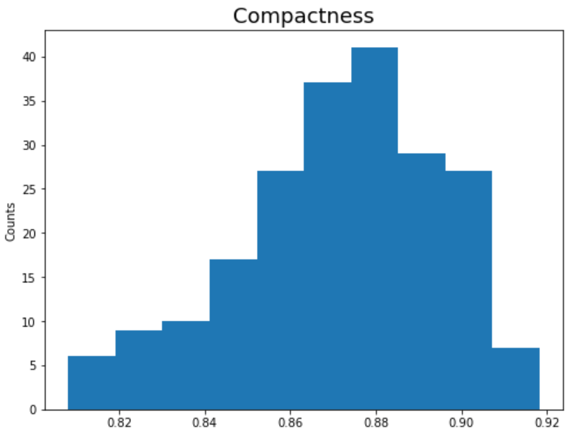
Question 2 :

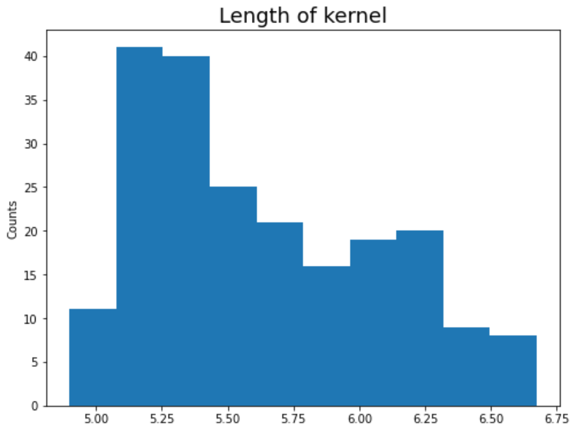
A. Visualising Distributions :

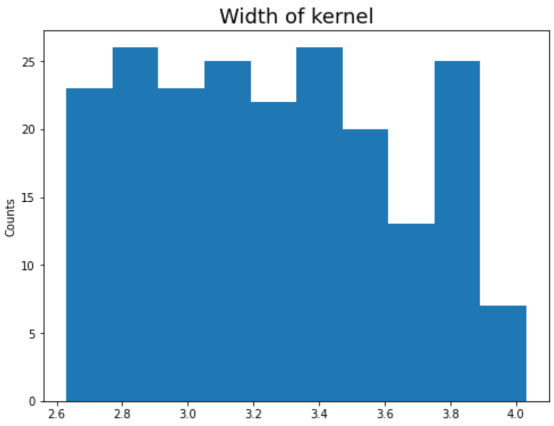
- Used histograms to visualise the distributions of the features. Six of them are shown here.

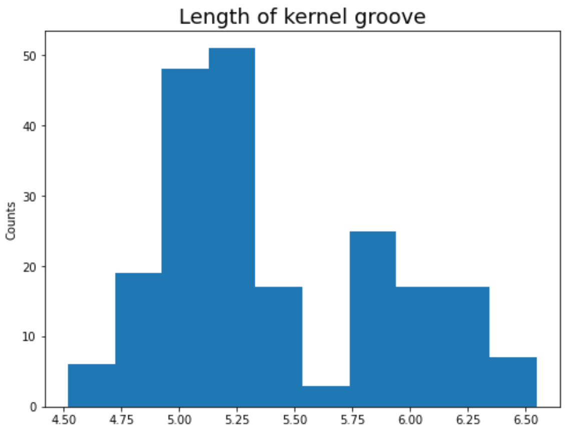






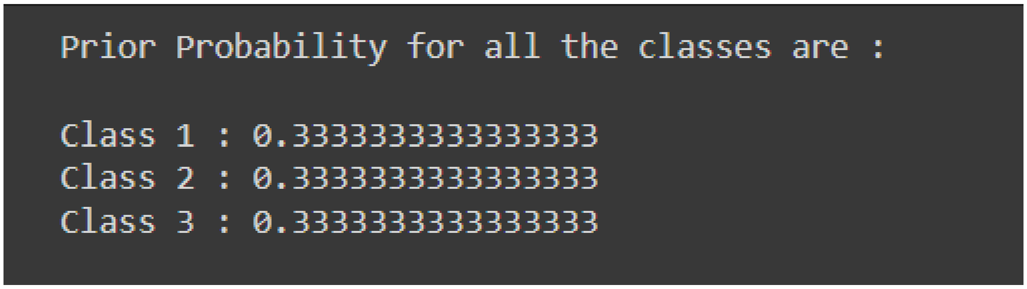






B. Prior Probabilities :

- Calculated the prior probabilities of all the classes :



```
Prior Probability for all the classes are :

Class 1 : 0.3333333333333333
Class 2 : 0.3333333333333333
Class 3 : 0.3333333333333333
```

C. Discretize into bins :

For all the features, we made 10 bins and converted continuous data into class-wise discrete data.

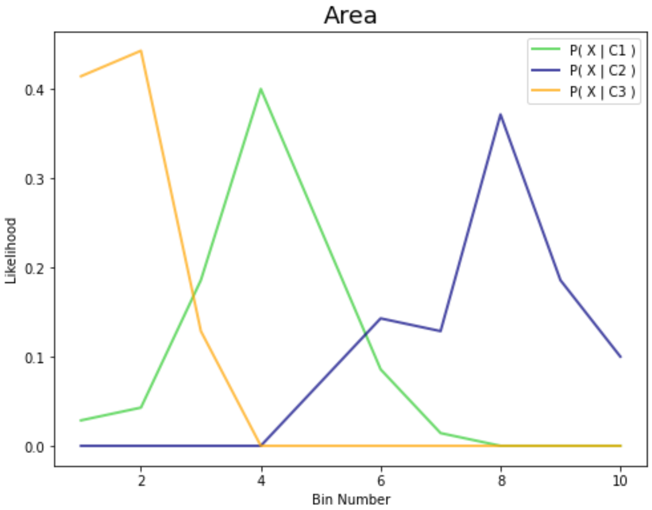
Took bin_size as follows :

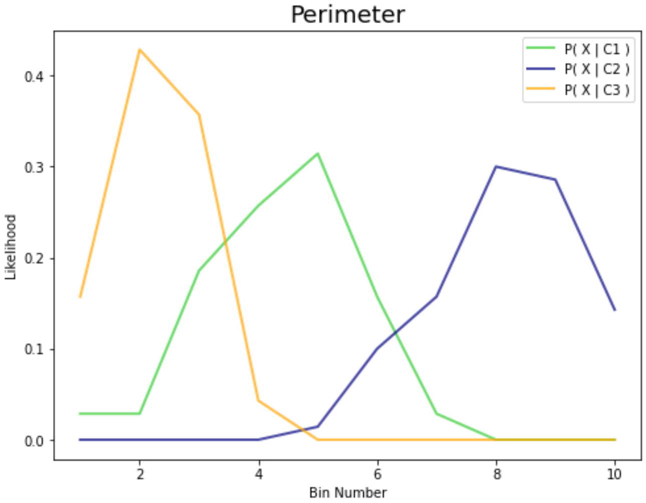
bin_size = (max_value - min_value) / 10

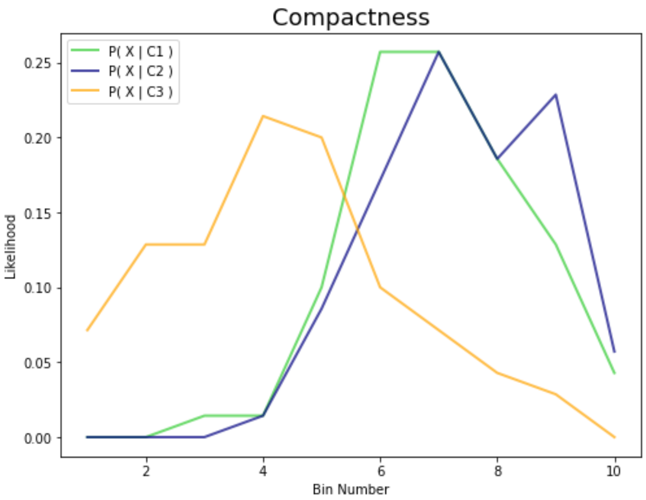
D. Likelihood Calculation :

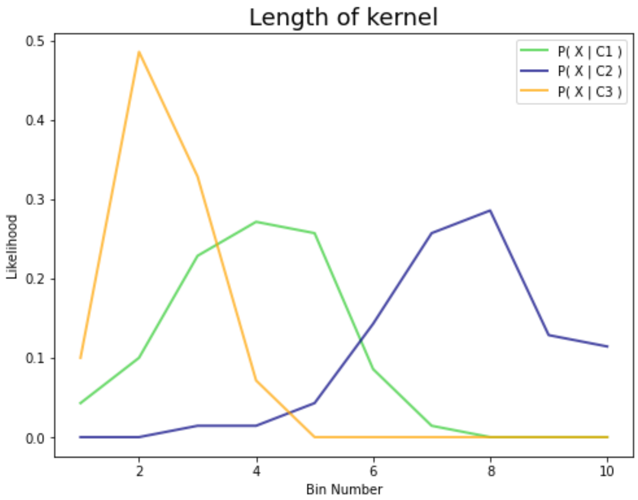
Calculated likelihood probabilities for all the features, bin-wise.

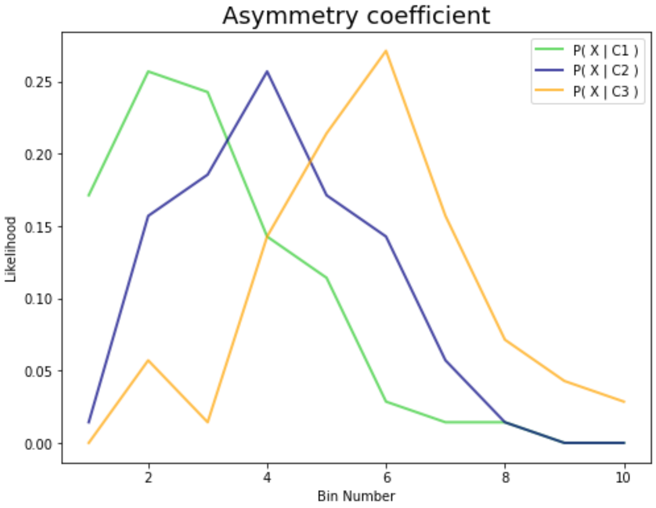
Here are the plots of 6 features :

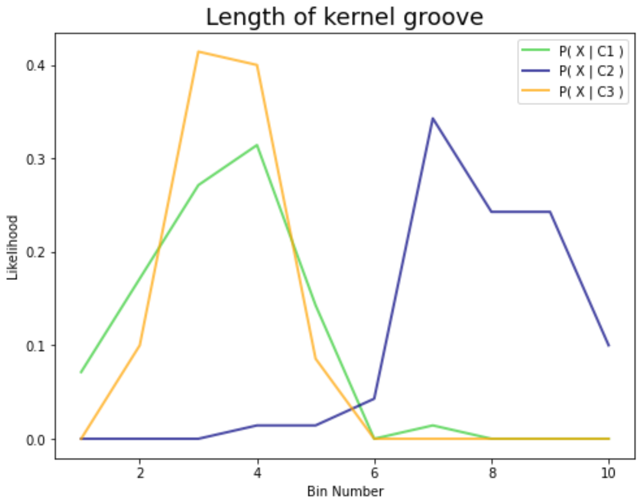






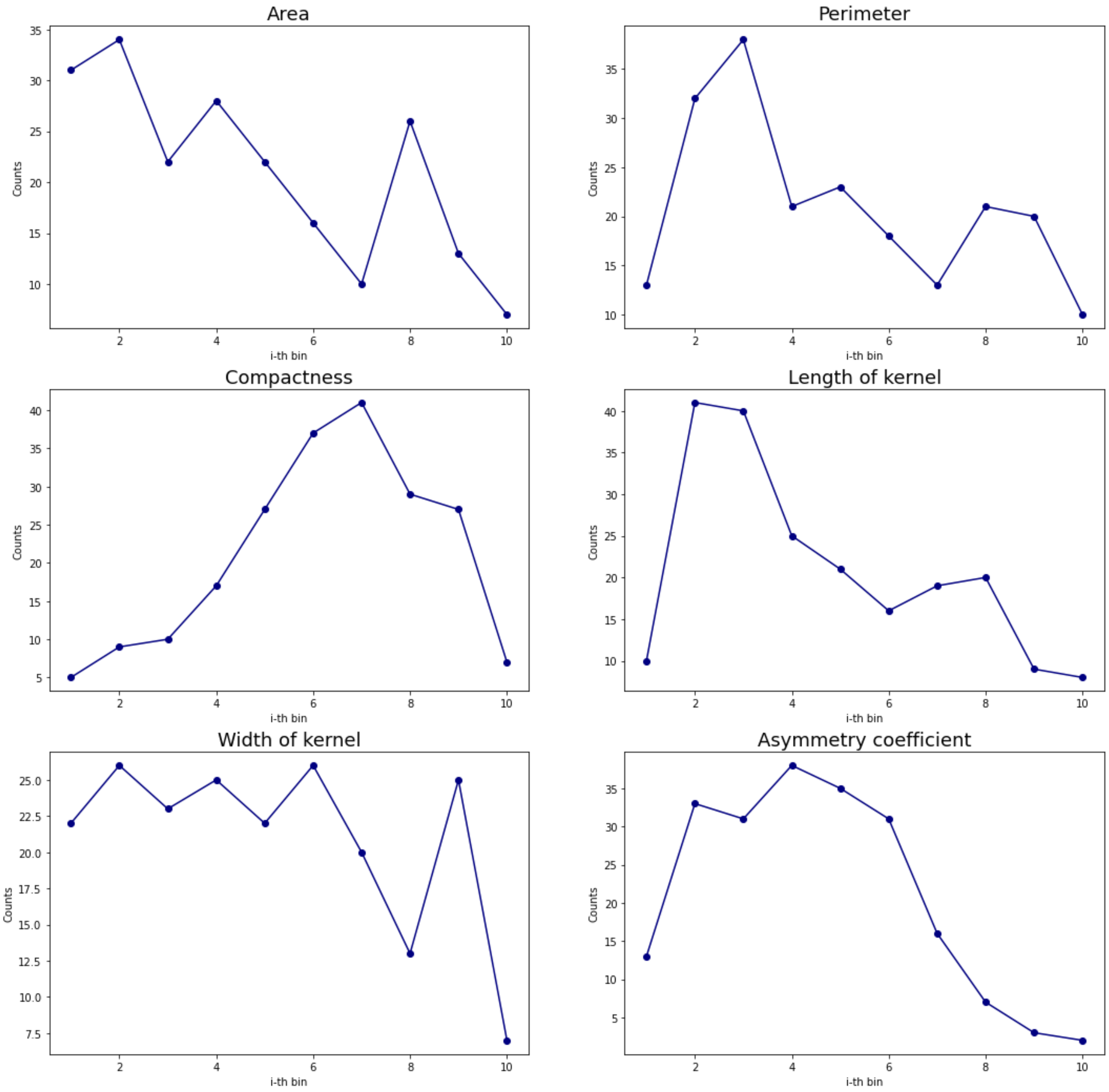






E. Count of uniques in each bins :

Calculated the number of unique values in each bin-class and plotted them feature-wise. Here are six of the plots for visualising.



F. Posterior Probabilities :

Calculated all the posterior probabilities from likelihood, prior and evidence as follows :

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{P(B)}$$

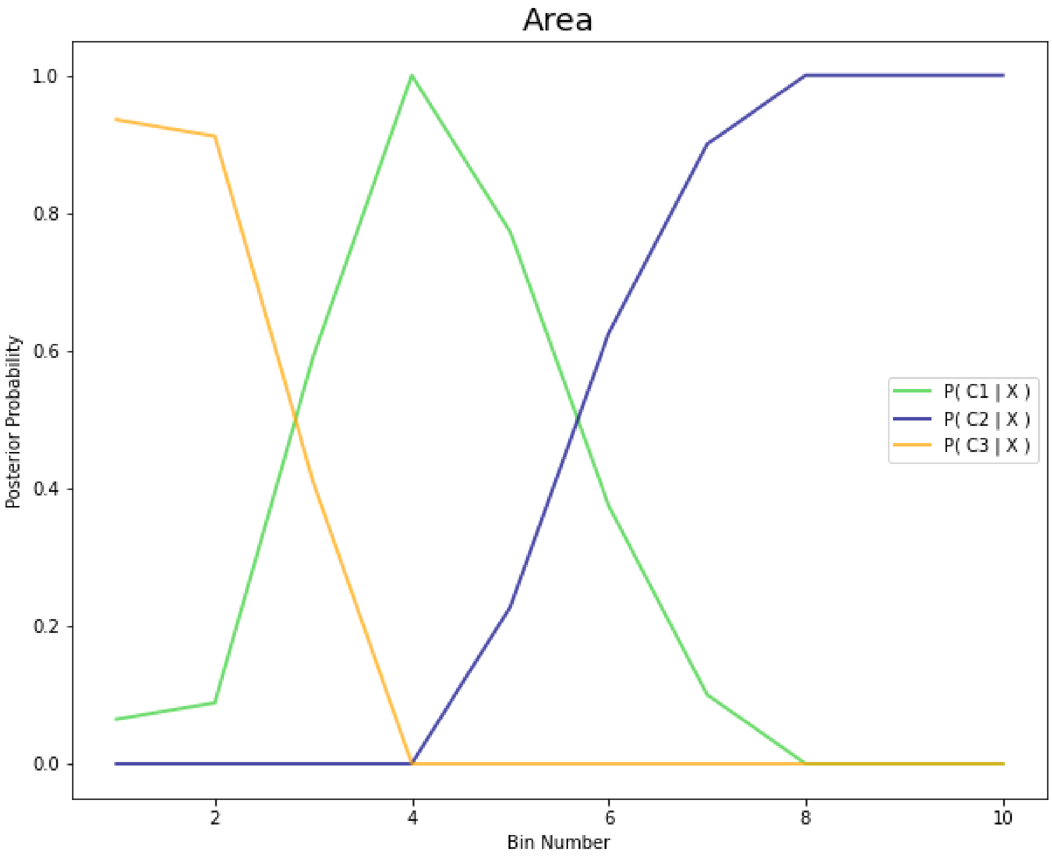
Posterior Probability = P(A | B)
Prior = P(A)

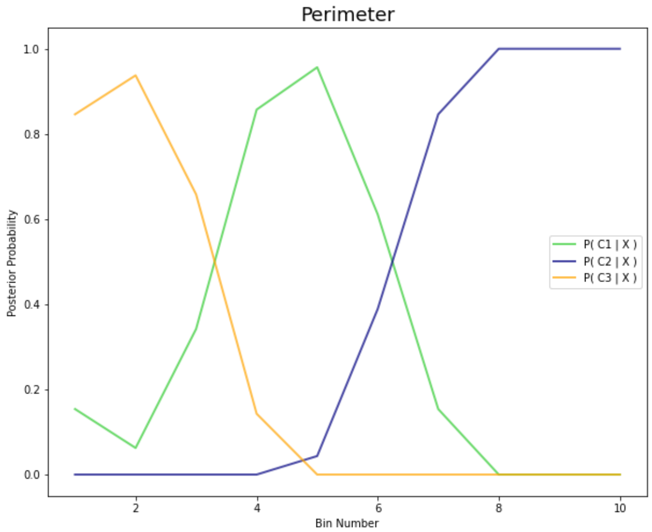
Likelihood = P(B | A)
Evidence = P(B)

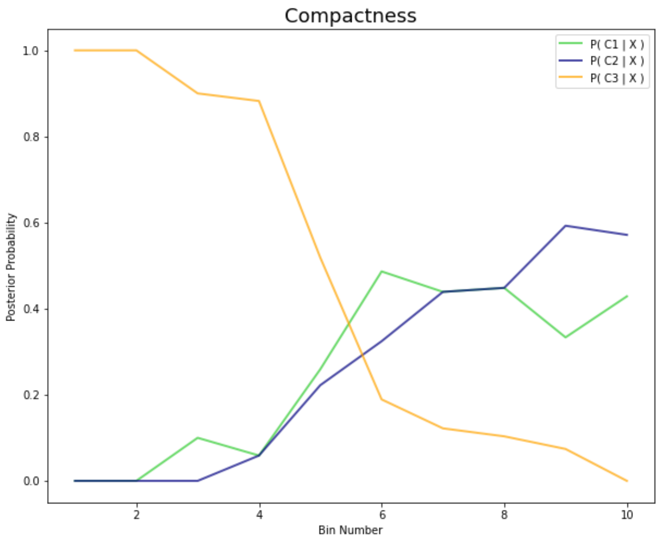
Here are the plots for posterior probabilities :
From plots we see that,

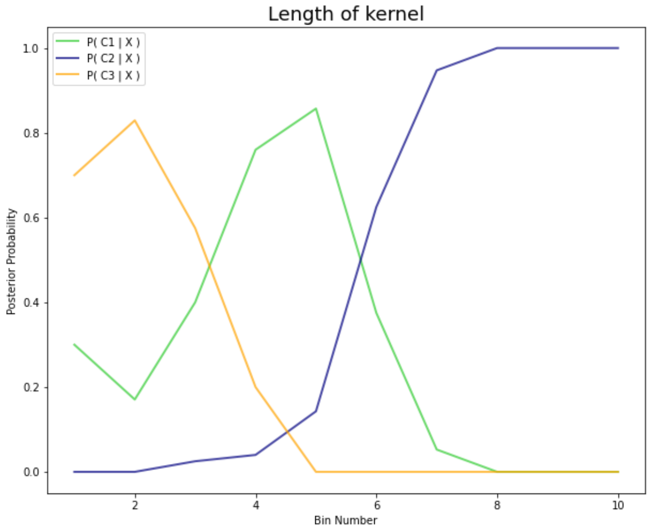
$$P(C1 \mid X) + P(C2 \mid X) + P(C3 \mid X) = 1$$

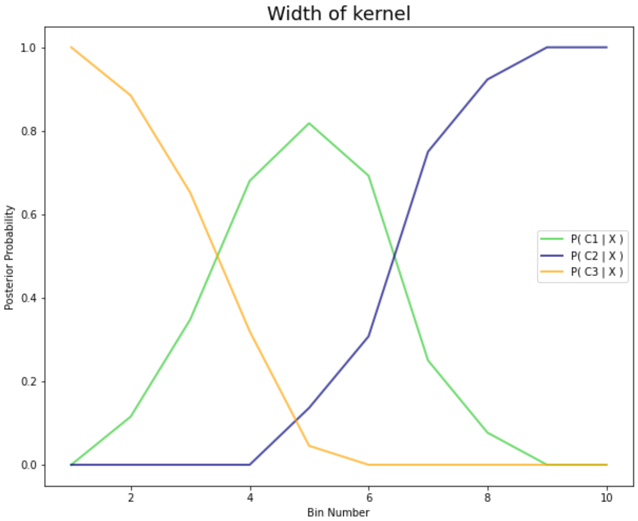
Hence we can say that our calculations are absolutely correct.

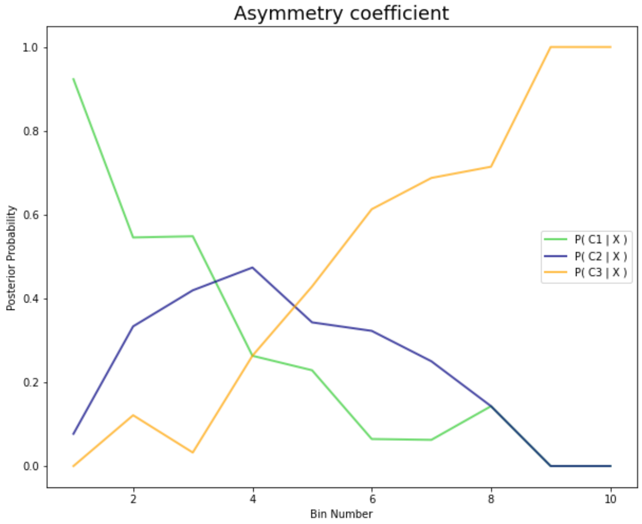


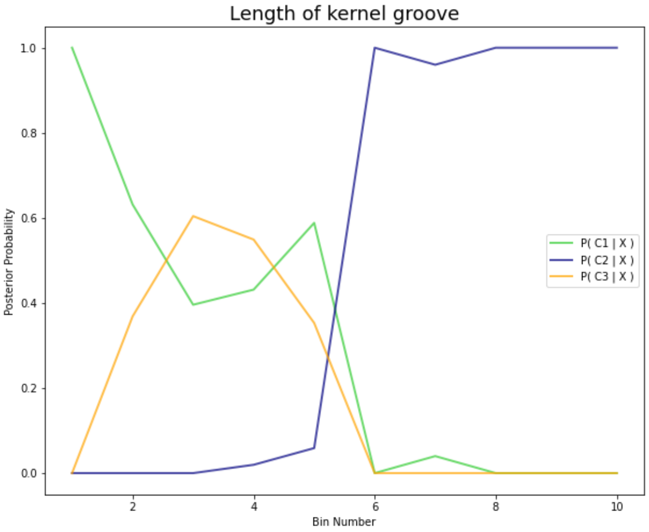





Length of kernel


Width of kernel


Asymmetry coefficient


Length of kernel groove

End of the Report !