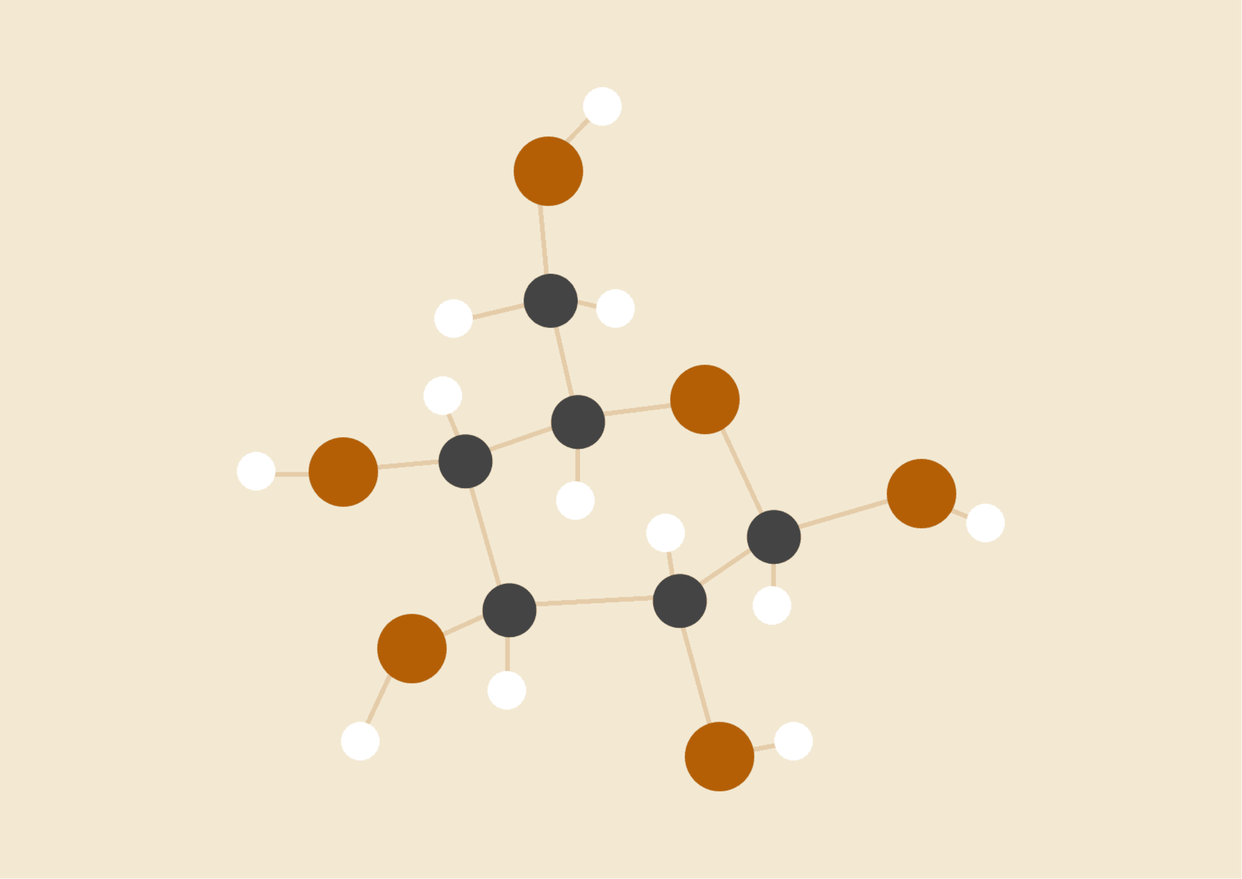
PRML LAB 3 REPORT



MANISH(B21CS044)

QUESTION 1:

a) In this part of preprocessing we are dropping the following column: Name,passengerId ,Ticket,Cabin because these don't have any relation with surviving data. Splitting of data is simply done in 80:20.

I have plotted all features in between two using classification as ‘Survived column’.

b) We are using ‘Gausssian Naive Bayes’ because all features in this are not Discrete, some are continuous like ‘Age’ , ‘Fare’ that’s why we can’t use other Naive Classifiers like ‘Binomial Naive Bayes and Multinomial Naive Bayes’.

c) In this part I am using ‘roc_auc score’ , ‘F1_score’ , ‘precision’ , ‘accuracy’ as the metrics to tell the performance of my model.

Following are the scores:

```
accuracy:0.8146067415730337,Roc-auc:0.853154821250534,Precisio
n:0.696969696969697,F1_score:0.7360000000000000
```

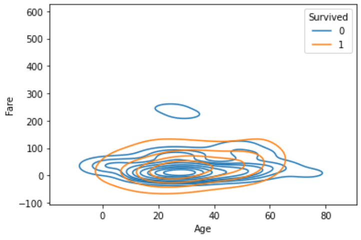
d) In this we are using KFold formula to calculate the Array of accuracy of 5-Fold validation here is the scores:

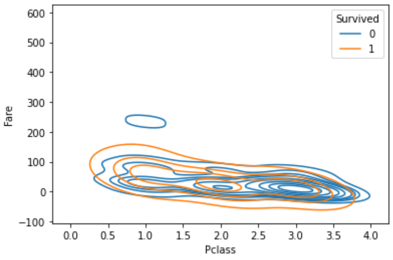
```
Array of acc:[0.81460674 0.78089888 0.80337079 0.79213483
0.70621469]

Mean acc: 0.7794451850441185
```

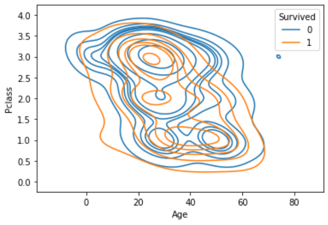
We have to find the probability of the top class for each row . We are using model.predict_probab function to give the probability of each data point.It will give us a flattened list in which we have to find the max probability of each list inside the main list.

e) For Contour plots we are using three features ‘Age’, ‘Fare’ , ‘Pclass’ in this we have seen that Plots are not elliptical at some points. So, this concludes that these features are not Independent but we are treating them as independently.

In these plots the elliptical curves should be concentric and should not intersect each other.If these are not it shows that features are not independent.







f) In this Dataset we are comparing Decision Tree Classifier and Gaussian Naive Classifier. So, In our case Decision Tree Classifier is performing better because As explained by contour plots features are not independent but Gaussian Naive Classifier treat it as Independent,but Decision Tree Classifier don’t treat dataset as Independent Features that’s why It is giving higher scores in 5-Fold Validation.

Following are the scores comparison:

0.7794451850441185 mean accuracy of Gaussian NB vs `0.7840093950358662 mean accuracy of Decision Tree Classifier`

It is not elliptical at some points so Score is slightly greater than Gaussian NB

In general, Gaussian Naive Bayes assumes independence between the features, while Decision Tree Classifier can handle non-independent features by creating partitions in the feature space.

QUESTION 2:

a) In this part we are plotting a histogram using the seaborn library of each individual feature.

b) In this part we are calculating the probability of each class which is the prior probability . Since, the count of all classes are the same that’s why the prior probability of each class is ‘⅓’.

c) In this part we are discretizing the data into 5 bins as data don’t have such a wide range so 5 bins will be enough.

For discretization I am using ‘max(feature)-min(feature)’ to calculate the range and dividing it with the No. of bins we want to give Bin width.After, that we have calculated the maximum value for each bin and then traversing through the whole column we are categories it into different bins.

I have performed this operation on each column of the data.

By this we get the numpy array for all the features and then I have concatenated it column wise and converted it to DataFrame for further use.

d) Likelihood/Class conditional Probability = P(X=x1,x2,x3…/Y)

It is said as probability of each features given Y

In this we are assuming all features as independent then we will calculate likelihood as the following

Likelihood = P(X=x1/Y)*P(X=x2/Y)*………

By using this formula we are calculating the likelihood of each bin for each class and we are returning a list of each bin's likelihood for each class.

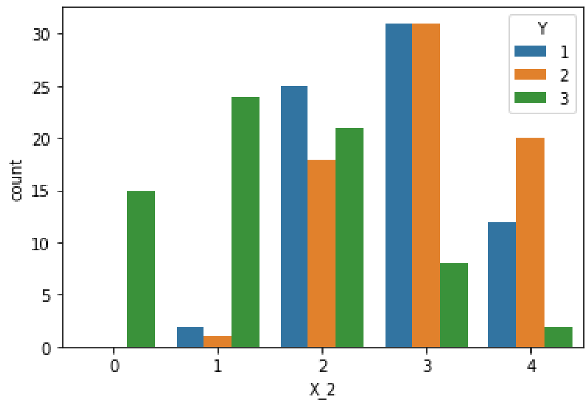
In these lists several values are 0 because they have a very less count and by dividing it with total

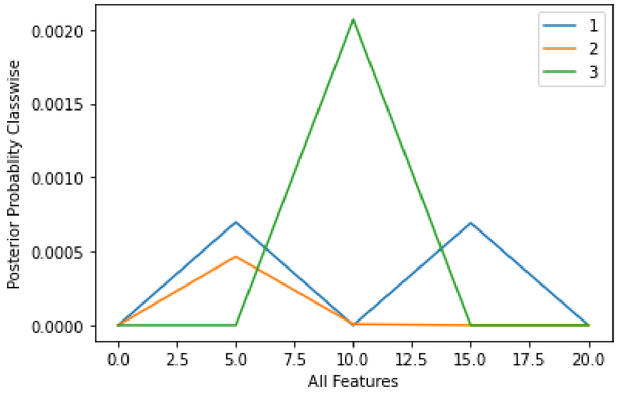
e) In this part we are plotting the unique element of each class using the seaborn library.
f) In this part we are calculating the Posterior probability by this formula:

Posterior probability = Likelihood*prior probability/evidence

This is because we are assuming the features as independent that's why evidence will be 1/7 and prior we have calculated earlier as ⅓ from this we can easily calculate the posterior probability.

Here is the plot for all classes:





After analyzing the plot we can say that for class 1 probability is higher in the range of (2.5,7.5) and (12.5,17.5) and maximum at 5 and 15.

For class 2 probability is higher in range of (2.5,7.5) and maximum at 5.

For class 3 probability is higher in range of(5,15) and maximum probability at 10.