Pattern Recognition and Machine Learning

Indian Institute of Technology, Jodhpur

LAB 3

Report

Harshil Kaneria

Department of Computer Science, IIT Jodhpur

Jan 29,2023

1)

1)

a) Pre-processing and Visualisation:

In pre-processing, I shuffled the dataset and filled the 'NAN' values in 'Age' column with mean of same column. I did this because if I removed all rows with 'NAN' values, 1/5 of the dataset is lost. Also, I dropped the 'Cabin' column because there are too many 'NAN' values in that column.
I plotted the correlation matrix of each column which each other. This showed Sex is moderately correlated with Survival. Age is very weakly correlated with survival.

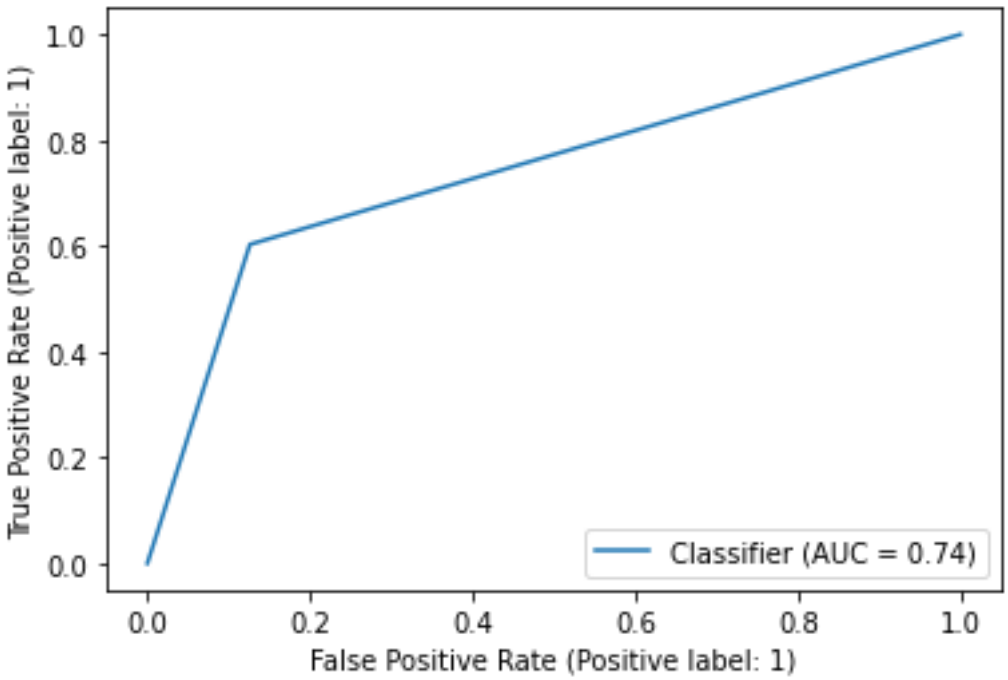
For visualisation, I plotted histogram of all features with the number. It shows that 'Passenger Id' and 'Name' are unique. It also showed that 'Fare' and 'Age' follows almost Gaussian distribution. Other features like 'Pclass', 'Sex', 'Embarked' are discrete.

b) Splitting Into Training and Testing data : Split data into training set(80%) and testing set(20%).

c) I dropped 'Name', 'Cabin', 'Passenger Id' and 'Ticket' these columns. I dropped these columns because these are unique values and doesn't matter in predicting whether person survived or not.

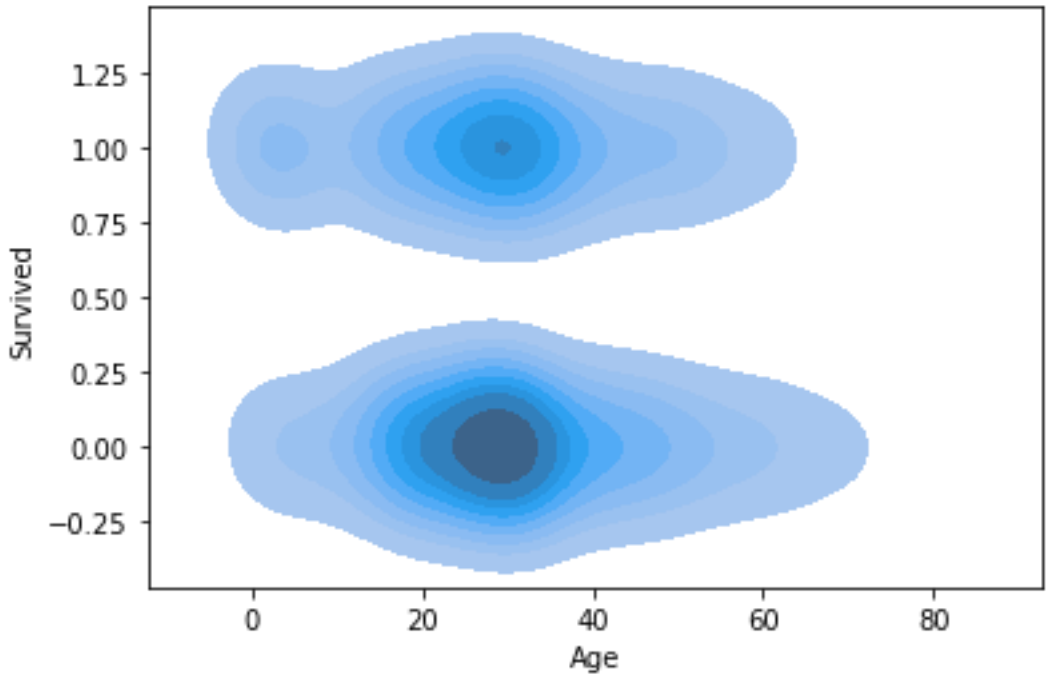
2) As seen in the Visualisation part, the distribution of 'Age' column almost follows Gaussian Distribution. So, we will use Gaussian Naïve Bayes Classifier here. The other Naïve Bayes Classifiers are Multinomial, Categorical and Bernoulli.

3) Implemented the Gaussian Naïve Bayes Classifier and calculated accuracy, ROC curve and AUC value. My data is being shuffled all the times. So, we will get different values with all the runs.
Accuracy: 77.095%
ROC Curve (with AUC value):



4) Performed a 5-fold Cross Validation on the dataset with Gaussian Naïve Bayes Classifier.
Average accuracy: 77.775% Maximum Accuracy: 83.2340%
Also, calculated all the top-class probabilities for each row.

5) Plotted the Contour plots for all features against Survived. Below shown is contour plot for Age vs Survived:

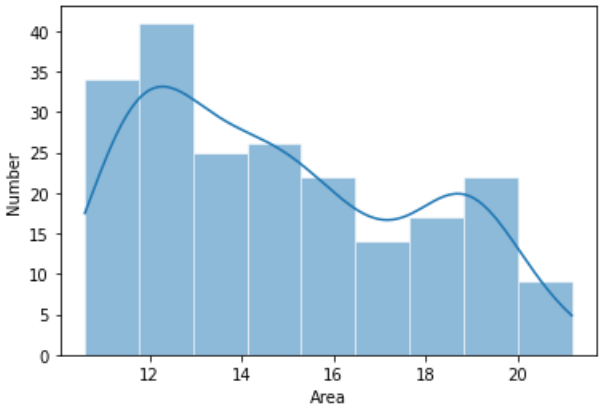


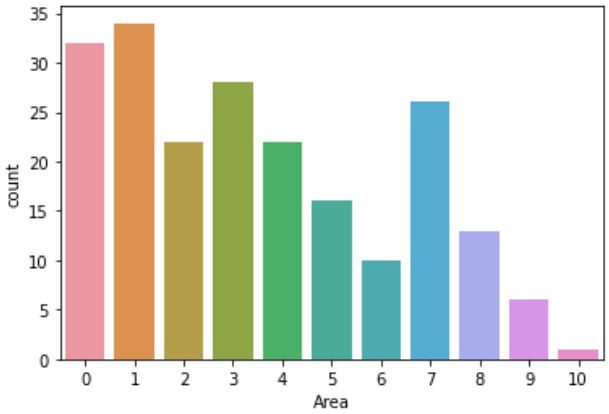
6) Performed a 5-fold Cross Validation on the dataset with Decision Tree Classifier.
Average accuracy: 77.441% Maximum Accuracy: 78.652%
We can see that accuracy of Gaussian Naïve Bayes Classifier has little more accuracy than Decision Tree Classifier. So, we can say that Decision Tree Classifier is better but not much difference is seen. Generally, for small data set Naïve Bayes Classifier is better and for large dataset Decision Tree Classifier is better. Also, Naïve Bayes Classifier assumes that the features are independent from each other but in real world scenarios this is not the case. Decision Tree Classifier makes no such assumptions. So, its better to use Decision Tree Classifier for real world dataset.

2)

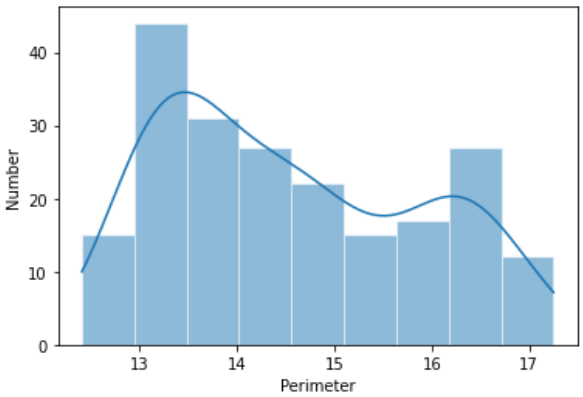
5) Plotted the count plot by seaborn library. Let's compare the distribution plot with count plots.

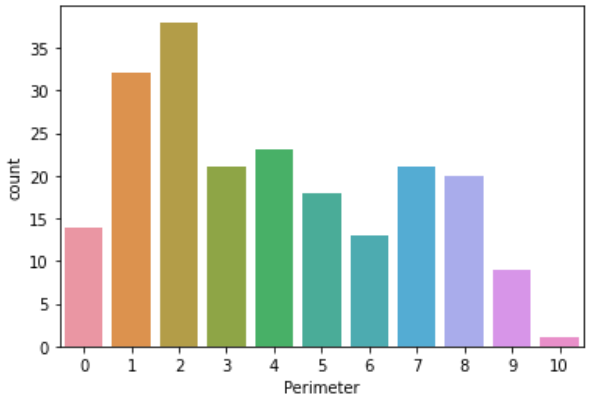
a) Area:



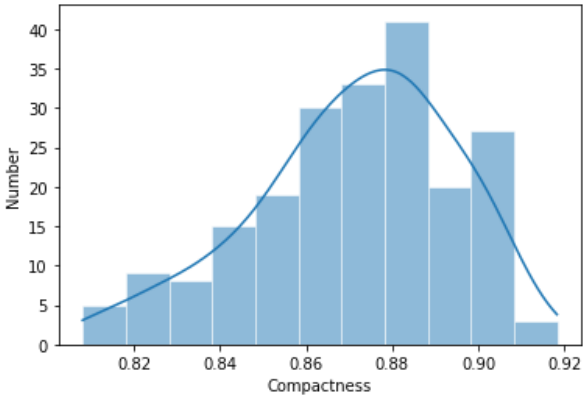


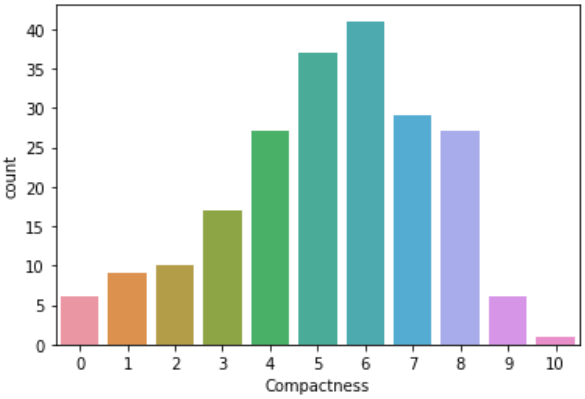
b) Perimeter:



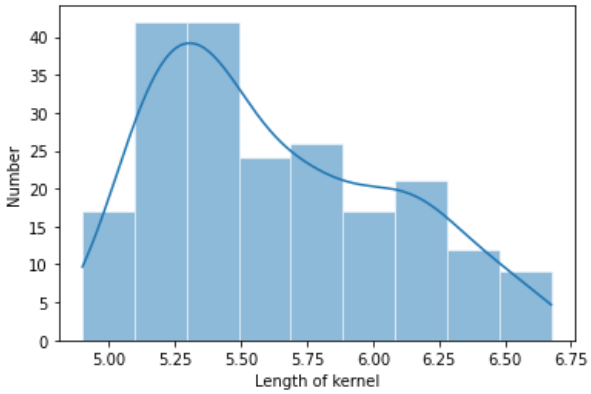


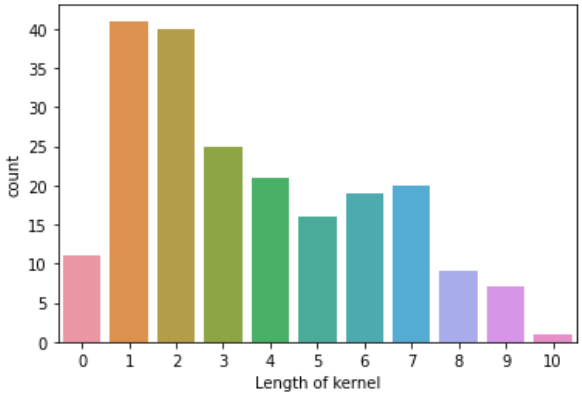
c) Compactness:



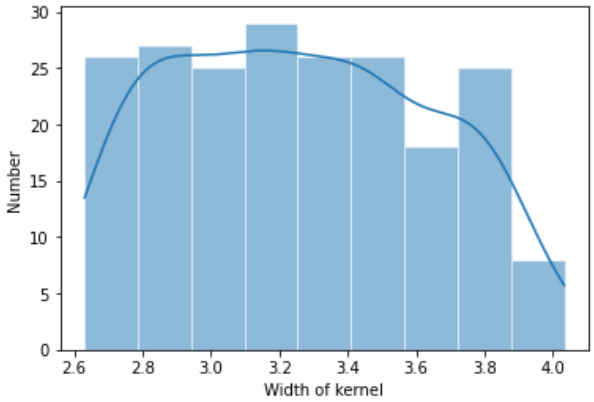


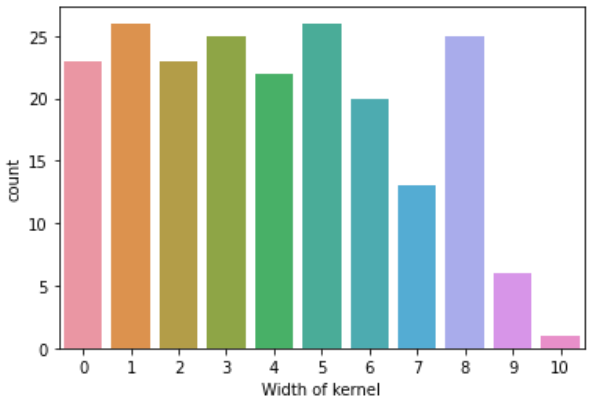
d) Length of Kernel:



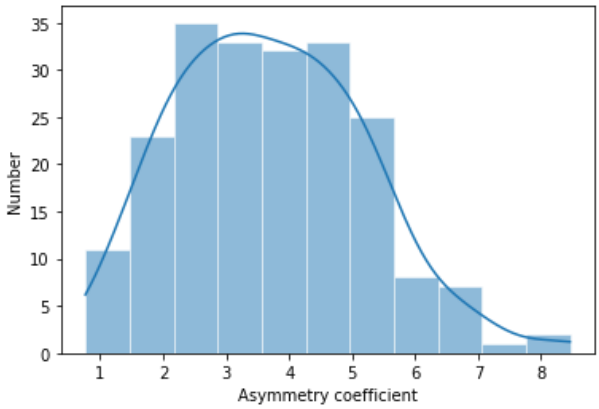


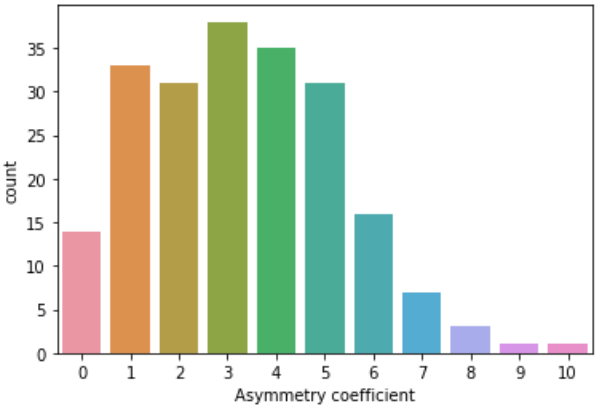
e) Width of Kernel:



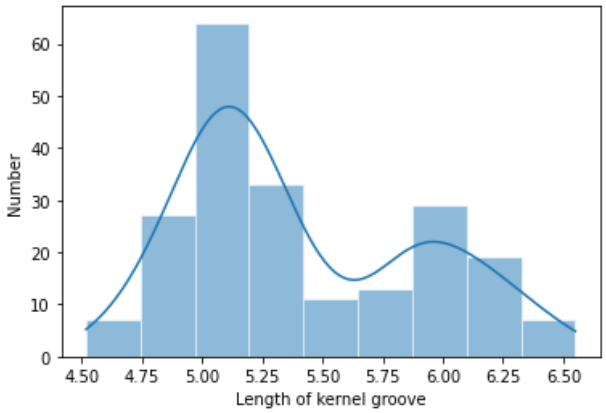


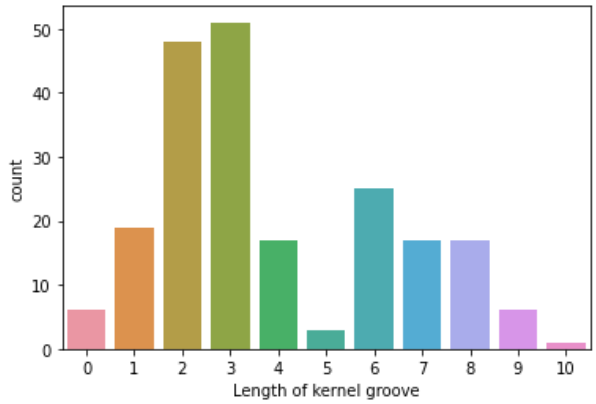
f) Asymmetry Coefficient:





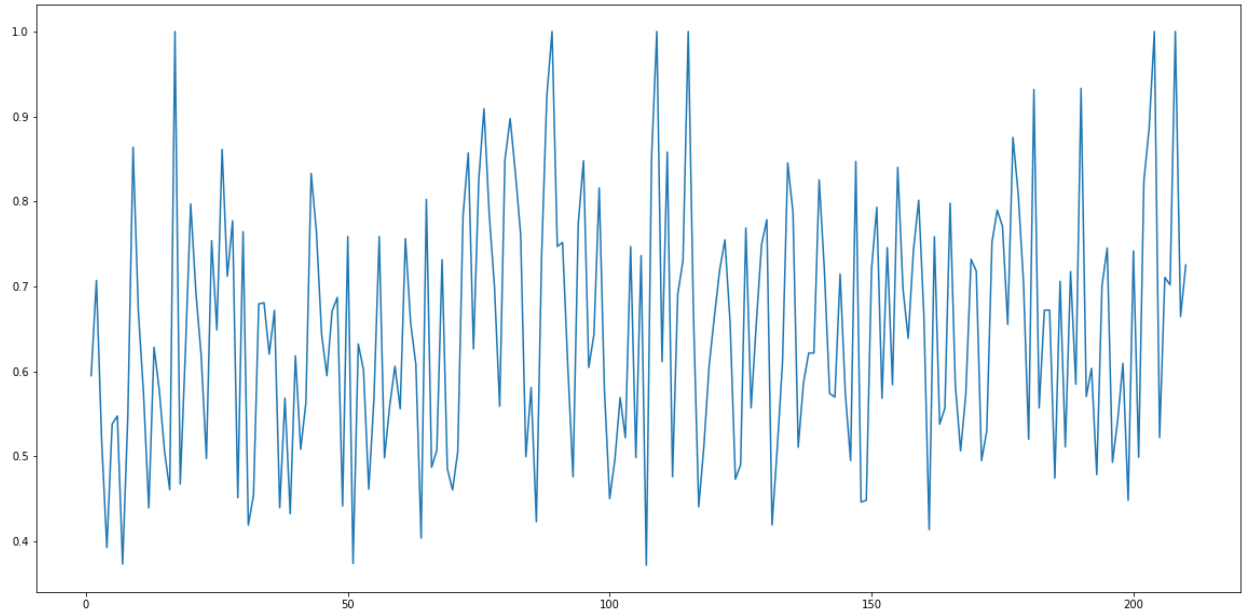
g) Length of Kernel groove:





As we can observe in all the above graphs, the distribution plot and count plots almost matches eachother. This shows that discretization and binning are a feasible action over a dataset.

6) Here I have calculated posterior probability from scratch. I got an accuracy of 55.714%. Then, plotted all the posterior probabilities for all the classes in the same graph. The graph is shown below:



Then plotted the labelled graph for each class. The labelled graph shows the in range (0 – around 70) there is only class 1 , range (70's – 150's) there is only class 2 and range(150's - 210) there is only class3.

So, we observe that all the classes are predicted with uniformity.