Homework 2

PROBLEM 1

Question 1

The principal assumption in Naive Bayes' model is that features of the class are independent of each other, and their importance is equal. Naive Bayes' model is most useful when our assumption of independence and equality of features holds true, in which case the classifier performs better as compared to other models.

Question 2

The quality of the predictions in KNN depends on the distance measure used in the classification. Therefore, the KNN algorithm is suitable for applications for which sufficient domain knowledge is available, as this knowledge supports the selection of an appropriate distance measures. KNN can also perform better in cases where the dataset is smaller which can help make predictions faster.

Question 3

$$Entropy = \frac{-150}{200}.\log(\frac{150}{200}) + \frac{-50}{200}.\log(\frac{50}{200})$$

Question 4

Domain A mean = 3.4
Domain A std. = 1.067
Domian B mean = 23.4
Domian B std. = 0.8
Domain A prior = 20/30 = 0.666
Domain B prior = 10/30 = 0.333

Question 5

From the table,

$$P(y=0) = \frac{1}{2}$$

$$P(y=1) = \frac{1}{2}$$

Using Gaussian distribution formula

$$P(x^t|y=y_i) = \frac{1}{\sigma_t\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-\mu_t)^2}{\sigma_t^2}}$$

For $x^{(1)}$,

$$\mu(x^1|y=1) = 6$$
$$\sigma(x^1|y=1) = 5.65$$
$$\mu(x^1|y=0) = 0$$
$$\sigma(x^1|y=0) = 5.65$$

$$P(x^{(1)}|y=1) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}}$$

$$P(x^{(1)}|y=0) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}}$$

For $x^{(2)}$,

$$\mu(x^2|y=1) = 7$$
$$\sigma(x^2|y=1) = 4.24$$
$$\mu(x^2|y=0) = 6$$
$$\sigma(x^2|y=0) = 1.41$$

$$P(x^{(2)}|y=1) = \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$

$$P(x^{(2)}|y=0) = \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}}$$

Therefore, using Naive Bayes,

$$P(y=0|x) = P(x^{(1)}|y=0) * P(x^{(2)}|y=0)$$

$$P(y=0|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}} * \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}}$$

Similarly,

$$P(y=1|x) = P(x^{(1)}|y=1) * P(x^{(2)}|y=1)$$

$$P(y=1|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}} * \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$

Using the above equations, we can get,

$$P(y=0|x).P(y=1|x) = \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x)^2}{5.65^2}} * \frac{1}{1.41\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{1.41^2}} * \frac{1}{5.65\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-6)^2}{5.65^2}} * \frac{1}{4.24\sqrt{2\pi}}.\exp^{-\frac{1}{2}.\frac{(x-7)^2}{4.24^2}}$$

PROBLEM 2

Question 1

$$InfoGain(S,A) = H(S) - \sum_{v \in V} \left(\frac{|S_v|}{|S|}\right) H(S_v)$$

Let's calcuate $H(S)$,

$$H(S) = \frac{-9}{16}\ln\left(\frac{9}{16}\right) - \frac{7}{16}\ln\left(\frac{7}{16}\right)$$

$$H(S) = 0.685314$$

$$InfoGain(S,"Color") = H(S) - \left(\frac{13}{16}\left(\frac{-8}{13}\ln\left(\frac{8}{13}\right) - \frac{5}{13}\ln\left(\frac{5}{13}\right)\right) + \frac{3}{16}\left(\frac{-1}{3}\ln\left(\frac{1}{3}\right) - \frac{2}{3}\ln\left(\frac{2}{3}\right)\right)\right)$$

$$InfoGain(S,"Color") = 0.02461$$

$$InfoGain(S,"Size") = H(S) - \left(\frac{8}{16}\left(\frac{-6}{8}\ln\left(\frac{6}{8}\right) - \frac{2}{8}\ln\left(\frac{2}{8}\right)\right) + \frac{8}{16}\left(\frac{-3}{8}\ln\left(\frac{3}{8}\right) - \frac{5}{8}\ln\left(\frac{5}{8}\right)\right)\right)$$

$$InfoGain(S,"Size") = 0.07336$$

$$InfoGain(S,"Shape") = H(S) - \left(\frac{12}{16}\left(\frac{-6}{12}\ln\left(\frac{6}{12}\right) - \frac{6}{12}\ln\left(\frac{6}{12}\right)\right) + \frac{4}{16}\left(\frac{-3}{4}\ln\left(\frac{3}{4}\right) - \frac{1}{4}\ln\left(\frac{1}{4}\right)\right)\right)$$

$$InfoGain(S,"Shape") = 0.02461$$

Thus we will choose Size attribute as the root of the node, since it has highest information gain.

Question 2

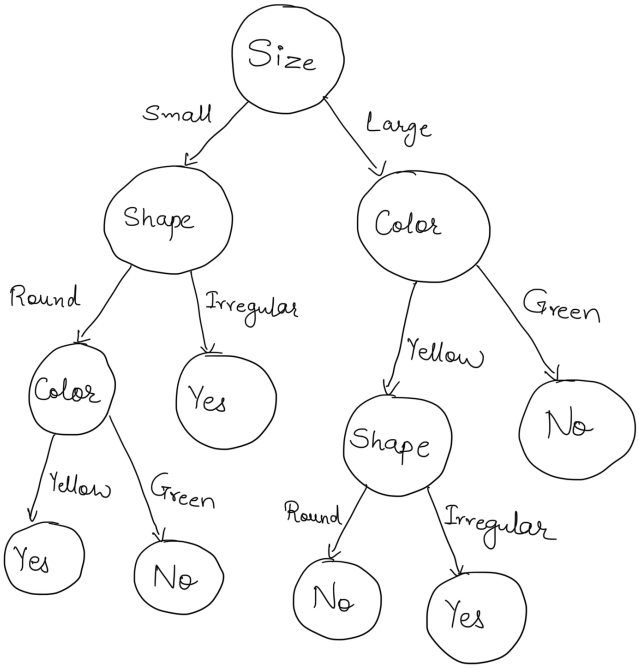

Figure 1: Decision Tree

Question 3

If we just use the numerical data in a feature as categorical data while creating a tree, then in that case each number would be a category. This would be a problem when all the numbers in the feature are unique. In such a case, it will create n different categories for n unique numbers. Here, while predicting if a the numerical value of that feature is not present in the trained decision tree then we wont be able to make a decision in such a case.

PROBLEM 3

Question 1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age
min	0	0	0	0	0	0	0.0780	21.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00
average	3.84	120.89	69.10	20.53	79.799	31.99	0.471	33.24
std deviation	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76

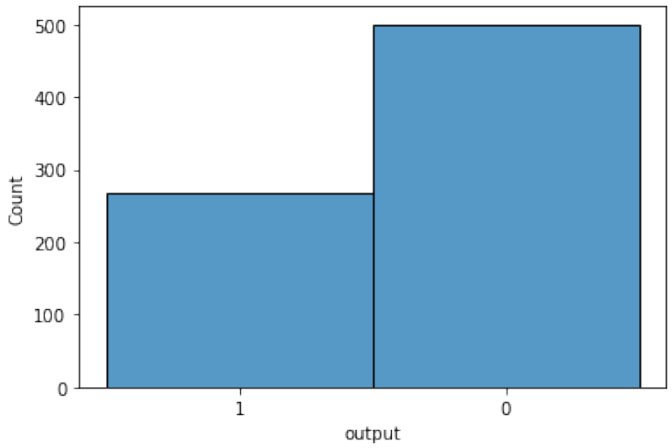

Figure 2: Histogram

Question 2

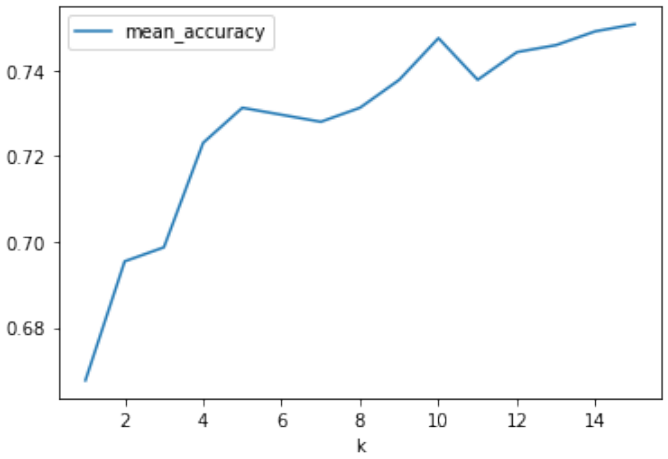

Figure 3: KNN accuracy vs k plot

As we can see the mean accuracy is highest when the number of neighbours are 15. We obtained an accuracy of 0.7507 at k=15 on the training data. Therefore I choose k=15 for this model.

Question 3

For the value of $k = 15$, we get the test error $rate = 0.24675$

Question 4

The accuracy for $k = 15$ has changed after centralization and standardization of the data, with $accuracy = 0.7442$. And now the highest accuracy can be seen for $k = 12$, with $accuracy = 0.759$. Therefore in our case centralization and standardization has impacted the accuracy for different values of k. This change is seen because the scale of features is different, and the distance calculation done in KNN uses feature values. When the one feature values are large than other, that feature will dominate the distance hence dominating the outcome of the KNN. In our statistics given above for the data, we can see that the scale of 'pregnancies' feature is very less compared to scale of 'insulin'. Therefore in our case 'insulin' feature will dominate the outcome prediction without centralization and standardization.

PROBLEM 4

Question 1

In this problem I have not standardized the input data.

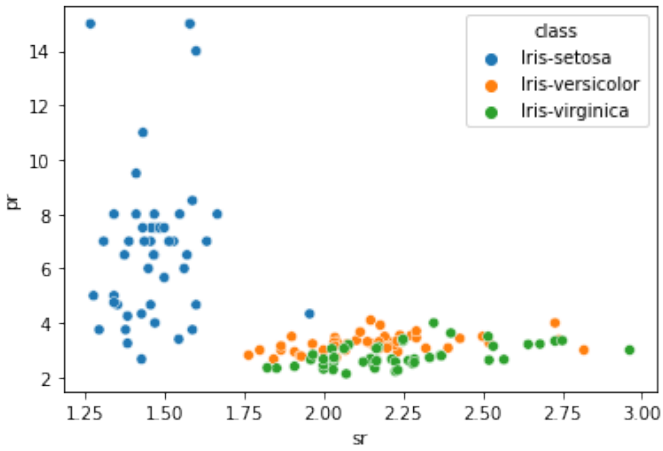

Figure 4: Cluster plot

Question 2

Implemented the code in notebook

Question 3

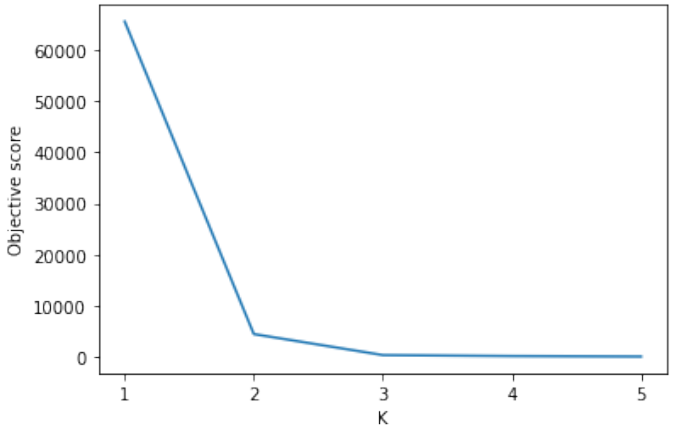

Figure 5: Accuracy vs. number of clusters

Question 4

I have chosen K=5 in this case, as it gives the minimum objective i.e. the minimum distance between the points and the centroid belonging to a cluster. We want to minimize the distance between points belonging to one cluster, thus having least objective score is desired.

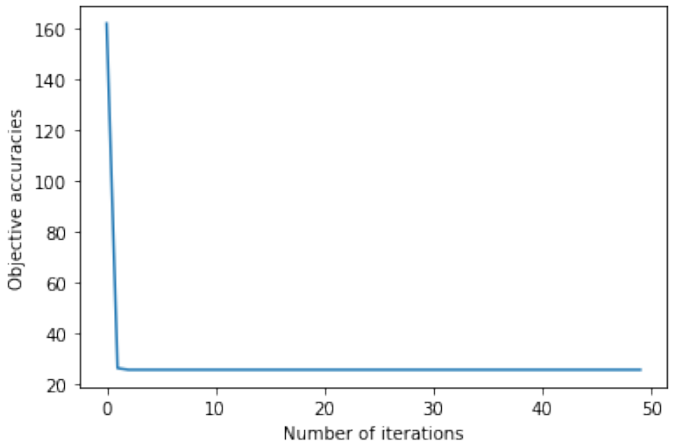

Figure 6: Accuracy change with number of iterations

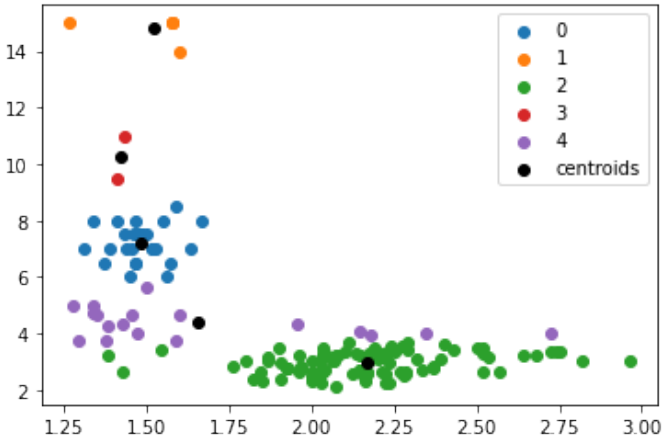

Figure 7: Clusters formed with K= 5 and their centroids