Q1

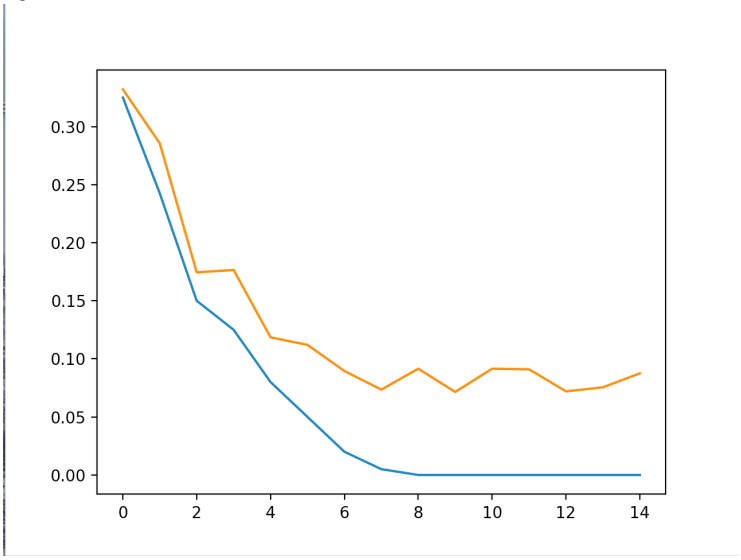

Figure 1 Training vs Testing error

1.1 Training error after a depth goes down to 0 however, testing error is still prevalent in the model.

1.2 To minimize the validation set error, decision tree depth of 3 should be chosen, however if we switch datasets, the depth changes to 6. We could depend on cross validation to add more reliability in our result.

Q2

Q2.1.1

1 $p(\text{spam}) = .6$

2 $p(\text{not spam}) = .4$

Q2.1.2

1 $\frac{1}{6}$

2 $\frac{5}{6}$

3 $\frac{2}{6}$

4 $\frac{4}{4}$

5 $\frac{1}{4}$

6 $\frac{3}{4}$

Q2.1.3

If $\mathrm{p}(\mathrm{y}=0|x_1=1,x_2=1,x_3=1) > \mathrm{p}(\mathrm{y}=1|x_1=1,x_2=1,x_3=1)$ then likely label in 0 otherwise the label is 1

$$\begin{gathered}\mathrm{p}(\mathrm{y}=1|x_1=1,x_2=1,x_3=1)\\ =\mathrm{p}(x_1=1,x_2=1,x_3=1)|p(y=1)\\ =p(x_1=1|y=1)p(x_2=1|y=1)p(x_3=1|y=1)p(y=1)\\ =\frac{1}{6}*\frac{5}{6}*\frac{2}{6}*\frac{6}{10}\\ =0.0278\end{gathered}$$

$$\begin{gathered}\mathrm{p}(\mathrm{y}=0|x_1=1,x_2=1,x_3=1)\\ =\mathrm{p}(x_1=1,x_2=1,x_3=1)|p(y=0)\\ =p(x_1=1|y=0)p(x_2=1|y=0)p(x_3=1|y=0)p(y=0)\\ =\frac{4}{4}*\frac{1}{4}*\frac{1}{4}*\frac{4}{10}\\ =0.075\end{gathered}$$

As observed, the more likely label is 0 as the probability of that is higher

Q2.1.4

0	0	1	$spam$
0	1	1	$spam$
0	1	1	$spam$
1	1	0	$spam$
0	1	0	$spam$
0	1	1	$spam$
1	0	0	$not\ spam$
1	1	0	$not\ spam$
1	0	1	$not\ spam$
1	0	0	$not\ spam$
0	1	0	$spam$
1	0	1	$not\ spam$
1	1	1	$spam$
0	0	0	$not\ spam$

Q2.2

1. Lunar
2. Bible, evidence, fact, Jesus, question, university
3. Talk. *

Q2.3

The validation error is .188 which is very similar compared to .187 of BernoulliNB.

Q2.4

Since there are 3 loops; One for features, second for objects and third for class labels hence the big O time is $O(dtk)$

Q3

Q3.2

$$k = 1; Training\ error: 0; Testing\ error: 0.0645;$$
$$k = 3; Training\ error: 0.0275; Testing\ error: 0.066;$$
$$k = 10; Training\ error: .0725; Testing\ error: 0.097;$$

Q3.3

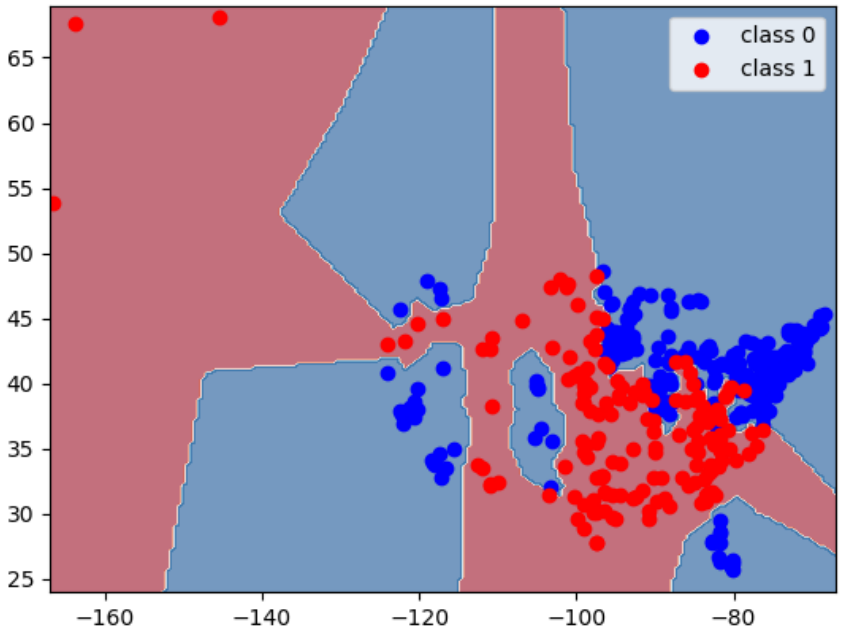

Figure 2 KNeighborsClassifier

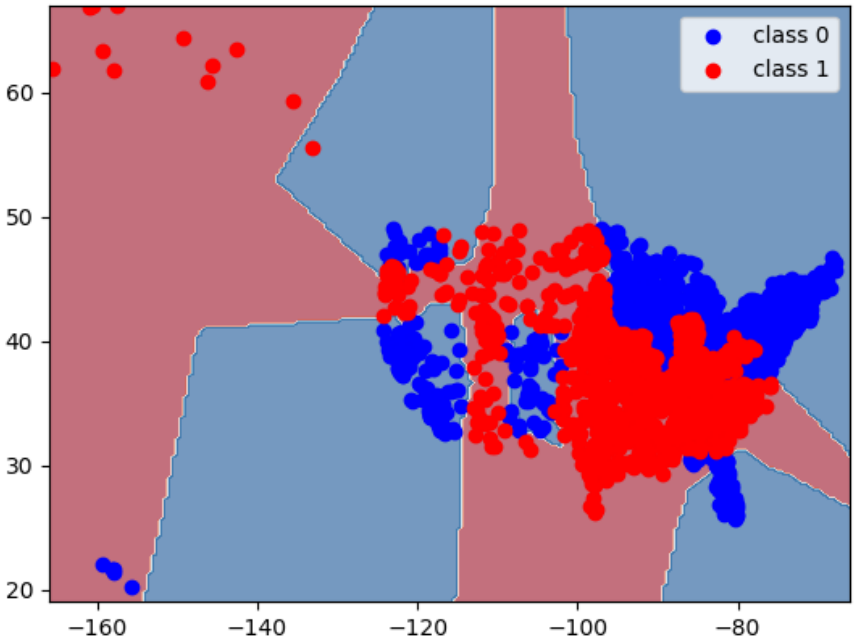

Figure 3 KNN

Q3.4

For K = 1, the training error is 0 because the data-point is its own neighbor when we are looking for 1 nearest neighbor.

Q3.5

I would have chosen ‘k’ using cross validation

Q4

Q4.1.1

Random Tree model does not have a training error of 0 because for the training error to be 0, the decision tree would have to go to infinity. Thus, making 0 an asymptote for the model.

Q4.1.3

The training error for 50 trees is 0 which is similar to training error of 0 for the decision tree model as well as random tree. The testing error for random forest is 0.193 which is much smaller compared to the testing error of 0.367 and 0.350 of decision tree and random tree respectively.

Q4.1.4

The accuracy for scikit’s implementation of RandomForestClassifier produces the same testing error of 0.193 however, it is extremely fast compared to random_forest.

Q5.1

Q5.1.2

It is observed that the error value continuously changes however, the minimum value achieved always is the same.

Q5.1.3

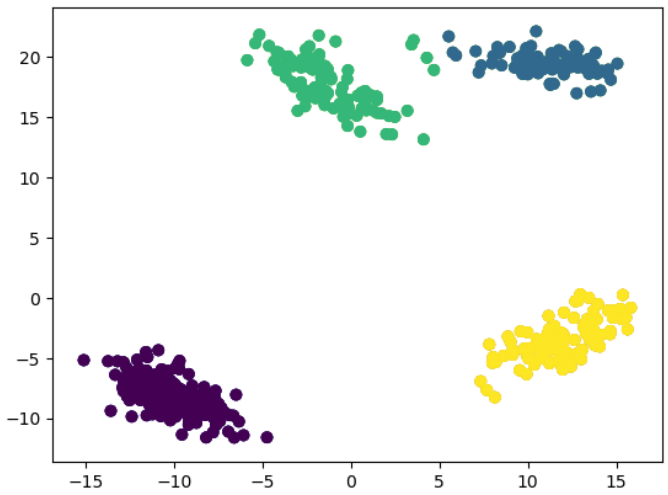

Figure 4 Clustering after 50 iterations of Kmeans

The minimum error found after 50 loops was 1062.014

Q5.1.4

n_clusters: The number of clusters you want the model to form
init: This is used to choose the method for initialization of the algorithem
n_init: The number of times the algorithem with run with different random seeds
max_iter: The maximum number of iterations of the algorithem

Q5.2

Q5.2.1

We cannot choose 'k' by taking the value that minimizes the error value as we potentially have many closest means. We should consider the k-mean value for each k. However, for large datasets, that is a lot of computation.

Q5.2.2

We would have the same problem as above wherein, we have potentially many closest means in test data. Additionally, we would not have a good prediction for k because test data is separate from training data and as k would increase the clusters would increase as well

Q5.2.3

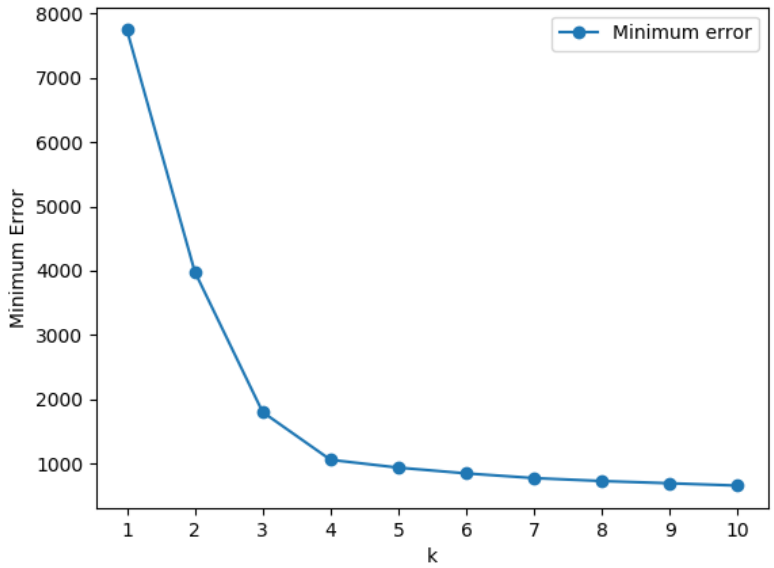

Figure 5 Minimum error VS K

Q5.2.4

Using the elbow method, the slope of k from 'k=1' to 'k=2' is the steepest, as there is a change of about 4000 points which is the biggest change visually interpreted from the graph hence, 'k=2' is most reasonable.

Q5.3

Q5.3.1

$eps = 2; minPts = 4;$

Q5.3.2

$eps = 5; minPts = 20;$

Q5.3.3

$eps = 15; minPts = 30;$

Q5.3.4

$eps = 25; minPts = 100;$

Q6

Q6.1

Boxplot is better compared to variance and mean as it shows the presence of outliers prominently whereas variance and mean do not and are heavily influenced from outliers.

Q6.2

The data may not be IID in the email spam filtering model because they are not independent of each other, the order in which the words are used matters

Q6.3

Validation set is where the machine learning model is trained and then testing dataset is used to test the accuracy of the model

Q6.4

We can't typically use training error to select a hyper-parameter to test the test dataset without any bias

Q6.5

As n in increased the optimization bias increases hence, n is directly proportional to optimization bias

Q6.6

Disadvantage of using a large k in k-fold cross validation is that it would create a bias however, an advantage it that it would create a good variance

Q6.7

We can ignore $p(x_i)$ when we use naïve Bayes because the model assumes that each element is conditionally independent of each other.

Q6.8

a) Parameter
b) Parameter
c) Hyperparameter

Q6.9

As K increases in KNN, the training error would decrease as the model would fit the data well and the approximation error would reduce as our scope scope or variance increases.

Q6.10

Make an artificial dataset mimicking the said audio along with no audio and divide into training dataset and validation dataset. Train your model with training dataset and check using validation dataset with final accuracy through testing dataset.

Q6.11

Supervised learning takes into the account the “pattern” that ‘x’ and ‘y’ form to make future predictions however clustering does not base its model on past “patterns”.

Q6.12

In comparison to picking ‘k’ from training dataset, validation set is better to reduce overfitting

Q6.13

No, they are not guaranteed to be convex areas as the hyperparameters are not same wherein, KNN finds the nearest “majority” and K-means finds similar or similar-ish data points