Data Science Project
Sub-Project IV
Data Classification using K-Nearest Neighbor Classifier and Bayes Classifier with Unimodal Gaussian Density

Student's Name: Anmol Bishnoi **Mobile No: 7042845211**

Roll Number: B19069 **Branch: CSE**

1 a.

	Prediction Outcome	
True Label	675	48
	47	6

Figure 1 KNN Confusion Matrix for K = 1

	Prediction Outcome	
True Label	708	15
	51	2

Figure 2 KNN Confusion Matrix for K = 3

		Prediction Outcome
True Label	716	7
	52	1

Figure 5 KNN Confusion Matrix for K = 5

b.

Table 1 KNN Classification Accuracy for K = 1, 3 and 5

K	Classification Accuracy (in %)
1	0.8775773195876289
3	0.9149484536082474
5	0.9239690721649485

Inferences:

1. The highest classification accuracy is obtained with K = 5
2. Increasing the value of K increases the prediction accuracy.
3. A small value of k means that noise will have a higher influence on the result.
4. As the classification accuracy increases with the increase in value of K, the number of diagonal elements increases for true positive and true negative.
5. Increase in accuracy means more correct predictions and less wrong predictions, thus increasing true positive and true negative frequencies.
6. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases.
7. Increase in accuracy means more correct predictions and less wrong predictions, thus decreasing false positive and false negative frequencies.

2 a.

	Prediction Outcome	
True Label	673	50
	42	11

Figure 6 KNN Confusion Matrix for K = 1 post data normalization

	Prediction Outcome	
True Label	704	19
	45	8

Figure 8 KNN Confusion Matrix for K = 3 post data normalization

	Prediction Outcome	
True Label	713	10
	49	4

Figure 10 KNN Confusion Matrix for K = 5 post data normalization

b.

Table 2 KNN Classification Accuracy for K = 1, 3 and 5 post data normalization

K	Classification Accuracy (in %)
1	0.8814432989690721
3	0.9175257731958762
5	0.9239690721649485

Inferences:

1. Data normalization increased classification accuracy for k = 1
2. The accuracy is increased because after normalization, the attributes on a bigger scale can no longer overpower and influence the results in their favour. This happens because Euclidean Distance is the total absolute distance along various axes and does not consider for the different ranges.
3. The highest classification accuracy is obtained with K = 5.
4. Increasing the value of K increases the prediction accuracy.
5. Increasing the value of K increases the prediction accuracy till a certain point as the no. of comparisons increases and thus noise is removed.
6. As the classification accuracy increases with the increase in value of K, the number of diagonal elements increases for true positive and true negative.
7. Increase in accuracy means more correct predictions and less wrong predictions, thus increasing true positive and true negative frequencies.
8. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases.
9. Increase in accuracy means more correct predictions and less wrong predictions, thus decreasing false positive and false negative frequencies.

3

		Prediction Outcome
True Label	675	48
	38	15

Figure 11 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 87.5%.

Table 3 Mean for Class 0

S. No.	Attribute Name	Mean
1.	seismic	1.335
2.	seismoacoustic	1.403
3.	shift	1.389
4.	genergy	76209.82
5.	gpuls	490.057
6.	gdenergy	12.082
7.	gdpuls	3.542
8.	ghazard	1.107
9.	energy	4941.741
10.	maxenergy	4374.600

Table 4 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.496
2.	seismoacoustic	1.445
3.	shift	1.1008
4.	genergy	198697.3
5.	gpuls	944.823
6.	gdenergy	17.202
7.	gdpuls	10.639
8.	ghazard	1.076
9.	energy	10278.99
10.	maxenergy	8246.22

Table 5 Covariance Matrix for Class 0

```
[[ 2.22222067e-01  1.31794149e-02 -6.17855056e-02 -1.40947235e+03
   5.85417882e+01  5.57472522e+00  4.08674920e+00  1.50325951e-02
   1.45616420e+03  1.24509156e+03]
 [ 1.31794149e-02  2.79879907e-01 -2.36710885e-02 -5.78262633e+02
   2.69605380e+01  7.63289220e+00  6.59136501e+00  8.48599792e-02
  -2.87157314e+02 -2.51016450e+02]
 [-6.17855056e-02 -2.36710885e-02  2.34197754e-01 -1.99957937e+04
  -1.10866726e+02 -4.04010477e+00 -3.26309491e+00 -1.00717687e-02
  -1.03011491e+03 -8.15375692e+02]
 [-1.40947235e+03 -5.78262633e+02 -1.99957937e+04  4.04697250e+10
   7.58153566e+07  9.03824856e+05  9.04843014e+05 -2.67430757e+03
   2.40370975e+08  1.69399016e+08]
 [ 5.85417882e+01  2.69605380e+01 -1.10866726e+02  7.58153566e+07
   2.73958983e+05  1.38396325e+04  1.36199114e+04  2.01716935e+01
   2.09138087e+06  1.76419083e+06]
 [ 5.57472522e+00  7.63289220e+00 -4.04010477e+00  9.03824856e+05
   1.38396325e+04  7.13695523e+03  4.40763936e+03  9.52190714e+00
   2.15128389e+05  2.09949916e+05]
 [ 4.08674920e+00  6.59136501e+00 -3.26309491e+00  9.04843014e+05
   1.36199114e+04  4.40763936e+03  4.16020877e+03  6.93936153e+00
   2.22839308e+05  2.13586207e+05]
 [ 1.50325951e-02  8.48599792e-02 -1.00717687e-02 -2.67430757e+03
   2.01716935e+01  9.52190714e+00  6.93936153e+00  1.22137036e-01
  -1.67798054e+02 -1.20172511e+02]
 [ 1.45616420e+03 -2.87157314e+02 -1.03011491e+03  2.40370975e+08
   2.09138087e+06  2.15128389e+05  2.22839308e+05 -1.67798054e+02
   4.24605926e+08  3.99238825e+08]
 [ 1.24509156e+03 -2.51016450e+02 -8.15375692e+02  1.69399016e+08
   1.76419083e+06  2.09949916e+05  2.13586207e+05 -1.20172511e+02
   3.99238825e+08  3.82891961e+08]]
```

Table 6 Covariance Matrix for Class 1

```
[[ 2.52136752e-01 -1.53256705e-02 -3.40406720e-02  6.82908267e+03
   1.00409151e+02  2.07574418e+00  1.56786030e+00  3.68405541e-04
   2.38646478e+03  2.10043472e+03]
 [-1.53256705e-02  3.00766284e-01 -2.87356322e-03  4.64772989e+03
  -1.32155172e+01  7.40708812e+00  6.97701149e+00  6.51340996e-02
   4.95593870e+02  2.16858238e+02]
 [-3.40406720e-02 -2.87356322e-03  9.28381963e-02 -1.70511737e+04
  -7.45523873e+01 -2.60234306e+00  6.21352785e-01 -2.21043324e-04
  -6.79034041e+02 -5.02243590e+02]
 [ 6.82908267e+03  4.64772989e+03 -1.70511737e+04  7.80414076e+10
   1.47052372e+08 -1.82727777e+06 -8.08656936e+05 -7.59846375e+03
   6.54976791e+08  6.10210589e+08]
 [ 1.00409151e+02 -1.32155172e+01 -7.45523873e+01  1.47052372e+08
   5.16572227e+05  2.05774271e+03  4.19101459e+03 -1.00537135e+01
   2.45802878e+06  2.37220524e+06]
 [ 2.07574418e+00  7.40708812e+00 -2.60234306e+00 -1.82727777e+06
   2.05774271e+03  4.57935028e+03  3.17468258e+03  2.68339228e+00
  -1.86144901e+05 -1.60879804e+05]
 [ 1.56786030e+00  6.97701149e+00  6.21352785e-01 -8.08656936e+05
   4.19101459e+03  3.17468258e+03  3.31814058e+03  3.78072502e+00
  -1.11248320e+05 -1.03548055e+05]
 [ 3.68405541e-04  6.51340996e-02 -2.21043324e-04 -7.59846375e+03
  -1.00537135e+01  2.68339228e+00  3.78072502e+00  7.88387857e-02
   4.29325081e+02  5.15443560e+02]
 [ 2.38646478e+03  4.95593870e+02 -6.79034041e+02  6.54976791e+08
   2.45802878e+06 -1.86144901e+05 -1.11248320e+05  4.29325081e+02
   3.41613286e+08  2.79957190e+08]
 [ 2.10043472e+03  2.16858238e+02 -5.02243590e+02  6.10210589e+08
   2.37220524e+06 -1.60879804e+05 -1.03548055e+05  5.15443560e+02
   2.79957190e+08  2.42985775e+08]]
```

Inferences:

1. Bayes Classifier accuracy = 89% which is less than the previous applied classifiers. This is because, when solving a problem which directly focusses on finding similarity between observations, K-NN does better because of its inherent nature to optimize locally. Also in the above example which involves just 2 clusters, KNN will give more accurate predictions than Bayes.
2. The diagonal elements of the covariance matrix denote the variance of the attribute with itself, that is, how much the data is spread out from the mean. From looking at the diagonal elements, we can infer the dispersion of the attribute and have an idea about the range of values in the attribute. For the given data, the attributes genergy and energy have the maximum covariance along the diagonal (for both the classes). This means that they have these attributes maximum spread and dispersion in their values. The attributes shift and ghazard have the minimum covariance which means that they have the least dispersion.
3. The off-diagonal elements indicate the covariance between the two attributes-how the attributes vary with respect to each other. Larger the value of covariance between 2 attributes, greater is the joint variability of the two variables.
   2 attributes with maximum covariance and joint variability:
   For class 0: maxenergy and energy, for class 1: genergy and energy.

   2 attributes with minimum covariance and joint variability:
   For class 0: shift and ghazard, for class 1: shift and ghazard.

4

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	92.40
2.	KNN on normalized data	92.40
3.	Bayes	88.92

Inferences:

1. Highest accuracy – 92.40% [KNN]
   Lowest accuracy – 87.50% [Bayes]
2. Bayes < KNN < KNN on normalized data
3. KNN gave 92.40% accuracy.

Bayes Classifier accuracy = 87.5% which is less than the previous applied classifiers. This is because, when solving a problem which directly focusses on finding similarity between observations, K-NN does better because of its inherent nature to optimize locally. Also in the above example which involves just 2 clusters, KNN will give more accurate predictions than Bayes.

The accuracy is increased relatively increased for knn on normalized data because after normalization, the attributes on a bigger scale can no longer overpower and influence the results in their favour. This happens because Euclidean Distance is the total absolute distance along various axes and does not consider for the different ranges.