IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – IV
Data Classification using K-Nearest Neighbor Classifier and Bayes Classifier with Unimodal Gaussian Density

Student's Name: Ankit **Mobile No: 9053437219**

Roll Number: B19236 **Branch: Mech**

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 3 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	**87.758**
3	**91.495**
5	**92.397**

Inferences:

1. The highest classification accuracy is obtained with **K =5**.
2. **Infer whether increasing the value of K increases/decreases the prediction accuracy.**
   The classification accuracy shows an upward trend with increasing K, from k=1 to 3, accuracy increases significantly and after k=3, accuracy remains closer to previous and does not jump too much.
3. **State a suitable reason why increasing the value of K increases/decreases the prediction accuracy.**
   This happens due to the spread of the data. Outliers/range of a class might be affecting the number of nearest Euclidean neighbors a test point has which leads to the readings.

4. **As the classification accuracy increases/decreases with the increase in value of K, infer does the number of diagonal elements increase/decrease.**
   The number of diagonal elements of the confusion matrix are increasing, implying a higher number of true positive and true negative (i.e. correct) classifications are done.
5. **State the reason for increase/decrease in diagonal elements.**
   Elements on the main diagonal represent True Positive and True Negative classifications.
   This is related to the classification accuracy. As noticed earlier, accuracy shows an upward trend with increasing 'k'. As $Accuracy \propto (TP + TN)$, with increasing number of 'k', the diagonal elements increase.
6. **As the classification accuracy increases/decreases with the increase in value of K, infer does the number of off-diagonal elements increase/decrease.**
   The number of off diagonal elements of the confusion matrix are decreasing, implying a lower number of false positive and false negative (i.e. wrong) classifications are done.

7. **State the reason for increase/decrease in off-diagonal elements.**
   Elements on the off diagonal represent False Positive and False Negative classifications.
   $$Accuracy \uparrow => (TP + TN) \uparrow => (FP + FN) \downarrow$$
   Therefore, the elements on the off-diagonal decrease.

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 = 1,2,3,4 and 5 post data normalization

K	Classification Accuracy (in %)
1	**88.144**
3	**91.753**
5	**92.397**

Inferences:

1. **Infer whether data normalization increases/decreases classification accuracy.**
   Compared to previous classes, data normalization increases classification accuracy with k=1 and k=3 but the maximum accuracy now obtained and earlier are almost same i.e. 92.397%
2. **State the reason for increase/decrease in classification accuracy after data normalization.**
   Earlier, data attributes with large ranges heavily influenced the Euclidean distance between test data point and the other data points. Data normalization has reduced the effect of the range of data attributes on final classification. The Euclidean distance now calculated has more "equal" (normalized) input from all the data attributes.
3. The highest classification accuracy is obtained with **K =5.**
4. **Infer whether increasing the value of K increases/decreases the prediction accuracy.**
   The classification accuracy shows an upward trend with increasing K, from k=1 to 3, accuracy increases significantly and after k=3, accuracy remains closer to previous and does not jump too much.
5. **State a suitable reason why increasing the value of K increases/decreases the prediction accuracy.**
   This happens due to the spread of the data. Outliers/range of a class might be affecting the number of nearest Euclidean neighbors a test point has which leads to the readings.
8. **As the classification accuracy increases/decreases with the increase in value of K, infer does the number of diagonal elements increase/decrease.**
   The number of diagonal elements of the confusion matrix are increasing, implying a higher number of true positive and true negative (i.e. correct) classifications are done.
9. **State the reason for increase/decrease in diagonal elements.**
   Elements on the main diagonal represent True Positive and True Negative classifications.
   This is related to the classification accuracy. As noticed earlier, accuracy shows an upward trend with increasing 'k'. As $Accuracy \propto (TP + TN)$, with increasing number of 'k', the diagonal elements increase.
10. **As the classification accuracy increases/decreases with the increase in value of K, infer does the number of off-diagonal elements increase/decrease.**
    The number of off diagonal elements of the confusion matrix are decreasing, implying a lower number of false positive and false negative (i.e. wrong) classifications are done.
11. **State the reason for increase/decrease in off-diagonal elements.**
    Elements on the off diagonal represent False Positive and False Negative classifications.
    $$Accuracy \uparrow => (TP + TN) \uparrow => (FP + FN) \downarrow$$
    Therefore, the elements on the off-diagonal decrease.

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 88.918%.

Table 3 Mean for Class 0

S. No.	Attribute Name	Mean
1.	seismic	1.333
2.	seismoacoustic	1.41
3.	shift	1.374
4.	genergy	76427.581
5.	gpuls	502.933
6.	gdenergy	12.928
7.	gdpuls	4.409
8.	ghazard	1.108
9.	energy	4726.257
10.	maxenergy	4107.096

Table 4 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.496
2.	seismoacoustic	1.444
3.	shift	1.103
4.	genergy	189497.179
5.	gpuls	939.923
6.	gdenergy	15.573
7.	gdpuls	9.744
8.	ghazard	1.085
9.	energy	8809.829
10.	maxenergy	6850.855

Table 5 Covariance Matrix for Class 0

Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.222	0.013	-0.062	-1409.47	58.542	5.575	4.087	0.015	1456.164	1245.092
seismoacoustic	0.013	0.28	-0.024	-578.263	26.961	7.633	6.591	0.085	-287.157	-251.016
shift	-0.062	-0.024	0.234	-19995.8	-110.867	-4.04	-3.263	-0.01	-1030.12	-815.376
genergy	-1409.47	-578.263	-19995.8	4.05E+10	75815357	903824.9	904843	-2674.31	2.4E+08	169399016
gpuls	58.542	26.961	-110.867	75815357	273959	13839.63	13619.91	20.172	2091381	1764190.8
gdenergy	5.575	7.633	-4.04	903824.9	13839.63	7136.955	4407.639	9.522	215128.4	209949.92
gdpuls	4.087	6.591	-3.263	904843	13619.91	4407.639	4160.209	6.939	222839.3	213586.21
ghazard	0.015	0.085	-0.01	-2674.31	20.172	9.522	6.939	0.122	-167.798	-120.173
energy	1456.164	-287.157	-1030.12	2.4E+08	2091381	215128.4	222839.3	-167.798	4.25E+08	399238825
maxenergy	1245.092	-251.016	-815.376	1.69E+08	1764191	209949.9	213586.2	-120.173	3.99E+08	382891961

Table 6 Covariance Matrix for Class 1

Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252	-0.015	-0.034	6829.083	100.409	2.076	1.568	0	2386.465	2100.435
seismoacoustic	-0.015	0.301	-0.003	4647.73	-13.216	7.407	6.977	0.065	495.594	216.858
shift	-0.034	-0.003	0.093	-17051.2	-74.552	-2.602	0.621	0	-679.034	-502.244
genergy	6829.083	4647.73	-17051.2	7.8E+10	1.47E+08	-1827278	-808657	-7598.46	6.55E+08	610210589
gpuls	100.409	-13.216	-74.552	1.47E+08	516572.2	2057.743	4191.015	-10.054	2458029	2372205.2
gdenergy	2.076	7.407	-2.602	-1827278	2057.743	4579.35	3174.683	2.683	-186145	-160879.8
gdpuls	1.568	6.977	0.621	-808657	4191.015	3174.683	3318.141	3.781	-111248	-103548.1
ghazard	0	0.065	0	-7598.46	-10.054	2.683	3.781	0.079	429.325	515.444
energy	2386.465	495.594	-679.034	6.55E+08	2458029	-186145	-111248	429.325	3.42E+08	279957190
maxenergy	2100.435	216.858	-502.244	6.1E+08	2372205	-160880	-103548	515.444	2.8E+08	242985775

Inferences:

1. Accuracy of Bayes Classifier = 88.918% and it is lesser than KNN classification accuracy.
   Bayes Classifier is a supervised learning algorithm, it assumes that all features are independent which is not usually the case in real life, so it makes the naïve Bayes less accurate.

2. The values along the diagonal of the covariance matrix is positive. The diagonal elements represent variance of corresponding attribute.

3. The values along the off-diagonal elements reflects w.r.t the diagonal.
   Max covariance: (genergy, genergy), (energy, genergy).
   Min covariance: (ghazard, seismic), (ghazard, shift).

4

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	92.397
2.	KNN on normalized data	92.397
3.	Bayes	88.918

Inferences:

1. Highest accuracy= KNN and KNN on normalized data.
   The lowest accuracy= Bayes classifier.
2. Arranging the classifiers in ascending order of classification accuracy.
   Bayes< KNN = KNN on normalized data.
   In terms of speed: Bayes>KNN.
3. KNN is an unsupervised classifier whereas Bayes is a supervised classifier.