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: Srishti Ginjala **Mobile No:** 9440000900

Roll Number: B19084 **Branch: CSE**

1 a.

		Prediction Outcome
True Label	671	54
	46	5

Figure 1 KNN Confusion Matrix for K = 1

		Prediction Outcome
True Label	707	18
	47	4

Figure 2 KNN Confusion Matrix for K = 3

	Prediction Outcome	
True Label	718	7
	46	5

Figure 3 KNN Confusion Matrix for K = 5

b.

Table 1 KNN Classification Accuracy for K = 1,2,3,4 and 5

K	Classification Accuracy (in %)
1	**0.871**
3	**0.916**
5	**0.932**

Inferences:

1. The highest classification accuracy is obtained with K =5.
2. Increasing the value of K increases the prediction accuracy.
3. Increasing the value of K increases the prediction accuracy as we choose the class which has highest number of frequencies among the 1st k ones.
4. As the classification accuracy increases with the increase in value of K the number of diagonal elements increase.
5. The diagonal elements increase because TP+TN increase because the classifier has more values to compare with.
6. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decrease.

7. The off-diagonal elements decrease because the number of false predictions decrease with increase in k(FP+FN).

2a.

	Prediction Outcome	
True Label	648	47
	42	9

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

	Prediction Outcome	
True Label	705	20
	44	7

Figure 7 KNN Confusion Matrix for K = 3 post data normalization

	Prediction Outcome	
True Label	718	7
	48	3

Figure 8 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	**0.885**
3	**0.918**
5	**0.929**

Inferences:

1. Infer whether data normalization increases classification accuracy.
2. The classification accuracy after data normalization because all the attributes are brought to a common scale and all of them equally influence the model.
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 as we choose the class which has highest number of frequencies among the 1st k ones.
6. As the classification accuracy increases with the increase in value of K, the number of diagonal elements increase.
7. The number of diagonal elements increase because the number of true predictions increase.
8. As the classification accuracy increases with the increase in value of K, infer does the number of off-diagonal elements decrease.
9. The off-diagonal elements decrease because the number of false predictions decrease with increase in k(FP+FN).

3.

	Prediction Outcome	
True	663	62

	35	16

Figure 11 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 87.5 %.

Table 3 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.495
2.	seismoacoustic	1.445
3.	shift	1.100
4.	genergy	198697.395
5.	gpuls	944.823
6.	gdenergy	1.720
7.	gdpuls	10.638
8.	ghazard	1.075
9.	energy	10278.991
10.	maxenergy	8246.218

Table 4 Mean for Class 0

S. No.	Attribute Name	Mean
1.	seismic	1.335
2.	seismoacoustic	1.403
3.	shift	1.388
4.	genergy	7620.982
5.	gpuls	490.05
6.	gdenergy	12.082
7.	gdpuls	3.542
8.	ghazard	1.107
9.	energy	4941.740
10.	maxenergy	4374.600

Table 5 Covariance Matrix for Class 0

Table 6 Covariance Matrix for Class 1

Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.223	0.016	-0.058	341.106	53.938	5.44	4.665	0.016	1306.739	1133.043
seismoacoustic	0.016	0.285	-0.018	2326.935	34.331	8.157	7.394	0.091	-34.79	5.745
shift	-0.058	-0.018	0.238	-20720.277	-108.223	-2.791	-2.712	-0.008	-967.727	-765.351
genergy	341.106	2326.935	-20720.277	43147695877	76016422.41	808600.411	1021197.415	-3538.72	343322901.7	271746897.3
gpuls	53.938	34.331	-108.223	76016422.41	253960.842	12700.784	13244.251	18.993	2346354.498	2013481.006
gdenergy	5.44	8.157	-2.791	808600.411	12700.784	6834.718	4165.206	8.992	279011.669	270563.881
gdpuls	4.665	7.394	-2.712	1021197.415	13244.251	4165.206	3928.186	6.55	278212.48	267202.824
ghazard	0.016	0.091	-0.008	-3538.72	18.993	8.992	6.55	0.124	-160.341	-120.558
energy	1306.739	-34.79	-967.727	343322901.7	2346354.498	279011.669	278212.48	-160.341	468144388.2	443099212.5
maxenergy	1133.043	5.745	-765.351	271746897.3	2013481.006	270563.881	267202.824	-120.558	443099212.5	426402725.3

Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252	0.006	-0.033	629.014	88.588	3.281	1.664	0.005	3384.233	2889.603
seismoacoustic	0.006	0.3	-0.011	-1728.237	-8.963	7.342	7.154	0.059	1681.47	1108.902
shift	-0.033	-0.011	0.091	-15394.057	-74.846	-3.444	-0.777	0.001	-539.389	-389.446
genergy	629.014	-1728.237	-15394.057	98499436799	180520099.7	-794559.64	69419.22	-8909.632	1436182.097	103759960.4
gpuls	88.588	-8.963	-74.846	180520099.7	615028.282	7514.434	9052.453	3.7	997000.499	1235626.022
gdenergy	3.281	7.342	-3.444	-794559.64	7514.434	4734.518	3430.124	6.315	-168083.863	-162052.621
gdpuls	1.664	7.154	-0.777	69419.22	9052.453	3430.124	3425.453	6.078	-127216.978	-136438.242
ghazard	0.005	0.059	0.001	-8909.632	3.7	6.315	6.078	0.071	805.84	854.102
energy	3384.233	1681.47	-539.389	1436182.097	997000.499	-168083.863	-127216.978	805.84	409162012.5	341912419.9
maxenergy	2889.603	1108.902	-389.446	103759960.4	1235626.022	-162052.621	-136438.242	854.102	341912419.9	300604879.6

Inferences:

1. the accuracy of Bayes Classifier is 87.5is generally greater than previous classification approaches with more number of classes.

1

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	93.2
2.	KNN on normalized data	92.9
3.	KNN on dimensionally reduced data	
4.	Bayes	87.5
5.	Bayes on dimensionally reduced data	

Inferences:

1. The classifiers with highest and lowest accuracy are Knn normalized and Bayes.

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**
- **Please fit a confusion matrix/covariance matrix /table in one page only.**
- **For making the covariance matrix you can use excel or online table generators and then paste as image.**