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: Tarun Singla **Mobile No:** 8872526396

Roll Number: B19198 **Branch:** EE

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 3 KNN Confusion Matrix for K = 3

	Prediction Outcome	
True Label	718	7
	46	5

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	87.11%
3	91.62%
5	93.17%

Inferences:

1. The highest classification accuracy is obtained with K =.5
2. Increasing the value of 'K' leads to increase in accuracy of prediction.
3. If the value of 'K' is small, then outliers and noise can affect the result. But as 'K' increases, their dominance decrease.
4. As the accuracy increases, number of diagonal elements also increase.
5. Diagonal elements refer to the number which are correctly predicted.

6. As the classification accuracy increases, the number of off-diagonal elements decrease.
7. Off-diagonal elements refer to the number which are incorrectly predicted.
8. As values of 'K' increases, the cost for computations also increases.

2 a.

	Prediction Outcome	
True Label	678	47
	42	9

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

	Prediction Outcome	
True Label	705	20
	44	7

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

		Prediction Outcome
True Label	718	7
	48	3

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	88.53%
3	91.75%
5	92.91%

Inferences:

1. After normalization, the accuracy increases for K=1,3 while for K=5 it decreases
2. In normalization, the dominancy of attribute over other due to large value decreases. In this way, accuracy (most likely) increases.
3. The highest classification accuracy is obtained with K =5.
4. On increasing the value of 'K', accuracy is increasing.
5. If the value of 'K' is small, then outliers and noise can affect the result. But as 'K' increases, their dominance decrease
6. As the accuracy increases, number of diagonal elements also increase
7. Diagonal elements refer to the number which are correctly predicted.
8. As the classification accuracy increases, the number of off-diagonal elements decrease
9. Off-diagonal elements refer to the number which are incorrectly predicted. It is nonlinear with k

3

	Prediction Outcome	
True Label	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 0

S. No.	Attribute Name	Mean
1.	seismic	1.335
2.	seismoacoustic	1.403
3.	shift	1.3889
4.	genergy	76209.828
5.	gpuls	490.056
6.	gdenergy	12.080
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	1.495
2.	seismoacoustic	1.445

3.	shift	1.100
4.	genergy	198697.394
5.	gpuls	944.823
6.	gdenergy	17.201
7.	gdpuls	10.638
8.	ghazard	1.075
9.	energy	10278.991
10.	maxenergy	8246.218

Table 5 Covariance Matrix for Class 0

	seismic	seismoacoustic	Shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.222943	0.015871	-0.05816	341.1062	53.9377	5.440415	4.665308	0.0162	1306.739	1133.043
seismoacoustic	0.015871	0.284611	-0.01831	2326.935	34.33133	8.156964	7.394355	0.090652	-34.7899	5.744762
Shift	-0.05816	-0.01831	0.237817	-20720.3	-108.223	-2.79092	-2.71227	-0.00794	-967.727	-765.351
genergy	341.1062	2326.935	-20720.3	4.31E+10	76016422	808600.4	1021197	-3538.72	3.43E+08	2.72E+08
gpuls	53.9377	34.33133	-108.223	76016422	253960.8	12700.78	13244.25	18.99331	2346354	2013481
gdenergy	5.440415	8.156964	-2.79092	808600.4	12700.78	6834.718	4165.206	8.99236	279011.7	270563.9
gdpuls	4.665308	7.394355	-2.71227	1021197	13244.25	4165.206	3928.186	6.550259	278212.5	267202.8
ghazard	0.0162	0.090652	-0.00794	-3538.72	18.99331	8.99236	6.550259	0.124173	-160.341	-120.558
energy	1306.739	-34.7899	-967.727	3.43E+08	2346354	279011.7	278212.5	-160.341	4.68E+08	4.43E+08
maxenergy	1133.043	5.744762	-765.351	2.72E+08	2013481	270563.9	267202.8	-120.558	4.43E+08	4.26E+08

Table 6 Covariance Matrix for Class 1

	seismic	seismoacoustic	Shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252101	0.006124	-0.03347	629.0144	88.58824	3.280516	1.663723	0.004558	3384.233	2889.603
seismoacoustic	0.006124	0.299957	-0.01139	-1728.24	-8.96311	7.341618	7.153824	0.059251	1681.47	1108.902
Shift	-0.03347	-0.01139	0.09144	-15394.1	-74.8465	-3.44424	-0.77681	0.000783	-539.389	-389.446
genergy	629.0144	-1728.24	-15394.1	9.85E+10	1.81E+08	-794560	69419.22	-8909.63	1436182	1.04E+08
gpuls	88.58824	-8.96311	-74.8465	1.81E+08	615028.3	7514.434	9052.453	3.6999	997000.5	1235626
gdenergy	3.280516	7.341618	-3.44424	-794560	7514.434	4734.518	3430.124	6.315126	-168084	-162053
gdpuls	1.663723	7.153824	-0.77681	69419.22	9052.453	3430.124	3425.453	6.078408	-127217	-136438
ghazard	0.004558	0.059251	0.000783	-8909.63	3.6999	6.315126	6.078408	0.070503	805.8396	854.102
energy	3384.233	1681.47	-539.389	1436182	997000.5	-168084	-127217	805.8396	4.09E+08	3.42E+08
maxenergy	2889.603	1108.902	-389.446	1.04E+08	1235626	-162053	-136438	854.102	3.42E+08	3.01E+08

Inferences:

1. The accuracy of Bayes classifier is 87.5%. Its accuracy is less as compare to other. This is because, this method is effective on large number of data set. Large data sets are more likely to follow gaussian distribution.
2. The values of diagonal elements of covariance matrix are positive. Most of them have very high values. This's because most of attributes are highly dispersed.
3. Off-diagonal element represent the correlation between the corresponding attributes. **'maxenergy'** and **'energy'** are highly correlated while **'ghazard'** and **'genergy'** are highly un-correlated.

4

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	93.170%
2.	KNN on normalized data	92.912%
3.	Bayes	87.500%

Inferences:

1. KNN (without normalization) has maximum accuracy while Bayes' classifier has minimum.
2. Bayes<KNN on normalized data ≈ KNN.
3. Bayes classifier is effective on large data points because large data set are more likely to follow gaussian distribution. So, here for relatively small data points It's quite ineffective.
4. Bayes classifier is faster than the KNN method.