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

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 3 KNN Confusion Matrix for K = 5

b.

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

K	**Classification Accuracy (in %)**
1	**0.878**
3	**0.915**
5	**0.924**

Inferences:

1. The highest classification accuracy is obtained with **K = 5**
2. Increasing the value of K, we can see that the prediction accuracy increases.
3. Increasing the value of K increase the accuracy of the model because when we increase the number of neighbors, we extract more features from individual classes and more features are compared with the test data hence giving better results
4. Since the diagonal elements of confusion matrix represent the True Positives, we can see that the number of diagonal elements increases with the increase of K. Hence, improving the Accuracy.
5. Since the accuracy of the model is directly proportional to the sum of number of Diagonal elements that represent true positive, we can say when the accuracy increases (with increase in K) the number of diagonal elements also increases.

$$Accuracy\ =\ Total\ True\ Positives\ (Diagonal\ Elements)\ /\ Total\ Test\ Cases$$

6. The number of off-diagonal elements decrease with increase in accuracy.
7. Since the off diagonal elements either represent the False Positives or the False Negative we can clearly see that they decrease the accuracy. But with increase in accuracy the count of these elements decreases.

2 a.

	Prediction Outcome	
True Label	673	50
	42	11

Figure 4 KNN Confusion Matrix for K = 1 post data normalization

	Prediction Outcome	
True Label	704	19
	45	8

Figure 5 KNN Confusion Matrix for K = 3 post data normalization

	Prediction Outcome	
True Label	713	10
	49	4

Figure 6 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	**0.881**
3	**0.918**
5	**0.924**

Inferences:

1. The classification accuracy decreases a lit bit after normalization, but it pretty much remains the same.
2. In KNN we use the Euclidean Distance and after normalization this distance changes and thus the normalized data might select different set of K neighbors and hence there is a difference in accuracy. But the accuracy may increase or decrease, and it cannot be inferred on which side it will fall.
3. The highest classification accuracy is obtained with **`K = 5`**.
4. Increasing the value of K increases the prediction accuracy as described in KNN without normalization.
5. When we increase the value of K, number of neighbors selected are higher which leads to selection of more unique features of a particular class and hence the prediction accuracy increases.
6. It is same as the previous statement that the diagonal Elements represent the number of True Positives and the accuracy is directly proportional to the number of diagonal elements. So, with increase in prediction accuracy the number of diagonal elements also increases.
7. Now we know that the off-diagonal elements represent the number of False Positives and False Negatives of a class and are inversely proportional to the accuracy. Thus, the increase in the prediction accuracy decreases the number of off-diagonal elements.

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 88.917%.

Table 3 Mean for Class 0

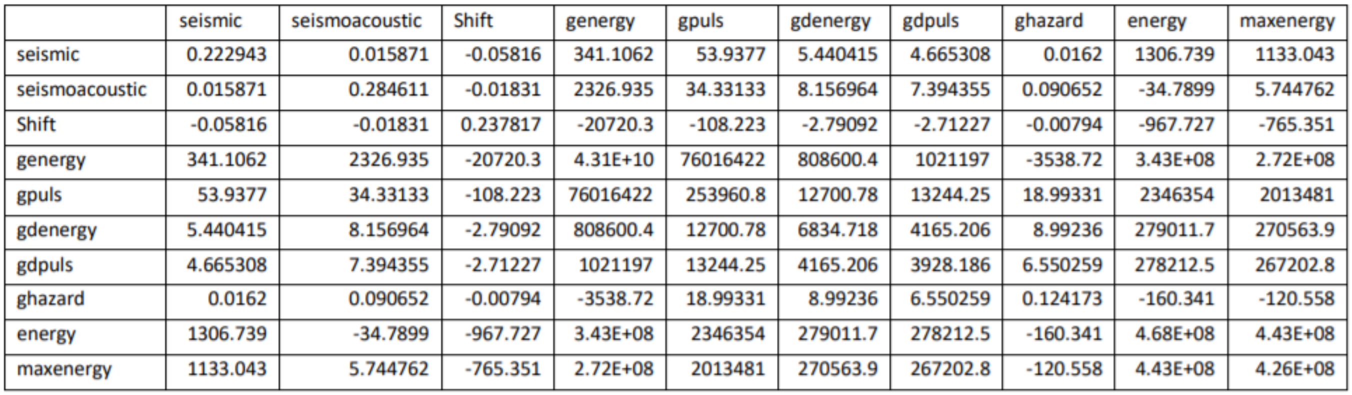
S. No.	Attribute Name	Mean
1.	seismic	1.33294
2.	seismoacoustic	1.40982
3.	shift	1.37374
4.	genergy	76427.5813
5.	gpuls	502.93318
6.	gdenergy	12.92844
7.	gdpuls	4.40923
8.	ghazard	1.10763
9.	energy	4726.25665
10.	maxenergy	4107.09639

Table 4 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.49573
2.	seismoacoustic	1.44444
3.	shift	1.10256
4.	genergy	189497.179
5.	gpuls	939.92308
6.	gdenergy	15.57265
7.	gdpuls	9.74359
8.	ghazard	1.08547
9.	energy	8809.82906

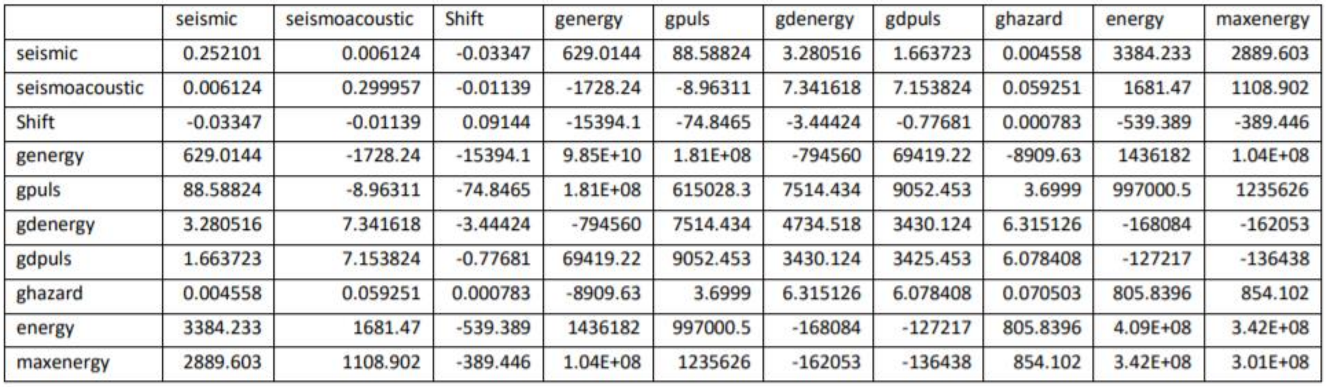
10.	maxenergy	6850.8547

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 88.917%. 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 is 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	92.396
2.	KNN on normalized data	92.396
3.	Bayes	88.917

Inferences:

1. KNN (without normalization) has maximum accuracy while Bayes' classifier has minimum.
2. Bayes Classifier < 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 is quite ineffective.
4. Bayes classifier is faster than the KNN method.