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: Aryan Garg **Mobile No: 08219383122**

Roll Number: B19153 **Branch: EE**

1 a.

	Prediction Outcome	
True Label	39	12
	21	30

Figure 1 KNN Confusion Matrix for K = 1

	Prediction Outcome	
True Label	45	6
	23	28

Figure 2 KNN Confusion Matrix for K = 2

	Prediction Outcome	
True Label	41	10
	18	33

Figure 3 KNN Confusion Matrix for K = 3

	Prediction Outcome	
True Label	45	6
	19	32

Figure 4 KNN Confusion Matrix for K = 4

	Prediction Outcome	
True Label	42	9
	16	35

Figure 5 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	**67.647**
2	**71.569**
3	**72.549**
4	**75.490**
5	**75.490**

Inferences:

1. The highest classification accuracy is obtained with K = 5.
2. Increasing the value of K increases the prediction accuracy.
3. Considering more data points in the vicinity of the test subject gives a better estimate for the test's class. Simply put, more data equates to better predictions.
4. As the classification accuracy increases with the increase in value of K, the number of diagonal elements increases, which is a direct consequence of 3.
5. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases because the total data under consideration(testing) is finite (102 cases here).

2 a.

	Prediction Outcome	
True Label	32	19
	14	37

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

	Prediction Outcome	
True Label	42	9
	19	32

Figure 7 KNN Confusion Matrix for K = 2 post data normalization

	Prediction Outcome	
True Label	38	13
	16	35

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

	Prediction Outcome	
True Label	44	7
	23	28

Figure 9 KNN Confusion Matrix for K = 4 post data normalization

	Prediction Outcome	
True Label	41	10
	20	31

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	**67.647**
2	**72.549**
3	**71.569**
4	**70.588**
5	**70.588**

Inferences:

1. Data normalization decreases overall classification accuracy.
2. Extreme values lose their significance on normalization thus leading to a slight information loss.
3. The highest classification accuracy is obtained with K = 2.
4. Accuracy improves initially, then declines to an almost constant value.
5. The reasoning of Q1 applies here too but the data normalization leads to some irregularity in our dataset, which is small and the significance of extreme values is clearly observed.
6. Count of diagonal values follow the same pattern as point 4.

3

	Prediction Outcome	
True Label	47	4
	18	33

Figure 11 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is **78.431 %**

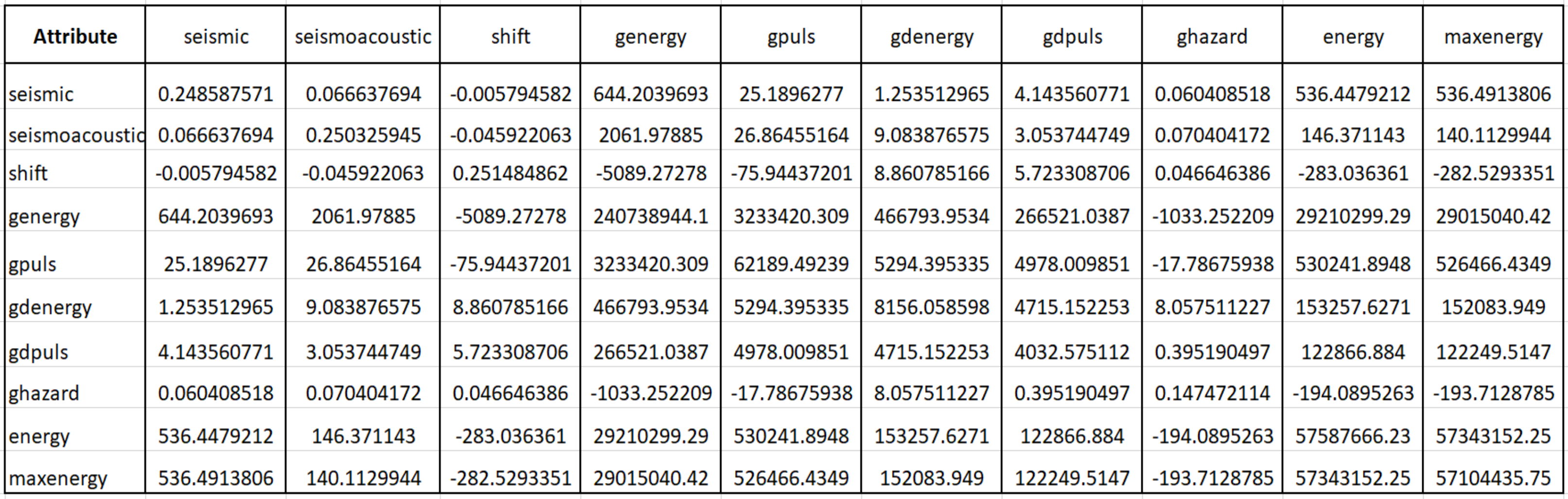
Table 3 Mean for Class 0

S. No.	Attribute Name	Mean
1.	seismic	1.42
2.	seismoacoustic	1.52
3.	shift	1.46
4.	genergy	22322.52
5.	gpuls	449.36
6.	gdenergy	24.52
7.	gdpuls	9.46
8.	ghazard	1.14
9.	energy	2176.47
10.	maxenergy	2163.02

Table 4 Mean for Class 1

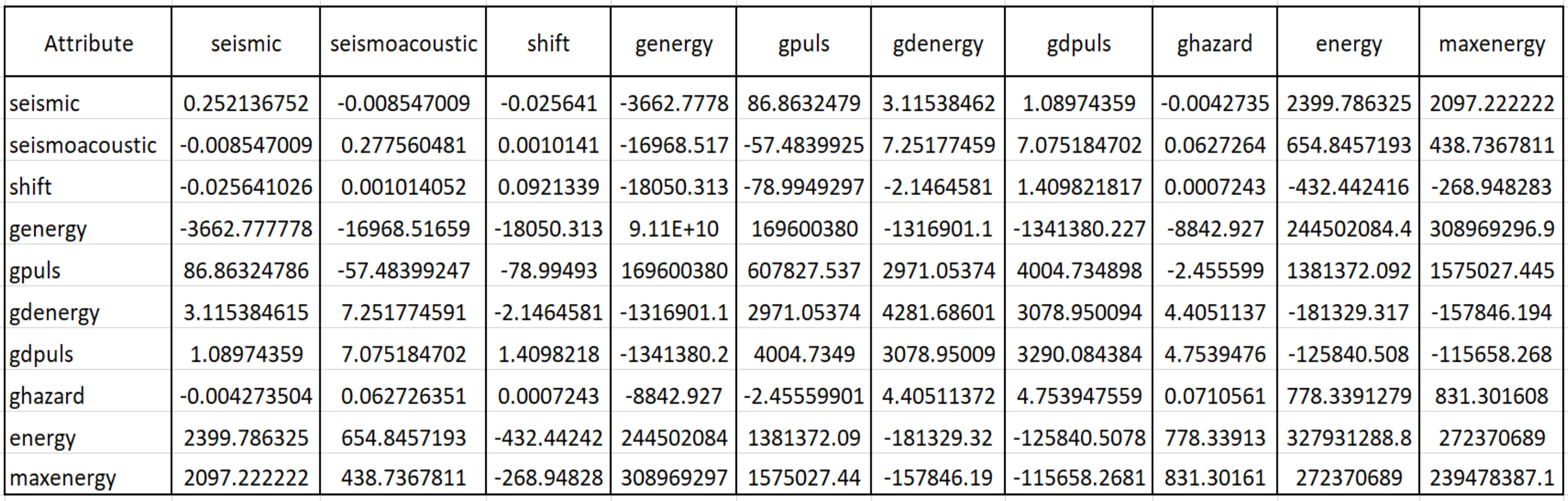
S. No.	Attribute Name	Mean
1.	seismic	1.48
2.	seismoacoustic	1.39
3.	shift	1.09
4.	genergy	190949.8
5.	gpuls	965.84
6.	gdenergy	17.03
7.	gdpuls	12.64
8.	ghazard	1.06
9.	energy	8333.19
10.	maxenergy	6504.62

Table 5 Covariance Matrix for Class 0



Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.248587571	0.066637694	-0.005794582	644.2039693	25.1896277	1.253512965	4.143560771	0.060408518	536.4479212	536.4913806
seismoacoustic	0.066637694	0.250325945	-0.045922063	2061.97885	26.86455164	9.083876575	3.053744749	0.070404172	146.371143	140.1129944
shift	-0.005794582	-0.045922063	0.251484862	-5089.27278	-75.94437201	8.860785166	5.723308706	0.046646386	-283.036361	-282.5293351
genergy	644.2039693	2061.97885	-5089.27278	240738944.1	3233420.309	466793.9534	266521.0387	-1033.252209	29210299.29	29015040.42
gpuls	25.1896277	26.86455164	-75.94437201	3233420.309	62189.49239	5294.395335	4978.009851	-17.78675938	530241.8948	526466.4349
gdenergy	1.253512965	9.083876575	8.860785166	466793.9534	5294.395335	8156.058598	4715.152253	8.057511227	153257.6271	152083.949
gdpuls	4.143560771	3.053744749	5.723308706	266521.0387	4978.009851	4715.152253	4032.575112	0.395190497	122866.884	122249.5147
ghazard	0.060408518	0.070404172	0.046646386	-1033.252209	-17.78675938	8.057511227	0.395190497	0.147472114	-194.0895263	-193.7128785
energy	536.4479212	146.371143	-283.036361	29210299.29	530241.8948	153257.6271	122866.884	-194.0895263	57587666.23	57343152.25
maxenergy	536.4913806	140.1129944	-282.5293351	29015040.42	526466.4349	152083.949	122249.5147	-193.7128785	57343152.25	57104435.75

Table 6 Covariance Matrix for Class 1



Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252136752	-0.008547009	-0.025641	-3662.7778	86.8632479	3.11538462	1.08974359	-0.0042735	2399.786325	2097.222222
seismoacoustic	-0.008547009	0.277560481	0.0010141	-16968.517	-57.4839925	7.25177459	7.075184702	0.0627264	654.8457193	438.7367811
shift	-0.025641026	0.001014052	0.0921339	-18050.313	-78.9949297	-2.1464581	1.409821817	0.0007243	-432.442416	-268.948283
genergy	-3662.777778	-16968.51659	-18050.313	9.11E+10	169600380	-1316901.1	-1341380.227	-8842.927	244502084.4	308969296.9
gpuls	86.86324786	-57.48399247	-78.99493	169600380	607827.537	2971.05374	4004.734898	-2.455599	1381372.092	1575027.445
gdenergy	3.115384615	7.251774591	-2.1464581	-1316901.1	2971.05374	4281.68601	3078.950094	4.4051137	-181329.317	-157846.194
gdpuls	1.08974359	7.075184702	1.4098218	-1341380.2	4004.7349	3078.95009	3290.084384	4.7539476	-125840.508	-115658.268
ghazard	-0.004273504	0.062726351	0.0007243	-8842.927	-2.45559901	4.40511372	4.753947559	0.0710561	778.3391279	831.301608
energy	2399.786325	654.8457193	-432.44242	244502084	1381372.09	-181329.32	-125840.5078	778.33913	327931288.8	272370689
maxenergy	2097.222222	438.7367811	-268.94828	308969297	1575027.44	-157846.19	-115658.2681	831.30161	272370689	239478387.1

Inferences:

1. The accuracy of Bayes Classifier is 78.431% and it is greater than previous classification approaches because this method uses all the data which is assumed to be normally distributed. The assumption is valid according to the Central Limit Theorem which states that all distributions can be approximated as normal distributions.
2. Diagonal elements of the covariance matrix tend to be 1 because a variable and itself are the same and hence perfectly correlated while the off-diagonal elements give the relationship between other attribute pairs.

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	75.490
2.	KNN on normalized data	72.549
3.	Bayes	78.431

Inferences:

1. The Bayes classifier had the highest accuracy while the KNN on normalized data the lowest
2. KNN on normalized data < KNN < Bayes Classifier
3. Normalizing data using Min-Max leads to slight data loss of the extreme values thus decreasing the accuracy to the lowest among the three while Bayes has the highest accuracy because it considers all the data and computes the likelihood based on the unimodal normal distribution of each attribute.