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: Sumit kumar **Mobile No: 7549233722**

Roll Number: B19118 **Branch:CSE**

1 a.

	Prediction Outcome	
True Label	671	46
	54	5

Figure 1 KNN Confusion Matrix for K = 1

	Prediction Outcome	
True Label	707	47
	18	4

Figure 3 KNN Confusion Matrix for K = 3

	Prediction Outcome	
True Label	718	46
	7	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	87.113
3	91.624
5	93.170

Inferences:

1. The highest classification accuracy is obtained with K = 5.
2. As the value of K increases, accuracy also increases.
3. The outliers and noise generally affect the result if the value of K is high. So on increasing the value of K , the contribution and dominancy of the outliers also increases, hence increasing the accuracy.
4. As the classification accuracy increases with the increase in value of K, the number of diagonal elements also increases.
5. The diagonal elements represent the number of correctly predicted values, therefore the accuracy increases.
6. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases.
7. The off – diagonal elements represent the number of falsely predicted values.

2 a.

	Prediction Outcome	
True Label	678	42
	47	9

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

	Prediction Outcome	
True Label	705	44
	20	7

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

	Prediction Outcome	
True Label	718	48
	7	3

Figure 9 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	88.531
3	91.753
5	92.912

Inferences:

1. For K = 1,3 the accuracy increases but for K = 5, accuracy decreases.
2. Since in KNN, we usually find the Euclidian distance to find the K nearest neighbors, so using normalized feature may select a different set of K neighbours than the ones chosen when unnormalized features is used therefore the accuracy is changed. After normalization, the outliers and noise are also confined to a certain range of values and the dominancy of the attribute over other due to the large values decreases, therefore the accuracy mostly increases.
3. The highest classification accuracy is obtained with K = 5.
4. The accuracy is increasing on increasing the value of K.
5. The outliers and noise generally affect the result if the value of K is high. So on increasing the value of K , the contribution and dominancy of the outliers also increases, hence increasing the accuracy.
6. As the classification accuracy increases with the increase in value of K, the number of diagonal elements also increases.
7. The diagonal elements represent the number of correctly predicted values, therefore the accuracy increases.
8. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases.
9. The off – diagonal elements represent the number of falsely predicted values.

3

	Prediction Outcome	
True Label	663	35
	62	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	1.388
4.	genergy	76209.828
5.	gpuls	490.056
6.	gdenergy	12.082
7.	gdpuls	3.542
8.	ghazard	1.107
9.	energy	4941.740
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	198697.39
5.	gpuls	944.82
6.	gdenergy	17.201
7.	gdpuls	10.638
8.	ghazard	1.075
9.	energy	10278.99
10.	maxenergy	8246.218

Table 5 Covariance Matrix for Class 0

Attribute	**seismic**	**seismoacoustic**	**shift**	**genergy**	**gpuls**	**gdenergy**	**gdpuls**	**ghazard**	**energy**	**maxenergy**
seismic	0.222943	0.0158	-0.058	341.10	53.93	5.4404	4.665	0.016200	1306.739	1133.04
seismoacoustic	0.015871	0.284	-0.018	2326.9	34.33	8.1569	7.394	0.090652	-34.789	5.744
shift	-0.05815	-0.0183	0.237	-20720.3	-108.22	-2.7909	-2.712	-0.00794	-967.727	-765.351
genergy	341.106	2326.935	-20720.3	4.314e+10	7.601e+07	808600.411	1.021e+06	-3538.71	3.433e+8	2.717e+08
gpuls	53.937	34.331	-108.22	7.601e+07	2.539e+05	12700.78	13244.25	18.99331	2.346e+6	2.013e+06
gdenergy	5.440	8.156	-2.791	8.086e+05	12700.78	6834.71	4165.206	8.9923	2.790e+5	2.705e+05
gdpuls	4.665	7.3943	-2.712	1.021e+06	13244.25	4165.205	3928.186	6.55025	2.782e+5	2.672e+05
ghazard	0.0162	0.0906	-0.0079	-3538.72	18.9e	8.992	6.5502	0.1241	-160.3407	-120.558
energy	1306.74	-34.7899	-967.72	3.433e+08	2.346e+06	279011.66	278212.5	-160.340	4.681e+8	4.430e+08
maxenergy	1133.04	5.744	-765.35	2.717e+08	2.013e+06	270563.880	267202.8	-120.558	4.430e+08	4.264e+08

Table 6 Covariance Matrix for Class 1

Attribute	seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252	0.006124	-0.033	629.01	88.58824	3.2805	1.6637	0.00455	3384.233	2889.603
seismoacoustic	0.0061	0.29995	-0.011	-1728.23	-8.96311	7.34161	7.153824	0.059251	1681.47	1108.902
shift	-0.0334	-0.0113	0.091	-15394.0	74.8464	-3.4442	-0.7768	0.000783	-539.389	-389.4459
genergy	629.01	-1728.23	-15394.05	9.849e+10	1.805e+08	-794559.639	69419.22019	-8909.631	1.436e+06	1.037e+08
gpuls	88.58	-8.9631	-74.846	1.805e+08	615028.3	7514.434	9052.4526	3.69990	997000.5	1.235e+06
gdenergy	3.280	7.3416	-3.444	-794559.6	7514.434	4734.518	3430.1243	6.3151	-168083.9	-162052.6
gdpuls	1.663	7.1538	-0.7768	69419.22	9052.453	3430.124	3425.4530	6.07840	-127217.0	-136438.2
ghazard	0.004	0.0592	0.0007	-8909.63	3.69990	6.3151	6.07840	0.0705	805.8396	854.1020
energy	3384.2	1681.46	-539.388	1.436e+06	997000.5	-168083.862	-127216.977	805.839	4.091e+08	3.419e+08
maxenergy	2889.6	1108.90	-389.4459	1.037e+08	1235626	-162052.6207	-136438.24	854.101	3.419e+08	3.006e+08

Inferences:

1. The accuracy of the bayes classifier is 87.5 %. The accuracy of the Bayes classifier is less as compared to the other two. This is because we are working on less number of dataset. Large number of dataset is likely to follow the gaussian distribution. So Bayes classifier is mostly effective in large number of dataset and for multiple class prediction.
2. The diagonal of the covariance matrix represent the variance of the particular attribute and the values are all positive. The variance of the attribute **‘genergy’** is very high. These values represents how the data is dispersed
3. **‘energy’** and **‘maxenergy’** are highly correlated whereas **‘ghazard’** and **‘shift’** are very less correlated.

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	87.50

Inferences:

1. The KNN classifier without normalization has the highest accuracy whereas the bayes classifier has the lowest accuracy.
2. The classifiers in ascending order of classification accuracy is :- Bayes < KNN on normalized data < KNN without normalization.
3. The Bayes classifier method is not very effective as compared to the others because Bayes method is effective for multiple class prediction and for large dataset but here we are working on less dataset with only two classes.