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: Prashant Kumar **Mobile No: 8700350173**

Roll Number: B19101 **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 = 1,3 and 5

K	Classification Accuracy (in %)
1	**87.113**
3	**91.623**
5	**93.170**

Inferences:

1. The highest classification accuracy is obtained with **K =5**.
2. On increasing the value of K, the prediction accuracy gets increases.
3. On increasing the value of K, the prediction accuracy increased because the number of neighbors get increase which result in getting more feature information of individual class.
4. On increasing the value of K, we observed that the diagonal elements of confusion matrix i.e. TP and TN get increased. Hence, accuracy get increased.
5. As we know accuracy = (TP+TN)/(TP+TN+FP+FN). Hence, on increasing accuracy, diagonal element also get increase.
6. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements get decrease.
7. Off-diagonal represents FP and FN elements and accuracy is directly proportional to TP and TN elements. Hence, on increasing accuracy off-diagonal elements get decrease.

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 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 = 1,3 and 5 post data normalization

K	Classification Accuracy (in %)
1	**88.530**
3	**91.752**
5	**92.912**

Inferences:

1. After normalization, classification accuracy decreases by small value but it is almost same as compared to data that was not normalized.
2. As K-NN use Euclidean distance to find that the sample belongs to which data set. But after normalization, distance get differ and then data may select different data set to decide class which may lead to increase or decrease in accuracy.
3. The highest classification accuracy is obtained with **K =5**.
4. On increasing the value of K, the prediction accuracy increased because the number of neighbors get increase which result in getting more feature information of individual class.
5. On increasing the value of K, we observed that the diagonal elements of confusion matrix i.e. TP and TN get increased. Hence, accuracy get increased.
6. As we know accuracy = (TP+TN)/(TP+TN+FP+FN). Hence, on increasing accuracy, diagonal element also get increase.
7. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements get decrease.
8. Off-diagonal represents FP and FN elements and accuracy is directly proportional to TP and TN elements. Hence, on increasing accuracy off-diagonal elements get decrease.

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	1.389
4.	genergy	76209.828
5.	gpuls	490.057
6.	gdenergy	12.082
7.	gdpuls	3.542
8.	ghazard	1.107
9.	energy	4941.741
10.	maxenergy	4374.6

Table 4 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.496
2.	seismoacoustic	1.445
3.	shift	1.101
4.	genergy	198697.395
5.	gpuls	944.824
6.	gdenergy	17.202
7.	gdpuls	10.639
8.	ghazard	1.076
9.	energy	10278.992
10.	maxenergy	8246.218

Table 5 Covariance Matrix for Class 0

Attributes	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	-20720.3	-108.223	-2.791	-2.712	-0.008	-967.727	-765.351
Genergy	341.106	2326.935	-20720.3	4.31E+10	76016422	808600.4	1021197	-3538.72	3.43E+08	2.72E+08
Gpuls	53.938	34.331	-108.223	76016422	253960.8	12700.78	13244.25	18.993	2346354	2013481
Gdenergy	5.44	8.157	-2.791	808600.4	12700.78	6834.718	4165.206	8.992	279011.7	270563.9
Gdpuls	4.665	7.394	-2.712	1021197	13244.25	4165.206	3928.186	6.55	278212.5	267202.8
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	3.43E+08	2346354	279011.7	278212.5	-160.341	4.68E+08	4.43E+08
maxenergy	1133.043	5.745	-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

Attributes	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	-1728.24	-8.963	7.342	7.154	0.059	1681.47	1108.902
shift	-0.033	-0.011	0.091	-15394.1	-74.846	-3.444	-0.777	0.001	-539.389	-389.446
Genergy	629.014	-1728.24	-15394.1	9.85E+10	1.81E+08	-794560	69419.22	-8909.63	1436182	1.04E+08
Gpuls	88.588	-8.963	-74.846	1.81E+08	615028.3	7514.434	9052.453	3.7	997000.5	1235626
Gdenergy	3.281	7.342	-3.444	-794560	7514.434	4734.518	3430.124	6.315	-168084	-162053
Gdpuls	1.664	7.154	-0.777	69419.22	9052.453	3430.124	3425.453	6.078	-127217	-136438
Ghazard	0.005	0.059	0.001	-8909.63	3.7	6.315	6.078	0.071	805.84	854.102
Energy	3384.233	1681.47	-539.389	1436182	997000.5	-168084	-127217	805.84	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. Accuracy of Bayes Classifier is 87.5%. Accuracy that we get from Bayes is low as compared to K-NN classification. Reason for this is that the Bayes Classifier used prior probability to decide the class of the test vector which is biased towards a particular class.
2. The diagonal elements of the covariance matrix represent the covariance of the column with the corresponding column. The values that are very large implies that they have very high variance as compared to other attributes which lead to shadow-off on other attributes. Due to this we normalized the data.
3. Off-diagonal elements represents the covariance of one attribute with the other attributes.

	Minimum Covariance	Maximum Covariance
Class 0	(Shift, G-hazard) – 0.008 (Seismic, Seismoacoustic) – 0.016	(Genergy, Gpuls) – 76016422 (Genergy, Gdenergy) – 808600.4
Class 1	(G-hazard, Seismoacoustic) – 0.059 (Shift, G-hazard) – 0.001	(Maxenergy, Genergy) – 1.04E+08 (Genergy, Gpuls) – 1.81E+088

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	87.5

Inferences:

1. Maximum Accuracy – K-NN Classifier (K = 5)
   Minimum Accuracy – Bayes Classifier
2. Bayes Classifier < KNN (k = 1) < KNN Normalized (k = 1) < KNN Normalized (k = 3) < KNN (k = 3) < KNN Normalized (k = 5) < KNN (k = 5).
3. Different models give different accuracy. For K-NN classifier accuracy increased with increasing value of k for some limit. And in Bayes classifier it depend upon the prior probability of class which is more biased toward a particular class.