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: Jai Prakash Yadav   **Mobile No: 9306871378**

Roll Number: B19247   **Branch: ME**

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 =. 3
2. On increasing the value of K , prediction accuracy increases.
3. With increase in k more elements are used which reduces the error caused by outliers of other classes.
4. On, increasing the value of K beyond a particular limit will result in the classifier to classifying all the test cases to the case with more training examples.
5. The number of diagonal elements increasing as the number of class included increases.
6. As the accuracy is directly proportional to diagonal elements (TP+TN).
7. The no of diagonal elements decreases as K decreases as accuracy decreases.

2   a.

	Prediction Outcome	
674		49
43		10

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

	Prediction Outcome	
708		15
47		6

Figure 7 KNN Confusion Matrix for K = 3 post data normalization

	Prediction Outcome	
716		7
52		1

Figure 8 KNN Confusion Matrix for K = 5 post data normalization

3

b.

Table 2 KNN Classification Accuracy for K = 1,3and 5 post data normalization

K	Classification Accuracy (in %)
1	0.881
3	0.918
5	0.924

Inferences:

1. Normalization increases classification accuracy.
2. There is increase in accuracy because normalization prevents one attribute from overshadowing other.
3. The highest classification accuracy is obtained with K = 5.
4. On increasing the value of a prediction accuracy decreases.
5. With increase in k more elements are used which reduces the error caused by outliers of other classes.
6. As the classification accuracy increases with the increase in K, the number of diagonal elements increases.
7. The number of diagonal elements is directly proportional to the number of diagonal elements.
8. As the classification accuracy decreases, the number of off diagonal elements decreases.
9. **The no of diagonal elements is directly proportional to the no of diagonal elements.**

3

	Prediction Outcome	
True Label	678	48
	38	15

Figure 11 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is   0.889 %.

Table 3 Mean for Class 0

S. No.	Attribute Name	Mean
1.	seismic	1.333
2.	seismoacoustic	1.410
3.	shift	1.374
4.	genergy	76427.581
5.	gpuls	502.933
6.	gdenergy	12.928
7.	gdpuls	4.409
8.	ghazard	1.108
9.	energy	4723.257
10.	maxenergy	4107.096

Table 4 Mean for Class 1

S. No.	Attribute Name	Mean
1.	seismic	1.496
2.	seismoacoustic	1.444
3.	shift	1.103
4.	genergy	189497.179
5.	gpuls	939.923
6.	gdenergy	15.573
7.	gdpuls	9.744
8.	ghazard	1.085
9.	energy	8809.829
10.	maxenergy	6850.855

Table 5 Covariance Matrix for Class 0

	seismic	seismoaco	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.222222	0.013179	-0.06179	-1409.47	58.54179	5.574725	4.086749	0.015033	1456.164	1245.092
seismoaco	0.013179	0.27988	-0.02367	-578.263	26.96054	7.632892	6.591365	0.08486	-287.157	-251.016
shift	-0.06179	-0.02367	0.234198	-19995.8	-110.867	-4.0401	-3.26309	-0.01007	-1030.11	-815.376
genergy	-1409.47	-578.263	-19995.8	4.05E+10	75815357	903824.9	904843	-2674.31	2.4E+08	1.69E+08
gpuls	58.54179	26.96054	-110.867	75815357	273959	13839.63	13619.91	20.17169	2091381	1764191
gdenergy	5.574725	7.632892	-4.0401	903824.9	13839.63	7136.955	4407.639	9.521907	215128.4	209949.9
gdpuls	4.086749	6.591365	-3.26309	904843	13619.91	4407.639	4160.209	6.939362	222839.3	213586.2
ghazard	0.015033	0.08486	-0.01007	-2674.31	20.17169	9.521907	6.939362	0.122137	-167.798	-120.173
energy	1456.164	-287.157	-1030.11	2.4E+08	2091381	215128.4	222839.3	-167.798	4.25E+08	3.99E+08
maxenergy	1245.092	-251.016	-815.376	1.69E+08	1764191	209949.9	213586.2	-120.173	3.99E+08	3.83E+08

Table 6 Covariance Matrix for Class 1

	seismic	seismoaco	shift	genergy	gpuls	gdenergy	gdpuls	ghazard	energy	maxenergy
seismic	0.252137	-0.01533	-0.03404	6829.083	100.4092	2.075744	1.56786	0.000368	2386.465	2100.435
seismoaco	-0.01533	0.300766	-0.00287	4647.73	-13.2155	7.407088	6.977011	0.065134	495.5939	216.8582
shift	-0.03404	-0.00287	0.092838	-17051.2	-74.5524	-2.60234	0.621353	-0.00022	-679.034	-502.244
genergy	6829.083	4647.73	-17051.2	7.8E+10	1.47E+08	-1827278	-808657	-7598.46	6.55E+08	6.1E+08
gpuls	100.4092	-13.2155	-74.5524	1.47E+08	516572.2	2057.743	4191.015	-10.0537	2458029	2372205
gdenergy	2.075744	7.407088	-2.60234	-1827278	2057.743	4579.35	3174.683	2.683392	-186145	-160880
gdpuls	1.56786	6.977011	0.621353	-808657	4191.015	3174.683	3318.141	3.780725	-111248	-103548
ghazard	0.000368	0.065134	-0.00022	-7598.46	-10.0537	2.683392	3.780725	0.078839	429.3251	515.4436
energy	2386.465	495.5939	-679.034	6.55E+08	2458029	-186145	-111248	429.3251	3.42E+08	2.8E+08
maxenergy	2100.435	216.8582	-502.244	6.1E+08	2372205	-160880	-103548	515.4436	2.8E+08	2.43E+08

Inferences:

1. Accuracy of Bayes classifier= 0.889.
2. Bayes Classifier is a supervised learning algorithm, it assumes that all features are independent which is not usually the case in real life so it makes the naïve Bayes less accurate.
3. The values along the diagonal (COVii) of the covariance matrix is positive. The diagonal elements represent variance of corresponding attribute (VARi).
4. The values along the off-diagonal elements reflects w.r.t the diagonal.
5. Max covariance: (genergy, genergy), (energy, genergy).
6. Min covariance: (ghazard, seismic), (ghazard, shift). (Magnitude).

Table 7 Comparison between Classifier based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	0.924
2.	KNN on normalized data	0.924
3.	Bayes	0.889

Inferences:

1. Highest accuracy= KNN and KNN on normalized data.
2. The lowest accuracy= Bayes classifier.
3. Arrange the classifiers in ascending order of classification accuracy.
4. Bayes< KNN = KNN Normalised.
5. In terms of speed Bayes>KNN.
6. KNN is an unsupervised classifier whereas Bayes is a supervised classifier.