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: Vivek jaiswal **Mobile No: 9887552039**

Roll Number: b20172 **Branch: DSE**

1 a.

		Prediction Outcome
True Label	93	25
	19	200

Figure 1 KNN Confusion Matrix for K = 1

		Prediction Outcome
True Label	92	26
	9	210

Figure 2 KNN Confusion Matrix for K = 3

		Prediction Outcome
True Label	92	26
	10	209

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	**0.869**
3	**0.896**
4	**0.893**

Inferences:

1.The highest accuracy is obtained with K=3.

2. Increasing the value of K increases the prediction accuracy and after k=3

3. By increasing the values of K, the value will be checked for greater number of sorted Euclidian distance, so due to that less effect of noise. For less value of K, the noise affects the prediction more.

4.As the accuracy increases with the increase in value of K the number of diagonal elements increases.

5. Increase in accuracy means more correct predictions and less wrong predictions, thus increasing true positive and true negative frequencies.

6. As the classification accuracy increases with the increase in value of K the number of off-diagonal elements decrease.

7. Increase in accuracy means more correct predictions and less wrong predictions, thus decreasing false positive and false negative frequencies.

2 a.

	Prediction Outcome	
True Label	115	3
	13	206

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

	Prediction Outcome	
True Label	116	2
	14	205

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

	Prediction Outcome	
True Label	116	2
	13	206

Figure 6 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	**95.22**
3	**95.23**
5	**95.5**

Inferences:

1.Data normalization increases accuracy as we see that without normalize data have max accuracy of 89percent but normalize data have accuracy about 95.5percent.

2. The accuracy is increased because after normalization, the attributes which have bigger ranges as compared to smaller range attribute, the bigger range attribute diminishes the smaller range value while calculating Euclidian distance and neglect the importance of smaller range values.

3. The highest accuracy is obtained with K = 5.

4. Increasing the value of K increases the prediction accuracy.

5. Increasing the value of K increases the prediction accuracy till a certain point as the no. of comparisons increases and at a certain point it almost becomes constant and decreasing for very high values of K.

6. As the classification accuracy increases with the increase in value of K the number of diagonal elements increase.

7. Increase in accuracy means more correct predictions and less wrong predictions, thus increasing true positive and true negative frequencies.

8. As the classification accuracy increases with the increase in value of K the number of off-diagonal elements decrease.

9. Increase in accuracy means more correct predictions and less wrong predictions, thus decreasing false positive and false negative frequencies

3

		Prediction Outcome
True Label	102	16
	3	216

Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is % 94.363.

Table 3 Mean for class 0 and class 1

S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Maximum	273.418	723.656
2.	Y_Maximum	1583169.659	1431588.69
3.	Pixels_Areas	7779.663	585.967
4.	X_Perimeter	393.835	54.491
5.	Y_Perimeter	273.183	45.658
6.	Sum_of_Luminosity	843350.275	62191.126
7.	Minimum_of_Luminosity	53.326	96.236
8.	Maximum_of_Luminosity	135.762	130.452
9.	Length_of_Conveyer	1382.762	1480.018
10.	Steel_Plate_Thickness	40.073	104.214
11.	Edges_Index	0.123	0.385
12.	Empty_Index	0.459	0.427
13.	Square_Index	0.592	0.513
14.	Outside_X_Index	0.108	0.02
15.	Edges_X_Index	0.550	0.608
16.	Edges_Y_Index	0.523	0.831
17.	Outside_Global_Index	0.288	0.608
18.	LogOfAreas	3.623	2.287
19.	Log_X_Index	2.057	1.227
20.	Log_Y_Index	1.848	1.318
21.	Orientation_Index	-0.314	0.136
22.	Luminosity_Index	-0.115	-0.116
23.	SigmoidOfAreas	0.925	0.543

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	46733.77	-6.1E+07	-320672	-15750.5	-12943.8	-3.3E+07	3686.073	2040.905	1237.644	16.734	25.36	-6.929	4.696	-1.516	16.654	22.505	30.839	-76.32	-47.782	-31.147	27.679	18.083	-30.093
2	-6.1E+07	1.82E+12	1.03E+09	83317353	1.6E+08	4.9E+10	-5669890	-6007837	-7505510	-114611	-47711.4	21948.27	-59251.3	4294.736	-19165.6	-35306.4	-86404.1	168069.8	111447.7	73014.36	-82046.9	-50711.2	73811.61
3	-320672	1.03E+09	1.05E+08	6692649	10371695	9.01E+09	-154934	6294.464	10070.21	547.01	-492.113	585.231	200.195	223.056	-1121.19	-354.573	556.075	3456.879	1427.026	2840.741	980.333	-300.211	575.04
4	-15750.5	83317353	6692649	442770.6	706256.5	5.57E+08	-7764.05	769.586	771.604	31.924	-24.093	38.161	10.596	10.994	-67.824	-13.284	45.342	183.057	68.412	169.129	72.436	-15.703	28.521
5	-12943.8	1.6E+08	10371695	706256.5	1206391	8.08E+08	-6894.47	1492.073	-1364.2	10.207	-17.571	44.182	-16.55	6.496	-65.417	13.411	63.25	176.64	44.055	207.792	105.12	-21.062	19.506
6	-3.3E+07	4.9E+10	9.01E+09	5.57E+08	8.08E+08	8.19E+11	-1.6E+07	777671.3	2214134	49759.91	-53267.3	58474.64	44601.85	25470.52	-123181	-50984.9	60033.13	361544.8	157340.8	278177.3	96509.49	-22290.5	62063.26
7	3686.073	-5669890	-154934	-7764.05	-6894.47	-1.6E+07	1458.213	439.236	-153.834	-1.973	3.932	-1.75	1.078	-1.455	3.739	4.623	4.759	-22.187	-12.861	-10.747	3.817	4.448	-6.557
8	2040.905	-6007837	6294.464	769.586	1492.073	777671.3	439.236	333.381	2.285	-0.791	1.769	-0.222	2.058	-0.353	-0.142	1.575	4.207	-5.859	-4.358	-1.529	4.136	2.716	-2.737
9	1237.644	-7505510	10070.21	771.604	-1364.2	2214134	-153.834	2.285	2521.557	-1.821	1.322	0.806	3.926	-0.192	-2.697	-0.534	4.536	2.03	-0.002	2.645	4.37	-0.485	0.211
10	16.734	-114611	547.01	31.924	10.207	49759.91	-1.973	-0.791	-1.821	0.73	-0.009	0.015	-0.015	0.019	0.003	-0.015	-0.021	0.041	0.041	0.019	-0.022	-0.008	0.005
11	25.36	-47711.4	-492.113	-24.093	-17.571	-53267.3	3.932	1.769	1.322	-0.009	0.029	-0.009	0.007	-0.006	0.015	0.022	0.026	-0.084	-0.054	-0.038	0.024	0.016	-0.028
12	-6.929	21948.27	585.231	38.161	44.182	58474.64	-1.75	-0.222	0.806	0.015	-0.009	0.015	0.005	0.005	-0.018	-0.012	0.003	0.052	0.03	0.036	0.005	-0.003	0.015
13	4.696	-59251.3	200.195	10.596	-16.55	44601.85	1.078	2.058	3.926	-0.015	0.007	0.005	0.064	-0.004	-0.036	-0.001	0.07	0.001	-0.02	0.023	0.069	0.016	-0.01
14	-1.516	4294.736	223.056	10.994	6.496	25470.52	-1.455	-0.353	-0.192	0.019	-0.006	0.005	-0.004	0.005	-0.002	-0.007	-0.01	0.029	0.021	0.014	-0.01	-0.004	0.007
15	16.654	-19165.6	-1121.19	-67.824	-65.417	-123181	3.739	-0.142	-2.697	0.003	0.015	-0.018	-0.036	-0.002	0.057	0.023	-0.039	-0.098	-0.039	-0.073	-0.045	0.003	-0.026
16	22.505	-35306.4	-354.573	-13.284	13.411	-50984.9	4.623	1.575	-0.534	-0.015	0.022	-0.012	-0.001	-0.007	0.023	0.031	0.025	-0.099	-0.063	-0.045	0.023	0.014	-0.031
17	30.839	-86404.1	556.075	45.342	63.25	60033.13	4.759	4.207	4.536	-0.021	0.026	0.003	0.07	-0.01	-0.039	0.025	0.203	-0.058	-0.073	0.019	0.138	0.033	-0.033
18	-76.32	168069.8	3456.879	183.057	176.64	361544.8	-22.187	-5.859	2.03	0.041	-0.084	0.052	0.001	0.029	-0.098	-0.099	-0.058	0.471	0.267	0.247	-0.044	-0.067	0.135
19	-47.782	111447.7	1427.026	68.412	44.055	157340.8	-12.861	-4.358	-0.002	0.041	-0.054	0.03	-0.02	0.021	-0.039	-0.063	-0.073	0.267	0.168	0.124	-0.066	-0.044	0.082
20	-31.147	73014.36	2840.741	169.129	207.792	278177.3	-10.747	-1.529	2.645	0.019	-0.038	0.036	0.023	0.014	-0.073	-0.045	0.019	0.247	0.124	0.157	0.029	-0.025	0.065
21	27.679	-82046.9	980.333	72.436	105.12	96509.49	3.817	4.136	4.37	-0.022	0.024	0.005	0.069	-0.01	-0.045	0.023	0.138	-0.044	-0.066	0.029	0.133	0.031	-0.028
22	18.083	-50711.2	-300.211	-15.703	-21.062	-22290.5	4.448	2.716	-0.485	-0.008	0.016	-0.003	0.016	-0.004	0.003	0.014	0.033	-0.067	-0.044	-0.025	0.031	0.027	-0.026
23	-30.093	73811.61	575.04	28.521	19.506	62063.26	-6.557	-2.737	0.211	0.005	-0.028	0.015	-0.01	0.007	-0.026	-0.031	-0.033	0.135	0.082	0.065	-0.028	-0.026	0.049

Figure 8: Covariance matrix for class 0

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2	256526.3	1.12E+08	-22254.6	1101.079	-1973.57	-2334976	-1224.81	-744.043	13220.08	-1932.62	8.914	-3.806	10.893	1.504	6.695	-5.018	-16.564	-13.781	5.306	-21.204	-25.896	-8.452	-14.221
3	1.12E+08	3.12E+12	3.23E+08	20351188	4659662	3.3E+10	-3631825	-43295.9	3999506	-3.6E+07	23556.3	-19251	-38009.7	13457.3	64532.97	-22198.8	-74705.2	15298.09	64300.31	-63426.8	-119870	-14717.9	-37674.9
4	-22254.6	3.23E+08	4714217	178492.1	129451.1	4.89E+08	-15632	-300.304	-23834.7	4262.208	-47.646	35.619	-90.634	52.909	-101.643	-96.057	55.178	653.051	330.779	355.115	65.419	-32.384	218.948
5	1101.079	20351188	178492.1	9807.203	5546.899	18662200	-570.116	30.15	-1446.88	282.113	-1.332	4.156	-7.318	3.972	-4.85	-9.176	-2.152	36.62	23.557	16.864	-3.758	-1.119	15.508
6	-1973.57	4659662	129451.1	5546.899	5000.647	13453353	-557.423	-79.146	-1139.31	438.56	-2.244	2.952	-6.496	1.204	-8.612	-2.367	7.11	29.028	10.681	21.025	11.045	-1.556	13.014
7	-2334976	3.3E+10	4.89E+08	18662200	13453353	5.09E+10	-1463161	84723.03	-2735155	343512.4	-4688.9	3985.075	-9652.58	5577.969	-10534.6	-10271.9	5462.295	67782.66	34740.29	36734.78	6364.119	-2282.38	22864.85
8	-1224.81	-3631825	-15632	-570.116	-557.423	-1463161	733.909	348.045	-993.311	-204.836	1.066	0.591	0.775	-0.151	0.427	-0.833	-2.224	-5.043	-1.299	-3.287	-2.503	3.684	-1.984
9	-744.043	-43295.9	-300.304	30.15	-79.146	84723.03	348.045	406.461	-381.093	-205.394	0.429	-0.025	-0.267	0.044	0.878	-1.09	-2.018	-1.504	0.678	-2.165	-2.874	2.786	-0.96
10	13220.08	3999506	-23834.7	-1446.88	-1139.31	-2735155	-993.311	-381.093	23100.77	1243.443	-0.09	-5.16	2.468	-0.698	6.591	1.971	-3.138	-7.953	-1.44	-10.567	-7.431	-4.547	-5.967
11	-1932.62	-3.6E+07	4262.208	282.113	438.56	343512.4	-204.836	-205.394	1243.443	5645.306	-1.331	0.699	-1.134	-0.165	-3.443	2.058	6.623	3.627	-1.376	5.403	7.846	-1.662	2.39
12	8.914	23556.3	-47.646	-1.332	-2.244	-4688.9	1.066	0.429	-0.09	-1.331	0.09	-0.001	0.011	0	0.008	-0.003	-0.017	-0.012	0.005	-0.017	-0.024	0.005	-0.004
13	-3.806	-19251	35.619	4.156	2.952	3985.075	0.591	-0.025	-5.16	0.699	-0.001	0.02	-0.002	0.001	-0.012	-0.011	-0.008	0.026	0.022	0.022	-0.004	0.002	0.024
14	10.893	-38009.7	-90.634	-7.318	-6.496	-9652.58	0.775	-0.267	2.468	-1.134	0.011	-0.002	0.082	-0.003	0.02	0.015	-0.016	-0.053	-0.021	-0.033	-0.021	0.001	-0.028
15	1.504	13457.3	52.909	3.972	1.204	5577.969	-0.151	0.044	-0.698	-0.165	0	0.001	-0.003	0.002	0.002	-0.005	-0.005	0.012	0.012	0.001	-0.008	0	0.005
16	6.695	64532.97	-101.643	-4.85	-8.612	-10534.6	0.427	0.878	6.591	-3.443	0.008	-0.012	0.02	0.002	0.065	-0.014	-0.068	-0.066	0.011	-0.086	-0.103	0.004	-0.045
17	-5.018	-22198.8	-96.057	-9.176	-2.367	-10271.9	-0.833	-1.09	1.971	2.058	-0.003	-0.011	0.015	-0.005	-0.014	0.049	0.064	-0.025	-0.058	0.024	0.086	-0.007	-0.017
18	-16.564	-74705.2	55.178	-2.152	7.11	5462.295	-2.224	-2.018	-3.138	6.623	-0.017	-0.008	-0.016	-0.005	-0.068	0.064	0.227	0.048	-0.073	0.113	0.229	-0.015	0.022
19	-13.781	15298.09	653.051	36.62	29.028	67782.66	-5.043	-1.504	-7.953	3.627	-0.012	0.026	-0.053	0.012	-0.066	-0.025	0.048	0.271	0.116	0.177	0.073	-0.019	0.147
20	5.306	64300.31	330.779	23.557	10.681	34740.29	-1.299	0.678	-1.44	-1.376	0.005	0.022	-0.021	0.012	0.011	-0.058	-0.073	0.116	0.119	0.017	-0.101	0	0.065
21	-21.204	-63426.8	355.115	16.864	21.025	36734.78	-3.287	-2.165	-10.567	5.403	-0.017	0.022	-0.033	0.001	-0.086	0.024	0.113	0.177	0.017	0.178	0.169	-0.017	0.103
22	-25.896	-119870	65.419	-3.758	11.045	6364.119	-2.503	-2.874	-7.431	7.846	-0.024	-0.004	-0.021	-0.008	-0.103	0.086	0.229	0.073	-0.101	0.169	0.302	-0.019	0.041
23	-8.452	-14717.9	-32.384	-1.119	-1.556	-2282.38	3.684	2.786	-4.547	-1.662	0.005	0.002	0.001	0	0.004	-0.007	-0.015	-0.019	0	-0.017	-0.019	0.025	-0.009
24	-14.221	-37674.9	218.948	15.508	13.014	22864.85	-1.984	-0.96	-5.967	2.39	-0.004	0.024	-0.028	0.005	-0.045	-0.017	0.022	0.147	0.065	0.103	0.041	-0.009	0.102

Figure 9: Covariance matrix for class 1

In Figure 8 and 9 representing covariance matrices for class 0 and class 1 and the column numbers and row numbers correspond to attribute with serial number as in Table 3.

Inferences:

1. The accuracy from Bayes classifier came is 94.362%. This is because, when solving a problem Bayes directly focusses on finding similarity between observations, KNN does better because of its inherent nature to optimize locally.
2. The diagonal elements of the covariance matrix denote the variance of the attribute with itself, that is, how much is the data spread across the median. From looking at the diagonal elements, we can infer the dispersion of the attribute and have an idea about the range of values in the attribute.

3. The off-diagonal elements indicate the covariance between the two attributes-how the attributes vary with respect to each other.2 pair of attributes with maximum covariance are Y_maximum & Sum_of_Luminosity and Sum_of_Luminosity and Pixel_Areas for both classes. 2 pair of attributes with minimum covariance are Square_Index & Edges_Y_Index and Square_Index & LogOfAreas for class 0 and for class 1 Luminosity_Index & Log_X_Index and Luminosity_Index & Outside_X_Index.

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.60
2.	KNN on normalized data	95.56
3.	Bayes	94.36

Inferences:

1. The classifier with the best accuracy is KNN on normalized data followed by Bayes Classifier. KNN shows the least accuracy among all.
2. The classifiers in ascending order of classification accuracy KNN < Bayes < KNN on normalized data.

3. KNN performs better when data is normalized because, the attributes on a bigger scale can no longer overpower and influence the results in their favor. This happens because Euclidean Distance is the total absolute distance along various axes and doesn't consider for the different ranges. The Bayes classifier directly focusses on finding similarity between observations, K-NN does better because of its inherent nature to optimize locally. Also, in the above example which involves just 2 clusters, KNN will give more accurate predictions than Bayes.