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: Abhay Vijayvargiya **Mobile No: 6377967485**

Roll Number: B20176 **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	86.94
3	89.61
5	89.31

Inferences:

1. The highest classification accuracy is obtained with K = 3.
2. On increasing the value of K, the accuracy increases but at a certain threshold if we further increase the K, then accuracy decreases.
3. On increasing the value of K (up to a certain point), the accuracy increases as we are comparing the test example with more no. adjacent training tuples as a result the accuracy increases.
4. As the classification accuracy increases there is increase in diagonal elements.
5. Since the diagonal elements represents true positives and false negatives which collectively represents the no. of test tuples correctly identified, so increasing the classification accuracy increases the sum of diagonal elements.
6. Since the off-diagonal elements represents true negatives and false positives which collectively represents the no. of test tuples incorrectly identified, so increasing the classification accuracy decreases the sum of off-diagonal elements.
7. We generally take K an odd number to minimize the ties.

2   a.

	Prediction Outcome	
True Label	111	7
	6	213

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

	Prediction Outcome	
True Label	113	5
	4	215

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

	Prediction Outcome	
True Label	109	9
	4	215

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	96.14
3	97.33
5	96.14

Inferences:

1. Data normalization leads to increase in classification accuracy.
2. Basically, in KNN classifier we measure the Euclidean distance between test tuple and training tuples so if on attribute has comparatively larger scale than other attributes, then the most clusters will be generated based on that attribute.
3. The highest classification accuracy is obtained with K = 3.
4. On increasing the value of K, the accuracy increases but at a certain threshold if we further increase the K, then accuracy decreases.
5. On increasing the value of K (up to a certain point), the accuracy increases as we are comparing the test example with more no. adjacent training tuples as a result the accuracy increases.
6. As the classification accuracy increases there is increase in diagonal elements.
7. Since the diagonal elements represents true positives and false negatives which collectively represents the no. of test tuples correctly identified, so increasing the classification accuracy increases the sum of diagonal elements.
8. Since the off-diagonal elements represents true negatives and false positives which collectively represents the no. of test tuples incorrectly identified, so increasing the classification accuracy decreases the sum of off-diagonal elements.

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 94.36 %.

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	0.55	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

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

	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	25.36021	-6.9293	4.696193	-1.51587	16.65354	22.50463	30.83904	-76.3196	-47.7816	-31.1473	27.67876	18.08286	-30.0931
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	547.0101	-492.113	585.2306	200.1953	223.0561	-1121.19	-354.573	556.0752	3456.879	1427.026	2840.741	980.3329	-300.211	575.0404
4	-15750.5	83317353	6692649	442770.6	706256.5	5.57E+08	-7764.05	769.5856	771.604	31.92388	-24.0928	38.16111	10.59581	10.99425	-67.8237	-13.284	45.34169	183.0575	68.41173	169.1286	72.43566	-15.7026	28.52111
5	-12943.8	1.6E+08	10371695	706256.5	1206391	8.08E+08	-6894.47	1492.073	-1364.2	10.20712	-17.5711	44.18238	-16.5502	6.495981	-65.4173	13.41058	63.25045	176.6405	44.05484	207.7917	105.1195	-21.062	19.50566
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	-1.9725	3.931511	-1.75004	1.077743	-1.45529	3.738841	4.623318	4.758855	-22.1867	-12.8607	-10.7472	3.816648	4.448267	-6.55741
8	2040.905	-6007837	6294.464	769.5856	1492.073	777671.3	439.236	333.3806	2.285014	-0.79132	1.768683	-0.22159	2.057703	-0.35296	-0.14245	1.57515	4.206583	-5.85939	-4.35841	-1.52924	4.136383	2.716174	-2.7371
9	1237.644	-7505510	10070.21	771.604	-1364.2	2214134	-153.834	2.285014	2521.557	-1.82073	1.321957	0.806365	3.925976	-0.19247	-2.69665	-0.53421	4.535627	2.03005	-0.00187	2.644925	4.369843	-0.4847	0.21099
10	16.734	-114611	547.0101	31.92388	10.20712	49759.91	-1.9725	-0.79132	-1.82073	0.729907	-0.00874	0.0147	-0.01549	0.019054	0.003184	-0.01538	-0.02114	0.041098	0.041366	0.019269	-0.02246	-0.0077	0.005483
11	25.36021	-47711.4	-492.113	-24.0928	-17.5711	-53267.3	3.931511	1.768683	1.321957	-0.00874	0.029323	-0.00928	0.007154	-0.00605	0.014692	0.022417	0.026357	-0.08402	-0.05352	-0.03759	0.024297	0.015975	-0.02755
12	-6.9293	21948.27	585.2306	38.16111	44.18238	58474.64	-1.75004	-0.22159	0.806365	0.0147	-0.00928	0.015302	0.00472	0.004944	-0.01766	-0.0116	0.003021	0.051673	0.030409	0.036164	0.005163	-0.00347	0.015267
13	4.696193	-59251.3	200.1953	10.59581	-16.5502	44601.85	1.077743	2.057703	3.925976	-0.01549	0.007154	0.00472	0.064486	-0.00411	-0.03633	-0.00065	0.070297	0.001334	-0.01967	0.023186	0.068654	0.016339	-0.0097
14	-1.51587	4294.736	223.0561	10.99425	6.495981	25470.52	-1.45529	-0.35296	-0.19247	0.019054	-0.00605	0.004944	-0.00411	0.004743	-0.00222	-0.00731	-0.00975	0.029154	0.020886	0.01388	-0.00952	-0.00376	0.007482
15	16.65354	-19165.6	-1121.19	-67.8237	-65.4173	-123181	3.738841	-0.14245	-2.69665	0.003184	0.014692	-0.01766	-0.03633	-0.00222	0.056908	0.022848	-0.03856	-0.09841	-0.03926	-0.07308	-0.04451	0.002776	-0.02567
16	22.50463	-35306.4	-354.573	-13.284	13.41058	-50984.9	4.623318	1.57515	-0.53421	-0.01538	0.022417	-0.0116	-0.00065	-0.00731	0.022848	0.030681	0.024941	-0.09928	-0.0626	-0.04465	0.023024	0.014378	-0.0311
17	30.83904	-86404.1	556.0752	45.34169	63.25045	60033.13	4.758855	4.206583	4.535627	-0.02114	0.026357	0.003021	0.070297	-0.00975	-0.03856	0.024941	0.202859	-0.05783	-0.07275	0.019258	0.138071	0.033017	-0.03252
18	-76.3196	168069.8	3456.879	183.0575	176.6405	361544.8	-22.1867	-5.85939	2.03005	0.041098	-0.08402	0.051673	0.001334	0.029154	-0.09841	-0.09928	-0.05783	0.471457	0.266901	0.246904	-0.04394	-0.06701	0.135218
19	-47.7816	111447.7	1427.026	68.41173	44.05484	157340.8	-12.8607	-4.35841	-0.00187	0.041366	-0.05352	0.030409	-0.01967	0.020886	-0.03926	-0.0626	-0.07275	0.266901	0.167866	0.124113	-0.06631	-0.04408	0.081643
20	-31.1473	73014.36	2840.741	169.1286	207.7917	278177.3	-10.7472	-1.52924	2.644925	0.019269	-0.03759	0.036164	0.023186	0.01388	-0.07308	-0.04465	0.019258	0.246904	0.124113	0.156846	0.029178	-0.02546	0.064575
21	27.67876	-82046.9	980.3329	72.43566	105.1195	96509.49	3.816648	4.136383	4.369843	-0.02246	0.024297	0.005163	0.068654	-0.00952	-0.04451	0.023024	0.138071	-0.04394	-0.06631	0.029178	0.133168	0.030895	-0.02766
22	18.08286	-50711.2	-300.211	-15.7026	-21.062	-22290.5	4.448267	2.716174	-0.4847	-0.0077	0.015975	-0.00347	0.016339	-0.00376	0.002776	0.014378	0.033017	-0.06701	-0.04408	-0.02546	0.030895	0.027438	-0.02644
23	-30.0931	73811.61	575.0404	28.52111	19.50566	62063.26	-6.55741	-2.7371	0.21099	0.005483	-0.02755	0.015267	-0.0097	0.007482	-0.02567	-0.0311	-0.03252	0.135218	0.081643	0.064575	-0.02766	-0.02644	0.049322

Figure 8: Covariance matrix for class 0

Data classification using K-nearest neighbor classifier and Bayes classifier with unimodal Gaussian density

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	256526.3	1.12E+08	-22254.6	1101.079	-1973.56	-2334976	-1224.81	-744.043	13220.08	-1932.62	8.913916	-3.8064	10.89266	1.504328	6.694786	-5.01836	-16.5642	-13.7813
2	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
3	-22254.6	3.23E+08	4714217	178492.1	129451.1	4.89E+08	-15632	-300.304	-23834.7	4262.208	-47.6455	35.6195	-90.6336	52.90864	-101.643	-96.0566	55.17783	653.0513
4	1101.079	20351188	178492.1	9807.203	5546.899	18662200	-570.116	30.14967	-1446.88	282.1131	-1.33167	4.155596	-7.3181	3.971901	-4.84985	-9.17608	-2.1516	36.6199
5	-1973.56	4659662	129451.1	5546.899	5000.647	13453353	-557.423	-79.1464	-1139.31	438.5596	-2.24421	2.951694	-6.49605	1.204469	-8.61151	-2.36737	7.109846	29.02755
6	-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
7	-1224.81	-3631825	-15632	-570.116	-557.423	-1463161	733.9089	348.0448	-993.311	-204.836	1.066368	0.591072	0.775182	-0.15145	0.427209	-0.83326	-2.22434	-5.04259
8	-744.043	-43295.9	-300.304	30.14967	-79.1464	84723.03	348.0448	406.4608	-381.093	-205.394	0.429118	-0.02454	-0.26703	0.04392	0.877571	-1.08968	-2.01841	-1.50427
9	13220.08	3999506	-23834.7	-1446.88	-1139.31	-2735155	-993.311	-381.093	23100.77	1243.443	-0.09047	-5.15952	2.468171	-0.69776	6.591052	1.97125	-3.13774	-7.95323
10	-1932.62	-3.6E+07	4262.208	282.1131	438.5596	343512.4	-204.836	-205.394	1243.443	5645.306	-1.3306	0.699194	-1.13384	-0.16545	-3.44259	2.058128	6.623469	3.626633
11	8.913916	23556.3	-47.6455	-1.33167	-2.24421	-4688.9	1.066368	0.429118	-0.09047	-1.3306	0.08965	-0.00063	0.010929	6.45E-05	0.008301	-0.00333	-0.01658	-0.01211
12	-3.8064	-19251	35.6195	4.155596	2.951694	3985.075	0.591072	-0.02454	-5.15952	0.699194	-0.00063	0.020283	-0.00202	0.001242	-0.01249	-0.01101	-0.00752	0.026336
13	10.89266	-38009.7	-90.6336	-7.3181	-6.49605	-9652.58	0.775182	-0.26703	2.468171	-1.13384	0.010929	-0.00202	0.082373	-0.00291	0.019744	0.014881	-0.01558	-0.05315
14	1.504328	13457.3	52.90864	3.971901	1.204469	5577.969	-0.15145	0.04392	-0.69776	-0.16545	6.45E-05	0.001242	-0.00291	0.002467	0.001752	-0.00529	-0.0052	0.011616
15	6.694786	64532.97	-101.643	-4.84985	-8.61151	-10534.6	0.427209	0.877571	6.591052	-3.44259	0.008301	-0.01249	0.019744	0.001752	0.065074	-0.01386	-0.06755	-0.06618
16	-5.01836	-22198.8	-96.0566	-9.17608	-2.36737	-10271.9	-0.83326	-1.08968	1.97125	2.058128	-0.00333	-0.01101	0.014881	-0.00529	-0.01386	0.049202	0.064322	-0.02518
17	-16.5642	-74705.2	55.17783	-2.1516	7.109846	5462.295	-2.22434	-2.01841	-3.13774	6.623469	-0.01658	-0.00752	-0.01558	-0.0052	-0.06755	0.064322	0.227474	0.047656
18	-13.7813	15298.09	653.0513	36.6199	29.02755	67782.66	-5.04259	-1.50427	-7.95323	3.626633	-0.01211	0.026336	-0.05315	0.011616	-0.06618	-0.02518	0.047656	0.270784
19	5.305991	64300.31	330.7791	23.55709	10.68092	34740.29	-1.29929	0.678254	-1.43972	-1.37643	0.004646	0.021686	-0.02053	0.011505	0.010977	-0.05805	-0.07282	0.116409
20	-21.2042	-63426.8	355.1146	16.86363	21.02465	36734.78	-3.28658	-2.16518	-10.5673	5.402716	-0.01652	0.021607	-0.03335	0.001317	-0.08629	0.023781	0.113361	0.177016
21	-25.8957	-119870	65.41943	-3.75763	11.04546	6364.119	-2.50299	-2.8738	-7.4308	7.846013	-0.02434	-0.00415	-0.02057	-0.00839	-0.10253	0.086409	0.229284	0.072903
22	-8.45195	-14717.9	-32.3838	-1.11861	-1.55636	-2282.38	3.683762	2.786478	-4.54679	-1.6621	0.004642	0.0021	0.001372	-0.00022	0.004337	-0.00723	-0.01479	-0.01936
23	-14.2211	-37674.9	218.948	15.50834	13.01395	22864.85	-1.98355	-0.96	-5.96676	2.390331	-0.00405	0.02383	-0.02827	0.004643	-0.04488	-0.01687	0.021824	0.147443

19	20	21	22	23
5.305991	-21.2042	-25.8957	-8.45195	-14.2211
64300.31	-63426.8	-119870	-14717.9	-37674.9
330.7791	355.1146	65.41943	-32.3838	218.948
23.55709	16.86363	-3.75763	-1.11861	15.50834
10.68092	21.02465	11.04546	-1.55636	13.01395
34740.29	36734.78	6364.119	-2282.38	22864.85
-1.29929	-3.28658	-2.50299	3.683762	-1.98355
0.678254	-2.16518	-2.8738	2.786478	-0.96
-1.43972	-10.5673	-7.4308	-4.54679	-5.96676
-1.37643	5.402716	7.846013	-1.6621	2.390331
0.004646	-0.01652	-0.02434	0.004642	-0.00405
0.021686	0.021607	-0.00415	0.0021	0.02383
-0.02053	-0.03335	-0.02057	0.001372	-0.02827
0.011505	0.001317	-0.00839	-0.00022	0.004643
0.010977	-0.08629	-0.10253	0.004337	-0.04488
-0.05805	0.023781	0.086409	-0.00723	-0.01687
-0.07282	0.113361	0.229284	-0.01479	0.021824
0.116409	0.177016	0.072903	-0.01936	0.147443
0.118643	0.017363	-0.10068	-0.0004	0.064663
0.017363	0.177852	0.168634	-0.01723	0.102501
-0.10068	0.168634	0.301511	-0.01872	0.041203
-0.0004	-0.01723	-0.01872	0.024525	-0.00898
0.064663	0.102501	0.041203	-0.00898	0.102273

Figure 9: Covariance matrix for class 1

Inferences:

1. The accuracy of the bayes classifier is come out to be greater than the KNN-classifier without normalized data but is little lesser than the KNN-classifier with normalized data.
2. The diagonals of covariance matrix represent the variances of the attributes.
3. For class = 1, maximum covariance (mag.) = 3.3×10^{10}, minimum covariance(mag.) = 6.45×10^{-5}

   For class = 0, maximum covariance (mag.) = 4.9×10^{10}, minimum covariance (mag.) = 0.0017

   Maximum = Sum_Of_Luminousity and Y_Maximum, Minimum = Outside_X_Index and Edges_Index.

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.61
2.	KNN on normalized data	97.33
3.	Bayes	94.36

Inferences:

1. **Highest** = KNN on normalized data, **Lowest** = KNN on unnormalized data.
2. On basis of classification accuracy: KNN on normalized > Bayes > KNN on unnormalized.
3. Since in KNN classifier we find the Euclidian distance for test tuple with every training tuple so it would give give best results if the ranges of all the attributes are same i.e., normalized data, it is observed that bayes classifier is also giving good accuracy. It indicates that the data is somewhat having gaussian distribution.