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: Amit Maindola **Mobile No: +91 7470985613**

Roll Number: B20079 **Branch: Computer Science & Engineering**

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	**86.944**
3	**89.614**
4	**89.318**

Inferences:

1. The highest classification accuracy is obtained with K = 3.
2. Increasing the value of K first significantly increases the accuracy , then accuracy become almost constant.
3. As we are increasing the value of K then we are predicting the value of K on the basis of more data, so accuracy increase.
4. As accuracy increase the number of diagonal elements increase since diagonal elements consist of True Negative and True Positive.
5. As diagonal elements are rightly predicted values the diagonal elements increases on increasing accuracy.
6. As accuracy increase the number of off-diagonal elements decreases since off-diagonal elements consist of False Negative and False Positive.
7. As off-diagonal elements are wrongly predicted values the diagonal elements decreases on increasing accuracy.

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	112	6
	4	215

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

	Prediction Outcome	
True	112	6

Label	3	216

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	**96.142**
3	**97.033**
5	**97.329**

Inferences:

1. Data normalisation increased the accuracy of prediction.
2. The accuracy is increased on data normalisation since after data normalisation all the attributes are.
3. The highest classification accuracy is obtained with K = 3.
4. Increasing the value of K significantly increases the accuracy.
5. As we are increasing the value of K then we are predicting the value of K on the basis of more data, so accuracy increase.
6. As accuracy increase the number of diagonal elements increase since diagonal elements consist of True Negative and True Positive.
7. As diagonal elements are rightly predicted values the diagonal elements increases on increasing accuracy.
8. As accuracy increase the number of off-diagonal elements decreases since off-diagonal elements consist of False Negative and False Positive.
9. As off-diagonal elements are wrongly predicted values the diagonal elements decreases on increasing accuracy.

3

	Prediction Outcome

True Label		
	100	200
	300	600

Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 94.362%.

Table 3 Mean for class 0 and class 1

S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Minimum	124.388	695.349
2.	X_Maximum	273.418	723.656
3.	Y_Minimum	1583013.432	1431553.131
4.	Y_Maximum	1583169.659	1431588.690
5.	Pixels_Areas	7779.663	585.967
6.	X_Perimeter	393.835	54.491
7.	Y_Perimeter	273.183	45.658
8.	Sum_of_Luminosity	843350.275	62191.126
9.	Minimum_of_Luminosity	53.326	96.236
10.	Maximum_of_Luminosity	135.762	130.452
11.	Length_of_Conveyer	1382.762	1480.018
12.	TypeOfSteel_A300	0.0	0.365
13.	TypeOfSteel_A400	1.0	0.635
14.	Steel_Plate_Thickness	40.073	104.214
15.	Edges_Index	0.123	0.385
16.	Empty_Index	0.459	0.426
17.	Square_Index	0.592	0.512
18.	Outside_X_Index	0.108	0.020
19.	Edges_X_Index	0.550	0.608
20.	Edges_Y_Index	0.523	0.831
21.	Outside_Global_Index	0.288	0.608
22.	LogOfAreas	3.623	2.287
23.	Log_X_Index	2.057	1.227

24.	Log_Y_Index	1.848	1.318
25.	Orientation_Index	-0.314	0.137
26.	Luminosity_Index	-0.115	-0.116
27.	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	46733.8	-6.1E+07	-320672	-15750.5	-12943.8	-3.3E+07	3686.07	2040.9	1237.64	16.734	25.3602	-6.9293	4.69619	-1.51587	16.6535	22.5046	30.839	-76.3196	-47.7816	-31.1473	27.6788	18.0829	-30.0931
2	-6.1E+07	1.8E+12	1E+09	8.3E+07	1.6E+08	4.9E+10	-5669890	-6007837	-7505510	-114611	-47711.4	21948.3	-59251.3	4294.74	-19165.6	-35306.4	-86404.1	168070	111448	73014.4	-82046.9	-50711.2	73811.6
3	-320672	1E+09	1E+08	6692649	1E+07	9E+09	-154934	6294.46	10070.2	547.01	-492.113	585.231	200.195	223.056	-1121.19	-354.573	556.075	3456.88	1427.03	2840.74	980.333	-300.211	575.04
4	-15750.5	8.3E+07	6692649	442771	706257	5.6E+08	-7764.05	769.586	771.604	31.9239	-24.0928	38.1611	10.5958	10.9942	-67.8237	-13.284	45.3417	183.057	68.4117	169.129	72.4357	-15.7026	28.5211
5	-12943.8	1.6E+08	1E+07	706257	1206391	8.1E+08	-6894.47	1492.07	-1364.2	10.2071	-17.5711	44.1824	-16.5502	6.49598	-65.4173	13.4106	63.2505	176.64	44.0548	207.792	105.12	-21.062	19.5057
6	-3.3E+07	4.9E+10	9E+09	5.6E+08	8.1E+08	8.2E+11	-1.6E+07	777671	2214134	49759.9	-53267.3	58474.6	44601.8	25470.5	-123181	-50984.9	60033.1	361545	157341	278177	96509.5	-22290.5	62063.3
7	3686.07	-5669890	-154934	-7764.05	-6894.47	-1.6E+07	1458.21	439.236	-153.834	-1.9725	3.93151	-1.75004	1.07774	-1.45529	3.73884	4.62332	4.75885	-22.1867	-12.8607	-10.7472	3.81665	4.44827	-6.55741
8	2040.9	-6007837	6294.46	769.586	1492.07	777671	439.236	333.381	2.28501	-0.79132	1.76868	-0.22159	2.0577	-0.35296	-0.14245	1.57515	4.20658	-5.85939	-4.35841	-1.52924	4.13638	2.71617	-2.7371
9	1237.64	-7505510	10070.2	771.604	-1364.2	2214134	-153.834	2.28501	2521.56	-1.82073	1.32196	0.80637	3.92598	-0.19247	-2.69665	-0.53421	4.53563	2.03005	-0.00187	2.64493	4.36984	-0.4847	0.21099
10	16.734	-114611	547.01	31.9239	10.2071	49759.9	-1.9725	-0.79132	-1.82073	0.72991	-0.00874	0.0147	-0.01549	0.01905	0.00318	-0.01538	-0.02114	0.0411	0.04137	0.01927	-0.02246	-0.0077	0.00548
11	25.3602	-47711.4	-492.113	-24.0928	-17.5711	-53267.3	3.93151	1.76868	1.32196	-0.00874	0.02932	-0.00928	0.00715	-0.00605	0.01469	0.02242	0.02636	-0.08402	-0.05352	-0.03759	0.0243	0.01598	-0.02755
12	-6.9293	21948.3	585.231	38.1611	44.1824	58474.6	-1.75004	-0.22159	0.80637	0.0147	-0.00928	0.0153	0.00472	0.00494	-0.01766	-0.0116	0.00302	0.05167	0.03041	0.03616	0.00516	-0.00347	0.01527
13	4.69619	-59251.3	200.195	10.5958	-16.5502	44601.8	1.07774	2.0577	3.92598	-0.01549	0.00715	0.00472	0.06449	-0.00411	-0.03633	-0.00065	0.0703	0.00133	-0.01967	0.02319	0.06865	0.01634	-0.0097
14	-1.51587	4294.74	223.056	10.9942	6.49598	25470.5	-1.45529	-0.35296	-0.19247	0.01905	-0.00605	0.00494	-0.00411	0.00474	-0.00222	-0.00731	-0.00975	0.02915	0.02089	0.01388	-0.00952	-0.00376	0.00748
15	16.6535	-19165.6	-1121.19	-67.8237	-65.4173	-123181	3.73884	-0.14245	-2.69665	0.00318	0.01469	-0.01766	-0.03633	-0.00222	0.05691	0.02285	-0.03856	-0.09841	-0.03926	-0.07308	-0.04451	0.00278	-0.02567
16	22.5046	-35306.4	-354.573	-13.284	13.4106	-50984.9	4.62332	1.57515	-0.53421	-0.01538	0.02242	-0.0116	-0.00065	-0.00731	0.02285	0.03068	0.02494	-0.09928	-0.0626	-0.04465	0.02302	0.01438	-0.0311
17	30.839	-86404.1	556.075	45.3417	63.2505	60033.1	4.75885	4.20658	4.53563	-0.02114	0.02636	0.00302	0.0703	-0.00975	-0.03856	0.02494	0.20286	-0.05783	-0.07275	0.01926	0.13807	0.03302	-0.03252
18	-76.3196	168070	3456.88	183.057	176.64	361545	-22.1867	-5.85939	2.03005	0.0411	-0.08402	0.05167	0.00133	0.02915	-0.09841	-0.09928	-0.05783	0.47146	0.2669	0.2469	-0.04394	-0.06701	0.13522
19	-47.7816	111448	1427.03	68.4117	44.0548	157341	-12.8607	-4.35841	-0.00187	0.04137	-0.05352	0.03041	-0.01967	0.02089	-0.03926	-0.0626	-0.07275	0.2669	0.16787	0.12411	-0.06631	-0.04408	0.08164
20	-31.1473	73014.4	2840.74	169.129	207.792	278177	-10.7472	-1.52924	2.64493	0.01927	-0.03759	0.03616	0.02319	0.01388	-0.07308	-0.04465	0.01926	0.2469	0.12411	0.15685	0.02918	-0.02546	0.06457
21	27.6788	-82046.9	980.333	72.4357	105.12	96509.5	3.81665	4.13638	4.36984	-0.02246	0.0243	0.00516	0.06865	-0.00952	-0.04451	0.02302	0.13807	-0.04394	-0.06631	0.02918	0.13317	0.0309	-0.02766
22	18.0829	-50711.2	-300.211	-15.7026	-21.062	-22290.5	4.44827	2.71617	-0.4847	-0.0077	0.01598	-0.00347	0.01634	-0.00376	0.00278	0.01438	0.03302	-0.06701	-0.04408	-0.02546	0.0309	0.02744	-0.02644
23	-30.0931	73811.6	575.04	28.5211	19.5057	62063.3	-6.55741	-2.7371	0.21099	0.00548	-0.02755	0.01527	-0.0097	0.00748	-0.02567	-0.0311	-0.03252	0.13522	0.08164	0.06457	-0.02766	-0.02644	0.04932

Fig. 8: Covariance matrix for class 0

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	256526	1.1E+08	-22254.6	1101.08	-1973.56	-2334976	-1224.81	-744.043	13220.1	-1932.62	8.91392	-3.8064	10.8927	1.50433	6.69479	-5.01836	-16.5642	-13.7813	5.30599	-21.2042	-25.8957	-8.45195	-14.2211
2	1.1E+08	3.1E+12	3.2E+08	2E+07	4659662	3.3E+10	-3631825	-43295.9	3999506	-3.6E+07	23556.3	-19251	-38009.7	13457.3	64533	-22198.8	-74705.2	15298.1	64300.3	-63426.8	-119870	-14717.9	-37674.9
3	-22254.6	3.2E+08	4714217	178492	129451	4.9E+08	-15632	-300.304	-23834.7	4262.21	-47.6455	35.6195	-90.6336	52.9086	-101.643	-96.0566	55.1778	653.051	330.779	355.115	65.4194	-32.3838	218.948
4	1101.08	2E+07	178492	9807.2	5546.9	1.9E+07	-570.116	30.1497	-1446.88	282.113	-1.33167	4.1556	-7.3181	3.9719	-4.84985	-9.17608	-2.1516	36.6199	23.5571	16.8636	-3.75763	-1.11861	15.5083
5	-1973.56	4659662	129451	5546.9	5000.65	1.3E+07	-557.423	-79.1464	-1139.31	438.56	-2.24421	2.95169	-6.49605	1.20447	-8.61151	-2.36737	7.10985	29.0276	10.6809	21.0247	11.0455	-1.55636	13.014
6	-2334976	3.3E+10	4.9E+08	1.9E+07	1.3E+07	5.1E+10	-1463161	84723	-2735155	343512	-4688.9	3985.08	-9652.58	5577.97	-10534.6	-10271.9	5462.3	67782.7	34740.3	36734.8	6364.12	-2282.38	22864.8
7	-1224.81	-3631825	-15632	-570.116	-557.423	-1463161	733.909	348.045	-993.311	-204.836	1.06637	0.59107	0.77518	-0.15145	0.42721	-0.83326	-2.22434	-5.04259	-1.29929	-3.28658	-2.50299	3.68376	-1.98355
8	-744.043	-43295.9	-300.304	30.1497	-79.1464	84723	348.045	406.461	-381.093	-205.394	0.42912	-0.02454	-0.26703	0.04392	0.87757	-1.08968	-2.01841	-1.50427	0.67825	-2.16518	-2.8738	2.78648	-0.96
9	13220.1	3999506	-23834.7	-1446.88	-1139.31	-2735155	-993.311	-381.093	23100.8	1243.44	-0.09047	-5.15952	2.46817	-0.69776	6.59105	1.97125	-3.13774	-7.95323	-1.43972	-10.5673	-7.4308	-4.54679	-5.96676
10	-1932.62	-3.6E+07	4262.21	282.113	438.56	343512	-204.836	-205.394	1243.44	5645.31	-1.3306	0.69919	-1.13384	-0.16545	-3.44259	2.05813	6.62347	3.62663	-1.37643	5.40272	7.84601	-1.6621	2.39033
11	8.91392	23556.3	-47.6455	-1.33167	-2.24421	-4688.9	1.06637	0.42912	-0.09047	-1.3306	0.08965	-0.00063	0.01093	6.5E-05	0.0083	-0.00333	-0.01658	-0.01211	0.00465	-0.01652	-0.02434	0.00464	-0.00405
12	-3.8064	-19251	35.6195	4.1556	2.95169	3985.08	0.59107	-0.02454	-5.15952	0.69919	-0.00063	0.02028	-0.00202	0.00124	-0.01249	-0.01101	-0.00752	0.02634	0.02169	0.02161	-0.00415	0.0021	0.02383
13	10.8927	-38009.7	-90.6336	-7.3181	-6.49605	-9652.58	0.77518	-0.26703	2.46817	-1.13384	0.01093	-0.00202	0.08237	-0.00291	0.01974	0.01488	-0.01558	-0.05315	-0.02053	-0.03335	-0.02057	0.00137	-0.02827
14	1.50433	13457.3	52.9086	3.9719	1.20447	5577.97	-0.15145	0.04392	-0.69776	-0.16545	6.5E-05	0.00124	-0.00291	0.00247	0.00175	-0.00529	-0.0052	0.01162	0.0115	0.00132	-0.00839	-0.00022	0.00464
15	6.69479	64533	-101.643	-4.84985	-8.61151	-10534.6	0.42721	0.87757	6.59105	-3.44259	0.0083	-0.01249	0.01974	0.00175	0.06507	-0.01386	-0.06755	-0.06618	0.01098	-0.08629	-0.10253	0.00434	-0.04488
16	-5.01836	-22198.8	-96.0566	-9.17608	-2.36737	-10271.9	-0.83326	-1.08968	1.97125	2.05813	-0.00333	-0.01101	0.01488	-0.00529	-0.01386	0.0492	0.06432	-0.02518	-0.05805	0.02378	0.08641	-0.00723	-0.01687
17	-16.5642	-74705.2	55.1778	-2.1516	7.10985	5462.3	-2.22434	-2.01841	-3.13774	6.62347	-0.01658	-0.00752	-0.01558	-0.0052	-0.06755	0.06432	0.22747	0.04766	-0.07282	0.11336	0.22928	-0.01479	0.02182
18	-13.7813	15298.1	653.051	36.6199	29.0276	67782.7	-5.04259	-1.50427	-7.95323	3.62663	-0.01211	0.02634	-0.05315	0.01162	-0.06618	-0.02518	0.04766	0.27078	0.11641	0.17702	0.0729	-0.01936	0.14744
19	5.30599	64300.3	330.779	23.5571	10.6809	34740.3	-1.29929	0.67825	-1.43972	-1.37643	0.00465	0.02169	-0.02053	0.0115	0.01098	-0.05805	-0.07282	0.11641	0.11864	0.01736	-0.10068	-0.0004	0.06466
20	-21.2042	-63426.8	355.115	16.8636	21.0247	36734.8	-3.28658	-2.16518	-10.5673	5.40272	-0.01652	0.02161	-0.03335	0.00132	-0.08629	0.02378	0.11336	0.17702	0.01736	0.17785	0.16863	-0.01723	0.1025
21	-25.8957	-119870	65.4194	-3.75763	11.0455	6364.12	-2.50299	-2.8738	-7.4308	7.84601	-0.02434	-0.00415	-0.02057	-0.00839	-0.10253	0.08641	0.22928	0.0729	-0.10068	0.16863	0.30151	-0.01872	0.0412
22	-8.45195	-14717.9	-32.3838	-1.11861	-1.55636	-2282.38	3.68376	2.78648	-4.54679	-1.6621	0.00464	0.0021	0.00137	-0.00022	0.00434	-0.00723	-0.01479	-0.01936	-0.0004	-0.01723	-0.01872	0.02452	-0.00898
23	-14.2211	-37674.9	218.948	15.5083	13.014	22864.8	-1.98355	-0.96	-5.96676	2.39033	-0.00405	0.02383	-0.02827	0.00464	-0.04488	-0.01687	0.02182	0.14744	0.06466	0.1025	0.0412	-0.00898	0.10227

Fig. 9: Covariance matrix for class 1

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.

Inferences:

1. Accuracy of Bayes classifier = 94.362, this is less than that of KNN classifier with normalization. The reason for this is that the we assumed that our data is coming from a normal or gaussian distribution which is not true for our case.
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, one 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. Greater the value of covariance between 2 attributes, greater is the joint variability of the two variables.
   2 attributes with maximum covariance are: **For class 0**: Y_Minimum and Y_Maximum, F**or class 1**: Y_Minimum and Y_Maximum.
   2 attributes with minimum covariance are: **For class 0**: TypeOfSteel_A300 and TypeOfSteel_A400, **For class 1**: Empty_Index and Steel_Plate_Thickness.

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.614
2.	KNN on normalized data	97.329
3.	Bayes	94.362

Inferences:

1. **Maximum accuracy**: KNN classifier with normalized data, **Minimum accuracy:** KNN classifier without normalization.
2. **Accuracy:** KNN classifier without normalization < Bayes Classifi <KNN classifier with normalized data.
3. KNN performs better with a lower number of features than a large number of features. You can say that when the number of features increases than it requires more data. So, in normalized data all data values lie within 0 and 1, so there will be less spread of data. Thus, the KNN on normalized data performs better than other KNNs. Solving a problem Bayes directly focusses on finding similarity between observations, K-NN does better because of its inherent nature to optimize locally according to the locations.