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: AKSHAR SINGH   **Mobile No:** 7428357700

Roll Number: B20147   **Branch:** Computer Science

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
5	89.318

Inferences:
1. The highest classification accuracy is obtained with K = 5.
2. The increasing value of K increases the prediction accuracy.
3. As the value of K increases as we take in account more number of neighbors hence its accuracy increasing.
4. As there is an increase in the value of K, the number of diagonal elements increases.
5. There is an increase in the diagonal elements because they represent the number of true values.
6. The number of off-diagonal elements decrease as the classification accuracy decreases.
7. There is a decrease in the off-diagonal elements because they represent the number of false values.

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 Label	112	6
	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	97.330

Inferences:

1. The data normalization increases the classification accuracy
2. There is increase in accuracy because through normalization ,no attribute outweighs the other in the calculation of Euclidean distance.
3. The highest classification accuracy is obtained with K =5.
4. The increasing value of K increases the prediction accuracy.
5. As the value of K increases as we take in account more number of neighbors hence its accuracy increasing.
6. As there is an increase in the value of K, the number of diagonal elements increases.
7. There is an increase in the diagonal elements because they represent the number of true values.
8. As the classification accuracy increases/decreases with the increase in value of K infer does the number of off-diagonal elements increase/decrease.
9. The number of off-diagonal elements decrease as the classification accuracy decreases.
10. There is a decrease in the off-diagonal elements because they represent the number of false values.

3

	Prediction Outcome	
True Label	109	9
	9	210

Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is   94.659%.

Table 3 Mean for class 0 and class 1

S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Minimum	-	-
2.	X_Maximum	273.418	723.656
3.	Y_Minimum	-	-
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	-	-
13.	TypeOfSteel_A400	-	-
14.	Steel_Plate_Thickness	40.073	104.214
15.	Edges_Index	0.123	0.385
16.	Empty_Index	0.459	0.427
17.	Square_Index	0.592	0.513
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	0.136
26.	Luminosity_Index	-0.115	-0.116
27.	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.
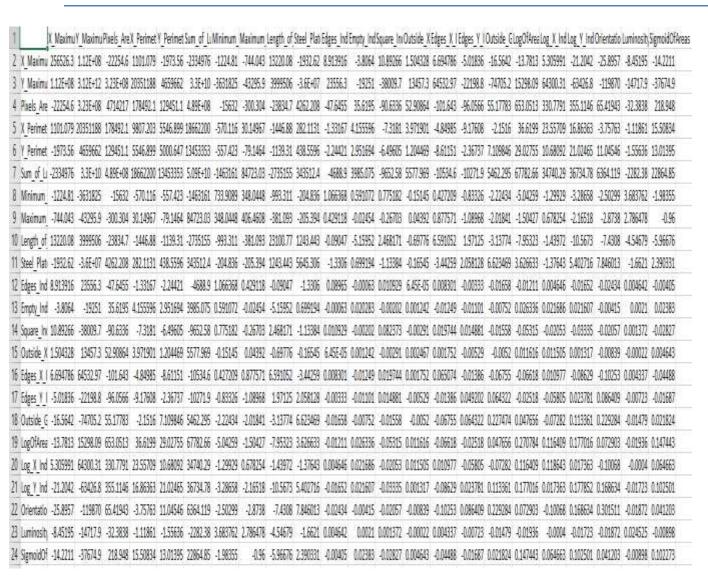
Covariance matrix for class0

	X_Maximi	Y_Maximi	Pixels_Are	X_Perimet	Y_Perimet	Sum_of_L	Minimum	Maximum	Length_of	Steel_Plat	Edges_Inc	Empty_Inc	Square_In	Outside_X	Edges_X_	Edges_Y_	Outside_(	LogOfAre	Log_X_Inc	Log_Y_Inc	Orientatio	Luminosit	SigmoidOfAreas
X_Maximi	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
Y_Maximi	-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
Pixels_Are	-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
X_Perimet	-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
Y_Perimet	-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
Sum_of_L	-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
Minimum	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
Maximum	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
Length_of	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
Steel_Plat	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
Edges_Inc	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
Empty_Inc	-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
Square_In	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
Outside_X	-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
Edges_X	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
Edges_Y	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
Outside_(	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
LogOfAre	-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
Log_X_Inc	-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
Log_Y_Inc	-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
Orientatio	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
Luminosit	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
SigmoidOf	-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

Covariance matrix for class1

	X_Maximu	Y_Maximu	Pixels_Are	X_Perimet	Y_Perimet	Sum_of_L	Minimum_	Maximum_	Length_of	Steel_Plat	Edges_Ind	Empty_Ind	Square_Inv	Outside_X	Edges_X_I	Edges_Y_I	Outside_G	LogOfArea	Log_X_Ind	Log_Y_Ind	Orientatio	Luminosit	SigmoidOfAreas
X_Maximu	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	5.305991	-21.2042	-25.8957	-8.45195	-14.2211
Y_Maximu	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
Pixels_Are	-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	330.7791	355.1146	65.41943	-32.3838	218.948
X_Perimet	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	23.55709	16.86363	-3.75763	-1.11861	15.50634
Y_Perimet	-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	10.60892	21.02465	11.04546	-1.55636	13.01395
Sum_of_Li	-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
Minimum_	-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	-1.29929	-3.28658	-2.50299	3.683762	-1.98355
Maximum_	-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	0.678254	-2.16518	-2.8738	2.786478	-0.96
Length_of	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	-1.43972	-10.5673	-7.4308	-4.54679	-5.96676
Steel_Plat	-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	-1.37643	5.402716	7.846013	-1.6621	2.390331
Edges_Ind	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	0.004646	-0.01652	-0.02434	0.004642	-0.00405
Empty_Ind	-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	0.021686	0.021607	-0.00415	0.0021	0.02383
Square_Inv	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	-0.02053	-0.03335	-0.02057	0.001372	-0.02827
Outside_X	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	0.011505	0.001317	-0.00839	-0.00022	0.004643
Edges_X_I	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	0.010977	-0.08629	-0.10253	0.004337	-0.04488
Edges_Y_I	-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	-0.05805	0.023781	0.086409	-0.00723	-0.01687
Outside_G	-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	-0.07282	0.113361	0.229284	-0.01479	0.021824
LogOfArea	-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	0.116409	0.177016	0.072903	-0.01936	0.147443
Log_X_Ind	5.305991	64300.31	330.7791	23.55709	10.60892	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	0.118643	0.017363	-0.10068	-0.0004	0.064663
Log_Y_Ind	-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	0.017363	0.177852	0.168634	-0.01723	0.102501
Orientatio	-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	-0.10068	0.168634	0.301511	-0.01872	0.041203
Luminosit	-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	-0.0004	-0.01723	-0.01872	0.024525	-0.00898
SigmoidOf	-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	0.064663	0.102501	0.041203	-0.00898	0.102273

Inferences:

1. The accuracy of Bayes Classifier is 94.362 and it is lesser than previous classification approach because it assumes normal distribution.
2. The nature of values along the diagonal is high for some attributes and low for some other attributes mainly because some attributes follow standard normal distribution.
3. The off-diagonal elements have varied values. The 2 pair of attribute with maximum covariance is (Sum_of_Luminosity,Y_Maximum) and (Pixels_Area,Sum_of_Luminosity) , the 2 pair of attribute with minimum covariance is (Outside_X, Empty_Index) and (Outside_X,Edges_X).

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.318
2.	KNN on normalized data	97.330
3.	Bayes	94.362

Inferences:

1. The classifier with highest accuracy is KNN on normalized data and lowest accuracy is KNN.
2. The classifiers in ascending order of classification accuracy. Classifier 1 < Classifier 3 < Classifier 2.
3. The reason for low accuracy in KNN classifier is due to the calculation of Euclidian distance and because of that one attribute outweighs the other. The reason for low accuracy of Bayes classifier than KNN on normalized data is because it assumes normal distribution.