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: Gaurav Kumar   **Mobile No: 8529143452**

Roll Number: B20197   **Branch:EE**

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	86.943
3	89.614
4	89.318

Inferences:

1. The highest classification accuracy is obtained with K = 3.
2. Increasing the value of K decreases the prediction accuracy.

3. By increasing the value of K we are actually considering the Euclidian distances of K points from the test point that's why we get low accuracy on increasing the K value.
4. As the classification accuracy decreases with the increase in value of K infer the number of diagonal element decreases.
5. As accuracy is percentage of sum of diagonal element and total no. of test sample. So, as accuracy decreases diagonal element also decreases.
6. As the classification accuracy decreases with the increase in value of K infer the number of off-diagonal elements increase.
7. Due to decrease in accuracy as K increases cause increase in off-diagonal element.

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. Data normalization increases classification accuracy.
2. Due to normalization, we are getting true Euclidian distance. Hence, K-NN classification becomes more accurate.
3. The highest classification accuracy is obtained with K =5.
4. Increasing the value of k actually decreases the prediction accuracy for large value of K. But, for lower value of k, increasing the value of k cause increase in classification accuracy.
5.
6. State a suitable reason why increasing the value of K increases/decreases the prediction accuracy.
7. As the classification accuracy increases/decreases with the increase in value of K infer does the number of diagonal elements increase/decrease.
8. State the reason for increase/decrease in diagonal elements.
9. As the classification accuracy increases/decreases with the increase in value of K infer does the number of off-diagonal elements increase/decrease.
10. State the reason for increase/decrease in off-diagonal elements.

3

	Prediction Outcome	
True Label	102	3
	16	214

Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is   94.362%.

4

LAB ASSIGNMENT – IV
Data classification using K-nearest neighbor classifier and Bayes classifier with unimodal Gaussian density

Table 3 Mean for class 0 and class 1

S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Minimum	124.388	695.350
2.	X_Maximum	273.418	723.656
3.	Y_Minimum	1583013.432	1431553.132
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	0.540

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.

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

	X_Maximu	Y_Maximu	Pixels_Are	X_Perimet	Y_Perimet	Sum_of_L	Minimum	Maximum	Length_of	Steel_Plat	Edges_Ind	Empty_Ind	Square_In	Outside_X	Edges_X_I	Edges_Y_I	Outside_C	LogOfArea	Log_X_Inc	Log_Y_Ind	Orientatic	Luminosit	SigmoidOfA
X_Maximu	46733.8	-6E+07	-320672	-15751	-12944	-3E+07	3686.07	2040.9	1237.64	16.734	25.3602	-6.9293	4.69619	-1.5159	16.6535	22.5046	30.839	-76.32	-47.782	-31.147	27.6788	18.0829	-30.093
Y_Maximu	-6E+07	1.8E+12	1E+09	8.3E+07	1.6E+08	4.9E+10	-6E+06	-6E+06	-8E+06	-114611	-47711	21948.3	-59251	4294.74	-19166	-35306	-86404	168070	111448	73014.4	-82047	-50711	73811.6
Pixels_Are	-320672	1E+09	1E+08	6692649	1E+07	9E+09	-154934	6294.46	10070.2	547.01	-492.11	585.231	200.195	223.056	-1121.2	-354.57	556.075	3456.88	1427.03	2840.74	980.333	-300.21	575.04
X_Perimet	-15751	8.3E+07	6692649	442771	706257	5.6E+08	-7764	769.586	771.604	31.9239	-24.093	38.1611	10.5958	10.9942	-67.824	-13.284	45.3417	183.057	68.4117	169.129	72.4357	-15.703	28.5211
Y_Perimet	-12944	1.6E+08	1E+07	706257	1206391	8.1E+08	-6894.5	1492.07	-1364.2	10.2071	-17.571	44.1824	-16.55	6.49598	-65.417	13.4106	63.2505	176.64	44.0548	207.792	105.12	-21.062	19.5057
Sum_of_L	-3E+07	4.9E+10	9E+09	5.6E+08	8.1E+08	8.2E+11	-2E+07	777671	2214134	49759.9	-53267	58474.6	44601.8	25470.5	-123181	-50985	60033.1	361545	157341	278177	96509.5	-22291	62063.3
Minimum	3686.07	-6E+06	-154934	-7764	-6894.5	-2E+07	1458.21	439.236	-153.83	-1.9725	3.93151	-1.75	1.07774	-1.4553	3.73884	4.62332	4.75885	-22.187	-12.861	-10.747	3.81665	4.44827	-6.5574
Maximum	2040.9	-6E+06	6294.46	769.586	1492.07	777671	439.236	333.381	2.28501	-0.7913	1.76868	-0.2216	2.0577	-0.353	-0.1424	1.57515	4.20658	-5.8594	-4.3584	-1.5292	4.13638	2.71617	-2.7371
Length_of	1237.64	-8E+06	10070.2	771.604	-1364.2	2214134	-153.83	2.28501	2521.56	-1.8207	1.32196	0.80637	3.92598	-0.1925	-2.6967	-0.5342	4.53563	2.03005	-0.0019	2.64493	4.36984	-0.4847	0.21099
Steel_Plat	16.734	-114611	547.01	31.9239	10.2071	49759.9	-1.9725	-0.7913	-1.8207	0.72991	-0.0087	0.0147	-0.0155	0.01905	0.00318	-0.0154	-0.0211	0.0411	0.04137	0.01927	-0.0225	-0.0077	0.00548
Edges_Ind	25.3602	-47711	-492.11	-24.093	-17.571	-53267	3.93151	1.76868	1.32196	-0.0087	0.02932	-0.0093	0.00715	-0.006	0.01469	0.02242	0.02636	-0.084	-0.0535	-0.0376	0.0243	0.01598	-0.0276
Empty_Ind	-6.9293	21948.3	585.231	38.1611	44.1824	58474.6	-1.75	-0.2216	0.80637	0.0147	-0.0093	0.0153	0.00472	0.00494	-0.0177	-0.0116	0.00302	0.05167	0.03041	0.03616	0.00516	-0.0035	0.01527
Square_In	4.69619	-59251	200.195	10.5958	-16.55	44601.8	1.07774	2.0577	3.92598	-0.0155	0.00715	0.00472	0.06449	-0.0041	-0.0363	-0.0007	0.0703	0.00133	-0.0197	0.02319	0.06865	0.01634	-0.0097
Outside_X	-1.5159	4294.74	223.056	10.9942	6.49598	25470.5	-1.4553	-0.353	-0.1925	0.01905	-0.006	0.00494	-0.0041	0.00474	-0.0022	-0.0073	-0.0098	0.02915	0.02089	0.01388	-0.0095	-0.0038	0.00748
Edges_X_I	16.6535	-19166	-1121.2	-67.824	-65.417	-123181	3.73884	-0.1424	-2.6967	0.00318	0.01469	-0.0177	-0.0363	-0.0022	0.05691	0.02285	-0.0386	-0.0984	-0.0393	-0.0731	-0.0445	0.00278	-0.0257
Edges_Y_I	22.5046	-35306	-354.57	-13.284	13.4106	-50985	4.62332	1.57515	-0.5342	-0.0154	0.02242	-0.0116	-0.0007	-0.0073	0.02285	0.03068	0.02494	-0.0993	-0.0626	-0.0447	0.02302	0.01438	-0.0311
Outside_C	30.839	-86404	556.075	45.3417	63.2505	60033.1	4.75885	4.20658	4.53563	-0.0211	0.02636	0.00302	0.0703	-0.0098	-0.0386	0.02494	0.20286	-0.0578	-0.0728	0.01926	0.13807	0.03302	-0.0325
LogOfArea	-76.32	168070	3456.88	183.057	176.64	361545	-22.187	-5.8594	2.03005	0.0411	-0.084	0.05167	0.00133	0.02915	-0.0984	-0.0993	-0.0578	0.47146	0.2669	0.2469	-0.0439	-0.067	0.13522
Log_X_Inc	-47.782	111448	1427.03	68.4117	44.0548	157341	-12.861	-4.3584	-0.0019	0.04137	-0.0535	0.03041	-0.0197	0.02089	-0.0393	-0.0626	-0.0728	0.2669	0.16787	0.12411	-0.0663	-0.0441	0.08164
Log_Y_Ind	-31.147	73014.4	2840.74	169.129	207.792	278177	-10.747	-1.5292	2.64493	0.01927	-0.0376	0.03616	0.02319	0.01388	-0.0731	-0.0447	0.01926	0.2469	0.12411	0.15685	0.02918	-0.0255	0.06457
Orientatic	27.6788	-82047	980.333	72.4357	105.12	96509.5	3.81665	4.13638	4.36984	-0.0225	0.0243	0.00516	0.06865	-0.0095	-0.0445	0.02302	0.13807	-0.0439	-0.0663	0.02918	0.13317	0.0309	-0.0277
Luminosit	18.0829	-50711	-300.21	-15.703	-21.062	-22291	4.44827	2.71617	-0.4847	-0.0077	0.01598	-0.0035	0.01634	-0.0038	0.00278	0.01438	0.03302	-0.067	-0.0441	-0.0255	0.0309	0.02744	-0.0264
SigmoidO	-30.093	73811.6	575.04	28.5211	19.5057	62063.3	-6.5574	-2.7371	0.21099	0.00548	-0.0276	0.01527	-0.0097	0.00748	-0.0257	-0.0311	-0.0325	0.13522	0.08164	0.06457	-0.0277	-0.0264	0.04932

Figure 8: Covariance matrix for class 0

	X_Maximu	Y_Maximu	Pixels_Are	X_Perimet	Y_Perimet	Sum_of_L	Minimum	Maximum	Length_of	Steel_Plat	Edges_Ind	Empty_Ind	Square_In	Outside_X	Edges_X_I	Edges_Y_I	Outside_C	LogOfArea	Log_X_Inc	Log_Y_Ind	Orientatic	Luminosit	SigmoidOfAreas
X_Maximu	256526	1.1E+08	-22255	1101.08	-1973.6	-2E+06	-1224.8	-744.04	13220.1	-1932.6	8.91392	-3.8064	10.8927	1.50433	6.69479	-5.0184	-16.564	-13.781	5.30599	-21.204	-25.896	-8.452	-14.221
Y_Maximu	1.1E+08	3.1E+12	3.2E+08	2E+07	4659662	3.3E+10	-4E+06	-43296	3999506	-4E+07	23556.3	-19251	-38010	13457.3	64533	-22199	-74705	15298.1	64300.3	-63427	-119870	-14718	-37675
Pixels_Are	-22255	3.2E+08	4714217	178492	129451	4.9E+08	-15632	-300.3	-23835	4262.21	-47.646	35.6195	-90.634	52.9086	-101.64	-96.057	55.1778	653.051	330.779	355.115	65.4194	-32.384	218.948
X_Perimet	1101.08	2E+07	178492	9807.2	5546.9	1.9E+07	-570.12	30.1497	-1446.9	282.113	-1.3317	4.1556	-7.3181	3.9719	-4.8499	-9.1761	-2.1516	36.6199	23.5571	16.8636	-3.7576	-1.1186	15.5083
Y_Perimet	-1973.6	4659662	129451	5546.9	5000.65	1.3E+07	-557.42	-79.146	-1139.3	438.56	-2.2442	2.95169	-6.496	1.20447	-8.6115	-2.3674	7.10985	29.0276	10.6809	21.0247	11.0455	-1.5564	13.014
Sum_of_L	-2E+06	3.3E+10	4.9E+08	1.9E+07	1.3E+07	5.1E+10	-1E+06	84723	-3E+06	343512	-4688.9	3985.08	-9652.6	5577.97	-10535	-10272	5462.3	67782.7	34740.3	36734.8	6364.12	-2282.4	22864.8
Minimum	-1224.8	-4E+06	-15632	-570.12	-557.42	-1E+06	733.909	348.045	-993.31	-204.84	1.06637	0.59107	0.77518	-0.1515	0.42721	-0.8333	-2.2243	-5.0426	-1.2993	-3.2866	-2.503	3.68376	-1.9836
Maximum	-744.04	-43296	-300.3	30.1497	-79.146	84723	348.045	406.461	-381.09	-205.39	0.42912	-0.0245	-0.267	0.04392	0.87757	-1.0897	-2.0184	-1.5043	0.67825	-2.1652	-2.8738	2.78648	-0.96
Length_of	13220.1	3999506	-23835	-1446.9	-1139.3	-3E+06	-993.31	-381.09	23100.8	1243.44	-0.0905	-5.1595	2.46817	-0.6978	6.59105	1.97125	-3.1377	-7.9532	-1.4397	-10.567	-7.4308	-4.5468	-5.9668
Steel_Plat	-1932.6	-4E+07	4262.21	282.113	438.56	343512	-204.84	-205.39	1243.44	5645.31	-1.3306	0.69919	-1.1338	-0.1655	-3.4426	2.05813	6.62347	3.62663	-1.3764	5.40272	7.84601	-1.6621	2.39033
Edges_Ind	8.91392	23556.3	-47.646	-1.3317	-2.2442	-4688.9	1.06637	0.42912	-0.0905	-1.3306	0.08965	-0.0006	0.01093	6.45E-05	0.0083	-0.0033	-0.0166	-0.0121	0.00465	-0.0165	-0.0243	0.00464	-0.004
Empty_Ind	-3.8064	-19251	35.6195	4.1556	2.95169	3985.08	0.59107	-0.0245	-5.1595	0.69919	-0.0006	0.02028	-0.002	0.00124	-0.0125	-0.011	-0.0075	0.02634	0.02169	0.02161	-0.0041	0.0021	0.02383
Square_In	10.8927	-38010	-90.634	-7.3181	-6.496	-9652.6	0.77518	-0.267	2.46817	-1.1338	0.01093	-0.002	0.08237	-0.0029	0.01974	0.01488	-0.0156	-0.0531	-0.0205	-0.0333	-0.0206	0.00137	-0.0283
Outside_X	1.50433	13457.3	52.9086	3.9719	1.20447	5577.97	-0.1515	0.04392	-0.6978	-0.1655	6.45E-05	0.00124	-0.0029	0.00247	0.00175	-0.0053	-0.0052	0.01162	0.0115	0.00132	-0.0084	-0.0002	0.00464
Edges_X_I	6.69479	64533	-101.64	-4.8499	-8.6115	-10535	0.42721	0.87757	6.59105	-3.4426	0.0083	-0.0125	0.01974	0.00175	0.06507	-0.0139	-0.0675	-0.0662	0.0109	-0.0863	-0.1025	0.00434	-0.0449
Edges_Y_I	-5.0184	-22199	-96.057	-9.1761	-2.3674	-10272	-0.8333	-1.0897	1.97125	2.05813	-0.0033	-0.011	0.01488	-0.0053	-0.0139	0.0492	0.06432	-0.0252	-0.058	0.02378	0.08641	-0.0072	-0.0169
Outside_C	-16.564	-74705	55.1778	-2.1516	7.10985	5462.3	-2.2243	-2.0184	-3.1377	6.62347	-0.0166	-0.0075	-0.0156	-0.0052	-0.0675	0.06432	0.22747	0.04766	-0.0728	0.11336	0.22928	-0.0148	0.02182
LogOfArea	-13.781	15298.1	653.051	36.6199	29.0276	67782.7	-5.0426	-1.5043	-7.9532	3.62663	-0.0121	0.02634	-0.0531	0.01162	-0.0662	-0.0252	0.04766	0.27078	0.11641	0.17702	0.0729	-0.0194	0.14744
Log_X_Inc	5.30599	64300.3	330.779	23.5571	10.6809	34740.3	-1.2993	0.67825	-1.4397	-1.3764	0.00465	0.02169	-0.0205	0.0115	0.0109	-0.058	-0.0728	0.11641	0.11864	0.01736	-0.1007	-0.0004	0.06466
Log_Y_Ind	-21.204	-63427	355.115	16.8636	21.0247	36734.8	-3.2866	-2.1652	-10.567	5.40272	-0.0165	0.02161	-0.0333	0.00132	-0.0863	0.02378	0.11336	0.17702	0.01736	0.17785	0.16863	-0.0172	0.1025
Orientatic	-25.896	-119870	65.4194	-3.7576	11.0455	6364.12	-2.503	-2.8738	-7.4308	7.84601	-0.0243	-0.0041	-0.0206	-0.0084	-0.1025	0.08641	0.22928	0.0729	-0.1007	0.16863	0.30151	-0.0187	0.0412
Luminosit	-8.452	-14718	-32.384	-1.1186	-1.5564	-2282.4	3.68376	2.78648	-4.5468	-1.6621	0.00464	0.0021	0.00137	-0.0002	0.00434	-0.0072	-0.0148	-0.0194	-0.0004	-0.0172	-0.0187	0.02452	-0.009
SigmoidO	-14.221	-37675	218.948	15.5083	13.014	22864.8	-1.9836	-0.96	-5.9668	2.39033	-0.004	0.02383	-0.0283	0.00464	-0.0449	-0.0169	0.02182	0.14744	0.06466	0.1025	0.0412	-0.009	0.10227

Figure 9: Covariance matrix for class 1

Inferences:

1. The accuracy of Bayes Classifier is 94.362% . Naive Bayes is an eager learning classifier. That is why its easy, fast and best accuracy classifier.
2. The nature of values along the diagonal for some attribute is high and for some other its low because some attribute follow standard normal distribution.

3. The off-diagonal values have varied values. The two pair of attributes having maximum covariance is (Y_Mximum, Sum of Luminosity) and (Y_Maximum,Pixel_Area). The two pair of attribute with minimum covariance is (Outside_X,Edges_X) and (Outside_X, Empty_index).

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.614
2.	KNN on normalized data	97.330
3.	Bayes	94.362

Inferences:

1. classifier with highest accuracy is Classifier KNN Bayes.

2. classifier with lowest accuracy is KNN Classifier.

3. Classifier KNN< Classifier Bayes < Classifier KNN on normalized data .