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: Vishal Sharma

Mobile No: 9540140310

Roll Number: B20239

Branch: Data Science and Engineering

1 a.

		Prediction Outcome
True Label	94	24
	57	397

Figure 1 KNN Confusion Matrix for K = 1

		Prediction Outcome
True Label	93	25
	31	423

Figure 2 KNN Confusion Matrix for K = 3

		Prediction Outcome
True Label	93	25
	31	423

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	85.839
3	90.209
5	90.209

Inferences:

1. The highest accuracy is obtained for k = 5
2. The value of accuracy increases as the value of k increases
3. For high value of k, the given point is checked for more numbers of nearest neighbors, thus decreasing the chances of noise.
4. As the classification accuracy increases, with increasing k, the value of diagonal elements also increases.
5. Increasing the value of k, helps in predicting more accurate data, thus chances for noise will be less.
6. As the classification accuracy increases, the value of off-diagonal elements decreases with increasing value of k.
7. Increasing the value of k, helps in achieving greater accuracy. Thus, chances for noise in data becomes less. Thus, chances of getting wrong values decreases.

2 a.

	Prediction Outcome	
True Label	114	4
	21	433

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

	Prediction Outcome	
True Label	114	4
	18	436

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

	Prediction Outcome	
True Label	113	5
	17	437

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	95.629
3	96.153
5	96.153

Inferences:

1. Data normalization increases the accuracy of the data.
2. Accuracy is increase after normalization, because some attributes having values too large than the other attributes, provides a dominating effect for that class. Thus, increasing the noise.
3. The highest classification accuracy is obtained with K =55.
4. Increasing the value of k, increases the accuracy.
5. For high value of k, the given point is checked for more numbers of nearest neighbors, thus decreasing the chances of noise.
6. As the classification accuracy increases, with increasing k, the value of diagonal elements also increases.
7. Increasing the value of k, helps in predicting more accurate data, thus chances for noise will be less.
8. As the classification accuracy increases, the value of off-diagonal elements decreases with increasing value of k.
9. Increasing the value of k, helps in achieving greater accuracy. Thus, chances for noise in data becomes less. Thus, chances of getting wrong values decreases.

3

		Prediction Outcome
True Label	74	44
	20	434

Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 88.811 %.

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	712.409
3.	Y_Minimum		
4.	Y_Maximum	1583169.659	1354749.591
5.	Pixels_Areas	7779.663	625.876
6.	X_Perimeter	393.835	55.175
7.	Y_Perimeter	273.183	45.522
8.	Sum_of_Luminosity	843350.275	66735.369
9.	Minimum_of_Luminosity	53.326	97.015
10.	Maximum_of_Luminosity	135.762	131.102
11.	Length_of_Conveyer	1382.762	1479.449
12.	TypeOfSteel_A300		
13.	TypeOfSteel_A400	1	0.631
14.	Steel_Plate_Thickness	40.073	104.522
15.	Edges_Index	0.123	0.396
16.	Empty_Index	0.459	0.434
17.	Square_Index	0.592	0.532
18.	Outside_X_Index	0.108	0.019
19.	Edges_X_Index	0.550	0.619
20.	Edges_Y_Index	0.523	0.825
21.	Outside_Global_Index	0.288	0.575
22.	LogOfAreas	3.623	2.269
23.	Log_X_Index	2.057	1.237

24.	Log_Y_Index	1.848	1.296
25.	Orientation_Index	-0.314	0.089
26.	Luminosity_Index	-0.115	-0.109
27.	SigmoidOfAreas	0.925	0.530

	X_Maximum	Y_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	Sum_of_Lun	Minimum_o	Maximum_c	Length_of_C	TypeOfStee	Steel_Plate_	Edges_Index	Empty_Inde	Square_Inde	Outside_X_I	Edges_X_In	Edges_Y_In	Outside_Glo	LogOfAreas	Log_X_Index	Log_Y_Index	Orientation_	Luminosity_	SigmoidOfA
X_Maximum	46734	-60848697	-320672	-15751	-12944	-32609925	3686	2041	1238	0	17	25	-7	5	-2	17	23	31	-76	-48	-31	28	18	-30
Y_Maximum	-60848697	1.822E+12	1.028E+09	83317353	160209449	4.9E+10	-5669890	-6007837	-7505510	0	-114611	-47711	21948	-59251	4295	-19166	-35306	-86404	168070	111448	73014	-82047	-50711	73812
Pixels_Areas	-320672	1.028E+09	104771843	6692649	10371695	9.008E+09	-154934	6294	10070	0	547	-492	585	200	223	-1121	-355	556	3457	1427	2841	980	-300	575
X_Perimeter	-15751	83317353	6692649	442771	706257	557116030	-7764	770	772	0	32	-24	38	11	11	-68	-13	45	183	68	169	72	-16	29
Y_Perimeter	-12944	160209449	10371695	706257	1206391	807551258	-6894	1492	-1364	0	10	-18	44	-17	6	-65	13	63	177	44	208	105	-21	20
Sum_of_Lun	-32609925	4.9E+10	9.008E+09	557116030	807551258	8.193E+11	-16498428	777671	2214134	0	49760	-53267	58475	44602	25471	-123181	-50985	60033	361545	157341	278177	96509	-22291	62063
Minimum_o	3686	-5669890	-154934	-7764	-6894	-16498428	1458	439	-154	0	-2	4	-2	1	-1	4	5	5	-22	-13	-11	4	4	-7
Maximum_c	2041	-6007837	6294	770	1492	777671	439	333	2	0	-1	2	0	2	0	0	2	4	-6	-4	-2	4	3	-3
Length_of_C	1238	-7505510	10070	772	-1364	2214134	-154	2	2522	0	-2	1	1	4	0	-3	-1	5	2	0	3	4	0	0
TypeOfStee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Steel_Plate_	17	-114611	547	32	10	49760	-2	-1	-2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_Index	25	-47711	-492	-24	-18	-53267	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Empty_Inde	-7	21948	585	38	44	58475	-2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Square_Inde	5	-59251	200	11	-17	44602	1	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Outside_X_I	-2	4295	223	11	6	25471	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_X_In	17	-19166	-1121	-68	-65	-123181	4	0	-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_Y_In	23	-35306	-355	-13	13	-50985	5	2	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Outside_Glo	31	-86404	556	45	63	60033	5	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LogOfAreas	-76	168070	3457	183	177	361545	-22	-6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Log_X_Index	-48	111448	1427	68	44	157341	-13	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Log_Y_Index	-31	73014	2841	169	208	278177	-11	-2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Orientation_	28	-82047	980	72	105	96509	4	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Luminosity_	18	-50711	-300	-16	-21	-22291	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SigmoidOfA	-30	73812	575	29	20	62063	-7	-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Class	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 8: Covariance matrix for class 0

	X_Maximum	Y_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	Sum_of_Lun	Minimum_o	Maximum_c	Length_of_C	TypeOfStee	Steel_Plate_	Edges_Index	Empty_Inde	Square_Inde	Outside_X_I	Edges_X_In	Edges_Y_In	Outside_Glo	LogOfAreas	Log_X_Index	Log_Y_Index	Orientation_	Luminosity_	SigmoidOfA
X_Maximum	46734	-60848697	-320672	-15751	-12944	-32609925	3686	2041	1238	0	17	25	-7	5	-2	17	23	31	-76	-48	-31	28	18	-30
Y_Maximum	-60848697	1.822E+12	1.028E+09	83317353	160209449	4.9E+10	-5669890	-6007837	-7505510	0	-114611	-47711	21948	-59251	4295	-19166	-35306	-86404	168070	111448	73014	-82047	-50711	73812
Pixels_Areas	-320672	1.028E+09	104771843	6692649	10371695	9.008E+09	-154934	6294	10070	0	547	-492	585	200	223	-1121	-355	556	3457	1427	2841	980	-300	575
X_Perimeter	-15751	83317353	6692649	442771	706257	557116030	-7764	770	772	0	32	-24	38	11	11	-68	-13	45	183	68	169	72	-16	29
Y_Perimeter	-12944	160209449	10371695	706257	1206391	807551258	-6894	1492	-1364	0	10	-18	44	-17	6	-65	13	63	177	44	208	105	-21	20
Sum_of_Lun	-32609925	4.9E+10	9.008E+09	557116030	807551258	8.193E+11	-16498428	777671	2214134	0	49760	-53267	58475	44602	25471	-123181	-50985	60033	361545	157341	278177	96509	-22291	62063
Minimum_o	3686	-5669890	-154934	-7764	-6894	-16498428	1458	439	-154	0	-2	4	-2	1	-1	4	5	5	-22	-13	-11	4	4	-7
Maximum_c	2041	-6007837	6294	770	1492	777671	439	333	2	0	-1	2	0	2	0	0	2	4	-6	-4	-2	4	3	-3
Length_of_C	1238	-7505510	10070	772	-1364	2214134	-154	2	2522	0	-2	1	1	4	0	-3	-1	5	2	0	3	4	0	0
TypeOfStee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Steel_Plate_	17	-114611	547	32	10	49760	-2	-1	-2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_Index	25	-47711	-492	-24	-18	-53267	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Empty_Inde	-7	21948	585	38	44	58475	-2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Square_Inde	5	-59251	200	11	-17	44602	1	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Outside_X_I	-2	4295	223	11	6	25471	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_X_In	17	-19166	-1121	-68	-65	-123181	4	0	-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Edges_Y_In	23	-35306	-355	-13	13	-50985	5	2	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Outside_Glo	31	-86404	556	45	63	60033	5	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LogOfAreas	-76	168070	3457	183	177	361545	-22	-6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Log_X_Index	-48	111448	1427	68	44	157341	-13	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Log_Y_Index	-31	73014	2841	169	208	278177	-11	-2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Orientation_	28	-82047	980	72	105	96509	4	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Luminosity_	18	-50711	-300	-16	-21	-22291	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SigmoidOfA	-30	73812	575	29	20	62063	-7	-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Class	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 9: Covariance matrix for class 1

Inferences:

1. The accuracy with the Bayes' classifier is 88.811%. Its focus is to similarity between the observations. While the KNN Classifier works better in this case due to its inherent property to optimize locally.

2. The diagonal elements in the covariance matrix denotes the variance of that attribute.
3. The off-diagonal elements indicate the covariance between those attributes. The attributes with maximum covariance are Y_maximum and Sum_of_Luminosity .

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	90.209
2.	KNN on normalized data	96.153
3.	Bayes	88.811

Inferences:

1. KNN classifier on normalized data has maximum accuracy, while the Bayes classifier has minimum accuracy.
2. The classifiers in ascending order are Bayes classifier < KNN Classifier < KNN Classifier on normalized data.
3. KNN performs better on normalized data. This is since KNN Classifier uses Euclidian distance between two points, thus its values depend on if the given values are too large or too small. But if the data is normalized, it will provide uniformity and thus, the accuracy will increase.