Indian Institute of Technology Mandi

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: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: CSE**

1 a.

	Prediction Outcome	
True Label	81	27
	27	201

Figure 1 KNN Confusion Matrix for K = 1

	Prediction Outcome	
True Label	83	12
	25	216

Figure 2 KNN Confusion Matrix for K = 3

	Prediction Outcome	
True Label	82	9
	26	219

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	**83.929**
3	**88.988**
5	**89.583**

Inferences:

1. The highest classification accuracy is obtained with K = 5.
2. Increasing the value of K increases the prediction accuracy.
3. As it more clear the surrounding the of the data hence K increases the prediction accuracy.
4. Classification accuracy increases with the increase in value of K infer does the number of diagonal elements increase.
5. Since the number of comparison increases and as we increase the value of k it will saturate to the real surrounding.
6. Increasing accuracy refers more approximate prediction and more near to the real data.
7. As the classification accuracy increases the with the increasing value of k, thus off-diagonal elements decreases.

2 a.

	Prediction Outcome	
True Label	104	9
	4	219

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

	Prediction Outcome	
True Label	104	6
	4	222

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

	Prediction Outcome	
True Label	103	7
	5	221

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	**96.131**
3	**97.024**
5	**96.429**

Inferences:

1. Data normalization increases classification accuracy.
2. Now the supression of values is not going to happen and each have same precidence.
3. The highest classification accuracy is obtained with K = 3.
4. Increasing the value of K increases the prediction accuracy.
5. As the surrounding data more clears with the real one.
6. The classification accuracy increases with the increase in value of K infer the number of diagonal elements increase.
7. Since our data is more to real one, the number of diagonal elements increase.
8. As the classification accuracy increases with the increase in value of K the number of off-diagonal elements decreases as the data goes more real so less wrong prediction.

3

	Prediction Outcome	
True Label	69	18
	39	210

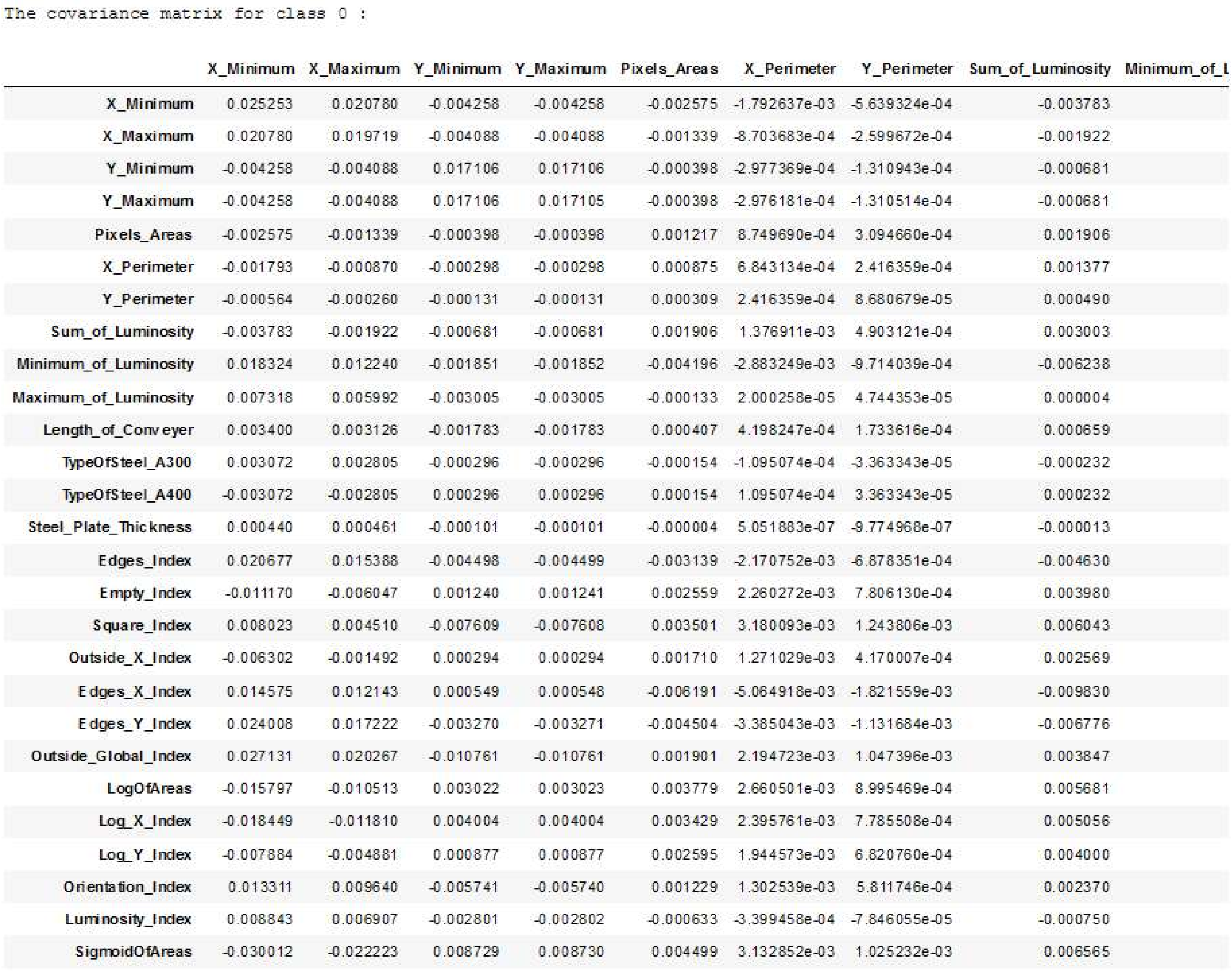
Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 83.03 %.

Table 3 Mean for class 0 and class 1

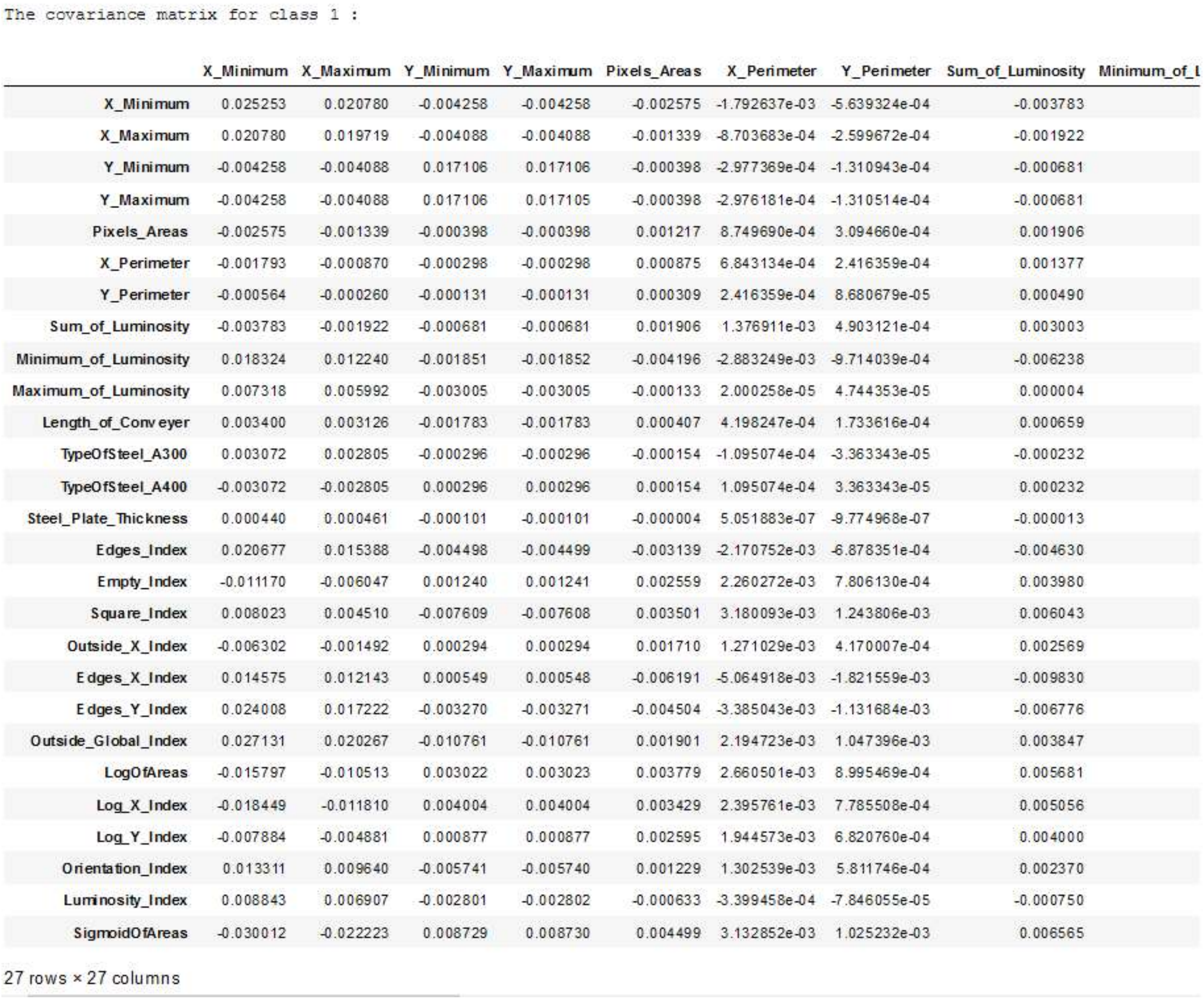
S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Minimum	0.08	0.42
2.	X_Maximum	0.16	0.43
3.	Y_Minimum	0.14	0.11
4.	Y_Maximum	0.05	0.11
5.	Pixels_Areas	0.03	0.00
6.	X_Perimeter	0.01	0.00
7.	Y_Perimeter	0.07	0.00
8.	Sum_of_Luminosity	0.26	0.00
9.	Minimum_of_Luminosity	0.46	0.46
10.	Maximum_of_Luminosity	0.32	0.43
11.	Length_of_Conveyer	0.01	0.53
12.	TypeOfSteel_A300	0.99	0.37
13.	TypeOfSteel_A400	0.00	0.62
14.	Steel_Plate_Thickness	0.13	0.23
15.	Edges_Index	0.48	0.40
16.	Empty_Index	0.59	0.44
17.	Square_Index	0.12	0.50
18.	Outside_X_Index	0.56	0.02
19.	Edges_X_Index	0.50	0.62
20.	Edges_Y_Index	0.29	0.82
21.	Outside_Global_Index	0.67	0.61
22.	LogOfAreas	0.62	0.40
23.	Log_X_Index	0.42	0.32
24.	Log_Y_Index	0.42	0.30
25.	Orientation_Index	0.33	0.56
26.	Luminosity_Index	0.54	0.53
27.	SigmoidOfAreas	0.90	0.46

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.



```
The covariance matrix for class 0 :
```

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	Sum_of_Luminosity	Minimum_of_L
X_Minimum	0.025253	0.020780	-0.004258	-0.004258	-0.002575	-1.792637e-03	-5.639324e-04	-0.003783	
X_Maximum	0.020780	0.019719	-0.004088	-0.004088	-0.001339	-8.703683e-04	-2.599672e-04	-0.001922	
Y_Minimum	-0.004258	-0.004088	0.017106	0.017106	-0.000398	-2.977369e-04	-1.310943e-04	-0.000681	
Y_Maximum	-0.004258	-0.004088	0.017106	0.017105	-0.000398	-2.976181e-04	-1.310514e-04	-0.000681	
Pixels_Areas	-0.002575	-0.001339	-0.000398	-0.000398	0.001217	8.749690e-04	3.094660e-04	0.001906	
X_Perimeter	-0.001793	-0.000870	-0.000298	-0.000298	0.000875	6.843134e-04	2.416359e-04	0.001377	
Y_Perimeter	-0.000564	-0.000260	-0.000131	-0.000131	0.000309	2.416359e-04	8.680679e-05	0.000490	
Sum_of_Luminosity	-0.003783	-0.001922	-0.000681	-0.000681	0.001906	1.376911e-03	4.903121e-04	0.003003	
Minimum_of_Luminosity	0.018324	0.012240	-0.001851	-0.001852	-0.004196	-2.883249e-03	-9.714039e-04	-0.006238	
Maximum_of_Luminosity	0.007318	0.005992	-0.003005	-0.003005	-0.000133	2.000258e-05	4.744353e-05	0.000004	
Length_of_Conveyer	0.003400	0.003126	-0.001783	-0.001783	0.000407	4.198247e-04	1.733616e-04	0.000659	
TypeOfSteel_A300	0.003072	0.002805	-0.000296	-0.000296	-0.000154	-1.095074e-04	-3.363343e-05	-0.000232	
TypeOfSteel_A400	-0.003072	-0.002805	0.000296	0.000296	0.000154	1.095074e-04	3.363343e-05	0.000232	
Steel_Plate_Thickness	0.000440	0.000461	-0.000101	-0.000101	-0.000004	5.051883e-07	-9.774968e-07	-0.000013	
Edges_Index	0.020677	0.015388	-0.004498	-0.004499	-0.003139	-2.170752e-03	-6.878351e-04	-0.004630	
Empty_Index	-0.011170	-0.006047	0.001240	0.001241	0.002559	2.260272e-03	7.806130e-04	0.003980	
Square_Index	0.008023	0.004510	-0.007609	-0.007608	0.003501	3.180093e-03	1.243806e-03	0.006043	
Outside_X_Index	-0.006302	-0.001492	0.000294	0.000294	0.001710	1.271029e-03	4.170007e-04	0.002569	
Edges_X_Index	0.014575	0.012143	0.000549	0.000548	-0.006191	-5.064918e-03	-1.821559e-03	-0.009830	
Edges_Y_Index	0.024008	0.017222	-0.003270	-0.003271	-0.004504	-3.385043e-03	-1.131684e-03	-0.006776	
Outside_Global_Index	0.027131	0.020267	-0.010761	-0.010761	0.001901	2.194723e-03	1.047396e-03	0.003847	
LogOfAreas	-0.015797	-0.010513	0.003022	0.003023	0.003779	2.660501e-03	8.995469e-04	0.005681	
Log_X_Index	-0.018449	-0.011810	0.004004	0.004004	0.003429	2.395761e-03	7.785508e-04	0.005056	
Log_Y_Index	-0.007884	-0.004881	0.000877	0.000877	0.002595	1.944573e-03	6.820760e-04	0.004000	
Orientation_Index	0.013311	0.009640	-0.005741	-0.005740	0.001229	1.302539e-03	5.811746e-04	0.002370	
Luminosity_Index	0.008843	0.006907	-0.002801	-0.002802	-0.000633	-3.399458e-04	-7.846055e-05	-0.000750	
SigmoidOfAreas	-0.030012	-0.022223	0.008729	0.008730	0.004499	3.132852e-03	1.025232e-03	0.006565	



```
The covariance matrix for class 1 :
```

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	Sum_of_Luminosity	Minimum_of_L
X_Minimum	0.025253	0.020780	-0.004258	-0.004258	-0.002575	-1.792637e-03	-5.639324e-04	-0.003783	
X_Maximum	0.020780	0.019719	-0.004088	-0.004088	-0.001339	-8.703683e-04	-2.599672e-04	-0.001922	
Y_Minimum	-0.004258	-0.004088	0.017106	0.017106	-0.000398	-2.977369e-04	-1.310943e-04	-0.000681	
Y_Maximum	-0.004258	-0.004088	0.017106	0.017105	-0.000398	-2.976181e-04	-1.310514e-04	-0.000681	
Pixels_Areas	-0.002575	-0.001339	-0.000398	-0.000398	0.001217	8.749690e-04	3.094660e-04	0.001906	
X_Perimeter	-0.001793	-0.000870	-0.000298	-0.000298	0.000875	6.843134e-04	2.416359e-04	0.001377	
Y_Perimeter	-0.000564	-0.000260	-0.000131	-0.000131	0.000309	2.416359e-04	8.680679e-05	0.000490	
Sum_of_Luminosity	-0.003783	-0.001922	-0.000681	-0.000681	0.001906	1.376911e-03	4.903121e-04	0.003003	
Minimum_of_Luminosity	0.018324	0.012240	-0.001851	-0.001852	-0.004196	-2.883249e-03	-9.714039e-04	-0.006238	
Maximum_of_Luminosity	0.007318	0.005992	-0.003005	-0.003005	-0.000133	2.000258e-05	4.744353e-05	0.000004	
Length_of_Conveyer	0.003400	0.003126	-0.001783	-0.001783	0.000407	4.198247e-04	1.733616e-04	0.000659	
TypeOfSteel_A300	0.003072	0.002805	-0.000296	-0.000296	-0.000154	-1.095074e-04	-3.363343e-05	-0.000232	
TypeOfSteel_A400	-0.003072	-0.002805	0.000296	0.000296	0.000154	1.095074e-04	3.363343e-05	0.000232	
Steel_Plate_Thickness	0.000440	0.000461	-0.000101	-0.000101	-0.000004	5.051883e-07	-9.774968e-07	-0.000013	
Edges_Index	0.020677	0.015388	-0.004498	-0.004499	-0.003139	-2.170752e-03	-6.878351e-04	-0.004630	
Empty_Index	-0.011170	-0.006047	0.001240	0.001241	0.002559	2.260272e-03	7.806130e-04	0.003980	
Square_Index	0.008023	0.004510	-0.007609	-0.007608	0.003501	3.180093e-03	1.243806e-03	0.006043	
Outside_X_Index	-0.006302	-0.001492	0.000294	0.000294	0.001710	1.271029e-03	4.170007e-04	0.002569	
Edges_X_Index	0.014575	0.012143	0.000549	0.000548	-0.006191	-5.064918e-03	-1.821559e-03	-0.009830	
Edges_Y_Index	0.024008	0.017222	-0.003270	-0.003271	-0.004504	-3.385043e-03	-1.131684e-03	-0.006776	
Outside_Global_Index	0.027131	0.020267	-0.010761	-0.010761	0.001901	2.194723e-03	1.047396e-03	0.003847	
LogOfAreas	-0.015797	-0.010513	0.003022	0.003023	0.003779	2.660501e-03	8.995469e-04	0.005681	
Log_X_Index	-0.018449	-0.011810	0.004004	0.004004	0.003429	2.395761e-03	7.785508e-04	0.005056	
Log_Y_Index	-0.007884	-0.004881	0.000877	0.000877	0.002595	1.944573e-03	6.820760e-04	0.004000	
Orientation_Index	0.013311	0.009640	-0.005741	-0.005740	0.001229	1.302539e-03	5.811746e-04	0.002370	
Luminosity_Index	0.008843	0.006907	-0.002801	-0.002802	-0.000633	-3.399458e-04	-7.846055e-05	-0.000750	
SigmoidOfAreas	-0.030012	-0.022223	0.008729	0.008730	0.004499	3.132852e-03	1.025232e-03	0.006565	

27 rows × 27 columns

Inferences:

1. Bayes has an accuracy of about 83.04, less than KNN as bayes work on similarity founding between the data values while KNN is more practical and near to inherent nature.
2. The diagonal values are positive means variable data.
3. The off-diagonal values are very low showing less relation between the other attribute.
4. Max covariance attributes are Y_maximum and sum_of_luminosity for both classes.

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.4
2.	KNN on normalized data	97.32
3.	Bayes	83.04

Inferences:

1. Highest is of KNN on normalization and lowest Is of bayes.
2. Bayes < KNN < KNN(normalized)
3. As Bayes work on similarity founding between the data values while KNN is more practical and near to inherent nature.