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: SARTHAK SAPTAMI KUMAR JHA **Mobile No: 8825319259**

Roll Number: B20317 **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	25
	12	216

Figure 2 KNN Confusion Matrix for K = 3

		Prediction Outcome
True Label	82	26
	9	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	**83.92**
3	**88.98**
4	**89.58**

Inferences:

1. The highest classification accuracy is obtained with K = 5.
2. With increase in k, classification accuracy increases.
3. As, we increase the value of k, more data is checked before assigning the class, thus, we have a smoother curve in the data, which leads to better accuracy
4. The values of diagonal elements increase, with increase in value of K.
5. The values of diagonal elements increase, with increase in value of K, as more data is now assigned to its correct class, by the algorithm.
6. The values of off-diagonal elements decrease, with increase in value of K.
7. The values of diagonal elements decrease, with increase in value of K, as now fewer data is predicted as being in wrong class.

2 a.

	Prediction Outcome	
True Label	104	4
	9	219

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

	Prediction Outcome	
True Label	105	3
	7	221

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

	Prediction Outcome	
True Label	104	4
	7	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	**96.13**
3	**97.02**
5	**96.76**

Inferences:

1. Data normalization increases the classification accuracy
2. After data normalization, classification accuracy increases, because, KNN is a distance-based algorithm, which means, that for attributes with higher ranges, will have more weight on the algorithm, and will thus give multiple error prone classifications
3. The highest classification accuracy is obtained with K = 3.
4. There is general trend of increase with minor decrease in K=3 to K=5.
5. While there is a general trend of increase in classification accuracy with increase in value of K, at high levels of accuracy, inherent variations in data can cause a minor decrease in the classification accuracy.
6. The values of diagonal elements increase, with increase in value of K.
7. The values of diagonal elements increase, with increase in value of K, as more data is now assigned to its correct class, by the algorithm.
8. The values of off-diagonal elements decrease, with increase in value of K.
9. The values of diagonal elements decrease, with increase in value of K, as now fewer data is predicted as being in wrong class.

3

		Prediction Outcome
True Label	101	7
	39	189

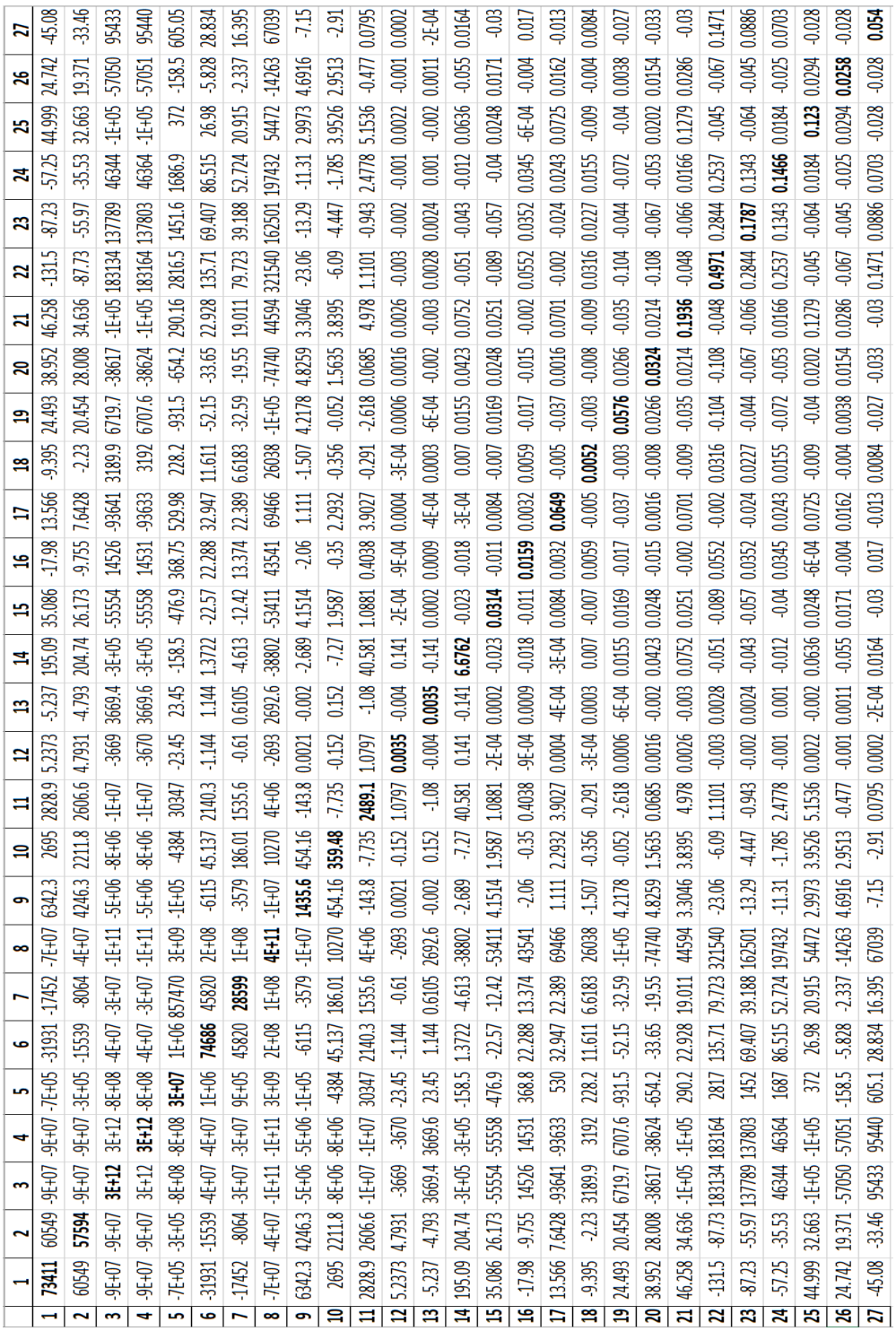
Figure 7 Confusion Matrix obtained from Bayes Classifier

The classification accuracy obtained from Bayes Classifier is 86.31%.

Table 3 Mean for class 0 and class 1

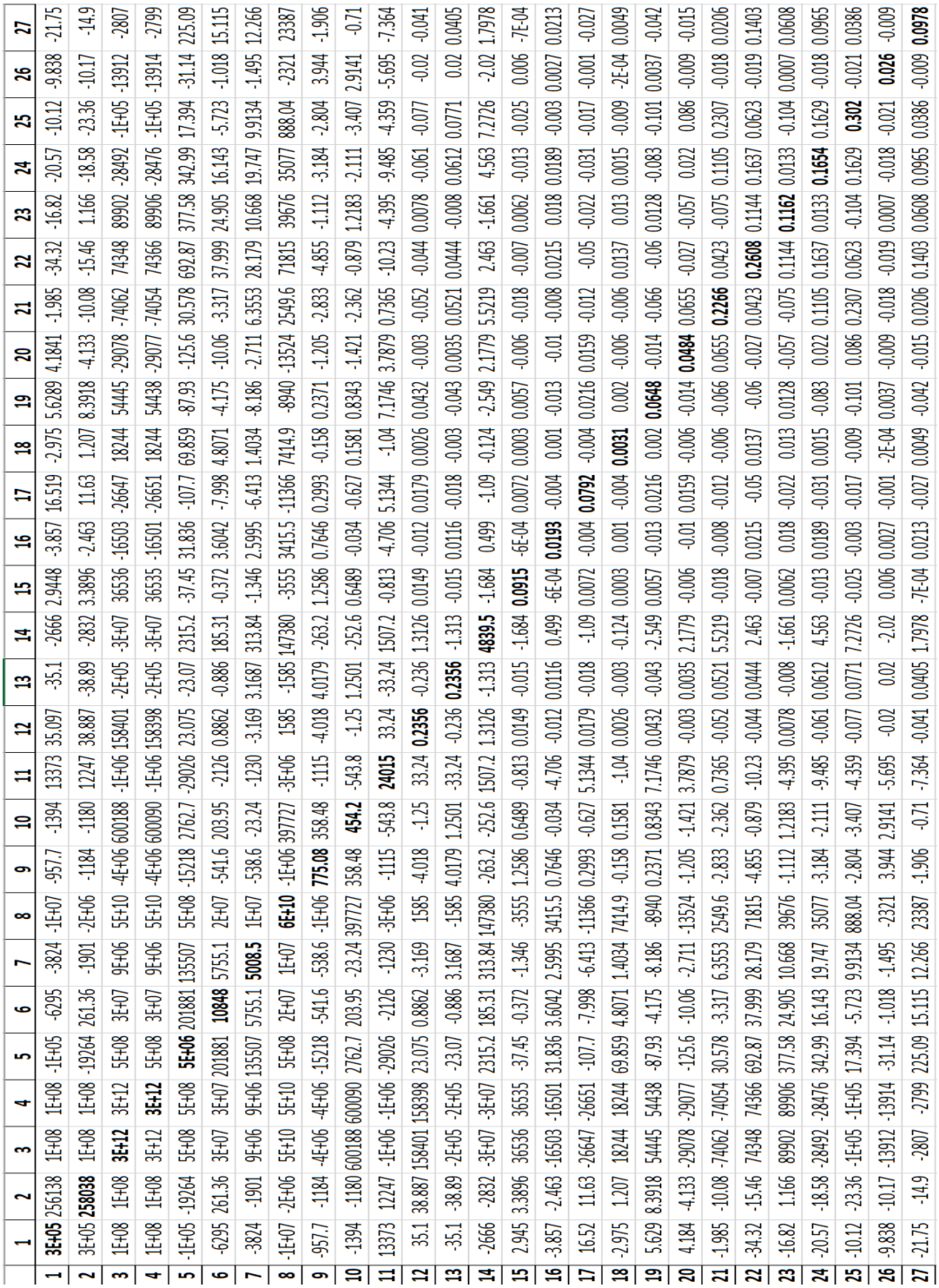
S. No.	Attribute Name	Mean	
		Class 0	Class 1
1.	X_Minimum	137.07	718.10
2.	X_Maximum	286.33	746.58
3.	Y_Minimum	1711388.84	1445930.35
4.	Y_Maximum	1711478.05	1445963.75
5.	Pixels_Areas	7268.03	583.51
6.	X_Perimeter	355.61	52.18
7.	Y_Perimeter	207.15	43.11
8.	Sum_of_Luminosity	808615.69	61552.41
9.	Minimum_of_Luminosity	53.40	94.80
10.	Maximum_of_Luminosity	135.85	130.18
11.	Length_of_Conveyer	1382.51	1486.63
12.	TypeOfSteel_A300	0.003	0.37
13.	TypeOfSteel_A400	0.996	0.62
14.	Steel_Plate_Thickness	40.24	100.43
15.	Edges_Index	0.12	0.38
16.	Empty_Index	0.44	0.41
17.	Square_Index	0.59	0.51
18.	Outside_X_Index	0.10	0.01
19.	Edges_X_Index	0.56	0.62
20.	Edges_Y_Index	0.52	0.83
21.	Outside_Global_Index	0.26	0.61
22.	LogOfAreas	3.59	2.26
23.	Log_X_Index	2.04	1.21
24.	Log_Y_Index	1.82	1.29
25.	Orientation_Index	-0.32	0.13
26.	Luminosity_Index	-0.10	-0.12
27.	SigmoidOfAreas	0.91	0.52

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.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	**73411**	60549	-9E+07	-9E+07	-7E+05	-31931	-17452	-7E+07	6342.3	2695	2828.9	5.2373	-5.237	195.09	35.086	-17.98	13.566	-9.395	24.493	38.952	46.258	-131.5	-87.23	-57.25	44.999	24.742	-45.08
2	60549	**57594**	-9E+07	-9E+07	-3E+05	-15539	-8064	-4E+07	4246.3	2211.8	2606.6	4.7931	-4.793	204.74	26.173	-9.755	7.6428	-2.23	20.454	28.008	34.636	-87.73	-55.97	-35.53	32.663	19.371	-33.46
3	-9E+07	-9E+07	**3E+12**	3E+12	-8E+08	-4E+07	-3E+07	-1E+11	-5E+06	-8E+06	-1E+07	-3669	3669.4	-3E+05	-55554	14526	-93641	3189.9	6719.7	-38617	-1E+05	183134	137789	46344	-1E+05	-57050	95433
4	-9E+07	-9E+07	3E+12	**3E+12**	-8E+08	-4E+07	-3E+07	-1E+11	-5E+06	-8E+06	-1E+07	-3670	3669.6	-3E+05	-55558	14531	-93633	3192	6707.6	-38624	-1E+05	183164	137803	46364	-1E+05	-57051	95440
5	-7E+05	-3E+05	-8E+08	-8E+08	**3E+07**	1E+06	857470	3E+09	-1E+05	-4384	30347	-23.45	23.45	-158.5	-476.9	368.75	529.98	228.2	-931.5	-654.2	290.16	2816.5	1451.6	1686.9	372	-158.5	605.05
6	-31931	-15539	-4E+07	-4E+07	1E+06	**74686**	45820	2E+08	-6115	45.137	2140.3	-1.144	1.144	1.3722	-22.57	22.288	32.947	11.611	-52.15	-33.65	22.928	135.71	69.407	86.515	26.98	-5.828	28.834
7	-17452	-8064	-3E+07	-3E+07	9E+05	45820	**28599**	1E+08	-3579	186.01	1535.6	-0.61	0.6105	-4.613	-12.42	13.374	22.389	6.6183	-32.59	-19.55	19.011	79.723	39.188	52.724	20.915	-2.337	16.395
8	-7E+07	-4E+07	-1E+11	-1E+11	3E+09	2E+08	1E+08	**4E+11**	-1E+07	10270	4E+06	-2693	2692.6	-38802	-53411	43541	69466	26038	-74740	-74740	44594	321540	162501	197432	54472	54472	67039
9	6342.3	4246.3	-5E+06	-5E+06	-1E+05	-6115	-3579	-1E+07	**1435.6**	454.16	-143.8	0.0021	-0.002	-2.689	4.1514	-2.06	1.111	-1.507	4.2178	4.8259	3.3046	-23.06	-13.29	-11.31	2.9973	4.6916	-7.15
10	2695	2211.8	-8E+06	-8E+06	-4384	45.137	186.01	10270	454.16	**359.48**	-7.735	-0.152	0.152	-7.27	1.9587	-0.35	2.2932	-0.356	-0.052	1.5635	3.8395	-6.09	-4.447	-1.785	3.9526	2.9513	-2.91
11	2828.9	2606.6	-1E+07	-1E+07	30347	2140.3	1535.6	4E+06	-143.8	-7.735	**2489.1**	1.0797	-1.08	40.581	1.0881	0.4038	3.9027	-0.291	-2.618	0.0685	4.978	1.1101	-0.943	2.4778	5.1536	-0.477	0.0795
12	5.2373	4.7931	-3669	-3670	-23.45	-1.144	-0.61	-2693	0.0021	-0.152	-0.004	**0.0035**	-0.004	0.141	-2E-04	-9E-04	0.0004	-3E-04	0.0006	0.0016	-0.003	-0.002	-0.002	-0.001	0.0022	-0.001	0.0002
13	-5.237	-4.793	3669.4	3669.6	23.45	1.144	0.6105	2692.6	-0.002	0.152	-1.08	-0.004	**0.0035**	-0.141	0.0002	0.0009	-4E-04	0.0003	-6E-04	-0.002	-0.003	0.0028	0.0024	0.001	-0.002	0.0011	-2E-04
14	195.09	204.74	-3E+05	-3E+05	-158.5	1.3722	-4.613	-38802	-2.689	-7.27	40.581	0.141	-0.141	**6.6762**	-0.023	-0.018	-3E-04	0.007	0.0155	0.0423	0.0752	-0.051	-0.043	-0.012	0.0636	-0.055	0.0164
15	35.086	26.173	-55554	-55558	-476.9	-22.57	-12.42	-53411	4.1514	1.9587	1.0881	-2E-04	0.0002	**0.0314**	-0.011	0.0084	-0.007	0.0169	0.0248	0.0251	-0.089	-0.089	-0.057	-0.04	0.0248	0.0171	-0.03
16	-17.98	-9.755	14526	14531	368.8	22.288	13.374	43541	-2.06	-0.35	0.4038	-9E-04	-0.0009	-0.018	-0.011	**0.0159**	0.0032	0.0059	-0.017	-0.015	-0.002	0.0552	0.0352	0.0345	-6E-04	-0.004	0.017
17	13.566	7.6428	-93641	-93633	530	32.947	22.389	69466	1.111	2.2932	3.9027	0.0004	-4E-04	-3E-04	0.0084	0.0032	**0.0649**	-0.005	-0.037	0.0016	0.0701	-0.002	-0.024	0.0243	0.0725	0.0162	-0.013
18	-9.395	-2.23	3189.9	3192	228.2	11.611	6.6183	26038	-1.507	-0.356	-0.291	-3E-04	0.0003	0.007	-0.007	0.0059	-0.005	**0.0052**	-0.003	-0.008	-0.009	0.0316	0.0227	0.0155	-0.009	-0.004	0.0084
19	24.493	20.454	6719.7	6707.6	-931.5	-52.15	-32.59	-1E+05	4.2178	-0.052	-2.618	0.0006	-6E-04	0.0155	0.0169	-0.017	-0.037	-0.003	**0.0576**	0.0266	-0.035	-0.104	-0.044	-0.072	-0.04	0.0038	-0.027
20	38.952	28.008	-38617	-38624	-654.2	-33.65	-19.55	-74740	4.8259	1.5635	0.0685	0.0016	-0.002	0.0423	0.0248	-0.015	0.0016	-0.008	0.0266	**0.0324**	0.0214	-0.108	-0.067	-0.053	0.0202	0.0154	-0.033
21	46.258	34.636	-1E+05	-1E+05	290.2	22.928	19.011	44594	3.3046	3.8395	4.978	0.0026	-0.003	0.0752	0.0251	-0.002	0.0701	-0.009	-0.035	0.0214	**0.1936**	-0.048	-0.066	0.0166	0.1279	0.0286	-0.03
22	-131.5	-87.73	183134	183164	2817	135.71	79.723	321540	-23.06	-6.09	1.1101	-0.003	0.0028	-0.051	-0.089	0.0552	-0.002	0.0316	-0.104	-0.108	-0.048	**0.4971**	0.2844	0.2537	-0.045	-0.067	0.1471
23	-87.23	-55.97	137789	137803	1452	69.407	39.188	162501	-13.29	-4.447	-0.943	-0.002	0.0024	-0.043	-0.057	0.0352	-0.024	0.0227	-0.044	-0.067	-0.066	0.2844	**0.1787**	0.1343	-0.064	-0.045	0.0886
24	-57.25	-35.53	46344	46364	1687	86.515	52.724	197432	-11.31	-1.785	2.4778	-0.001	0.001	-0.012	-0.04	0.0345	0.0243	0.0155	-0.072	-0.053	0.0166	0.2537	0.1343	**0.1466**	0.0184	-0.025	0.0703
25	44.999	32.663	-1E+05	-1E+05	372	26.98	20.915	54472	2.9973	3.9526	5.1536	0.0022	-0.002	0.0636	0.0248	-6E-04	0.0725	-0.009	-0.04	0.0202	0.1279	-0.045	-0.064	0.0184	**0.123**	0.0294	-0.028
26	24.742	19.371	-57050	-57051	-158.5	-5.828	-2.337	-14263	4.6916	2.9513	-0.477	-0.001	0.0011	-0.055	0.0171	-0.004	0.0162	-0.004	0.0038	0.0154	0.0286	-0.067	-0.045	-0.025	0.0294	**0.0258**	-0.028
27	-45.08	-33.46	95433	95440	605.1	28.834	16.395	67039	-7.15	-2.91	0.0795	0.0002	-2E-04	0.0164	-0.03	0.017	-0.013	0.0084	-0.027	-0.033	-0.03	0.1471	0.0886	0.0703	-0.028	-0.028	**0.054**

Figure 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	24	25	26	27
1	**3E+05**	256138	1E+08	1E+08	-1E+05	-6295	-3824	-1E+07	-957.7	-1394	13373	35.097	-35.1	-2666	2.9448	-3.857	16.519	-2.975	5.6289	4.1841	-1.985	-34.32	-16.82	-20.57	-10.12	-9.838	-21.75
2	3E+05	**258038**	1E+08	1E+08	-19264	261.36	-1901	-2E+06	-1184	-1180	12247	38.887	-38.89	-2832	3.3896	-2.463	11.63	1.207	8.3918	-4.133	-10.08	-15.46	1.166	-18.58	-23.36	-10.17	-14.9
3	1E+08	1E+08	**3E+12**	3E+12	5E+08	3E+07	9E+06	5E+10	-4E+06	600188	-1E+06	158401	-2E+05	-26647	36536	-16503	-26647	18244	54445	-29078	-74062	74348	89902	-28492	-1E+05	-13912	-2807
4	1E+08	1E+08	3E+12	**3E+12**	5E+08	3E+07	9E+06	5E+10	-4E+06	600090	-1E+06	158398	-2E+05	-3E+07	36535	-16501	-26651	18244	54438	29077	-74054	74366	89906	-28476	-1E+05	-13914	-2799
5	-1E+05	-19264	5E+08	5E+08	**5E+06**	201881	135507	5E+08	-15218	2762.7	-29026	23.075	-23.07	2315.2	-37.45	31.836	-107.7	69.859	-87.93	-125.6	30.578	692.87	377.58	342.99	17.394	-31.14	225.09
6	-6295	261.36	3E+07	3E+07	201881	**10848**	5755.1	2E+07	-541.6	203.95	-2126	0.8862	-0.886	185.31	-0.372	3.6042	-7.998	4.8071	-4.175	-10.06	-3.317	37.999	24.905	16.143	-5.723	-1.018	15.115
7	-3824	-1901	9E+06	9E+06	135507	5755.1	**5008.5**	1E+07	-538.6	-23.24	-1230	-3.169	3.1687	313.84	-1.346	2.5995	-6.413	1.4034	-8.186	-2.711	6.3553	28.179	10.668	19.747	9.9134	-1.495	12.266
8	-1E+07	-2E+06	5E+10	5E+10	5E+08	2E+07	1E+07	**6E+10**	-1E+06	397727	-3E+06	1585	-1585	147380	-3555	3415.5	-11366	7414.9	-8940	-13524	2549.6	71815	39676	35077	888.04	-2321	23387
9	-957.7	-1184	-4E+06	-4E+06	-15218	-541.6	-538.6	-1E+06	**775.08**	358.48	-1115	-4.018	4.0179	-263.2	1.2586	0.7646	0.2993	-0.158	0.2371	-1.205	-2.833	-4.855	-1.112	-3.184	-2.804	3.944	-1.906
10	-1394	-1180	600188	600090	2762.7	203.95	-23.24	397727	358.48	**454.2**	-543.8	-1.25	1.2501	-252.6	0.6489	-0.034	-0.627	0.1581	0.8343	-1.421	-2.362	-0.879	1.2183	-2.111	-3.407	2.9141	-0.71
11	13373	12247	-1E+06	-1E+06	-29026	-2126	-1230	-3E+06	-1115	-543.8	**24015**	33.24	-33.24	1507.2	-0.813	-4.706	5.1344	-1.04	7.1746	3.7879	0.7365	-10.23	-4.395	-9.485	-4.359	-5.695	-7.364
12	35.1	38.887	158401	158398	23.075	0.8862	-3.169	1585	-4.018	-1.25	33.24	**0.2356**	-0.236	1.3126	0.0149	-0.012	0.0179	0.0026	0.0432	-0.003	-0.052	-0.044	0.0078	-0.061	-0.077	-0.02	-0.041
13	-35.1	-38.89	-2E+05	-2E+05	-23.07	-0.886	3.1687	-1585	4.0179	1.2501	-33.24	-0.236	**0.2356**	-1.313	-0.015	0.0116	-0.018	-0.003	-0.043	0.0035	0.0521	0.0444	-0.008	0.0612	0.0771	0.02	0.0405
14	-2666	-2832	-3E+07	-3E+07	2315.2	185.31	313.84	147380	-263.2	-252.6	1507.2	1.3126	-1.313	**4839.5**	-1.684	0.499	-1.09	-0.124	-2.549	2.1779	5.5219	2.463	-1.661	4.563	7.2726	-2.02	1.7978
15	2.945	3.3896	36536	36535	-37.45	-0.372	-1.346	-3555	1.2586	0.6489	-0.813	0.0149	-0.015	-1.684	**0.0915**	-6E-04	0.0072	0.0003	0.0057	-0.006	-0.018	-0.007	0.0062	-0.013	-0.025	0.006	-7E-04
16	-3.857	-2.463	-16503	-16501	31.836	3.6042	2.5995	3415.5	0.7646	-0.034	-4.706	-0.012	0.0116	0.499	-6E-04	**0.0193**	-0.004	0.001	-0.013	-0.01	-0.008	0.0215	0.018	0.0189	-0.003	0.0027	0.0213
17	16.52	11.63	-26647	-26651	-107.7	-7.998	-6.413	-11366	0.2993	-0.627	5.1344	0.0179	-0.018	-1.09	0.0072	-0.004	**0.0792**	-0.004	0.0216	-0.014	-0.012	-0.05	-0.022	-0.031	-0.017	-0.001	-0.027
18	-2.975	1.207	18244	18244	69.859	4.8071	1.4034	7414.9	-0.158	0.1581	-1.04	0.0026	-0.003	-0.124	0.0003	0.001	-0.004	**0.0031**	0.002	-0.006	-0.006	0.0137	0.013	0.0015	-0.009	-2E-04	0.0049
19	5.629	8.3918	54445	54438	-87.93	-4.175	-8.186	-8940	0.2371	0.8343	7.1746	0.0432	-0.043	-2.549	0.0057	-0.013	0.0216	0.002	**0.0648**	-0.014	-0.066	-0.06	0.0128	-0.083	-0.101	0.0037	-0.042
20	4.184	-4.133	-29078	-29077	-125.6	-10.06	-2.711	-13524	-1.205	-1.421	3.7879	-0.003	0.0035	2.1779	-0.006	-0.01	0.0159	-0.006	-0.014	**0.0484**	0.0655	-0.027	-0.057	0.022	0.086	-0.009	-0.015
21	-1.985	-10.08	-74062	-74054	30.578	-3.317	6.3553	2549.6	-2.833	-2.362	0.7365	-0.052	0.0521	5.5219	-0.018	-0.008	-0.012	-0.006	-0.066	0.0655	**0.2266**	0.0423	-0.075	0.1105	0.2307	-0.018	0.0206
22	-34.32	-15.46	74348	74366	692.87	37.999	28.179	71815	-4.855	-0.879	-10.23	-0.044	0.0444	2.463	-0.007	0.0215	-0.05	0.0137	-0.06	-0.027	0.0423	**0.2608**	0.1144	0.1637	0.0623	-0.019	0.1403
23	-16.82	1.166	89902	89906	377.58	24.905	10.668	39676	-1.112	1.2183	-4.395	0.0078	-0.008	-1.661	0.0062	0.018	-0.022	0.013	0.0128	-0.057	-0.075	0.1144	**0.1162**	0.0133	-0.104	0.0007	0.0608
24	-20.57	-18.58	-28492	-28476	342.99	16.143	19.747	35077	-3.184	-2.111	-9.485	-0.061	0.0612	4.563	-0.013	0.0189	-0.031	0.0015	-0.083	0.022	0.1105	0.1637	0.0133	**0.1654**	0.1629	-0.018	0.0965
25	-10.12	-23.36	-1E+05	-1E+05	17.394	-5.723	9.9134	888.04	-2.804	-3.407	-4.359	-0.077	0.0771	7.2726	-0.025	-0.003	-0.017	-0.009	-0.101	0.086	0.2307	0.0623	-0.104	0.1629	**0.302**	-0.021	0.0386
26	-9.838	-10.17	-13912	-13914	-31.14	-1.018	-1.495	-2321	3.944	2.9141	-5.695	-0.02	0.02	-2.02	0.006	0.0027	-0.001	-2E-04	0.0037	-0.009	-0.018	-0.019	0.0007	-0.018	-0.021	**0.026**	-0.009
27	-21.75	-14.9	-2807	-2799	225.09	15.115	12.266	23387	-1.906	-0.71	-7.364	-0.041	0.0405	1.7978	-7E-04	0.0213	-0.027	0.0049	-0.042	-0.015	0.0206	0.1403	0.0608	0.0965	0.0386	-0.009	**0.0978**

Figure 9: Covariance matrix for class 1

Inferences:

1. The accuracy of bayes classifier is 86.31%
2. The diagonal elements are the variance of a particular attributes for that class, when we calculate the covariance matrix, the diagonal elements are variance of the attribute
3. The off-diagonal elements give the covariance between two attributes for that class. The maximum covariance between any two attributes is for Y_Maximum and Y_Minimum, and Y_Maximum and Pixels_Areas. The minimum covariance is between Edges_Index and SigmoidOfAreas, TypeOfSteel_A300 and Empty_Index.

4

Table 4 Comparison between classifiers based upon classification accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.58
2.	KNN on normalized data	97.02
3.	Bayes	86.31

Inferences:

1. KNN on normalized data has highest accuracy and Bayes Classifier has lowest
2. Accuracy: Bayes < KNN < KNN on normalized data
3. Bayes classifier has poor accuracy when compared to KNN and normalized KNN, which maybe because of the inherent characteristics of the data i.e., the data may not be a gaussian distribution. The normalized data lead to much higher accuracy in KNN than in non-normalized data, because, normalization removes any inherent high range values in the data.