MULTI-TABULAR MEDICAL DATA CLASSIFICATION (MICCAI DATASET)

Rajat Singh
Computer Science and Engineering
IIT Delhi
rajat.singh@cse.iitd.ac.in

Srikanta Bedathur
Computer Science and Engineering
IIT Delhi
srikanta@cse.iitd.ac.in

1 ABOUT MICCAI DATASET

For this study, I am using the MICCAI Dataset, a medical dataset for this study. This dataset contains records for 1,500 patients. Each patient has three distinct test reports, including Blood (Figure-1), Radiology (Figure-2), and Temperature (Figure-4), which are tested at different points in time. Based on these three reports, we are given the outcome (Figure-3) associated with each patient, i.e., the risk level of each patient.

Unnamed: 0	Albumin	Calcium	Immature Granulocytes	Lymphocytes	Neutrophils	Glucose	M.C.H. (Hb/Rbc)	M.C.H.C. (Hb/Ht)	Magnesium	Sodium	Thrombocytes	Urea	Leukocytes	Chloride	Phosphate	Reticulocytes	Segmentc. granulocytes	Erythroblasts	IL-6
7	34.497163	2.368936	0.197166	2.45237	6.001551	9.194186	1.761328	20.024231	0.678005	142.040465	355.452799	3.619929	13.793344	97.115492	1.174293	58.875736	21.658552	0.008558	156.440498
10	31.334999	2.531898	0.229335	2.603475	7.524609	8.028247	1.8104	20.229186	0.700896	142.479725	276.647957	4.453052	12.257076	97.00779	1.102597	63.338522	9.506187	0.008558	109.723829
12	30.643057	2.347358	0.076279	2.27732	5.591308	9.050117	1.808832	20.059522	0.72783	142.521524	235.504536	4.163227	10.920605	95.796102	1.035573	45.215933	14.062417	0.003646	147.250777
42	35.616692	2.412769	0.18881	1.427464	6.368001	6.355795	1.873926	20.277585	0.787974	139.788598	228.645944	5.663475	7.189219	101.200039	1.140113	66.726903	6.285339	0.00899	64.840375
44	36.263341	2.331255	0.103533	0.772011	7.415897	7.326521	1.872507	20.572529	0.798774	139.480754	280.380615	5.40023	6.612034	100.861831	1.104198	61.207002	5.308238	0.00899	139.165586
46	37.770787	2.327411	0.103533	1.256424	7.415263	5.979774	1.959948	20.356812	0.800547	139.495446	215.808253	5.851741	8.190891	101.560378	1.080399	71.697956	6.799655	0.00899	177.503641
52	38.88325	2.364771	0.080393	0.876853	8.517093	5.67434	1.948567	19.98714	0.830986	139.339321	183.083438	7.336493	4.797979	103.437966	1.101844	76.91175	8.790052	0.027558	184.330667
58	38.857037	2.336425	0.101693	1.468864	9.243405	5.99598	1.902171	20.325576	0.872287	139.096402	265.299401	7.731982	4.923367	103.912899	1.080445	59.709935	8.790052	0.02824	200.101195
59	39.096946	2.402311	0.266764	0.875981	7.85542	5.499987	1.900093	20.214468	0.813448	138.661584	357.617776	8.700301	7.516471	99.996925	1.263494	76.290877	3.68746	0.058584	111.64954
70	41.287431	2.384144	0.143656	0.830549	5.976302	5.231216	1.829425	20.047761	0.878752	136.849148	403.184904	9.70962	9.910267	102.970721	1.109948	75.221429	8.018615	0.066573	164.861455
77	45.084195	2.452909	0.045512	0.225675	7.516749	4.949276	1.727367	20.122269	0.8679	135.555865	423.83329	10.249626	14.865103	100.470953	1.107599	55.381978	10.703966	0.030387	111.948935
78	47.899036	2.371165	0.181725	0.604488	8.239821	4.914569	1.736362	19.908446	0.896811	134.868553	330.271055	10.183995	12.765309	100.856586	1.089156	61.346938	10.151339	0.104054	171.231424
80	49.452496	2.325125	0.109092	0.959186	4.915385	4.184117	1.691108	20.192073	0.878842	135.446724	358.242961	9.566475	17.437359	101.824137	1.042385	65.405356	8.045958	0.082502	122.494858
84	50.215488	2.408427	0.186595	0.95432	11.051539	5.303672	1.700271	19.978072	0.877795	134.556087	422.803207	10.554754	15.087143	98.821192	1.104724	62.707673	15.781897	0.110099	141.592552

Figure 1: Blood report of patient

	atelectasis	fatty_liver	renal_cysts	liver_cysts	kidney_stones	bladder_stones	distended_bowels	ground_glass
39	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
75	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
78	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE

Figure 2: Radiology report of patient

	risk_1	risk_2	risk_3
PT00355	FALSE	FALSE	FALSE

Figure 3: Outcome of patient

	temperature
3	34.573900263836
5	34.6557123841078
6	34.985888958684
7	35.3658587269648
13	33.5367088971278
14	36.3863577895124
16	38.9130269976883
22	36.9391464850053
25	37.213128304642
26	35.3945756527164
27	35.4548939669479
28	37.303172148895
32	38.0212420129074
35	36.7675870999218
44	35.4308350796133
47	38.2955947431586
48	37.0814540154629
52	37.0747072280464
53	36.3837701592789
57	35.6720595870546
60	36.3773795279455
61	34.5331412366137
65	38.6072711314763
67	35.9261562131681
69	37.416021585672
73	34.9256536534339
75	37.3728736347748
78	35.8786208318693
79	36.0164187788006
85	36.8677769668845
88	37.9025442081049

Figure 4: Temperature report of patient

1.1 Data Distribution

The distribution of the train, test, and Validation set of the MICCAI dataset is shown in Figure-5, -6, and -7. The mapping of the classes with the risk value in outcome.csv is shown in Table-1.

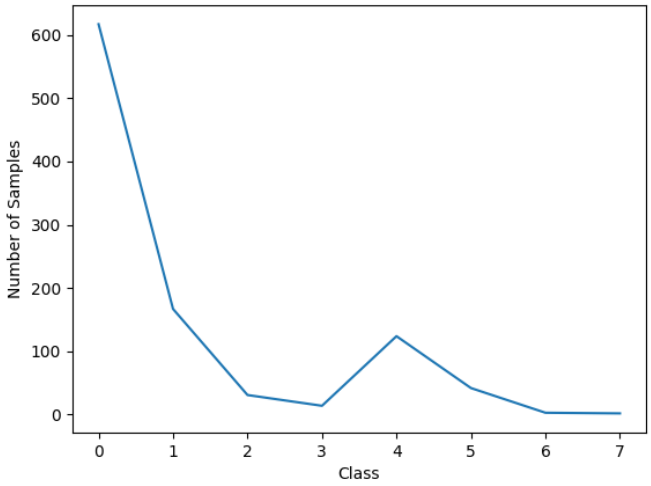

Figure 5: Train data class distribution. There are 1000 samples in the train set.

Risk_1	Risk_2	Risk_3	Class
False	False	False	0
False	False	True	1
False	True	False	2
False	True	True	3
True	False	False	4
True	False	True	5
True	True	False	6
True	True	True	7

Table 1: Risk level to class mapping

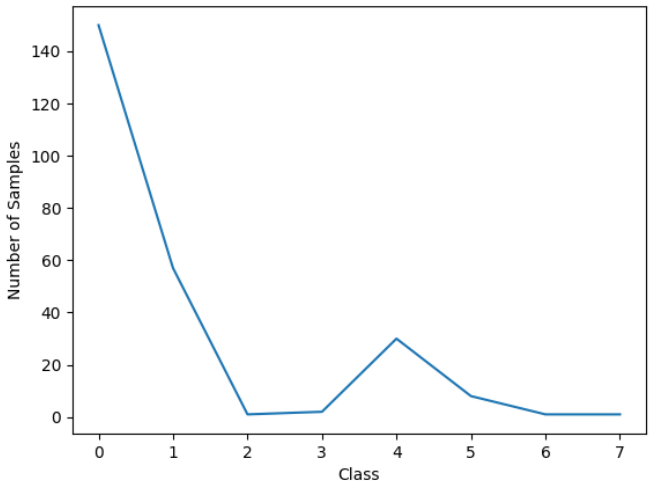

Figure 6: Test data class distribution. There are 250 samples in the test set.

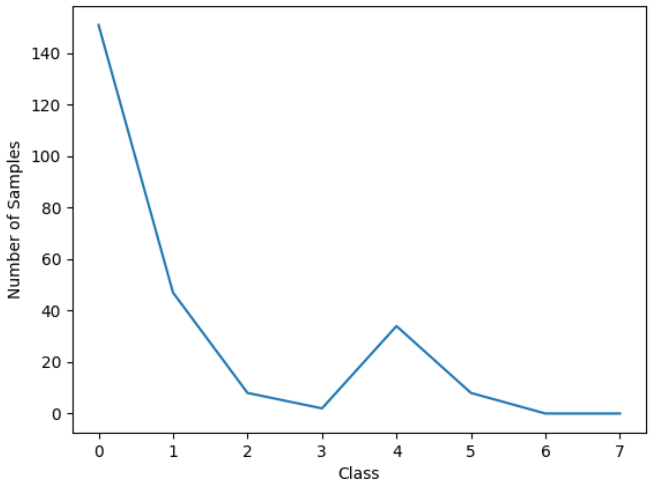

Figure 7: Validation data class distribution. There are 250 samples in the validation set.

2 MODELS

This section provides an abstract description of the various models used in this assignment.

2.1 XGBOOST

I have used the default XGBoost Chen & Guestrin (2016) model. It is a boosted decision tree model. For this model, I have taken the input as the average value of each patient table. I have skipped the cells that do not contain any value in the Temperature and Blood report table and taken the median in the Radiology report table, and then combined all the readings. This approach does not utilize the temporal component of the dataset.

2.2 XGBOOST-WIMV

It stands for XGBoost With Imputed Missing Values, where I have imputed the missing values by taking an average of previous and next readings for that sample.

2.3 SVM

SVMCortes & Vapnik (1995) stands for Support Vector Machines, a supervised machine learning model for classification, regression, and outliers detection. An SVM outputs the hyperplane that best separates the data samples. This plane represents the decision boundary.

3 RESULTS

Approach	Accuracy
XGBoost	0.564
XGBoost-WIMV	0.576
SVM	0.556

Table 2: Results

REFERENCES

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 1995.