Assignment 2

Matthias
Bennett

February 15th, 2024

1 Experiment Results

index	Normalization	Max Depth	Splitter	Accuracy %
0	No Preprocessing	3	best	0.9416666666666668
1	No Preprocessing	3	random	0.9333333333333333
2	No Preprocessing	5	best	0.9333333333333332
3	No Preprocessing	5	random	0.95
4	No Preprocessing	7	best	0.9333333333333332
5	No Preprocessing	7	random	0.9416666666666668
6	No Preprocessing	9	best	0.9
7	No Preprocessing	9	random	0.9583333333333334
8	Z-Score	3	best	0.9333333333333332
9	Z-Score	3	random	0.925
10	Z-Score	5	best	0.9
11	Z-Score	5	random	0.9166666666666666
12	Z-Score	7	best	0.9333333333333332
13	Z-Score	7	random	0.9583333333333334
14	Z-Score	9	best	0.9166666666666667
15	Z-Score	9	random	0.9583333333333334
16	MinMax	3	best	0.9416666666666668
17	MinMax	3	random	0.95
18	MinMax	5	best	0.9333333333333332
19	MinMax	5	random	0.9333333333333332
20	MinMax	7	best	0.9333333333333332
21	MinMax	7	random	0.925
22	MinMax	9	best	0.9083333333333332
23	MinMax	9	random	0.9333333333333333

2 Discussion

Initially when the data was processed, before cross validation, it included outliers and inconsistent results at various depths. This inconsistency is shown in the fluctuating accuracy scores. For example, in some cases MinMax "random" at depth 9 proved the to be the best configuration, then Z-score "random" at depth 9 proved to be the new best configuration with variance from as low as 76% accuracy all the way up to 99% accuracy.

One step that helped get better performance was cross validation. Running through all the processes multiple times across different folds and getting our average results helped produce better accuracy percentages across each technique. This step helped in mitigating the variability initially observed and in identifying the most effective configurations.

In analyzing the impact of decision tree depth, it became apparent in our final test in the report that depth 7 seems to give the most accurate results. From our observations, using a depth of 7 seems to be the best in avoiding overfitting and identifying patterns in the data.

As for the splitter, the "best" splitter seemed to provide more consistent results whereas the "random" splitter introduced more variability in our results. As evident in our table, the choice of the most appropriate splitting technique depends on the preprocessing technique and tree depth.

Overall, the best configuration seems to be Z-score "best" at depth 7. This configuration consistently provided high accuracy across runs.