COMP8320
Data Mining and Knowledge Discovery

Assessment 2: Practical Data Analysis

Luke Kenny
lgk7

Word count: 500

Part A: Default J48 (Weka) parameters resulted in an accuracy of 98.1% for the training set and 95.1% for cross-validation (CV) (**Table 1**). The aggressiveness of the pruning process, determined by the confidenceFactor (-C), yielded the highest CV accuracy (95.3%) within the range of 0.3-0.5 (**Figure 1**). A lower value of -C (0.3) was considered optimal across runs #2-10 for its interpretability, resulting in 16 leaves.

Adjustments to minNumObjects (-M) in the range of 1-30 were tested against -C = 0.25 (default) and 0.3 (**Figure 2**). Deviating from the default -M value did not improve accuracy in either the training sets or CV results.

Pruned trees (-U = 'TRUE') with varied -M values showed lower CV accuracy overall, whilst the highest training set accuracy (99.9%) was achieved with -M = 1, i.e., overfitting.

ReducedErrorPruning (-R = 'TRUE') and adjusting numFolds (-N) within the range of 1-10 improved generalisation. -N = 4 splits the dataset into four folds (three for training, one for testing) and achieved the joint highest CV accuracy (95.3%) with four leaves making use of two attributes (**Figure 3**).

Across most runs, consistently high accuracy rates suggest low bias and minimal noise, whilst an overly complex decision boundary is not required, as indicated by the requirement for 16 leaves (at most) to support the maximum CV accuracy of 95.3%. The exceptions are the unpruned decision trees, which comprise as many as 32 leaves and markedly higher accuracy rates for the training sets versus CV results, indicating overfitting and high variance. However, this can be mitigated through pruning to improve the performance of the derived decision trees.

Part B: Of the runs achieving the highest CV accuracy, -C 0.3 -M 2 produced the larger tree (**Figure 4**). 'Uniformity of cell size' is considered at the root and on a further two occasions in sub-trees. Although this may be a significant predictor of the class variable, the narrow numerical ranges generated for this repeated attribute instead hint at redundant splits and the possibility of overfitting. Repetition of other attributes ('clump thickness' and 'bare nuclei') indicates that, even with a moderately interpretable tree size, overfitting leads to a lack of generalisation with this decision tree.

In contrast, reducedErrorPruning creates an interpretable and balanced decision tree (**Figure 5**) with four leaves considering only two attributes creating a simple decision boundary. 'Bare nuclei' forms the root node, whilst the remaining two nodes split on 'uniformity of cell size'. Both attributes were used in the larger decision tree; however, this demonstrates minimal additional value is achieved in growing the tree further through repeated use of the same attributes.

When pathologists find breast cancer, a crucial indicator of malignancy is 'how much the cells look like normal breast cells'[1], i.e., the nuclear grade, which relates to the 'bare nuclei' attribute. Similarly, pathologists will also assess size and shape of the nuclei[2] in the tumour cells to determine abnormality. These are both clearly important attributes in real-world diagnosis indicating their inclusion in **Figure 5** is sensible.

References

[1] ‘Understanding Your Pathology Report: Breast Cancer’. Accessed: Mar. 26, 2024. [Online]. Available: https://www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/breast-pathology/breast-cancer-pathology.html

[2] ‘Breast Cancer Treatment - NCI’. Accessed: Mar. 26, 2024. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq

Appendix

Table 1: Summary of Parameter Selection and Accuracy Results. Highlighted values to be interpreted as follows: green (maximum within range), red (minimum within range), and purple (value different from default parameter).

Run	minNumObj (-M)	unpruned (-U)	confFactor (-C)	reducedErrorPruning (-R)	numFolds (-N)	No. Leaves	Accuracy		
							Training Set	Cross Validation	Net Change
1	2	FALSE	0.25	FALSE	-	8	98.1%	95.1%	-3.0%
2	2	FALSE	0.10	FALSE	-	9	97.3%	94.6%	-2.7%
3	2	FALSE	0.20	FALSE	-	14	98.1%	95.0%	-3.1%
4	2	FALSE	0.30	FALSE	-	16	98.4%	95.3%	-3.1%
5	2	FALSE	0.40	FALSE	-	16	98.4%	95.3%	-3.1%
6	2	FALSE	0.50	FALSE	-	22	99.0%	95.3%	-3.7%
7	2	FALSE	0.60	FALSE	-	22	99.0%	95.0%	-4.0%
8	2	FALSE	0.70	FALSE	-	22	99.0%	95.0%	-4.0%
9	2	FALSE	0.80	FALSE	-	22	99.0%	95.0%	-4.0%
10	2	FALSE	0.90	FALSE	-	22	99.0%	95.0%	-4.0%
11	1	FALSE	0.25	FALSE	-	18	98.7%	94.7%	-4.0%
12	2	FALSE	0.25	FALSE	-	See Run #1			
13	3	FALSE	0.25	FALSE	-	12	97.9%	94.7%	-3.1%
14	4	FALSE	0.25	FALSE	-	12	97.7%	94.4%	-3.3%
15	5	FALSE	0.25	FALSE	-	10	97.3%	94.6%	-2.7%
16	6	FALSE	0.25	FALSE	-	9	97.0%	94.8%	-2.1%
17	7	FALSE	0.25	FALSE	-	9	96.9%	95.0%	-1.9%
18	8	FALSE	0.25	FALSE	-	6	96.4%	95.1%	-1.3%
19	9	FALSE	0.25	FALSE	-	6	96.4%	94.8%	-1.6%
20	10	FALSE	0.25	FALSE	-	6	96.4%	94.6%	-1.9%
21	15	FALSE	0.25	FALSE	-	4	95.1%	94.0%	-1.1%
22	20	FALSE	0.25	FALSE	-	4	95.1%	94.0%	-1.1%
23	30	FALSE	0.25	FALSE	-	3	92.8%	93.0%	0.1%
24	1	FALSE	0.30	FALSE	-	20	99.0%	94.7%	-4.3%
25	2	FALSE	0.30	FALSE	-	See Run #4			
26	3	FALSE	0.30	FALSE	-	12	97.9%	95.0%	-2.9%
27	4	FALSE	0.30	FALSE	-	12	97.7%	94.4%	-3.3%
28	5	FALSE	0.30	FALSE	-	10	97.3%	94.8%	-2.4%
29	6	FALSE	0.30	FALSE	-	9	97.0%	94.8%	-2.1%
30	7	FALSE	0.30	FALSE	-	9	96.9%	95.0%	-1.9%
31	8	FALSE	0.30	FALSE	-	6	96.4%	95.1%	-1.3%
32	9	FALSE	0.30	FALSE	-	6	96.4%	94.8%	-1.6%

Run	minNumObj (-M)	unpruned (-U)	confFactor (-C)	reducedErrorPruning (-R)	numFolds (-N)	No. Leaves	Accuracy		
							Training Set	Cross Validation	Net Change
33	10	FALSE	0.30	FALSE	-	6	96.4%	95.0%	-1.4%
34	15	FALSE	0.30	FALSE	-	4	95.1%	94.0%	-1.1%
35	20	FALSE	0.30	FALSE	-	4	95.1%	94.0%	-1.1%
36	30	FALSE	0.30	FALSE	-	3	92.8%	93.0%	0.1%
37	1	TRUE	-	-	-	32	99.9%	94.7%	-5.2%
38	2	TRUE	-	-	-	22	99.0%	95.0%	-4.0%
39	3	TRUE	-	-	-	18	98.3%	95.1%	-3.1%
40	4	TRUE	-	-	-	16	98.0%	94.8%	-3.1%
41	5	TRUE	-	-	-	12	97.6%	94.8%	-2.7%
42	6	TRUE	-	-	-	9	97.0%	95.0%	-2.0%
43	7	TRUE	-	-	-	9	96.9%	94.8%	-2.0%
44	8	TRUE	-	-	-	8	96.4%	94.8%	-1.6%
45	9	TRUE	-	-	-	8	96.4%	94.6%	-1.9%
46	10	TRUE	-	-	-	7	96.3%	94.8%	-1.4%
47	15	TRUE	-	-	-	5	94.7%	93.6%	-1.1%
48	20	TRUE	-	-	-	5	94.7%	94.3%	-0.4%
49	30	TRUE	-	-	-	3	92.8%	93.3%	0.4%
50	2	-	-	TRUE	2	9	96.7%	93.3%	-3.4%
51	2	-	-	TRUE	3	8	96.9%	92.6%	-4.3%
52	2	-	-	TRUE	4	4	96.0%	95.3%	-0.7%
53	2	-	-	TRUE	5	8	96.0%	94.1%	-1.9%
54	2	-	-	TRUE	6	10	96.1%	93.4%	-2.7%
55	2	-	-	TRUE	7	7	96.6%	93.7%	-2.9%
56	2	-	-	TRUE	8	4	94.4%	93.0%	-1.4%
57	2	-	-	TRUE	9	3	94.3%	93.4%	-0.9%
58	2	-	-	TRUE	10	10	96.1%	93.7%	-2.4%

Figure 1: Percentage of correct classifications in training/cross-validation sets and no. leaves vs confidence factor (-C). Correct classification of CV tests peaked for -C in range of 0.3 to 0.5. For training set data, correct classifications plateaued for -C = 0.5 onwards. Higher values of -C generates a more detailed and nuanced tree, resulting in a larger tree/more leaves.

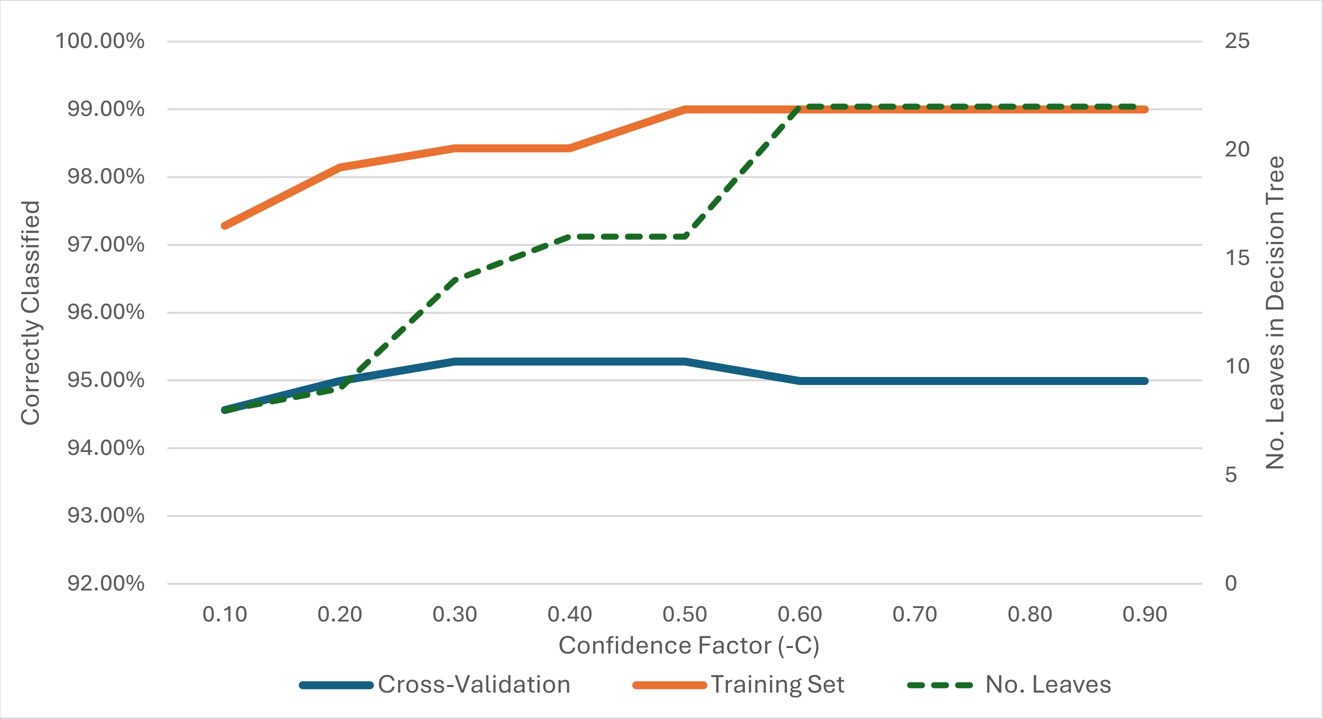
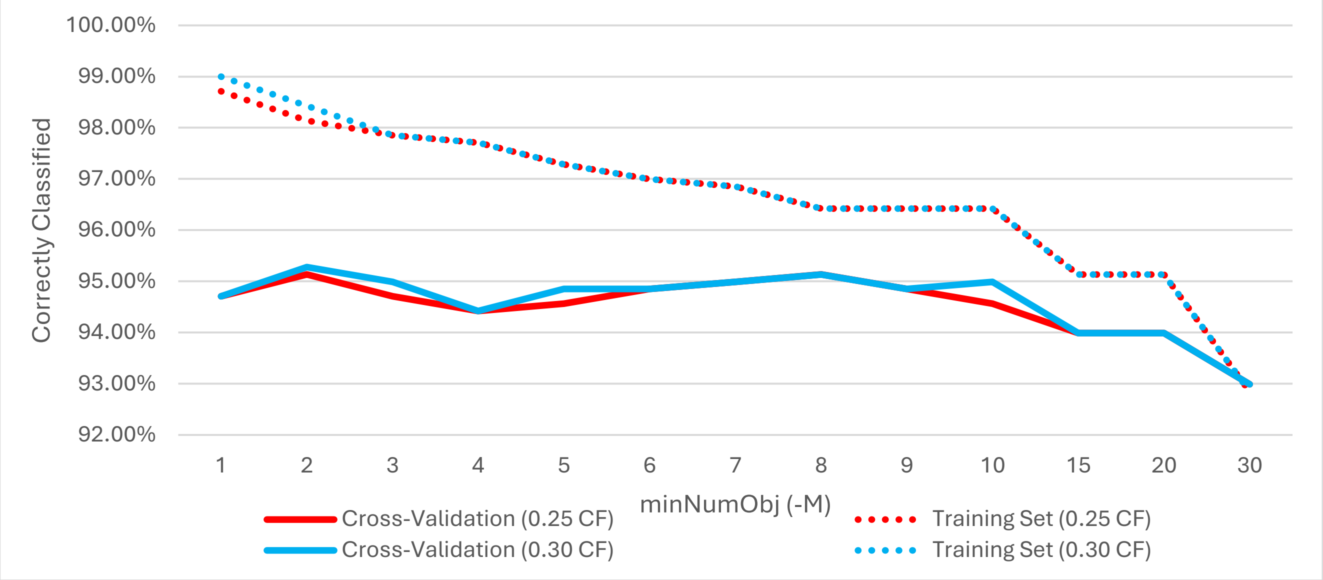


Figure 2: Percentage of correct classifications for (1) -C = 0.25 and (2) -C = -0.30 vs minNumObjects (-M). CV performs consistently for -M in range of 1-10. Beyond -M =2, training set accuracy is equal for both values of -C. As -M increases, training set and CV results converge with CV results outperforming at M = 30.



Note: When interpreting this data, it should be noted that the x-axis is not uniformly spaced.

Figure 3: Percentage of correct classifications and no. leaves vs numFolds (-N) using reducedErrorPruning (-R = 'True'). There is a weak correlation between -N and accuracy for both training set and CV results, whilst no. leaves also fluctuates noticeably for -N values. Maximum accuracy achieved for -N = 4, which simultaneously results in low leaf count.

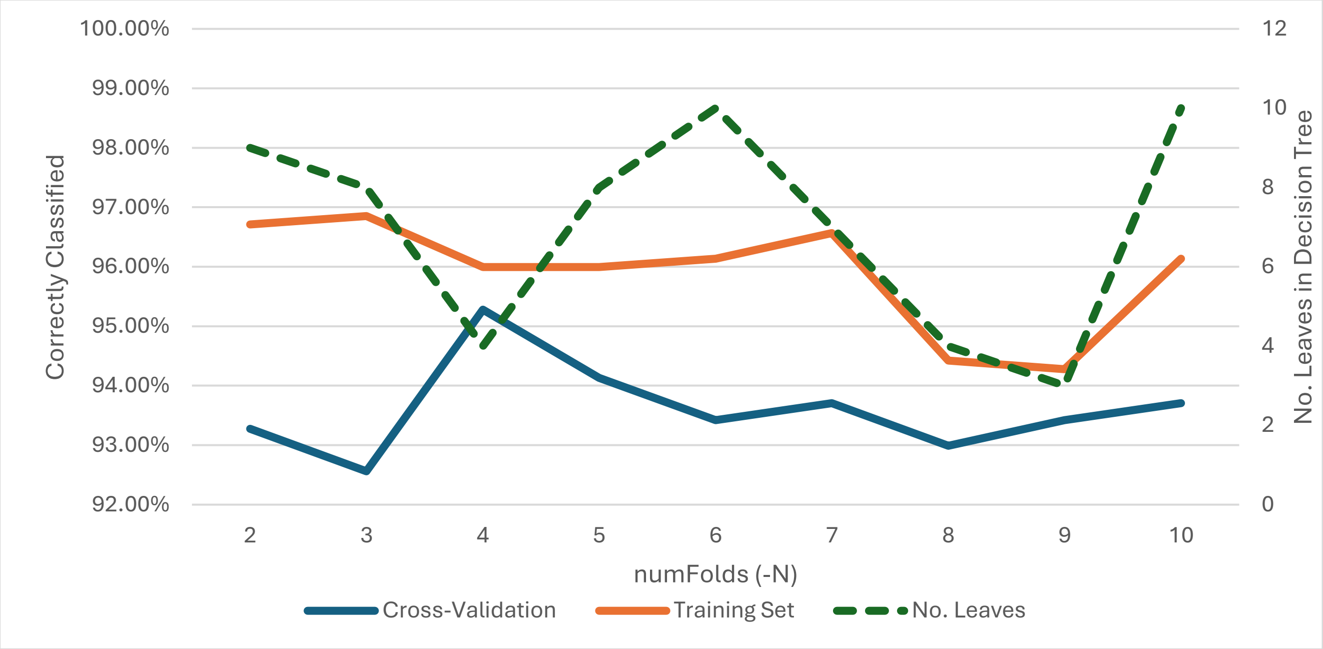


Figure 4: J48 Decision Tree for Run #4 (-C 0.3 -M 2) with 16 leaves and a tree size of 31. Shows repeated attributes ('uniformity of cell size', 'clump thickness', and 'bare nuclei' with narrow numerical ranges. Whilst moderately interpretable, this indicates overfitting and a lack of generalisation.

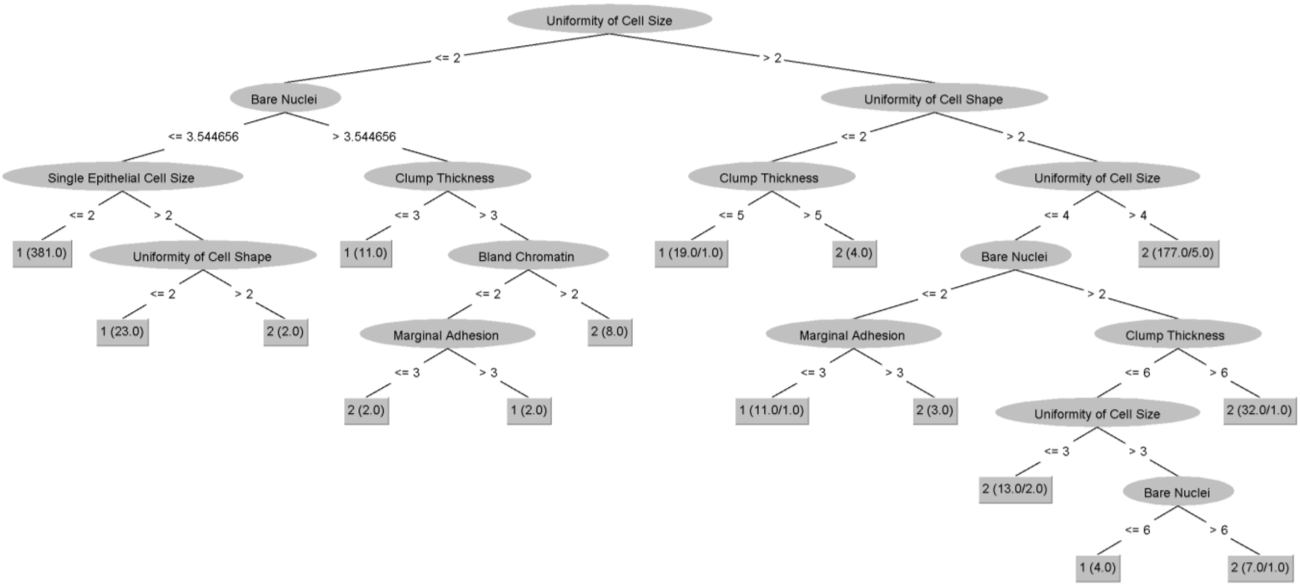
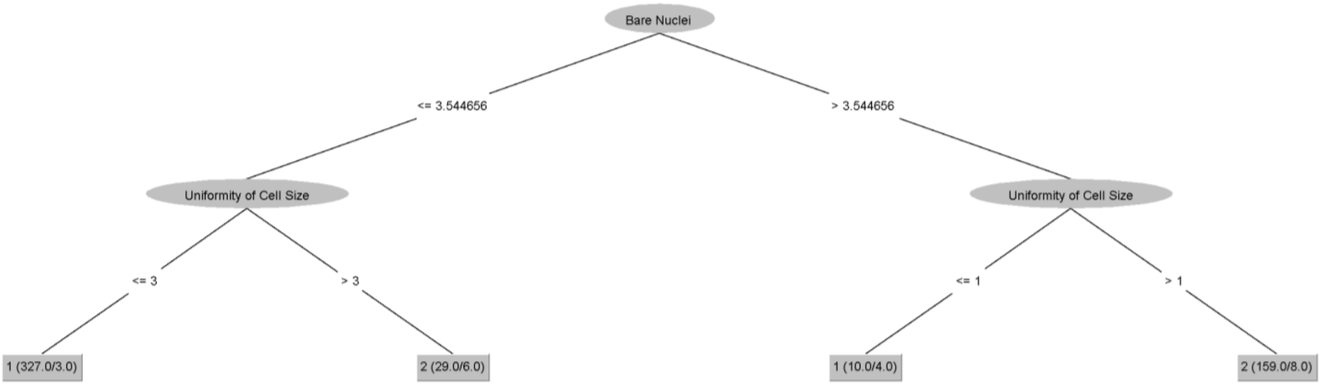


Figure 5: J48 Decision Tree for Run #52 (-R -N 4 -Q 1 -M -2) with 4 leaves and a tree size of 7. Illustrates simple decision boundary formed by attributes considered important in real-world diagnostics of malignancy.


Bare Nuclei
<= 3.544656
> 3.544656
Uniformity of Cell Size
<= 3
> 3
1 (327.0/3.0)
2 (29.0/6.0)
Uniformity of Cell Size
<= 1
> 1
1 (10.0/4.0)
2 (159.0/8.0)