Submitted by: Harsh Vinay Shah
M12401397

02-05-2018

CLASSIFICATION OF ORTHOPEDIC PATIENTS USING DECISION TREES

BACKGROUND

This data is taken from the UCI Machine Learning Repository and contains biomechanical features of some patients. The task is to predict whether the patient is normal or abnormal. The second dataset splits the abnormal group into two different diagnoses. Both datasets have same 310 feature vectors, six features, and a class column

Tools

i.   MATLAB
ii. Microsoft Excel

SOLUTION 1.A

Multiple decision tree models were created on randomly selected 210 records as training dataset from the first dataset, each having Minimum leaf size of 5, 10, 15, 20, 25 respectively. The decision tree with minimum leaf size of 25 is attached below
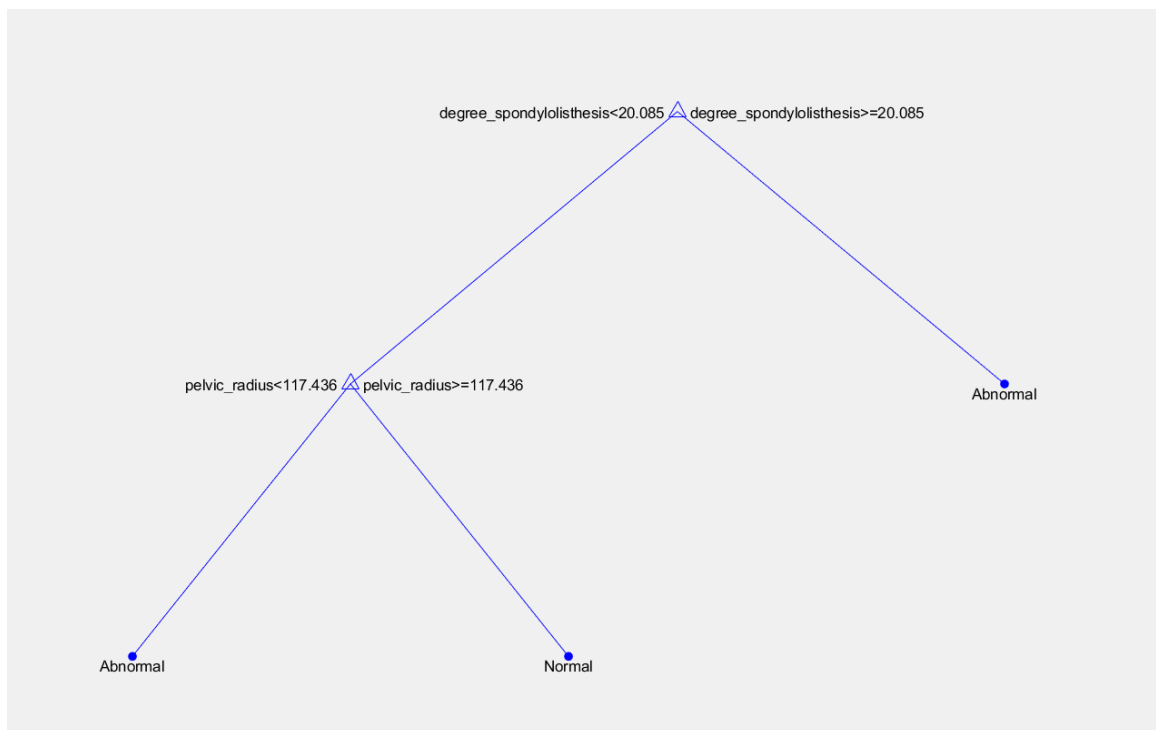
Observations in five trees

- The pruning level decreases with the increase in minimum leaf nodes. The pruning level is 6 for tree with minimum 5 leaf nodes and 3 for tree with 25 leaf nodes.
- This also causes the height of tree to increase gradually with the increase in minimum records per leaf nodes
- The threshold values of the attributes that predict the class remain constant throughout. Only additional attributes are added for segregating the tree when the minimum records per leaf node are reduced.
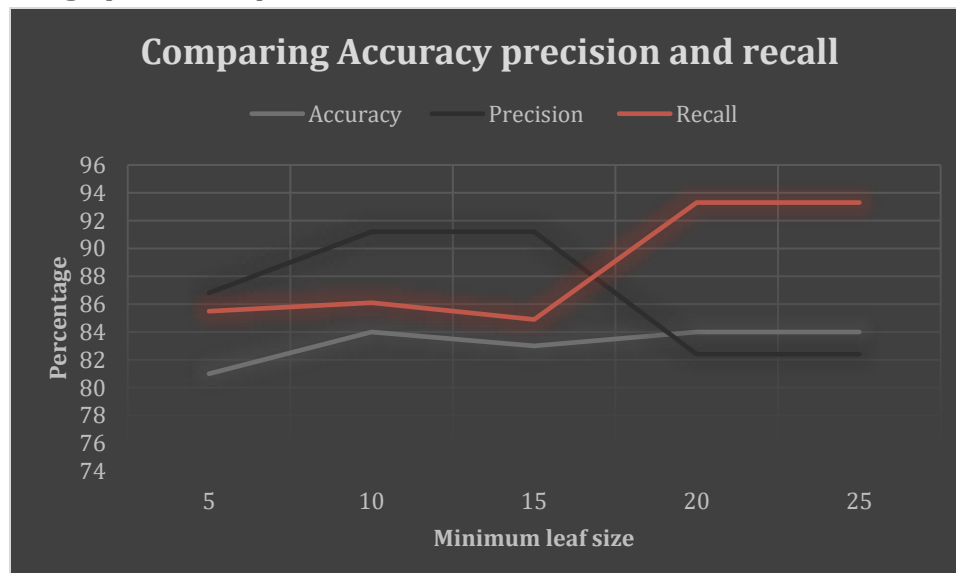
SOLUTION 1.B

Accuracy, Precision and Recall was calculated plotting confusion matrix for each decision tree. The values are listed below

Min leaf size	Accuracy	Precision	Recall
5	81	86.8	85.5
10	84	91.2	86.1
15	83	91.2	84.9
20	84	82.4	93.3
25	84	82.4	93.3

- While the accuracy of all the trees remains consistent between 83 and 84, the precision and the recall values vary
- Trees with minimum leaf size of 10 and 15 have maximum precision of 91.2% and trees with minimum leaf size of 20 and 25 have maximum recall of 93.3%

The graph below explains this further

**Comparing Accuracy precision and recall**

2

SOLUTION 1.C

Now, we take the model with minimum leaf size 10 and run it multiple time over randomly selected data. Accuracy, precision and recall for all these trails have been listed below. (All the values in percentage)

Sample	Accuracy	Precision	Recall
1	80	93.4	78.1
2	81	93.4	79.2
3	86	90.9	88.2
4	83	83.3	92.3
5	83	88.7	87.5
avg	82.6	89.94	85.06
sd	2.302173	4.197976	6.144347

- The Accuracy, precision and recall vary for each trail even though the minimum leaf size is kept constant at 10. This is because the data is randomized for each trail.
- Accuracy has the least variance among the other measures with an average of 82.9
- Precision has a standard deviation of 4.19 and average of 89.94
- Recall has most variance with a standard deviation of 6.14 and average of 85.6

SOLUTION 2.A
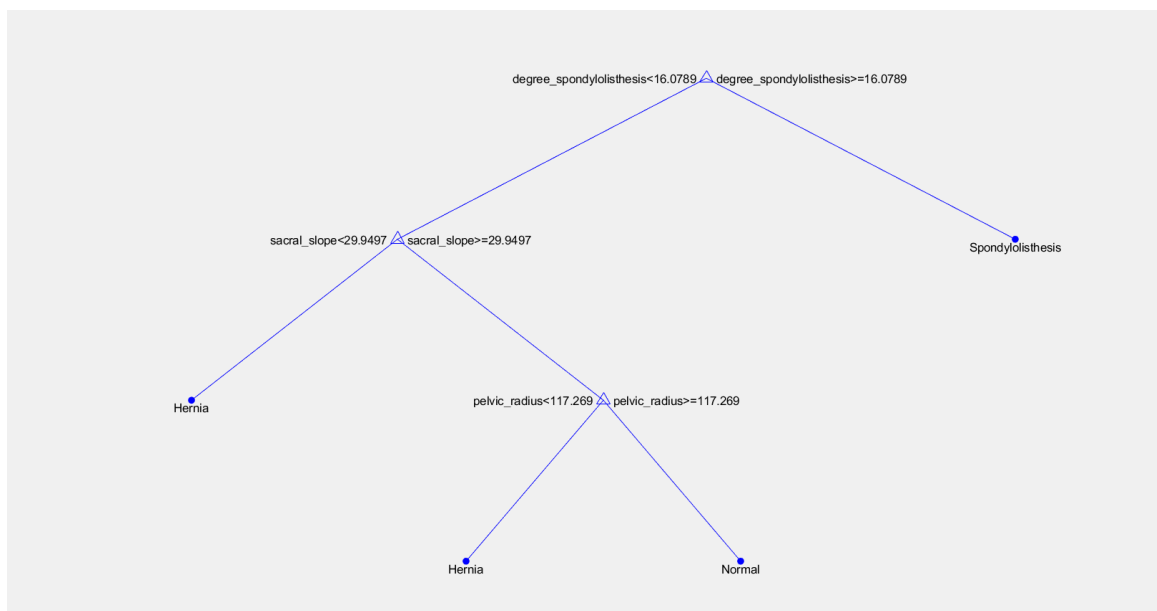
The decision tree with min leaf size of 25 is shown above.

Observations:

- As before, The pruning level decreases with the increase in minimum leaf nodes. The pruning level is 6 for tree with minimum 5 leaf nodes and 3 for tree with 25 leaf nodes.
- This also causes the height of tree to increase gradually with the increase in minimum records per leaf nodes
- The threshold values of the attributes that predict the class remain constant throughout. Only additional attributes are added for segregating the tree when the minimum records per leaf node are reduced.
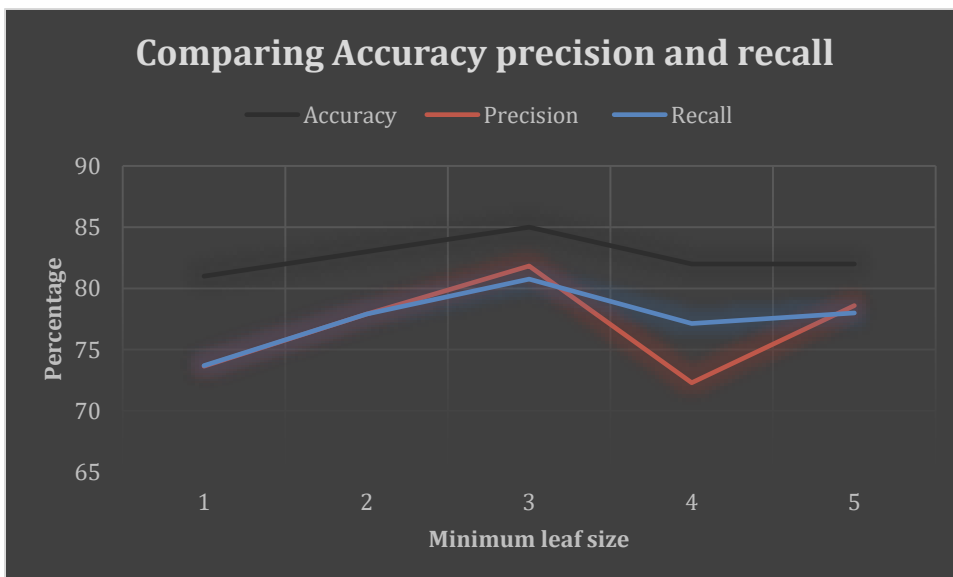
SOLUTION 2.B

Accuracy, precision and recall for trees with different minimum leaf size is mentioned below

Min leaf size	Accuracy	Precision	Recall
5	81	73.65	73.7
10	83	77.9	77.9
15	85	81.83	80.76
20	82	72.3	77.13
25	82	78.6	78

The combination of Accuracy, precision and recall is best for tree with leaf size 15.

Below is a chart for comparison

SOLUTION 2.C

4

Now, we take the model with minimum leaf size 10 and run it multiple time over randomly selected data. Accuracy, precision and recall for all these trails have been listed below. (All the values in percentage)

Sample	Accuracy	Precision	Recall
1	77	71.73333	75.33
2	78	72.03333	72.43
3	82	76.43333	78.5
4	83	78.26667	78
5	87	84.86667	83.87
Average	81.4	76.66667	77.63
Standard Deviation	4.03733	5.377267	4.245

- The accuracy gradually increases with every trail and has an average of 81.4% and standard deviation of 4.03
- The precision and recall also increases gradually with every trail and has an average of 76.67 and 77.3 respectively and standard deviation of 5.37 and 4.24 respectively

SOLUTION 2.D

- Comparing 2.C and 1.C we observe that the accuracy is pretty much the same around 82%
- The average precision and recall drop greatly from 89% to 76% and 85% to 77%

SOLUTION 3.A

	pelvic_incidence	pelvic_tilt numeric	lumbar_lordosis_angle	sacral_slope	pelvic_radius	degree_spondylolisthesis
0	26.1479 – 46.4264	-6.5549 - 10.6599	14 - 37	13.3669 - 33.3407	70.0826 - 110.7090	-11.0582 - 1.5947
1	46.4421 - 58.7825	10.6887 - 16.4175	37 - 49.7755	33.3664 - 42.4362	110.7099 - 118.3433	1.6307 - 12.0726
2	58.8284- 70.9527	16.4208 - 22.1818	50 - 63	42.4471 - 52.7447	118.3634 - 125.4802	12.3826 - 41.4228
3	73.6360 - 129.8340	22.2185 - 49.4319	63.0147 - 125.7424	52.8831 - 121.4296	125.5936 - 163.0710	41.6855 - 418.5431

- The first column indicates the discrete intervals

- All the rows are sorted in ascending order

SOLUTION 3.B

Checking accuracy, precision and recall for 5 trails.

Sample	Accuracy	Precision	Recall
1	88	91.8	91.8
2	85	93	86.8
3	87	93.9	87.3
4	80	84.1	84.1
5	83	90.1	86.5
Average	84.6	90.58	87.3
Standard Deviation	3.20936	3.891915	2.801

- The difference in accuracy, precision and recall is very less. Hence the standard deviation is comparatively lesser.
- The average values are also similar to actual corresponding values.

SOLUTION 3.C

Comparison with 1.C results in the following

- The standard deviation of these values is smaller and consistent in comparison to those obtained in 1.C
- The average values are like average values in 1.C
- As the values are discretized, data elements have similar values and hence the standard deviation is low.