# Implementation details

## Splitting criteria

I used information-gain as the splitting criteria. For continuous attributes, I try to split the data into 2 subsets, trying all the values mid-way between 2 consecutive values in the samples available at that node.

## Missing attributes

I have implemented the code to handle missing attributes, however it wasn't tested since all the datasets I used did not have missing attributes in them.

## Pruning methods

### Reduced-error pruning

This pruning method always resulted in increase in classification accuracy in the test data.

### Pessimistic pruning

Pessimistic pruning resulted in greater reduction in the tree size, however the pruned-accuracy marginally reduced (~1%) for few cases and increased for few cases. The possible reasoning for the reduction in accuracy could be that this method allows for some error (standard-error) while pruning the nodes.

# Sample tree

Indentations are used to represent sub-trees; each node is represented as (split attribute index) = (value) for that sub-tree. '-->' is used to represent a class if classification was successful.

20 = f

9 = f

32 = f

31 = f

5 = f

34 = f

14 = b

0 = f

-->nowin

0 = t

13 = f

-->won

13 = t

-->nowin

14 = n

0 = f

30 = f

-->nowin

30 = t

19 = f

11 = f

4 = f

-->nowin

4 = t

-->nowin

11 = t

-->nowin

19 = t

-->nowin

0 = t

13 = f

-->won

13 = t

-->nowin

14 = w

-->won

34 = t

14 = b

0 = f

-->nowin

0 = t

-->won

14 = n

22 = f

11 = f

-->won

11 = t

12 = g

29 = f

8 = f

25 = f

-->won

25 = t

35 = n

-->won

35 = t

-->nowin

8 = t

-->won

29 = t

-->nowin

12 = l

-->won

22 = t

16 = f

-->nowin

16 = t

-->won

14 = w

-->won

5 = t

-->won

31 = t

1 = f

-->nowin

1 = t

17 = f

23 = f

-->won

23 = t

-->nowin

17 = t

-->nowin

32 = t

17 = f

6 = f

26 = f

15 = f

3 = f

8 = f

24 = f

29 = f

2 = f

19 = f

-->won

19 = t

-->nowin

2 = t

-->nowin

29 = t

-->nowin

24 = t

-->nowin

8 = t

-->nowin

3 = t

-->nowin

15 = t

-->nowin

26 = t

-->nowin

6 = t

-->nowin

17 = t

33 = f

6 = f

3 = f

15 = f

2 = f

-->won

2 = t

-->nowin

15 = t

-->nowin

3 = t

-->nowin

6 = t

-->nowin

33 = t

-->nowin

9 = t

-->nowin

20 = t

-->won

# Results

Datasets were chosen to have a good mix based on the following characteristics:

1. the number of samples available
2. the number of attributes
3. continuous vs discrete attributes

## Accuracies with reduced-error pruning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy before pruning (%)** | **Accuracy with pruning (%)** | **95% confidence interval** | **Accuracy of majority classifier (%)** | **Average tree size before pruning** | **Average tree size after pruning** |
| **Car Evaluation Data Set** | 90.9063 | 89.1379 | 0.8618 to 0.9208 | 70.3278 | 265 | 261.6 |
| **Iris Data Set** | 84.2302 | 86.6429 | 0.7566 to 0.9762 | 47.5931 | 121.8 | 121 |
| **Chess (King-Rook vs. King-Pawn) Data Set** | 99.1217 | 99.3371 | 0.9877 to 0.9989 | 52.0712 | 87.6 | 85.6 |
| **Tic-Tac-Toe Endgame Data Set** | 83.9206 | 85.7298 | 0.8125 to 0.902 | 63.6265 | 203.7 | 196.9 |
| **Wine Data Set** | 59.4167 | 64.3674 | 0.5012 to 0.7861 | 47.7652 | 195.8 | 194.6 |

95% confidence interval (in the context of our problem) is a range of probability values for an arbitrary sample to get correctly classified; the probability (for classification) will be in this range with a probability of 95%.

## Accuracies with pessimistic pruning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy before pruning (%)** | **Accuracy with pruning (%)** | **Accuracy of majority classifier (%)** | **Average tree size before pruning** | **Average tree size after pruning** |
| **Car Evaluation Data Set** | 91.0162 | 91.0649 | 70.3278 | 265 | 157.4 |
| **Iris Data Set** | 84.2302 | 82.119 | 47.5931 | 121.8 | 108.4 |
| **Chess (King-Rook vs. King-Pawn) Data Set** | 99.1217 | 98.9531 | 52.0712 | 87.6 | 74.2 |
| **Tic-Tac-Toe Endgame Data Set** | 83.9877 | 83.3055 | 63.6265 | 203.7 | 168 |
| **Wine Data Set** | 59.4167 | 63.2083 | 47.7652 | 195.8 | 182.6 |