**ASSIGNMENT I**

**Comparison of two decision tree induction methods on UCI**

**datasets using (stratified) 10- fold cross-validation.**

**SUBMITTED TO**

**DR BYRON GAO**

**DEPARTMENT OF COMPUTER SCIENCE**

**TEXAS STATE UNIVERSITY**

**1.Data Sets:**

1. Balance Evaluations
2. Breast-cancer
3. Chronic\_Kidney\_disease
4. Contact\_Lenses
5. Glass
6. Ionosphere
7. Iris
8. Seismic-Bumps

9. Soybean

10. Thoracic\_Surgery

11. EEG Eye State Data Set

**2.Table of Results:**

1. Balance Evaluations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 625  625 | 5  5 | 5  5 | 2  4 | Correctly -32.8  Incorrectly-67.2  Correctly -6.08  Incorrectly-93.92 |

1. Breast-cancer

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 286  286 | 10  10 | 2  2 | 2  3 | Correctly -75.52  Incorrectly-24.47  Correctly -66.78  Incorrectly-33.21 |

1. Chronic\_Kidney\_disease

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 400  400 | 25  25 | 2  2 | 5  3 | Correctly -99  Incorrectly-1  Correctly -95.55  Incorrectly-4.5 |

1. Contact\_Lenses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 24  24 | 5  5 | 3  3 | 3  4 | Correctly -83.33  Incorrectly-16.67  Correctly -70.83  Incorrectly-29.17 |

1. Glass

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 214  214 | 10  10 | 7  7 | 10  13 | Correctly -67.28  Incorrectly-32.71  Correctly -70.09  Incorrectly-29.90 |

1. Ionosphere

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 351  351 | 35  35 | 2  2 | 10  11 | Correctly -91.45  Incorrectly-8.547  Correctly -87.74  Incorrectly-12.25 |

1. Iris

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 150  150 | 5  5 | 3  3 | 4  5 | Correctly -96  Incorrectly-4  Correctly -92  Incorrectly-8 |

1. Seismic-Bumps

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height | Average  Accuracy(%)/  Classified |
| C4.5  Random | 2584  2584 | 19  19 | 2  2 | 14  25 | Correctly -90.23  Incorrectly-9.75  Correctly -88.74  Incorrectly11.24 |

9.Soybean

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height | Average  Accuracy(%)/  Classified |
| C4.5  Random | 683  683 | 36  36 | 19  19 | 13  14 | Correctly -91.36  Incorrectly-8.68  Correctly -84.04  Incorrectly-15.96 |

10.Thoracic\_Surgery

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 470  470 | 17  17 | 2  2 | 12  18 | Correctly -80.42  Incorrectly-19.57  Correctly -74.89  Incorrectly-25.01 |

11 EEG Eye State Data Set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Number of  Instances | Number of Attributes | Number of Classes | Average  Height  (Manually calculated) | Average  Accuracy(%)/  Classified |
| C4.5  Random | 14980  14980 | 15  15 | 2  2 | 33  28 | Correctly -84.38  Incorrectly-15.61  Correctly -82.77  Incorrectly-17.23 |

**3.Supporting Evidences:**

Implementation:

**WEKA:**

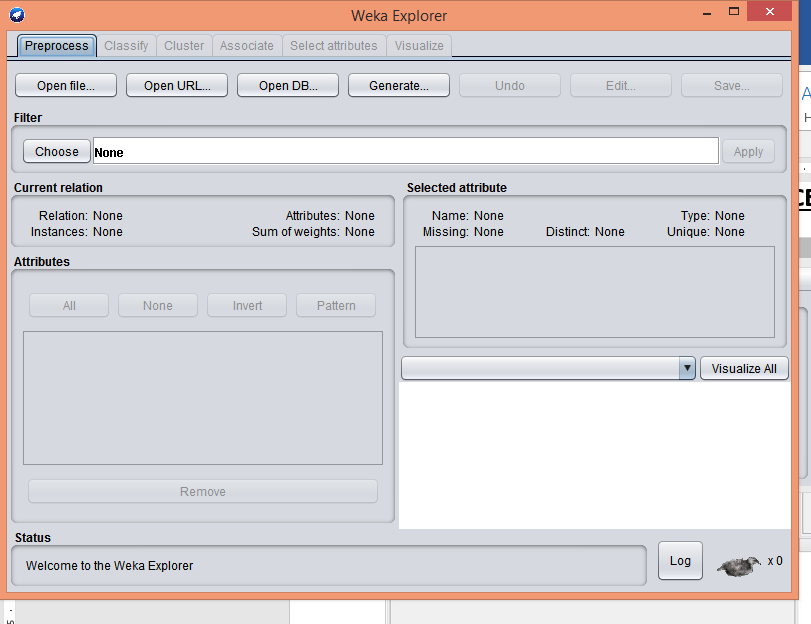
The Waikato Environment for the Knowledge Analysis Weka emerged following the perceived need for a unified workbench that would allow the researchers an easy access to the state of the art techniques in machine learning. This software came from the Waikato university of New Zealand. The Weka project aimed to provide the researchers and practitioners alike with a comprehensive collection of machine learning algorithms and data preprocessing tools. The workbench includes the algorithms for regression, classification, clustering, association rule mining and attribute selection. The preliminary exploration of data is well catered by data visualization facilities and many preprocessing tools. Weka has several graphical user interfaces that enable an easy access to the underlying functionality. The main graphical user interface is the Explorer. It has a panel-based interface, where different panels correspond to different data mining tasks. In the first panel, called the Preprocess panel, the data can be loaded and transformed by using the Weka’s data preprocessing tools called the filters. The data can be loaded from various sources, including the files, URLs, and databases. The supported file formats include the Weka’s own ARFF format and the others such as CSV, LibSVMs, C4.5s

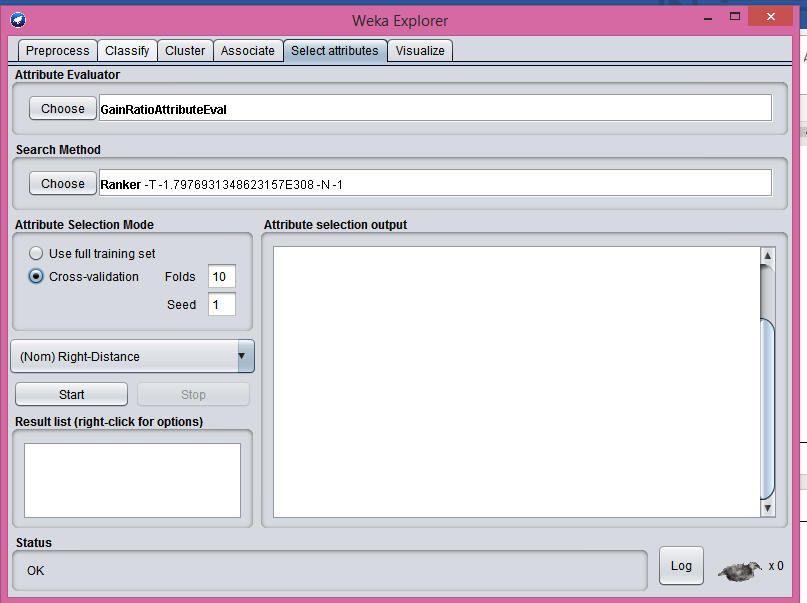
**C.45**

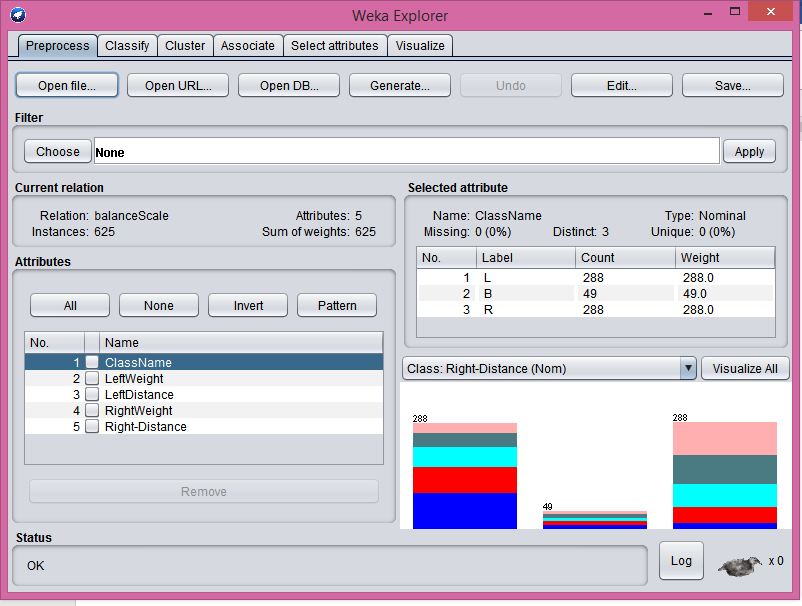
J48 is an open source and a key benefit of Weka is it’s Java implementation of the C4.5 algorithm for creating decision trees in the weka data mining tool, especially regarding tree pruning. The C4.5 algorithm is exceptionally beneficial in generating accurate decision trees. The second panel in the Explorer gives access to Weka’s classification and regression algorithms. That is called Classify because the regression techniques are viewed as the predictors continuous classes. In the classify panel we can find various types of classifier. We can click the trees entry to reveal its subentries and click J48 to choose the classifier. By double clicking on the algorithm’s name, we can change the tree options like pruning and numfolds.

**Implemented in WEKA.**

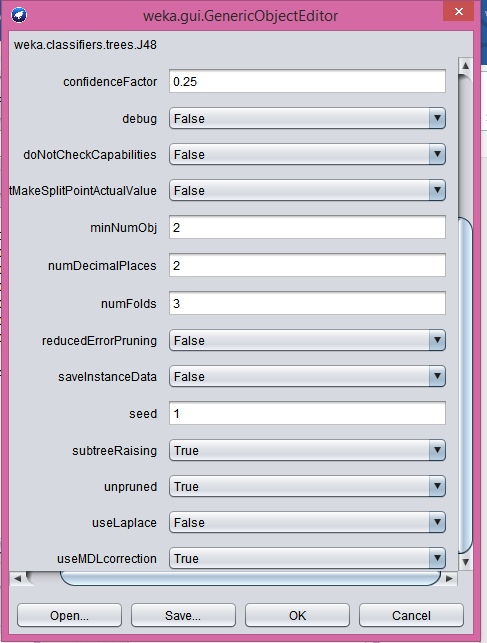




Attribute SelectionLoading DataSet

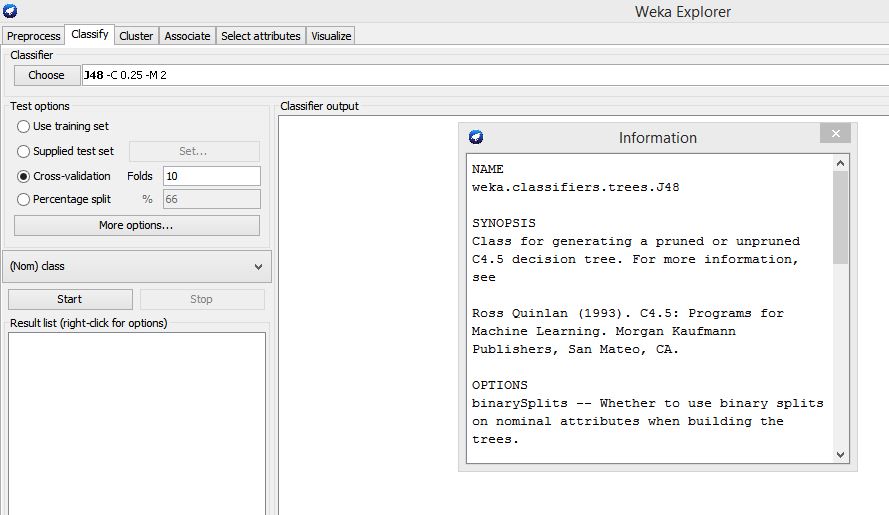


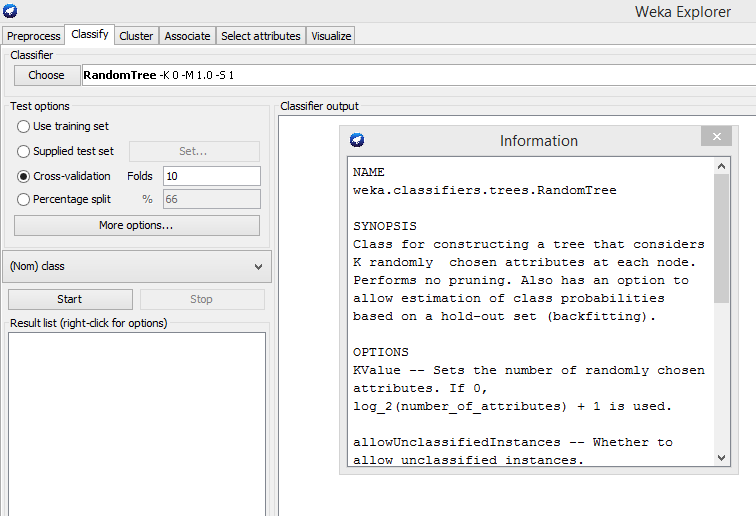
Pruning Selection



Unpruned :True

C4.5 Algorithm Selection:

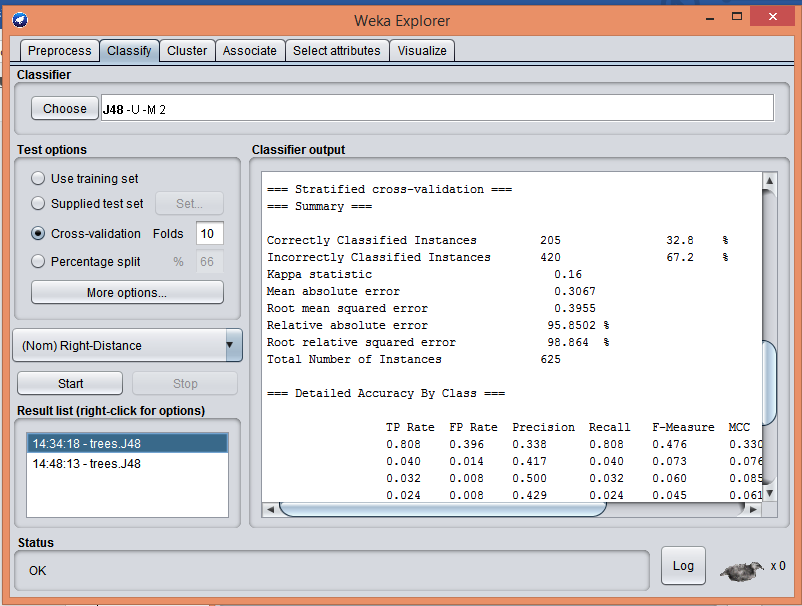


Random Tree Algorithm Selection: 

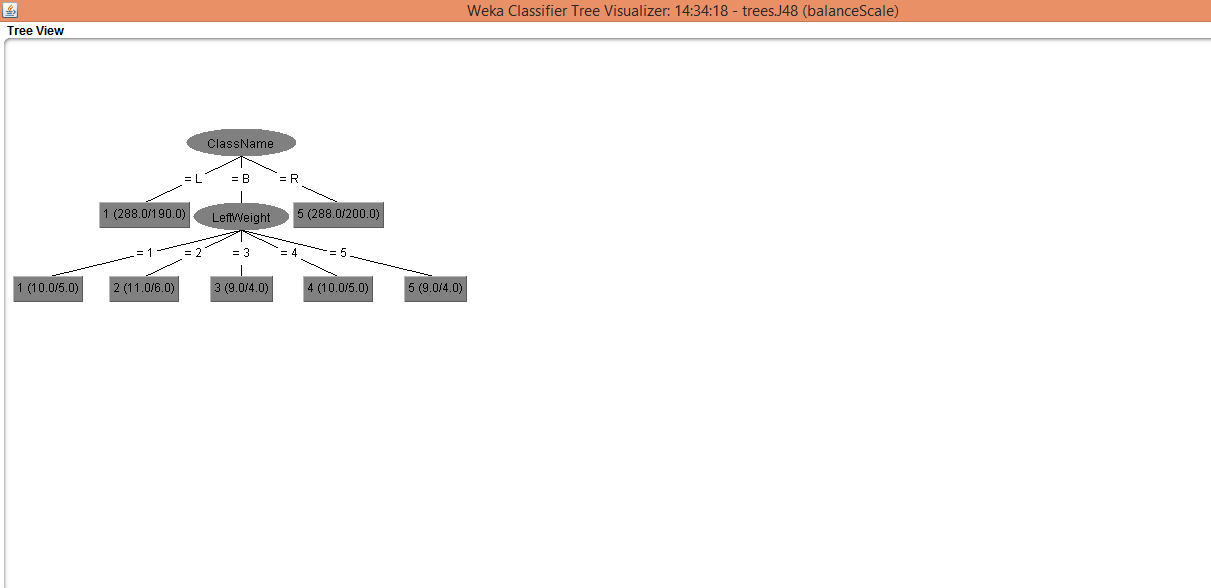
1. Balance Evaluations

Weka Environment:

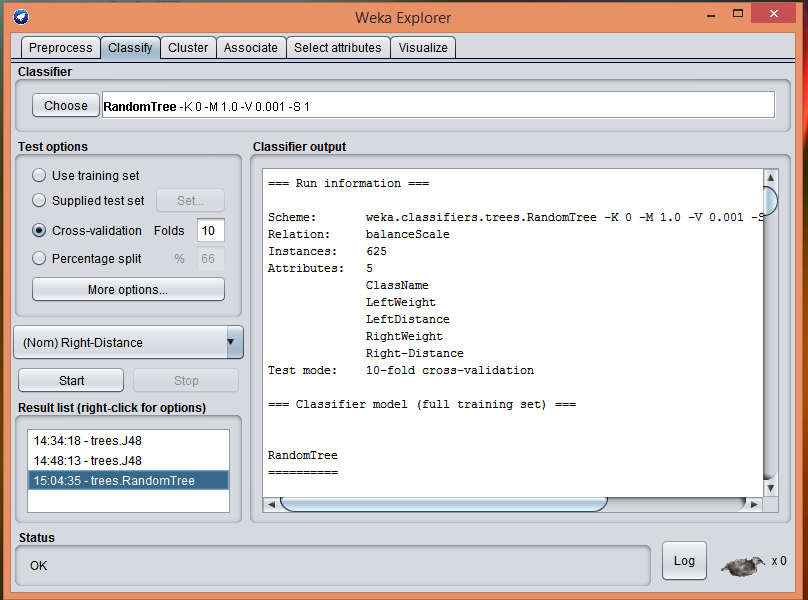
C4.5 Algorithm Selection:

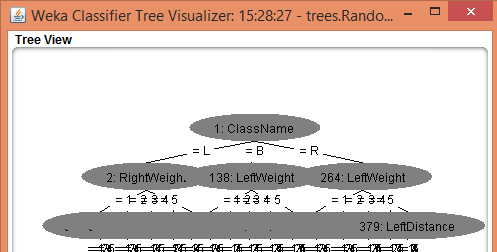


Tree:



Random Tree Algorithm Selection:



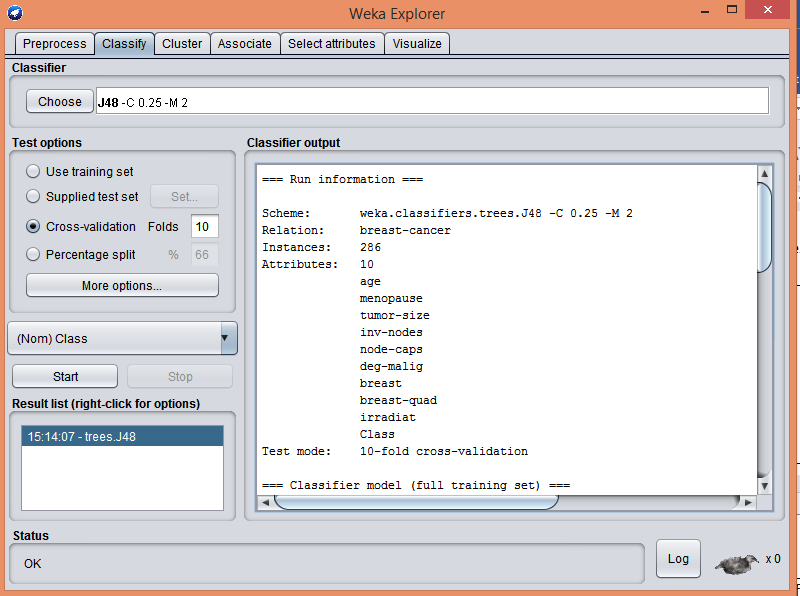


1. Breast-cancer

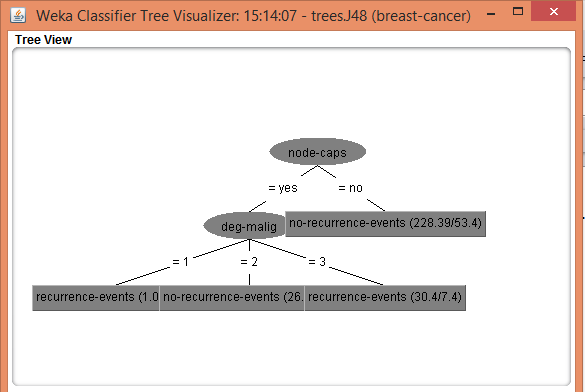
Implemented in WEKA.

Weka Environment:

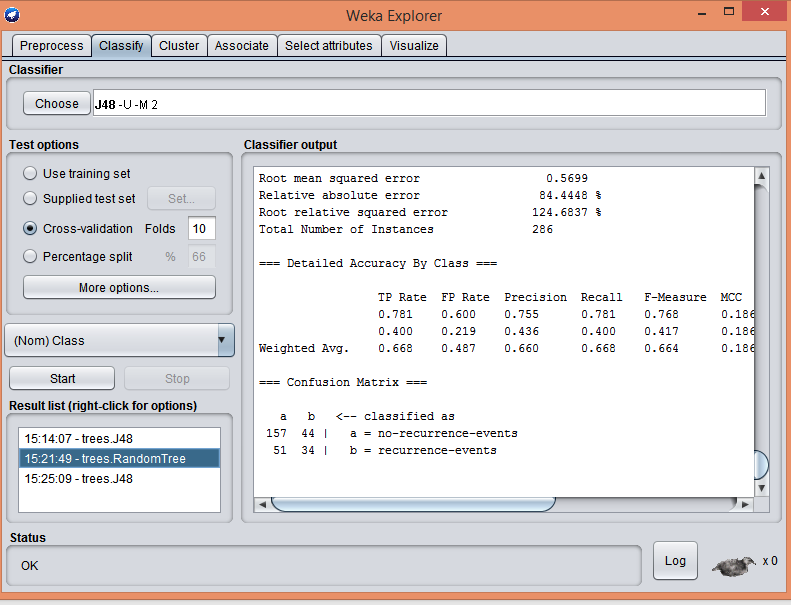
C4.5 Algorithm Selection:



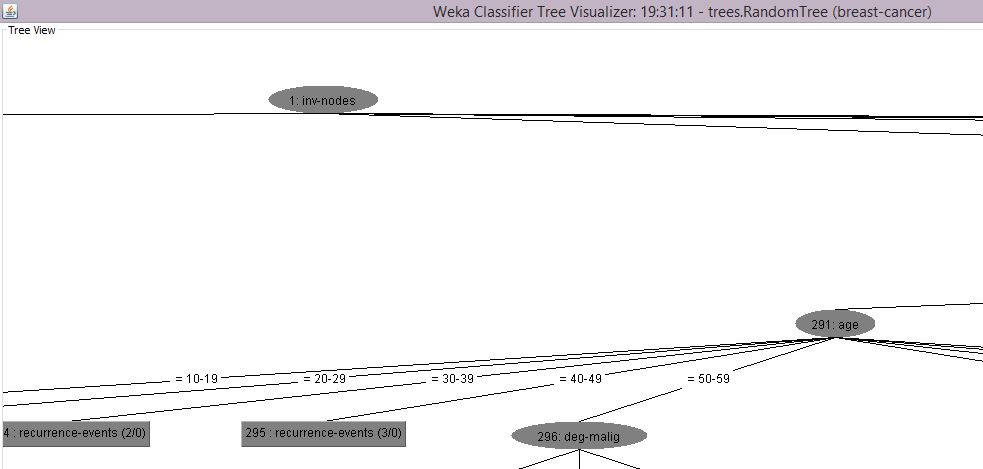
Tree



Random Tree Algorithm Selection:



Tree:

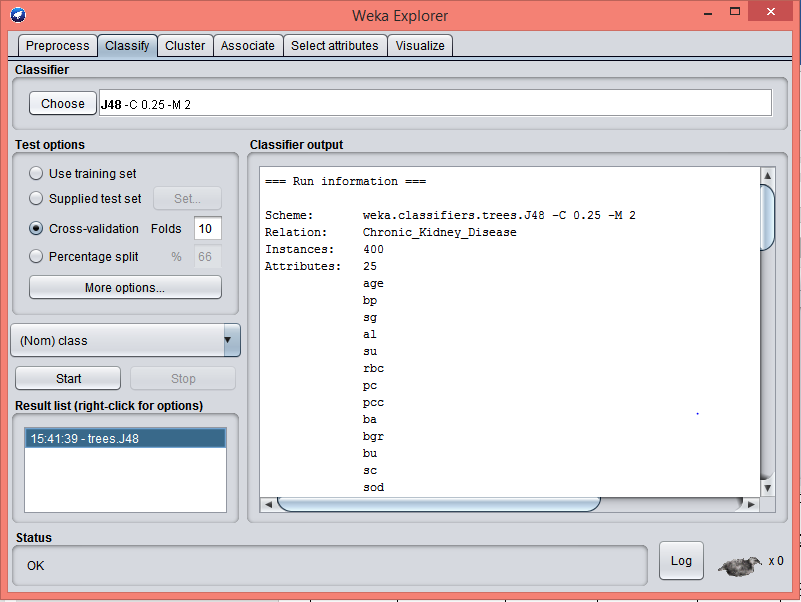


3.Chronic\_Kidney\_disease

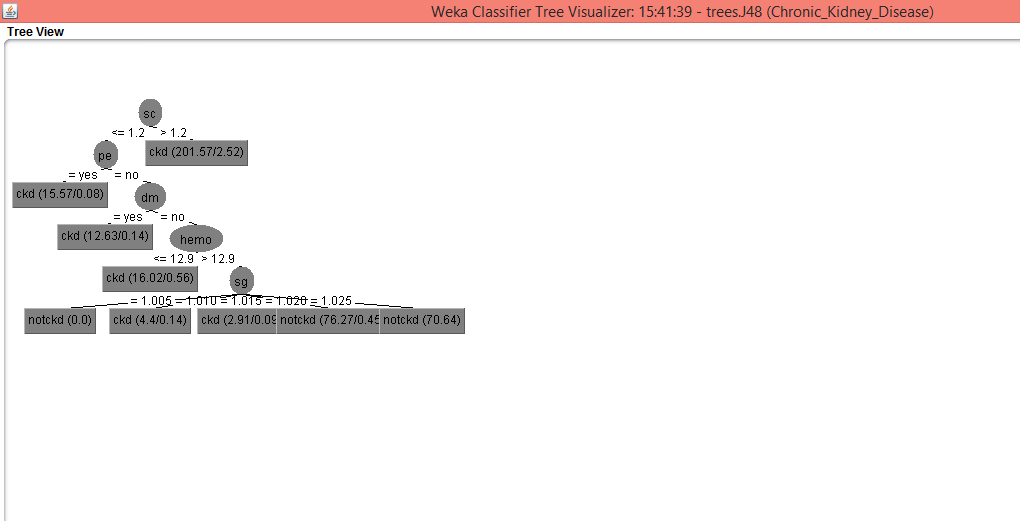
Implemented in WEKA.

Weka Environment:

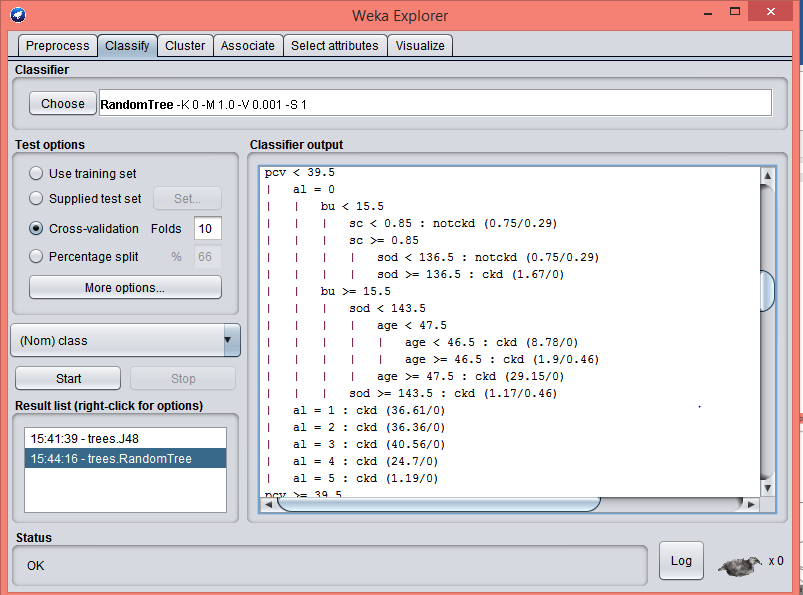
C4.5 Algorithm Selection:



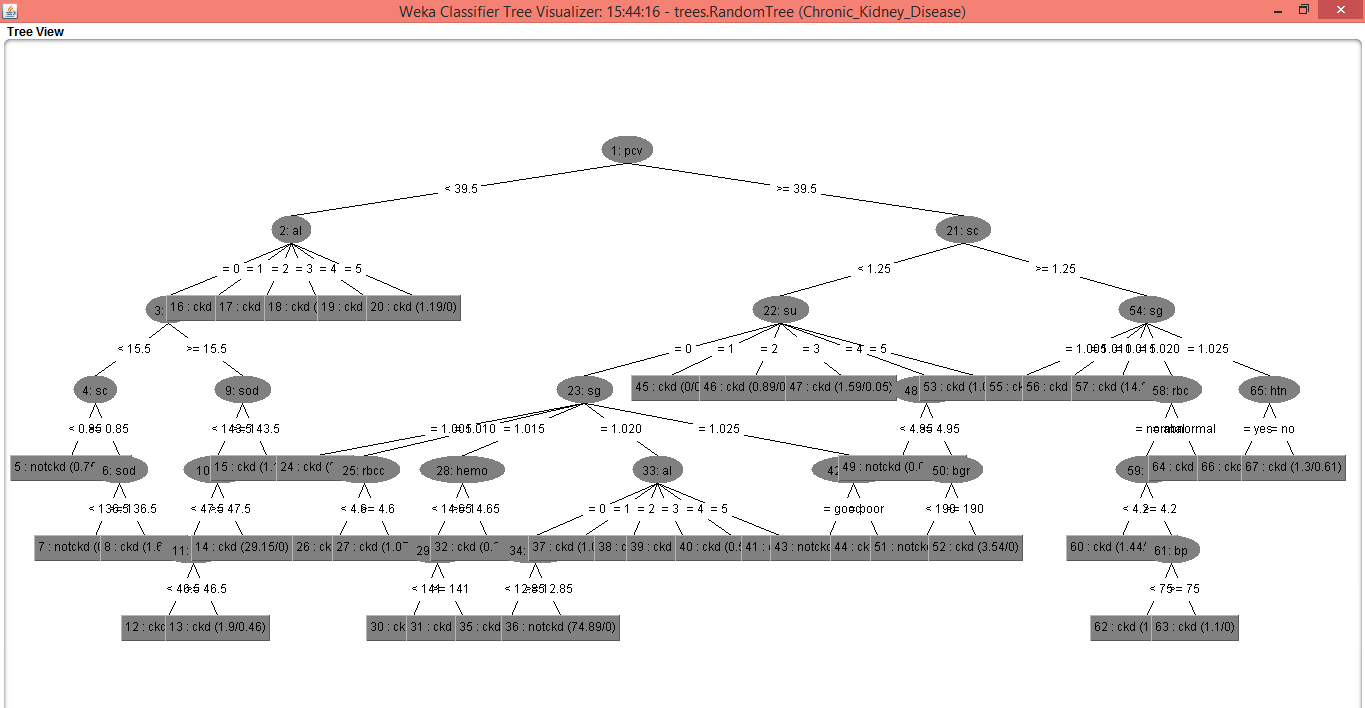
Tree



Random Tree Algorithm Selection:



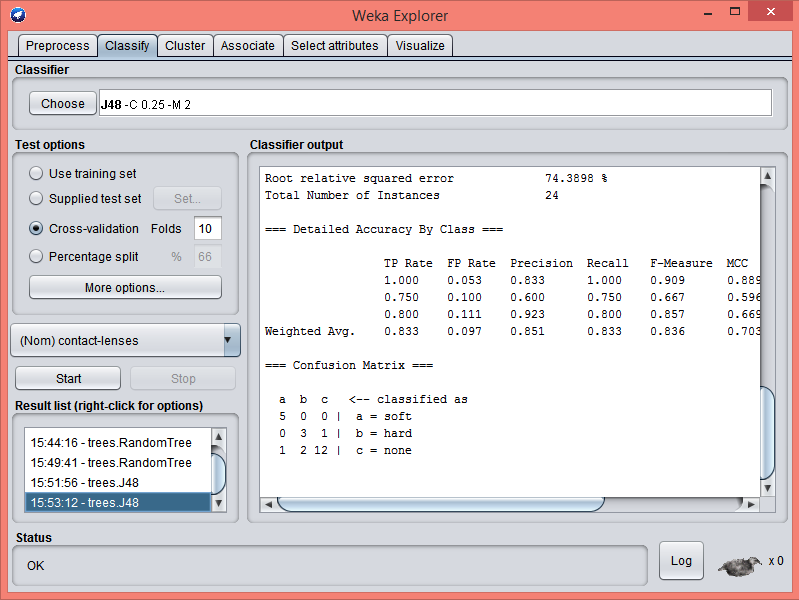
Tree



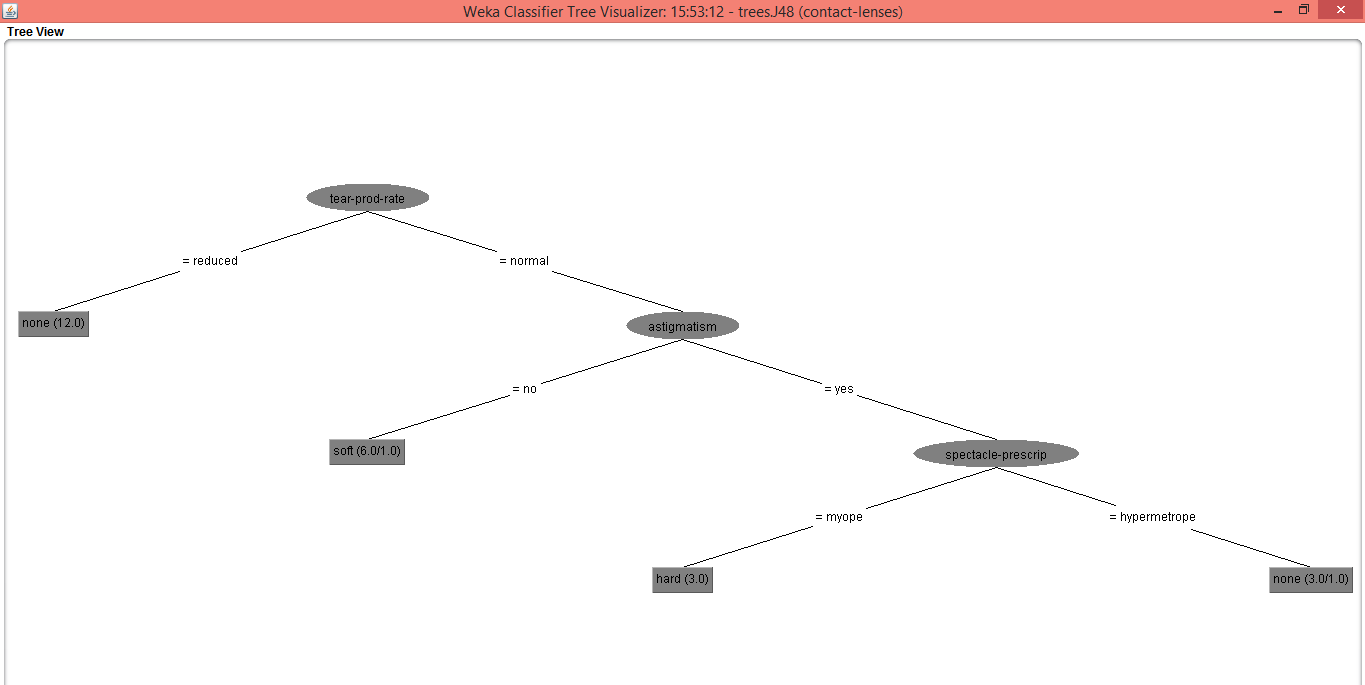
4.Contact\_Lenses

Weka Environment:

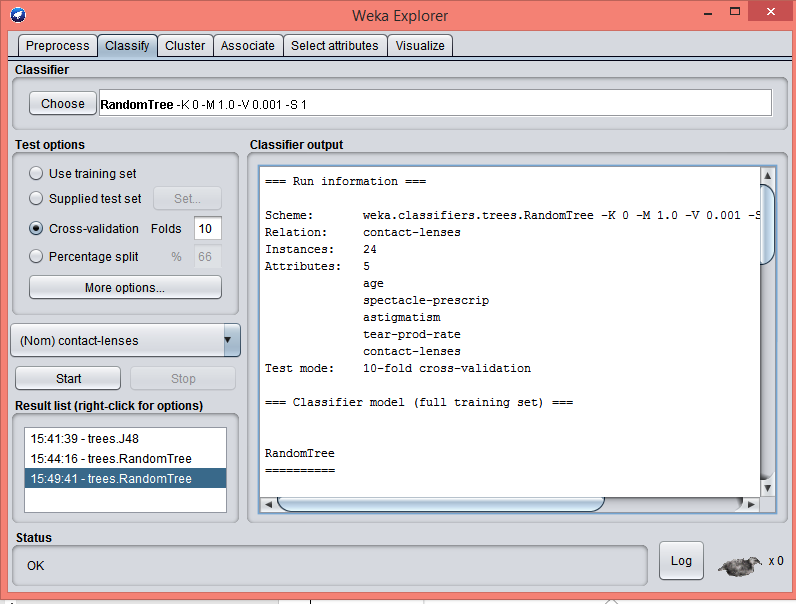
C4.5 Algorithm Selection:



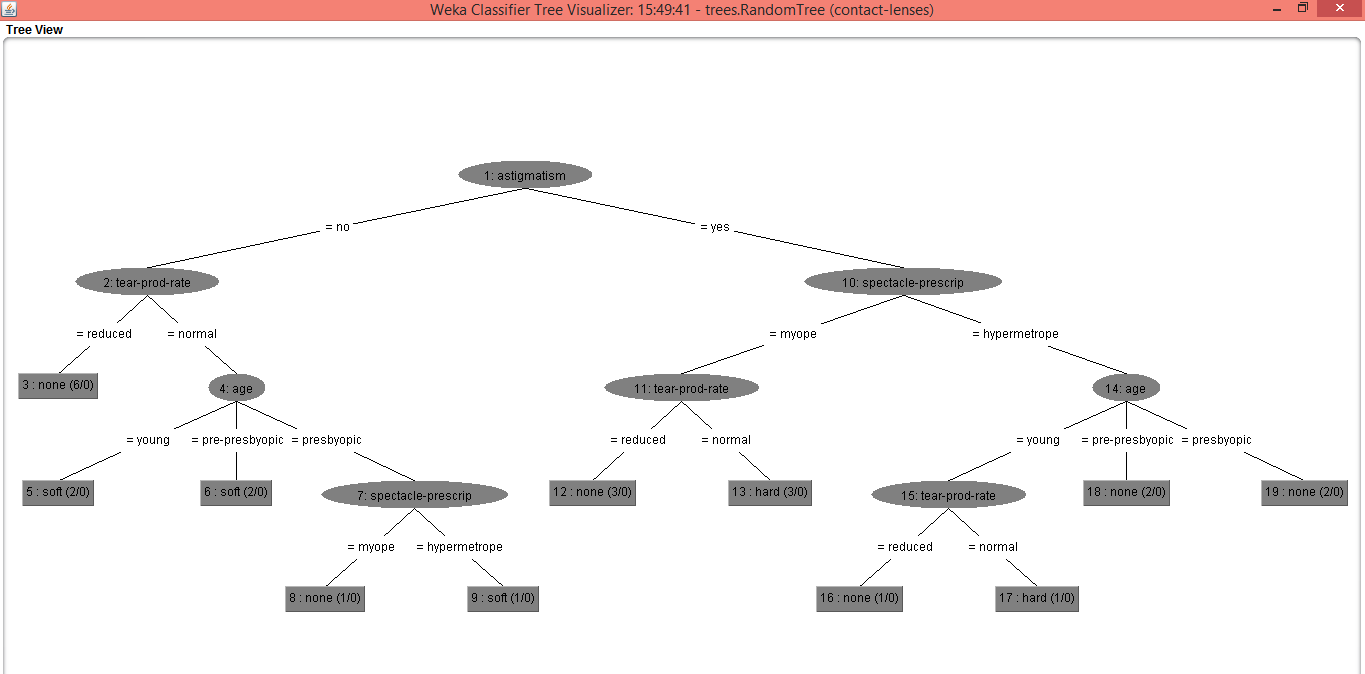
Tree



Random Tree Algorithm Selection:



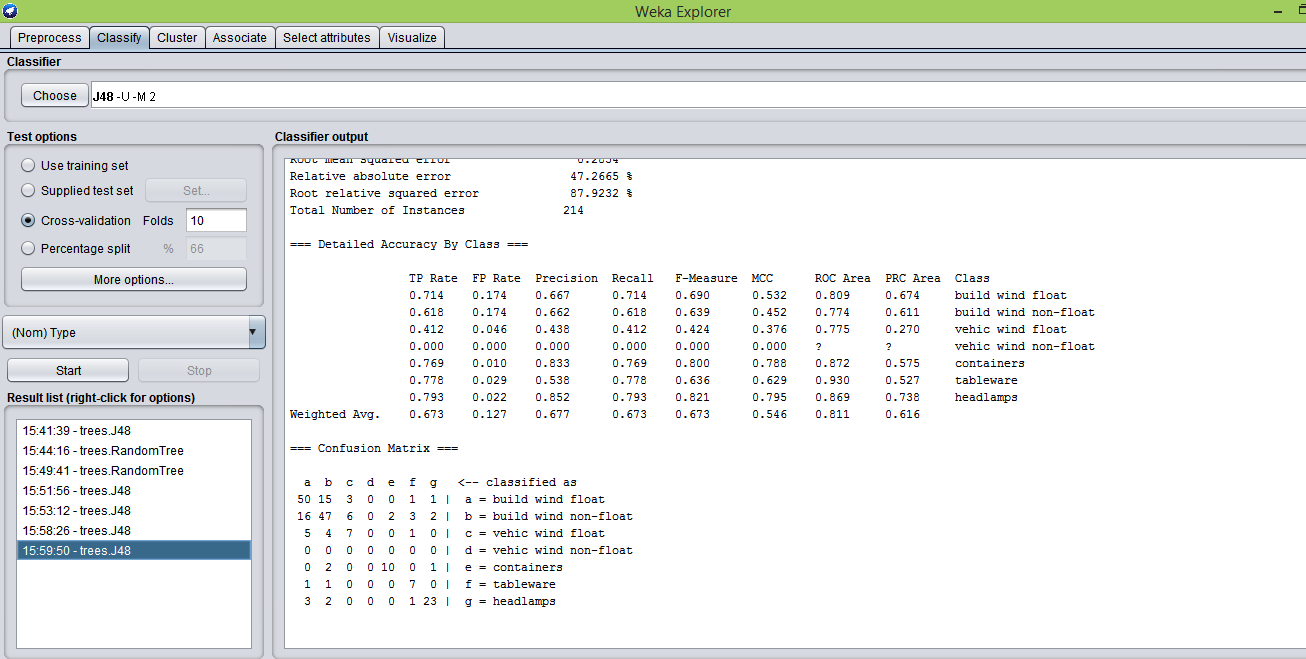
Tree



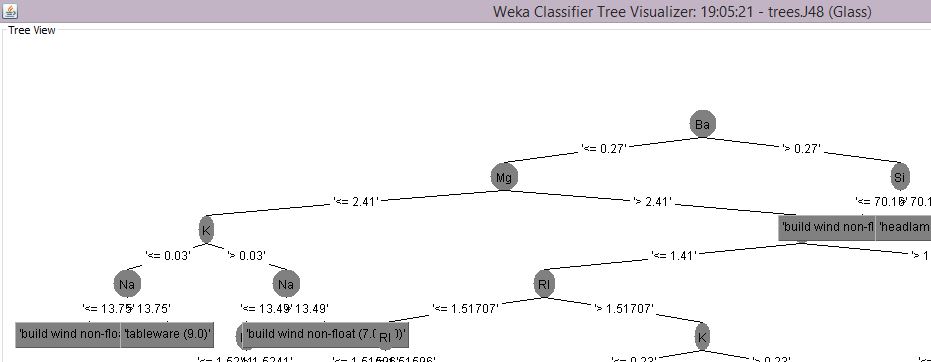
5.Glass

Weka Environment:

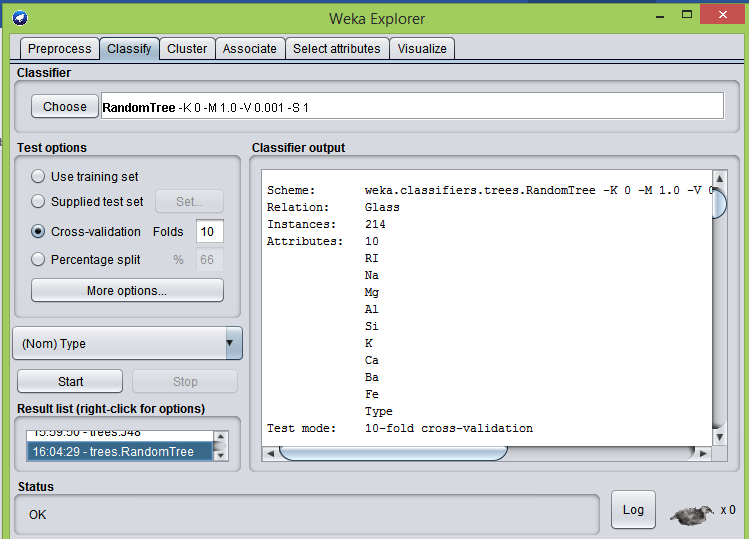
C4.5 Algorithm Selection:



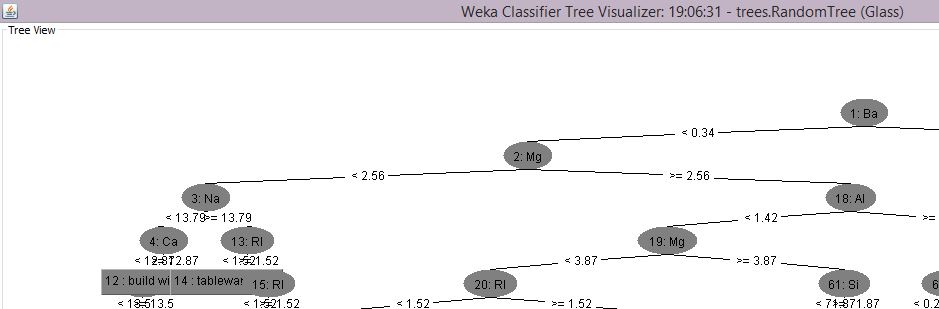
Tree



Random Tree Algorithm Selection:



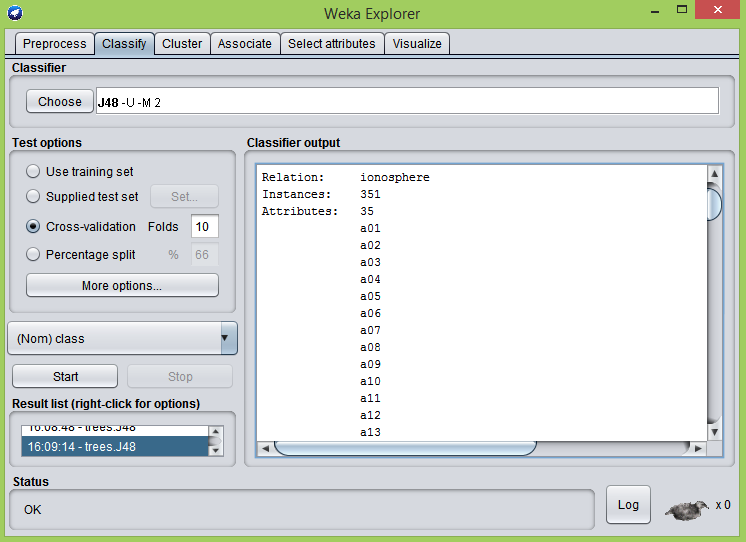
Tree

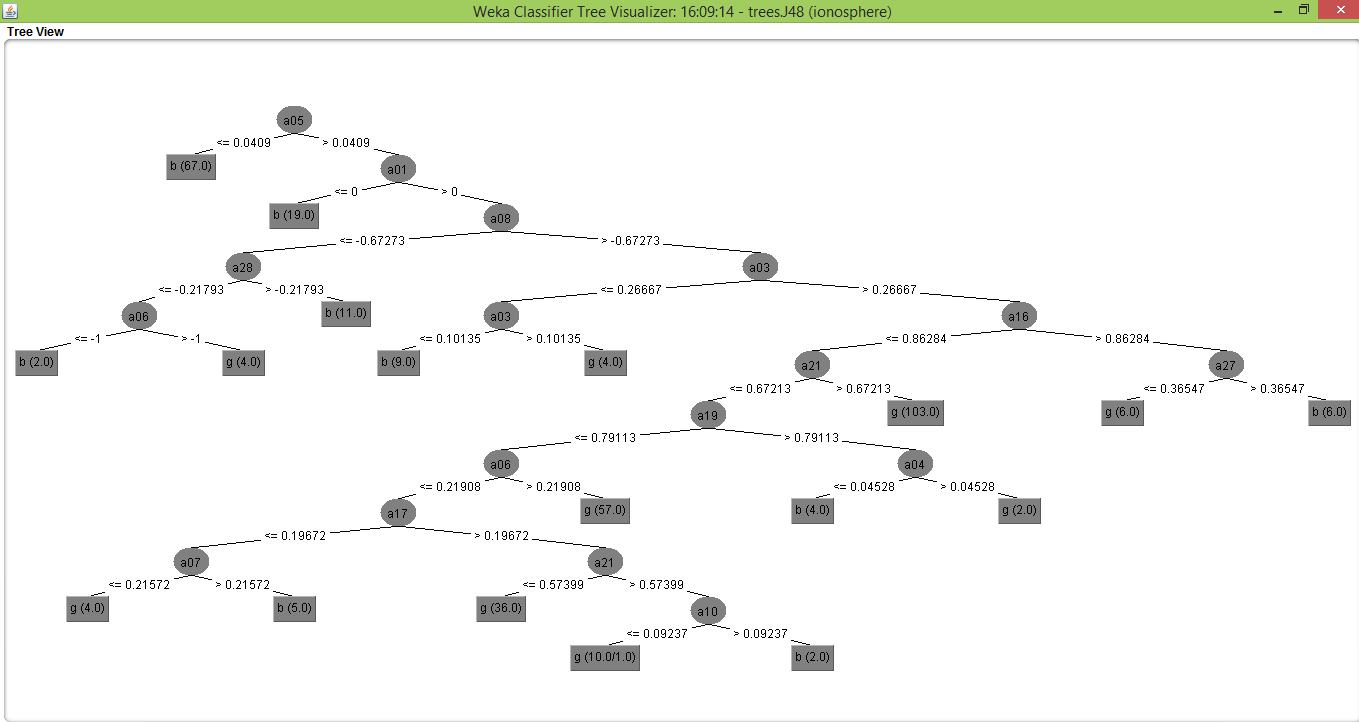


6. Ionosphere

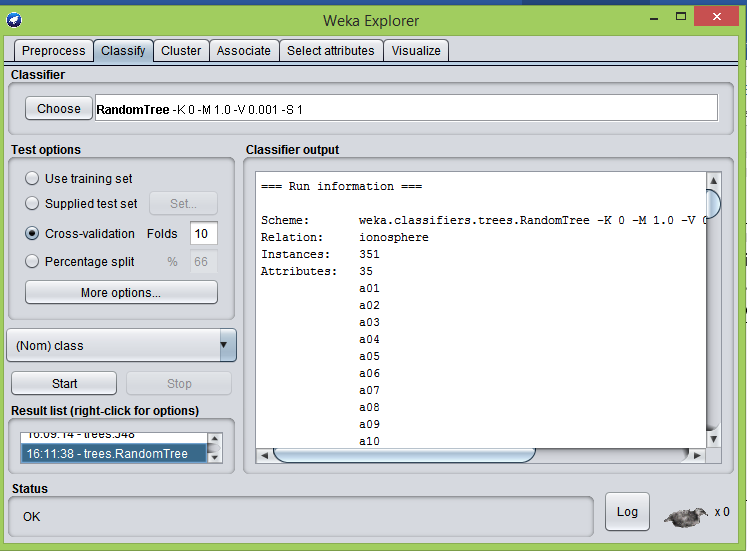
Weka Environment:

C4.5 Algorithm Selection:

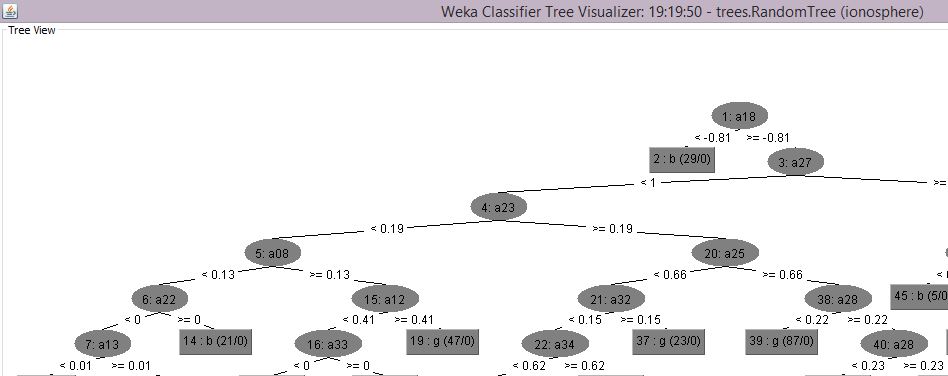


Tree

Random Tree Algorithm Selection:



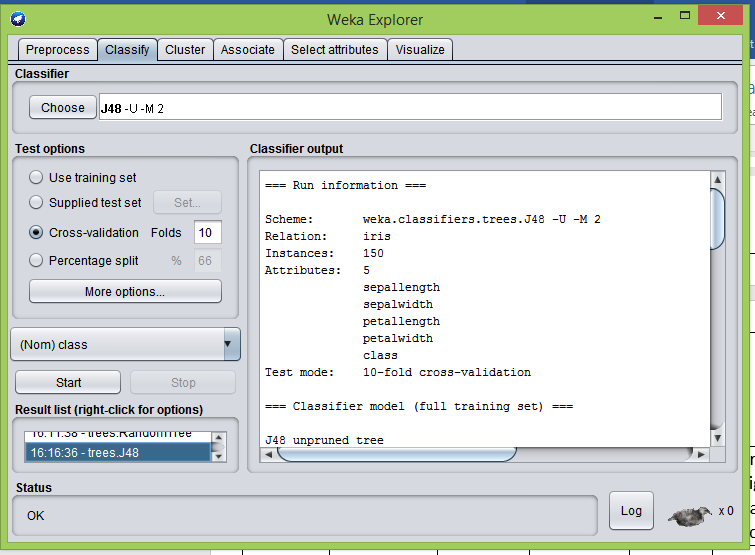
Tree



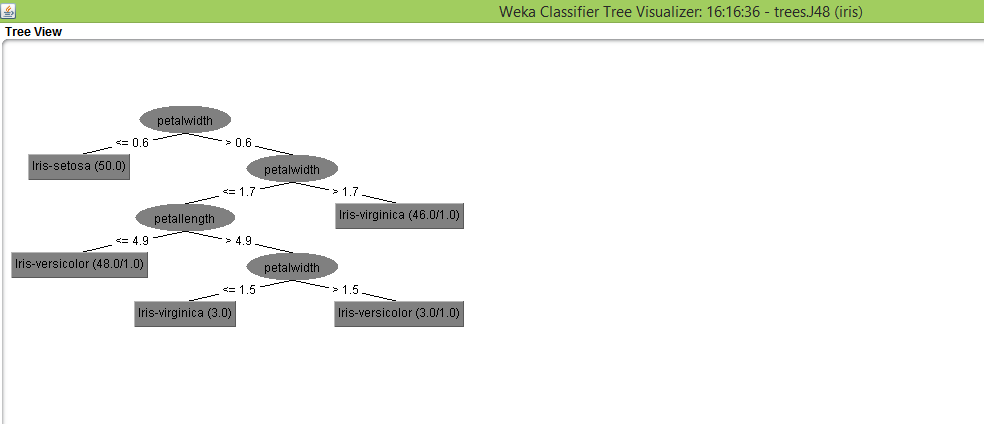
7.Iris

Weka Environment:

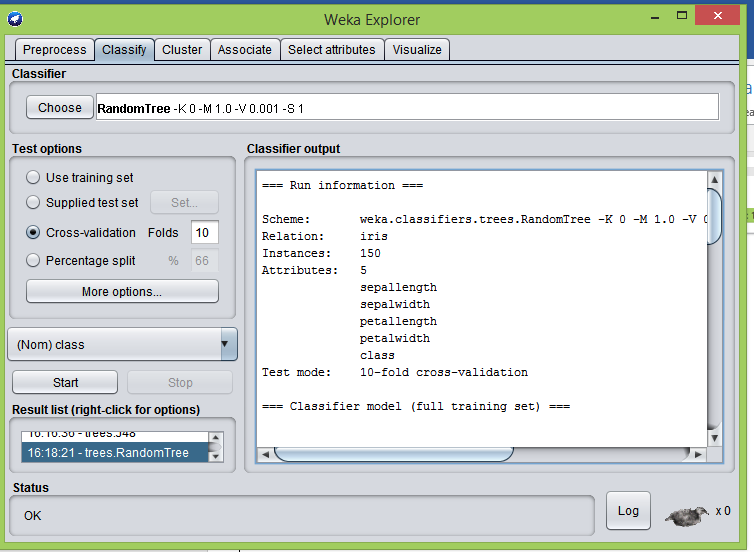
C4.5 Algorithm Selection:



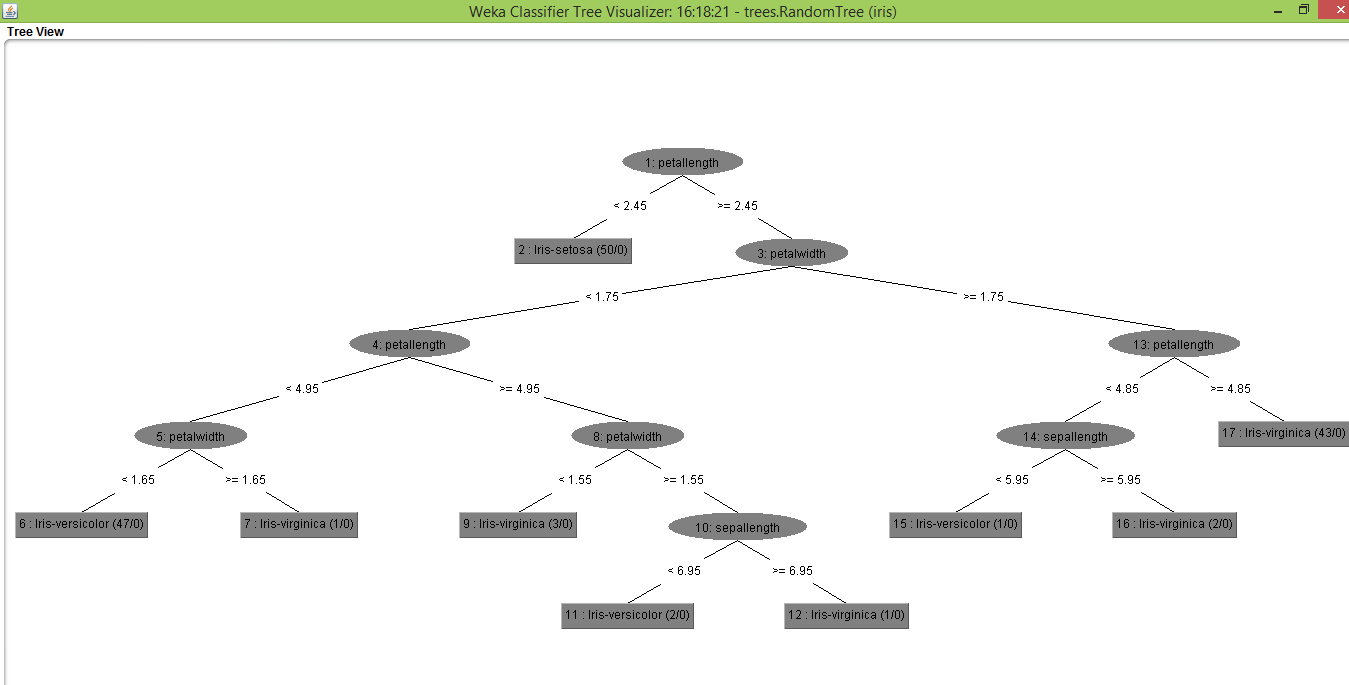
Tree



Random Tree Algorithm Selection:



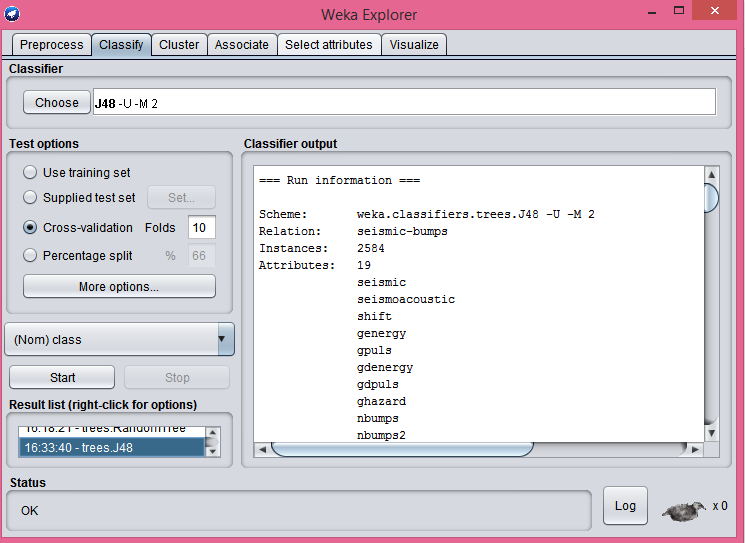
Tree



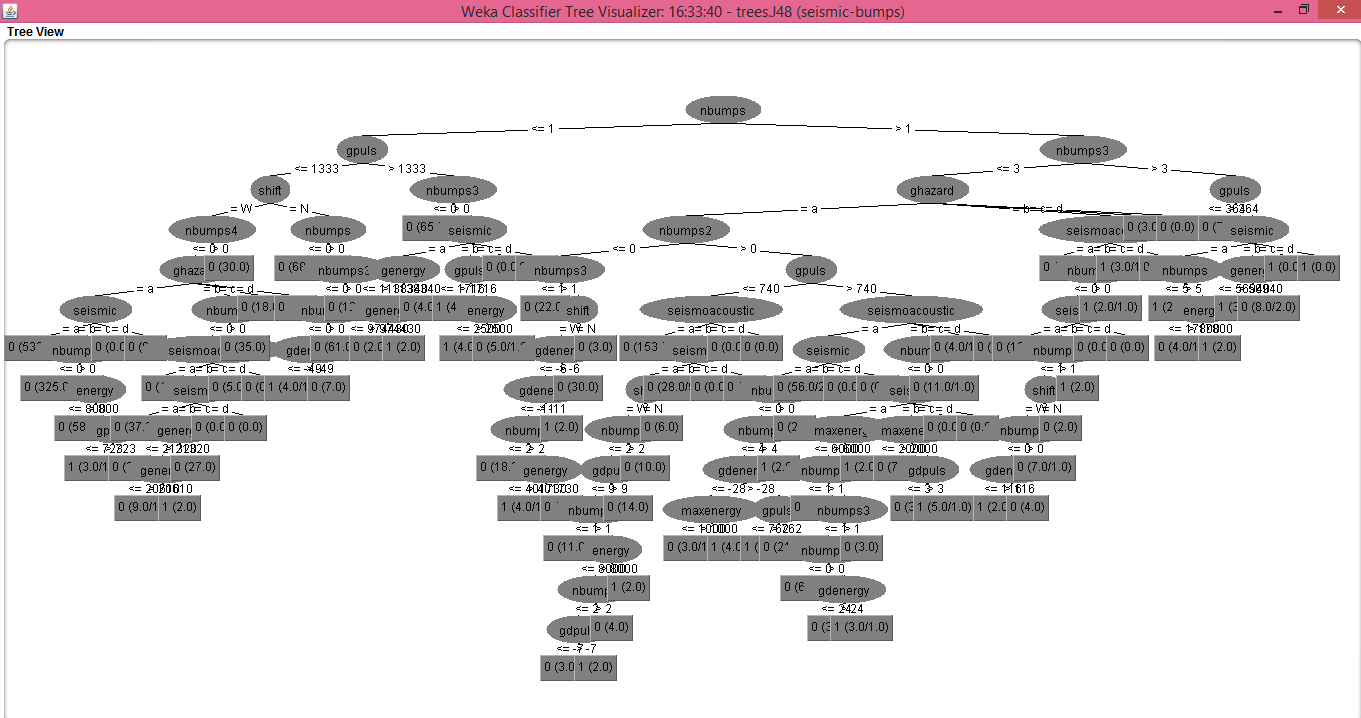
8.Seismic-Bumps

Weka Environment:

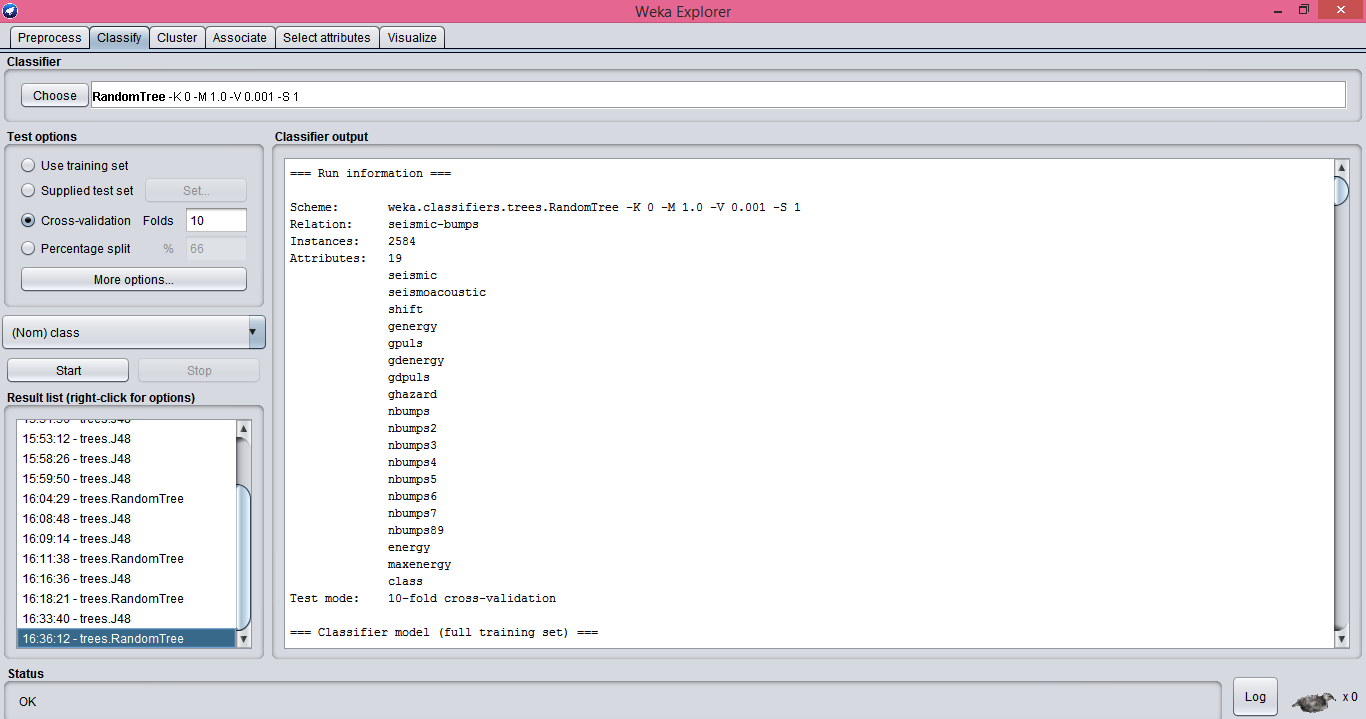
C4.5 Algorithm Selection:



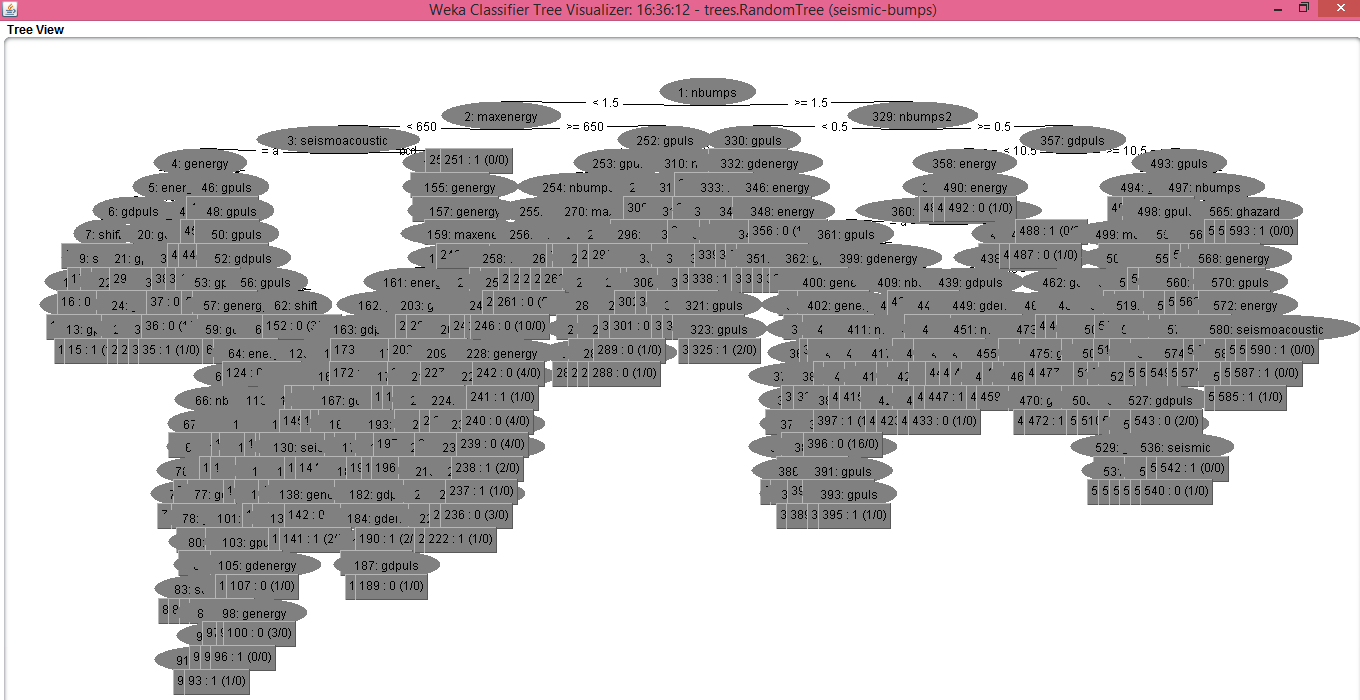
Tree



Random Tree Algorithm Selection:



Tree

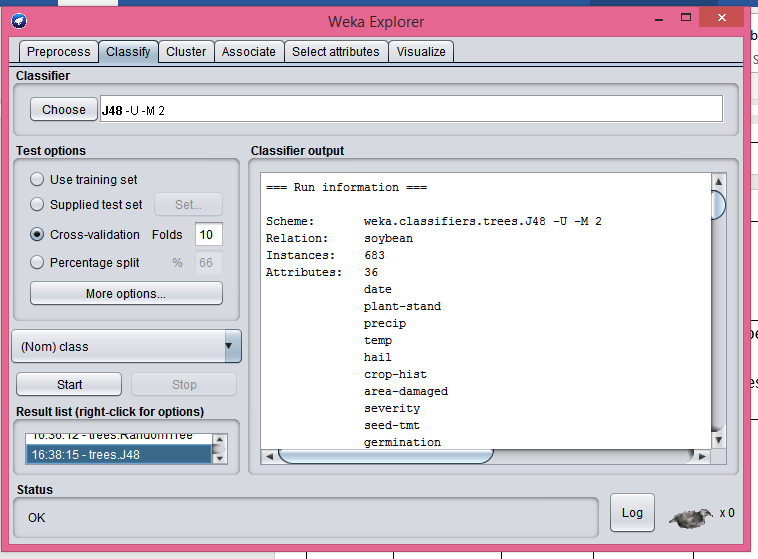


9.Soybean

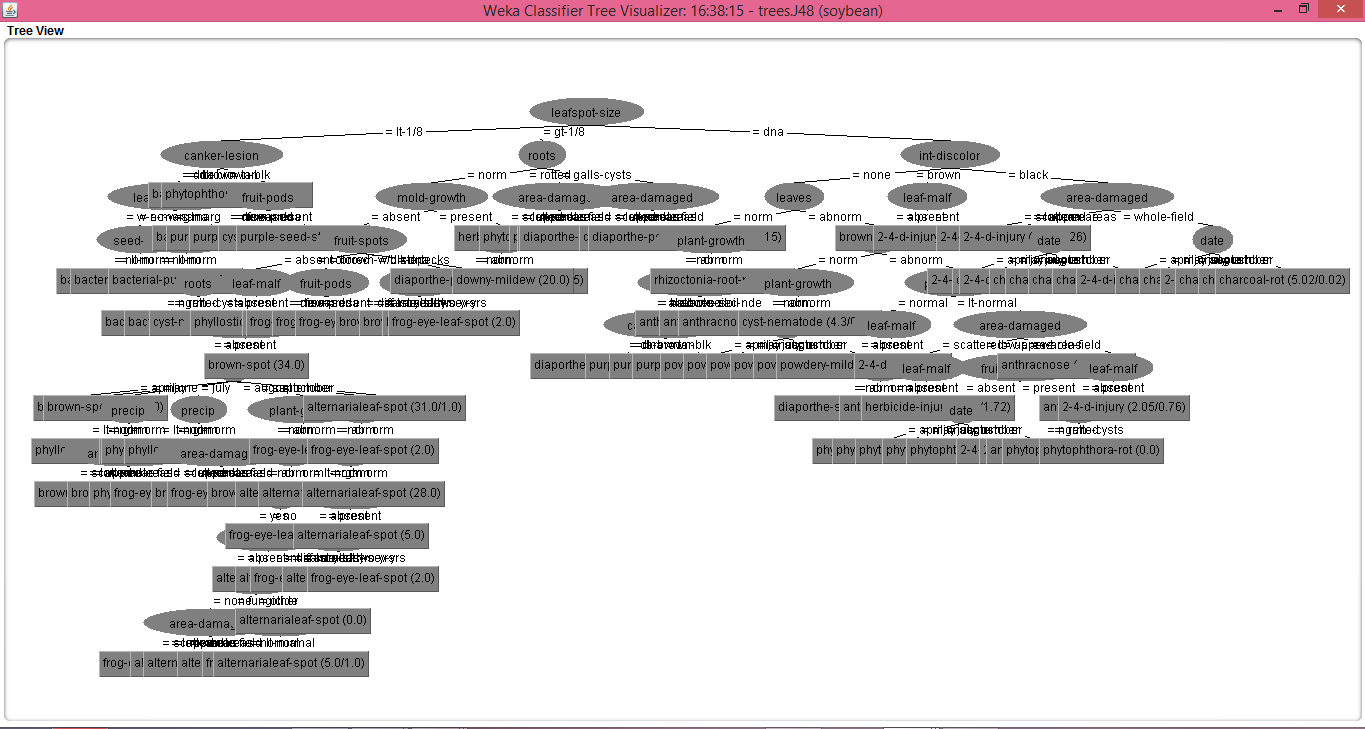
.

Weka Environment:

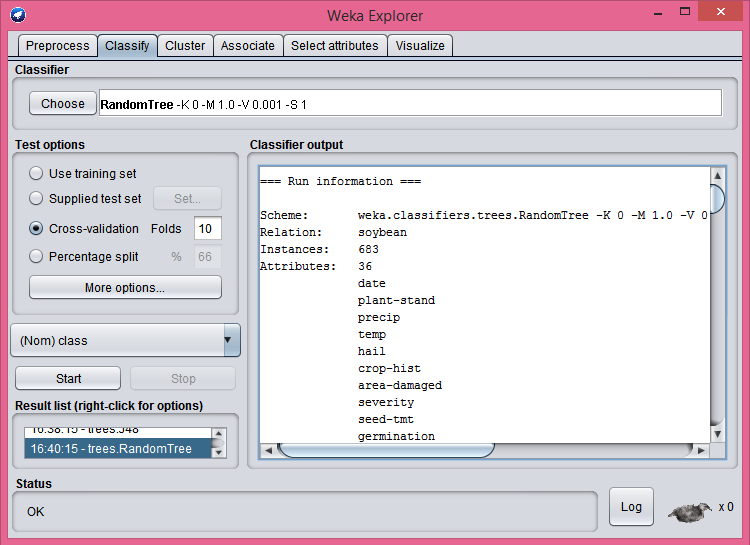
C4.5 Algorithm Selection:



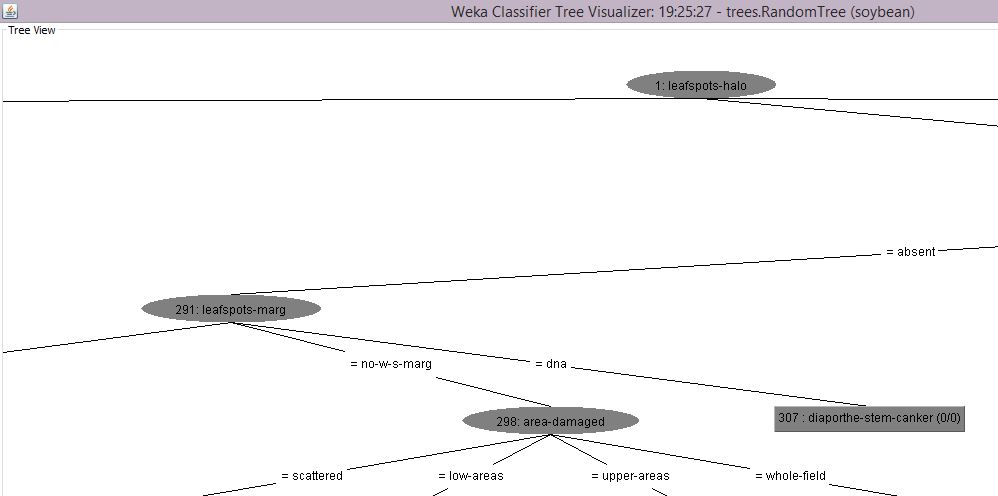
Tree

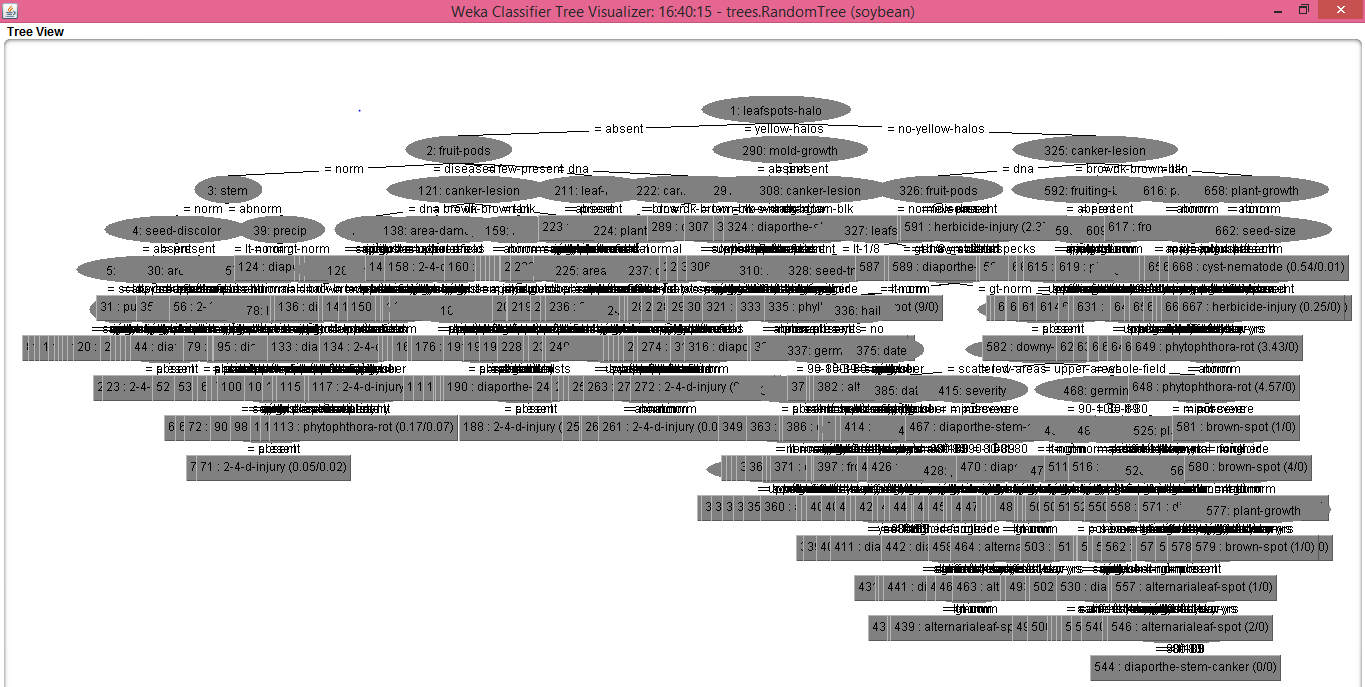


Random Tree Algorithm Selection:



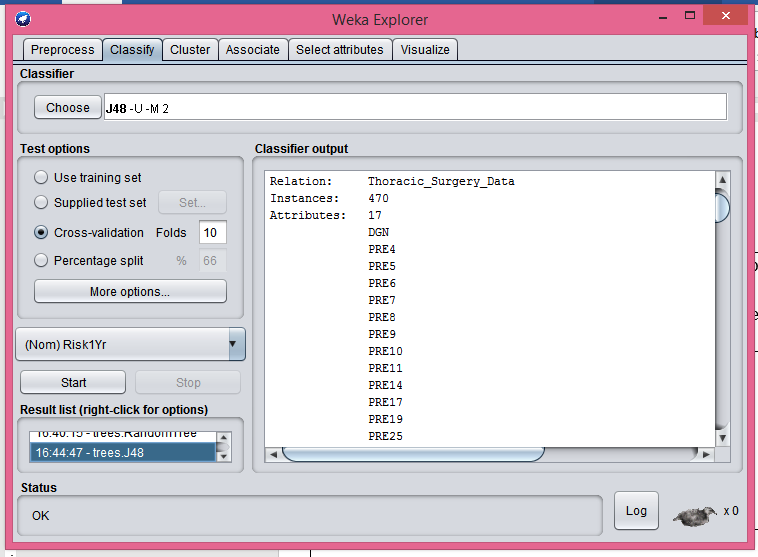
Tree



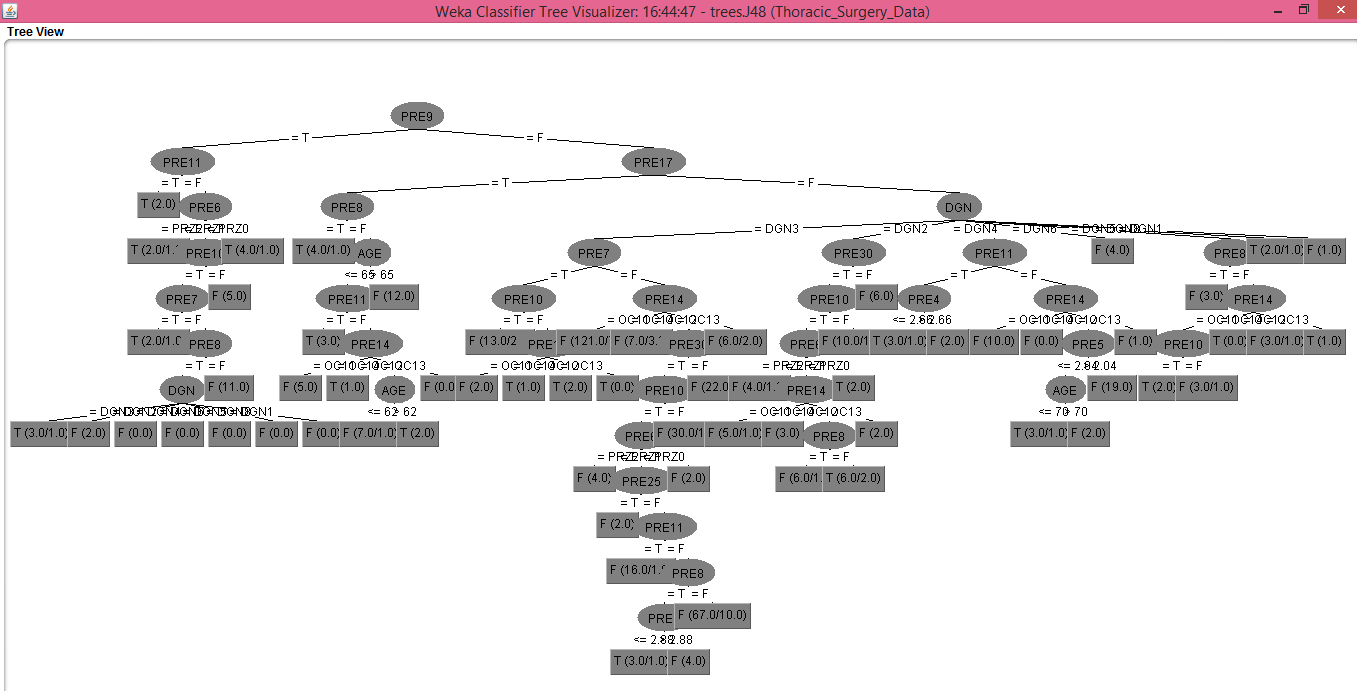


10. Thoracic\_Surgery.

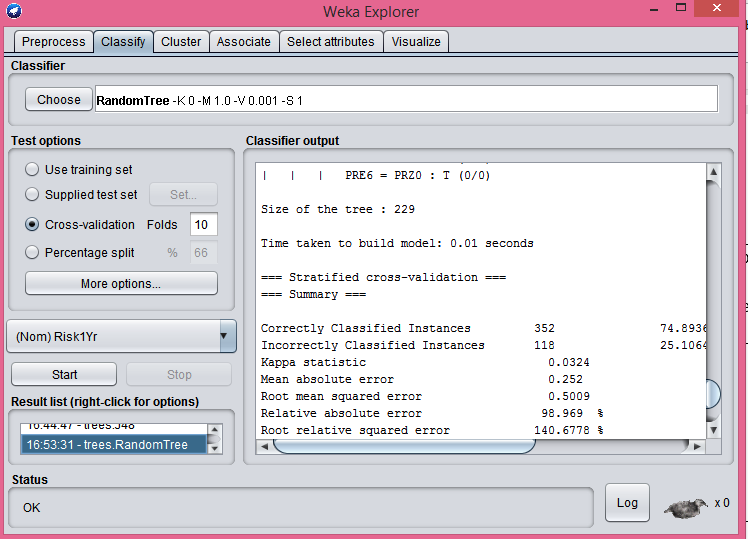
Weka Environment:

C4.5 Algorithm Selection: 

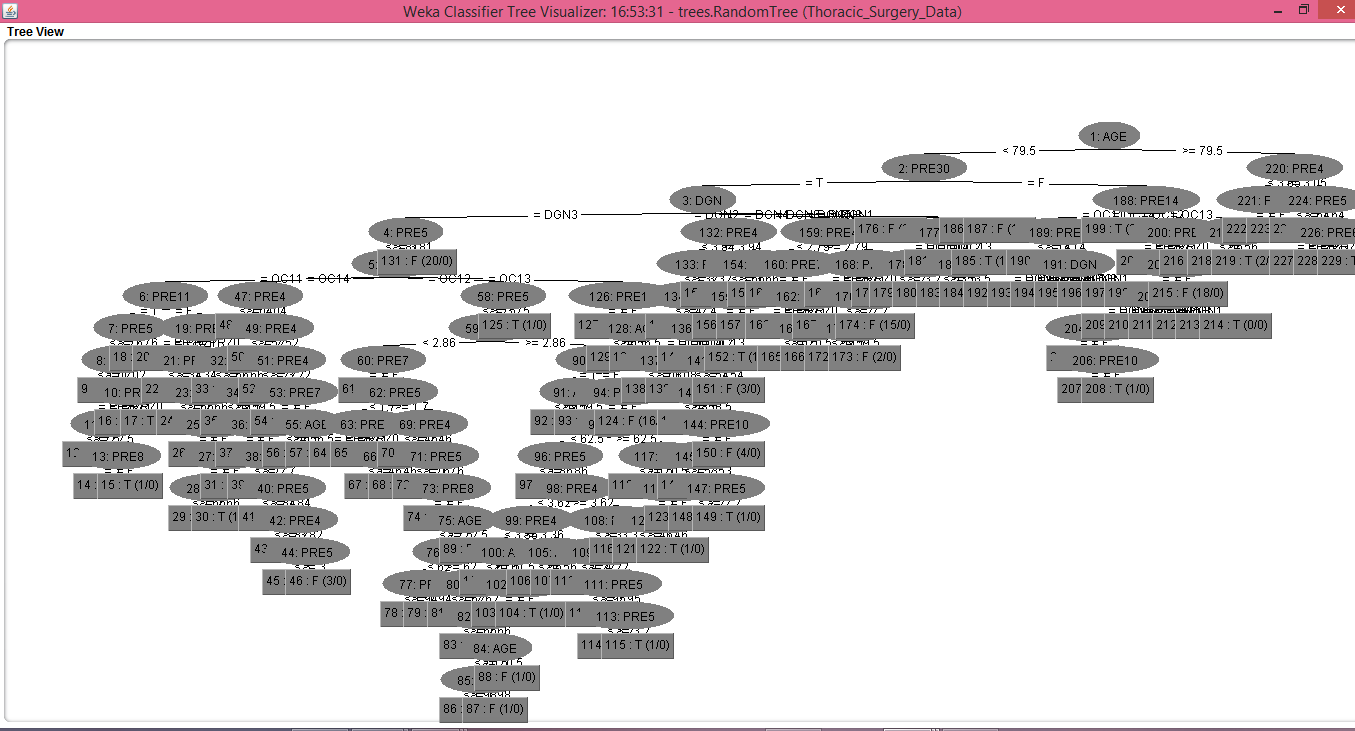
Tree



Random Tree Algorithm Selection:



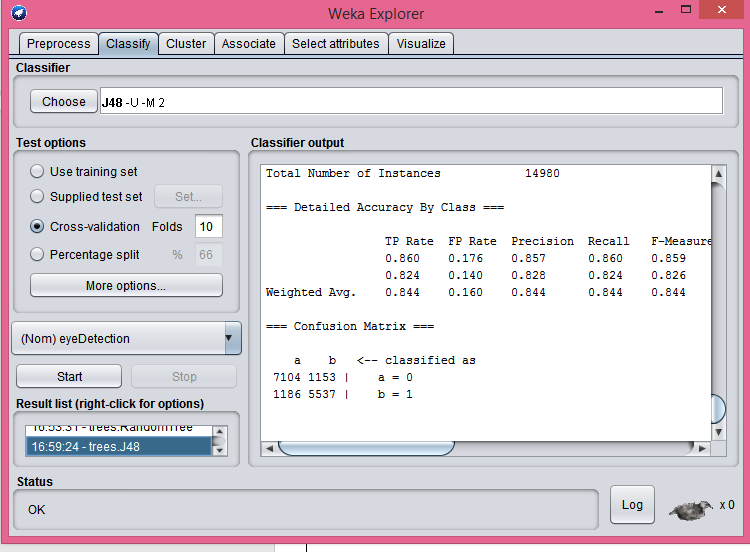
Tree



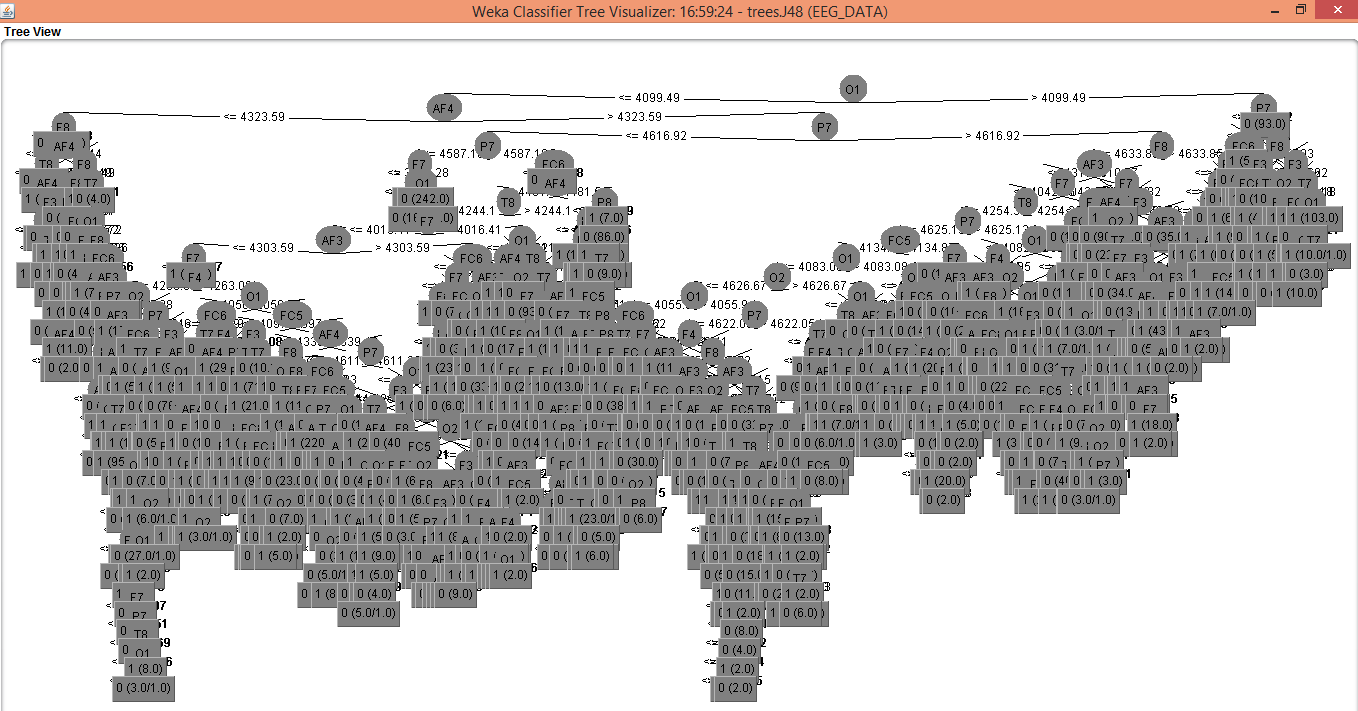
11 EEG Eye State Data Set

Weka Environment:

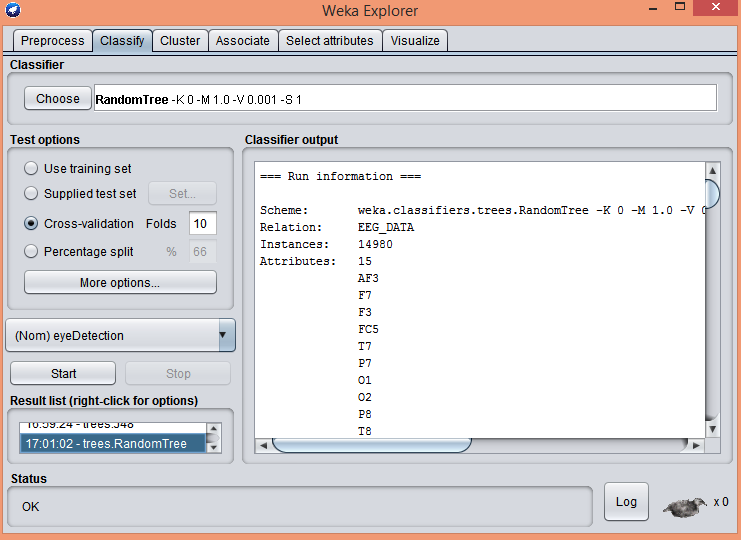
C4.5 Algorithm Selection:



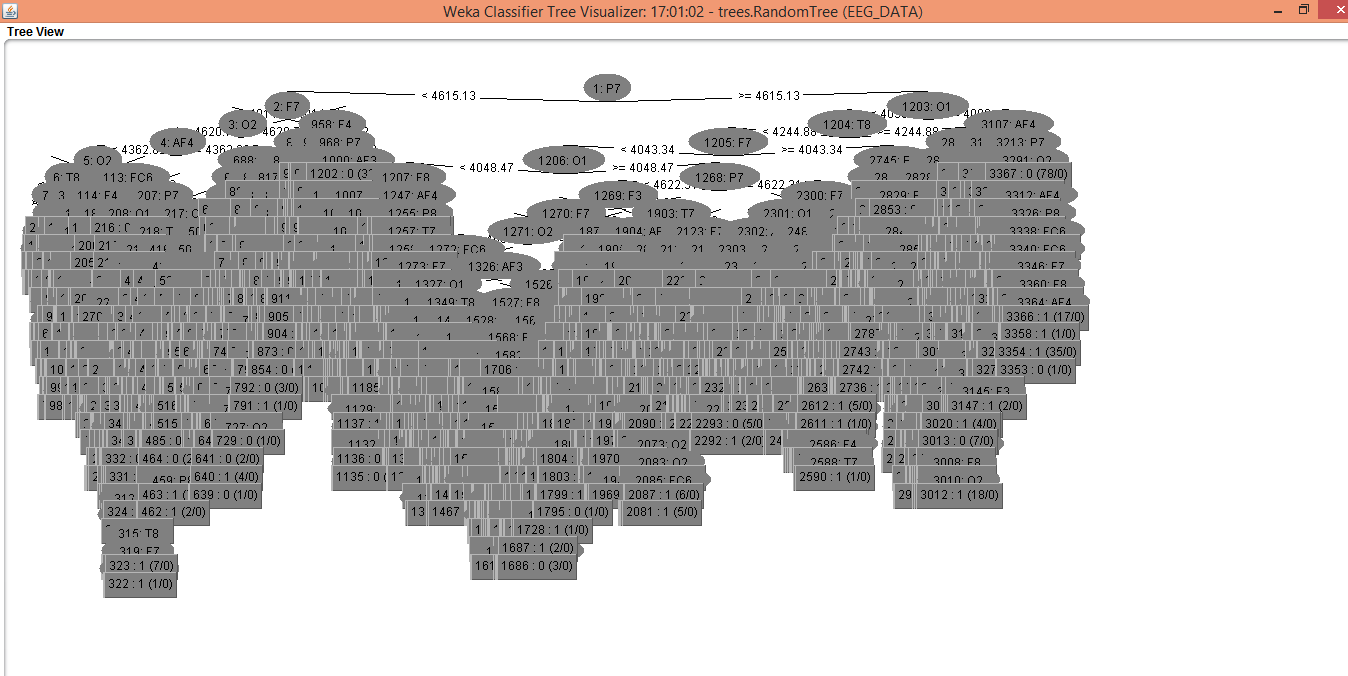
Tree

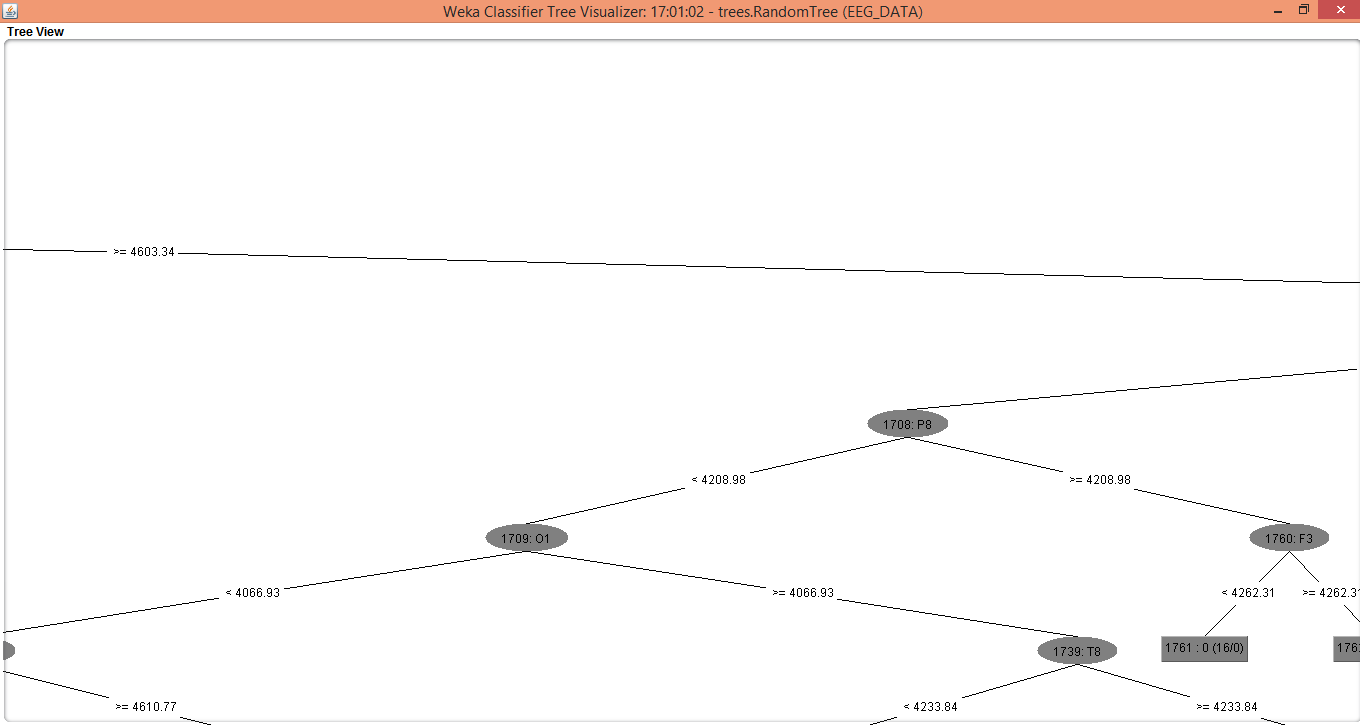


Random Tree Algorithm Selection:



Tree





**4.Summary:**

Would C4.5 generate shorter trees than Random on average? Which is more accurate on

average?

Yes, on average C4.5 algorithm generated shorter trees than Random Tree   
 algorithm. On average C4.5 algorithm is more accurate than Random Tree algorithm.

According to references and study:

1. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized [information gain](https://en.wikipedia.org/wiki/Information_gain) (difference in [entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory))). The attribute with the highest normalized information gain is chosen to make the decision.

2.As decision trees use the “divide and conquer” method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present. One of the reasons for this is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree.

3.Most of the algorithms (like ID3 and C4.5) require that the target attribute will have only discrete values.

4.Random Algorithm is over–sensitivity to the training set, to irrelevant attributes and to noise

**5.References:**

Google

Videos

Different Research Papers, Articles.

Included used datasets, Data Models, Supporting evidences, report.

Thank you !