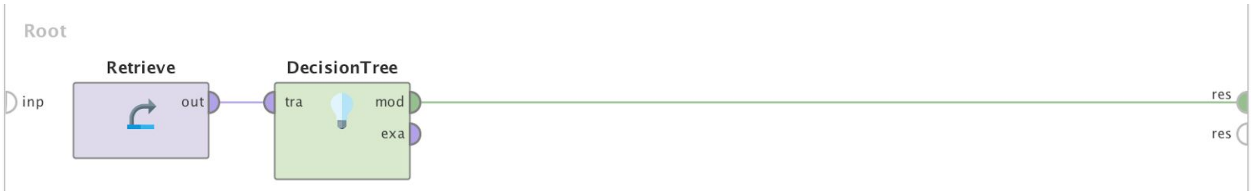
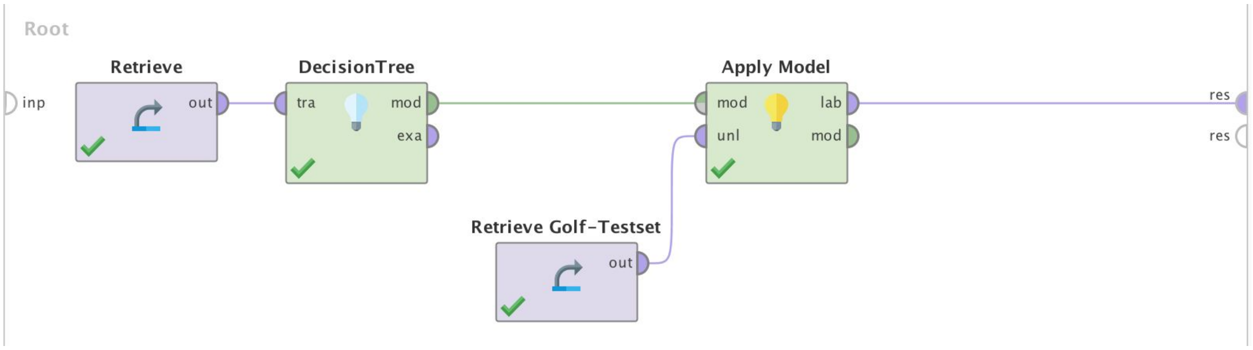
Assignment 7 – Introduction to Data Mining

Problem 1:

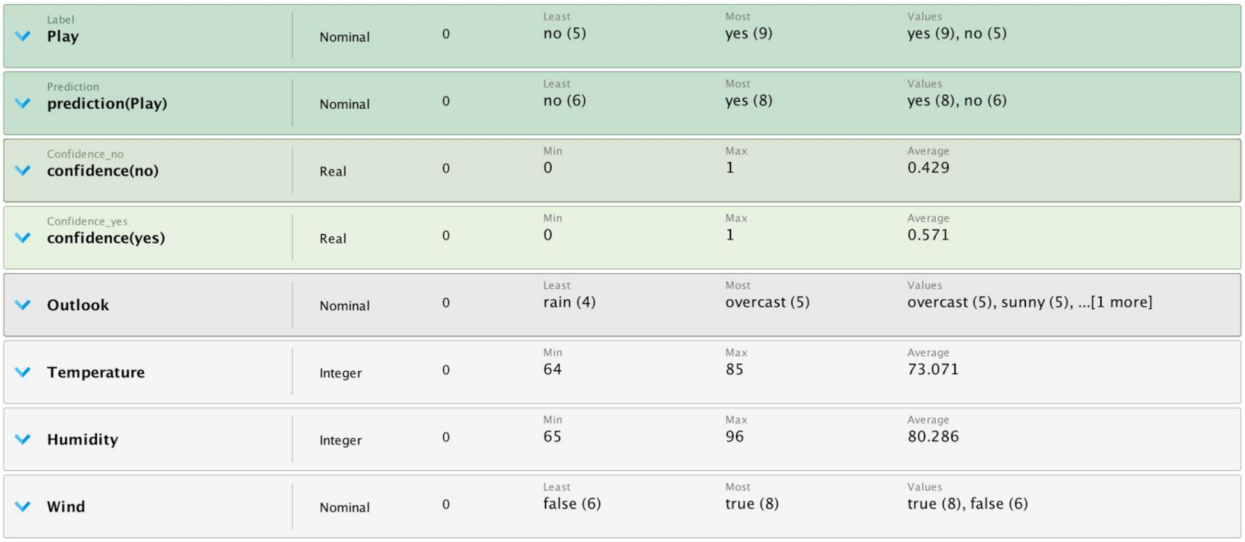


The data is initially loaded, and then a learning step is performed by implementing a decision tree learner also able to handle numerical values. In this example, the initial operator "Input" does not demand input and delivers an example set as output. The example set is then taken in by the learner, who then delivers the final output to the learned model.

Problem 2:



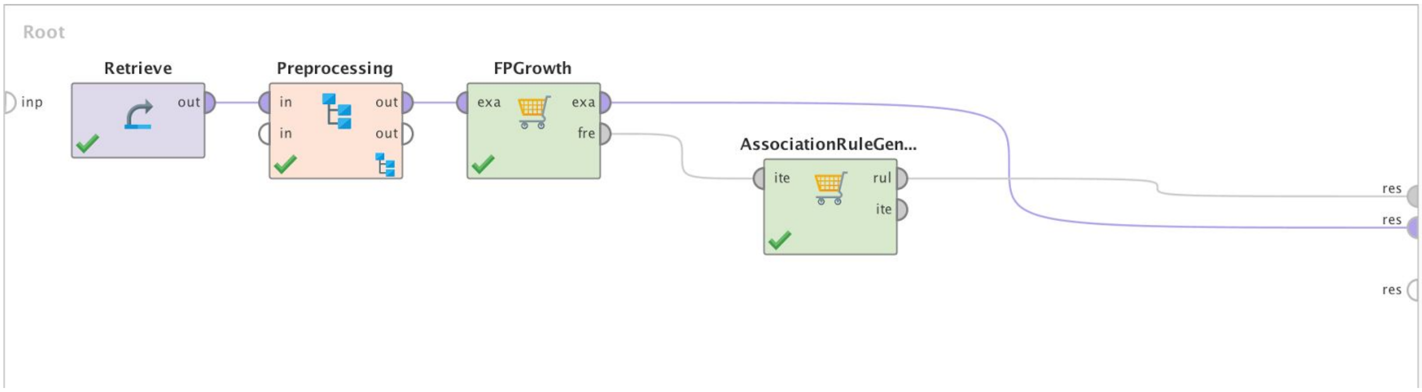
The results tell us that for 'Play' there are 9 for yes and 5 for no. For 'prediction' there are 8 for yes and 6 for no. The average for 'confidence(no)' is 0.429, and the average for 'confidence(yes)' is 0.571. For 'Outlook' there are 4 for rain, 5 for overcast, and 5 for sunny. For 'Temperature' the minimum is 64, the maximum is 85, and the average is 73.071. For 'Humidity' the minimum is 65, the maximum is 96, and the average is 80.286. Finally, for 'Wind' there are 6 for false, and 8 for true.



Label Play	Nominal	0	Least no (5)	Most yes (9)	Values yes (9), no (5)
Prediction prediction(Play)	Nominal	0	Least no (6)	Most yes (8)	Values yes (8), no (6)
Confidence_no confidence(no)	Real	0	Min 0	Max 1	Average 0.429
Confidence_yes confidence(yes)	Real	0	Min 0	Max 1	Average 0.571
Outlook	Nominal	0	Least rain (4)	Most overcast (5)	Values overcast (5), sunny (5), ...[1 more]
Temperature	Integer	0	Min 64	Max 85	Average 73.071
Humidity	Integer	0	Min 65	Max 96	Average 80.286
Wind	Nominal	0	Least false (6)	Most true (8)	Values true (8), false (6)

Problem 3:

There are two preprocessing operators in this process. The first is the frequency discretization operator, which discretizes numerical attributes by putting the values into bins of equal sizes. The second is the filter operator nominal to binominal creates for each possible nominal value of a polynomial attribute a new binomial (binary) feature which is true if the example had the particular nominal value. The preprocessing operators are necessary since particular learning schemes can not handle attributes of certain value types.



Association Rules:

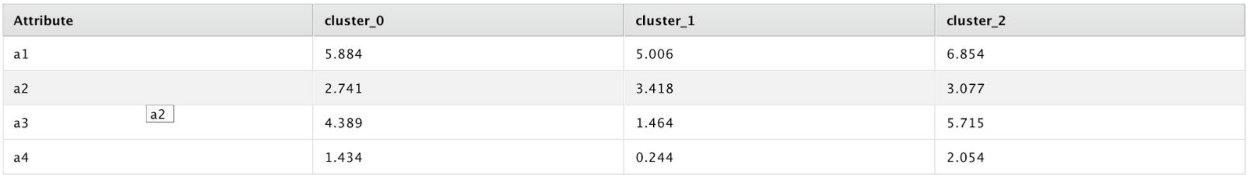
[a3 = range5 [5.350 - ∞]] --> [a4 = range5 [1.950 - ∞]] (confidence: 0.700)
[a3 = range5 [5.350 - ∞], a4 = range5 [1.950 - ∞]] --> [a1 = range5 [6.550 - ∞]] (confidence: 0.714)
[a4 = range5 [1.950 - ∞]] --> [a3 = range5 [5.350 - ∞]] (confidence: 0.724)
[a3 = range1 [-∞ - 1.550]] --> [a4 = range1 [-∞ - 0.250]] (confidence: 0.730)
[a3 = range5 [5.350 - ∞], a1 = range5 [6.550 - ∞]] --> [a4 = range5 [1.950 - ∞]] (confidence: 0.750)
[a2 = range2 [2.750 - 3.050], a4 = range3 [1.150 - 1.550]] --> [a3 = range3 [3.950 - 4.650]] (confidence: 0.762)
[a4 = range1 [-∞ - 0.250], a1 = range1 [-∞ - 5.050]] --> [a3 = range1 [-∞ - 1.550]] (confidence: 0.773)
[a4 = range1 [-∞ - 0.250]] --> [a3 = range1 [-∞ - 1.550]] (confidence: 0.794)
[a3 = range1 [-∞ - 1.550], a1 = range1 [-∞ - 5.050]] --> [a4 = range1 [-∞ - 0.250]] (confidence: 0.810)
[a3 = range3 [3.950 - 4.650]] --> [a4 = range3 [1.150 - 1.550]] (confidence: 0.862)
[a1 = range5 [6.550 - ∞], a4 = range5 [1.950 - ∞]] --> [a3 = range5 [5.350 - ∞]] (confidence: 0.882)
[a2 = range2 [2.750 - 3.050], a3 = range3 [3.950 - 4.650]] --> [a4 = range3 [1.150 - 1.550]] (confidence: 1.000)

Description of Association Rules:

Problem 4:

In many cases, no target attribute (label) can be defined and the data should be automatically grouped. This procedure is called “Clustering”. In this process, the well-known Iris data is loaded (the label is loaded, too, but it is only used for visualization and comparison and not for building the cluster itself). One of the most simple clustering schemes, namely KMeans, is then applied to this data set. Afterwards, a dimensionality reduction is performed in order to better support the visualization of the data set in two dimensions.

Centroid Table:



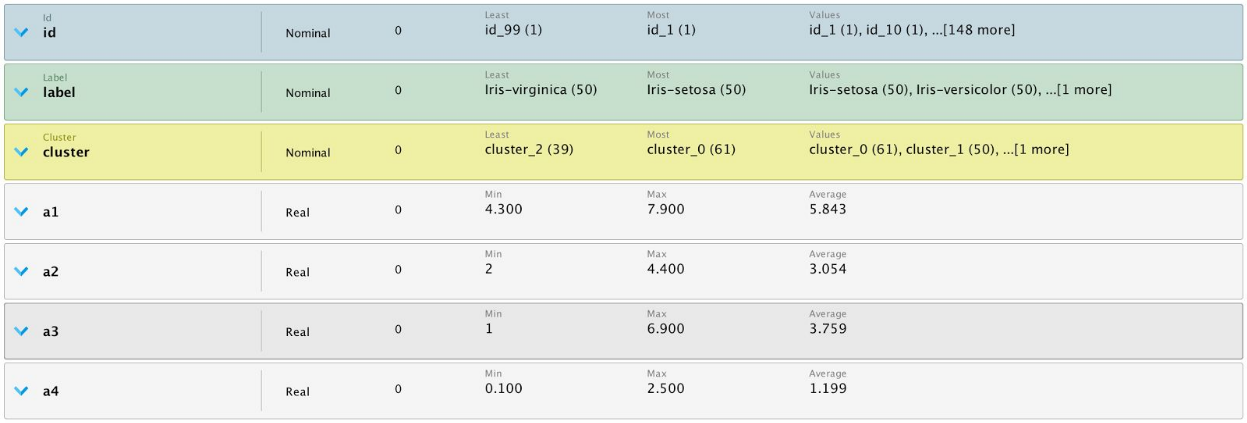
Attribute	cluster_0	cluster_1	cluster_2
a1	5.884	5.006	6.854
a2	2.741	3.418	3.077
a3	4.389	1.464	5.715
a4	1.434	0.244	2.054

Description of Centroid Table Results:

Cluster Model:

Cluster 0: 61 items
Cluster 1: 50 items
Cluster 2: 39 items
Total number of items: 150

Statistics:



Id id	Nominal	0	Least id_99 (1)	Most id_1 (1)	Values id_1 (1), id_10 (1), ...[148 more]
Label label	Nominal	0	Least Iris-virginica (50)	Most Iris-setosa (50)	Values Iris-setosa (50), Iris-versicolor (50), ...[1 more]
Cluster cluster	Nominal	0	Least cluster_2 (39)	Most cluster_0 (61)	Values cluster_0 (61), cluster_1 (50), ...[1 more]
a1	Real	0	Min 4.300	Max 7.900	Average 5.843
a2	Real	0	Min 2	Max 4.400	Average 3.054
a3	Real	0	Min 1	Max 6.900	Average 3.759
a4	Real	0	Min 0.100	Max 2.500	Average 1.199