Breast-Cancer Dataset

Xavier Cho

Data Mining– CSCE 462

Dr. Kirk Scott

1. **Background Information on the Problem Domain and the Data Set**

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, and Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at [[Web Link]](http://www.cs.wisc.edu/~street/images/) Separating plane described above was obtained using Multi surface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree; Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.   
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].   
This database is also available through the UW CS ftp server:   
ftp ftp.cs.wisc.edu   
cd math-prog/cpo-dataset/machine-learn/WDBC/

**I.A. Random Information Drawn from the Online Data Files Posted with the Data Set**

This data consists of cases that convey diagnostics that are evaluated malignant or benign as well as the different stages (1-3) of breast cancer. Stage 4 is not included based on these cases are determining whether if a patient is operable or inoperable based on the diagnosis. Since this disease is still being studied there are still areas that need more information to determine a more efficient diagnosis.

5. Number of Instances: 286

6. Number of Attributes: 9 + the class attribute

7. Attribute Information:

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. Menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,

45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,

27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. Breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

10. Irradiat: yes, no.

8. Missing Attribute Values: (denoted by "?")

Attribute #: Number of instances with missing values:

6. 8

9. 1.

9. Class Distribution:

1. no-recurrence-events: 201 instances

2. recurrence-events: 85 instances

Num Instances: 286

Num Attributes: 10

Num Continuous: 0 (Int 0 / Real 0)

Num Discrete: 10

Missing values: 9 / 0.3%

Name type enum ints real missing distinct (1)

1 'age' Enum 100% 0% 0% 0 / 0% 6 / 2% 0%

2 'menopause' Enum 100% 0% 0% 0 / 0% 3 / 1% 0%

3 'tumor-size' Enum 100% 0% 0% 0 / 0% 11 / 4% 0%

4 'inv-nodes' Enum 100% 0% 0% 0 / 0% 7 / 2% 0%

5 'node-caps' Enum 97% 0% 0% 8 / 3% 2 / 1% 0%

6 'deg-malig' Enum 100% 0% 0% 0 / 0% 3 / 1% 0%

7 'breast' Enum 100% 0% 0% 0 / 0% 2 / 1% 0%

8 'breast-quad' Enum 100% 0% 0% 1 / 0% 5 / 2% 0%

9 'irradiat' Enum 100% 0% 0% 0 / 0% 2 / 1% 0%

10 'Class' Enum 100% 0% 0% 0 / 0% 2 / 1% 0%

**I.B. Contents of the Data File**

'40-49','premeno','15-19','0-2','yes','3','right','left\_up','no','recurrence-events'

'50-59','ge40','15-19','0-2','no','1','right','central','no','no-recurrence-events'

'50-59','ge40','35-39','0-2','no','2','left','left\_low','no','recurrence-events'

'40-49','premeno','35-39','0-2','yes','3','right','left\_low','yes','no-recurrence-events'

'40-49','premeno','30-34','3-5','yes','2','left','right\_up','no','recurrence-events'

'50-59','premeno','25-29','3-5','no','2','right','left\_up','yes','no-recurrence-events'

'50-59','ge40','40-44','0-2','no','3','left','left\_up','no','no-recurrence-events'

'40-49','premeno','10-14','0-2','no','2','left','left\_up','no','no-recurrence-events'

'40-49','premeno','0-4','0-2','no','2','right','right\_low','no','no-recurrence-events'

'40-49','ge40','40-44','15-17','yes','2','right','left\_up','yes','no-recurrence-events'

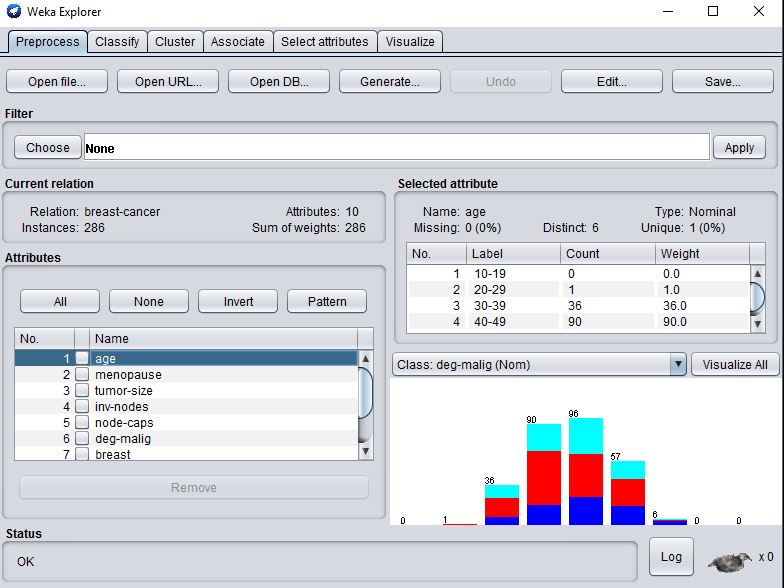
'50-59','premeno','25-29','0-2','no','2','left','left\_low','no','no-recurrence-events'

**I.C. Summary of Background Information**

The breast cancer classification is to diagnose a patient for cancer to be malignant or benign. There are 286 Number of Instances in the data set regarding 10 attributes including the class attribute, 4 numeric attributes and 6 nominal attributes. This data set includes 201 instances of one class and 85 instances of another class. This is a topic that is still in research since this data is from the 1970 – 1988 further data mining in contrast to data from now is still being tested.

**I.D. Screen Shot of Open File**

**What this shows:** The age attribute is highlighted on the bottom left. The varying values are on the upper right. In the lower right shows the proportion of the degree (1-3) of how malignant the cancer is based on the instances of the different age groups.



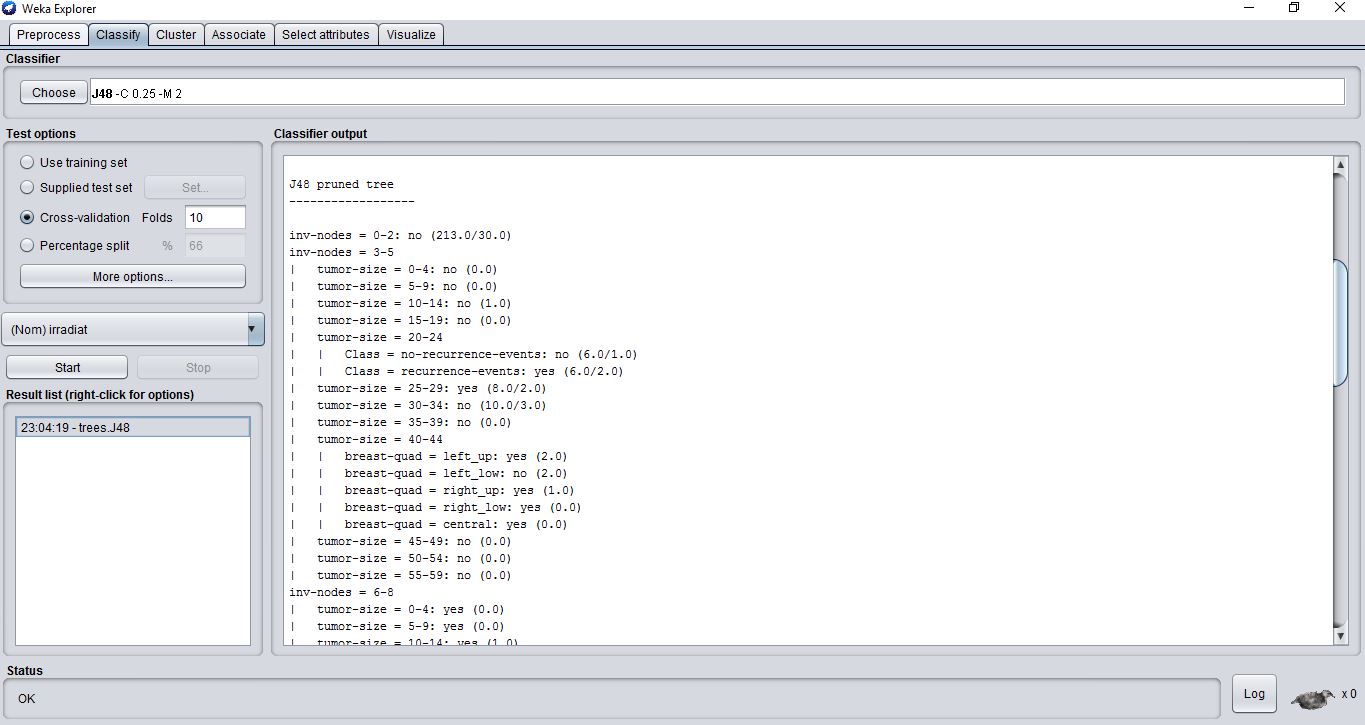
**II. Applications of Data Mining Algorithms to the Data Set**

**II. Case 1. This Needs to Be a Classification Algorithm**

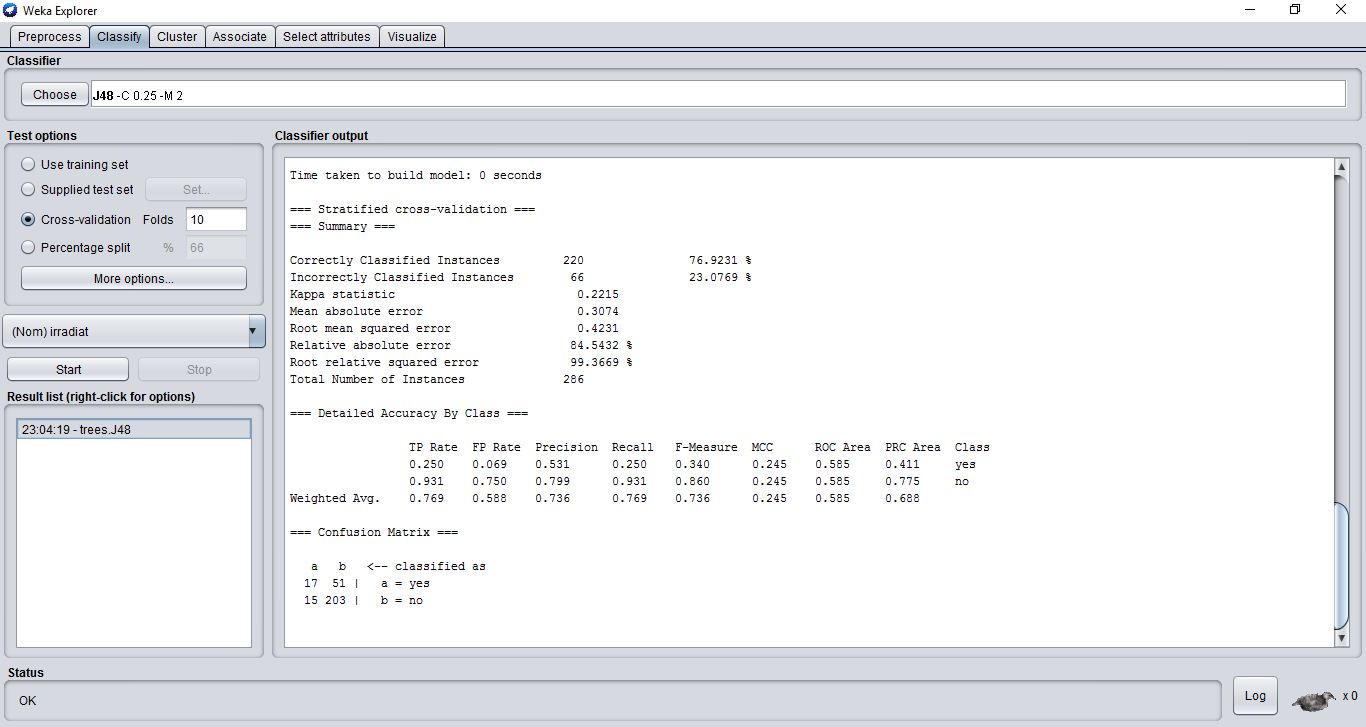
**Name of Algorithm: J48**

**i. Output Results**

What this shows: This shows the classifier tree generated by the J48 algorithm.



What this shows: This gives the analysis of the output of the algorithm. The most notable thing that should focus on the decision tree has classified correctly is 76.9% versus the incorrect classified being 23%.

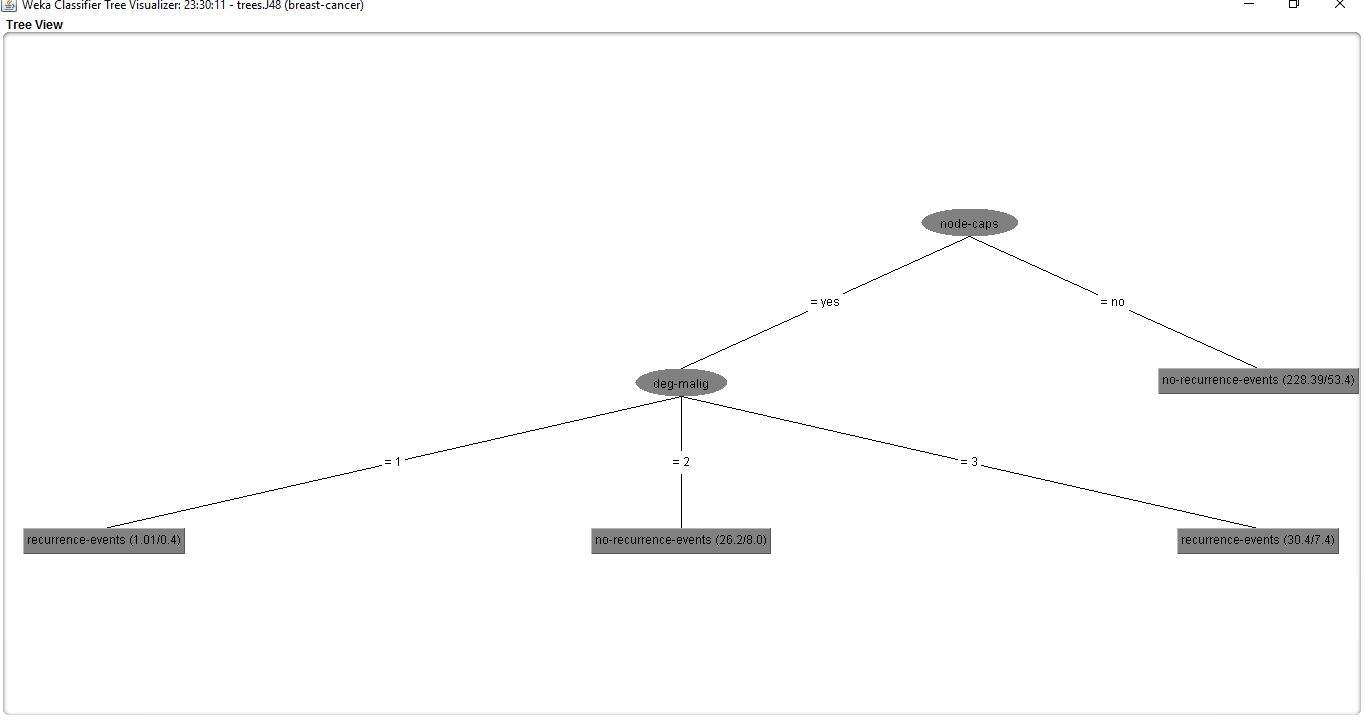


**ii. Explanation of Item**

Choosing the confusion matrix we can see that there are 51 false positives and 15 false negatives. It is shocking to see how many false positives have occurred based on the representation of this value based on the other attributes of either no-recurrence versus recurrence of the cancer diagnosis.

**iii. Graphical or Other Special Purpose Additional Output**

What this shows: this is the result of the tree after running j48 on the dataset showing a visualization of the branching that occurs.

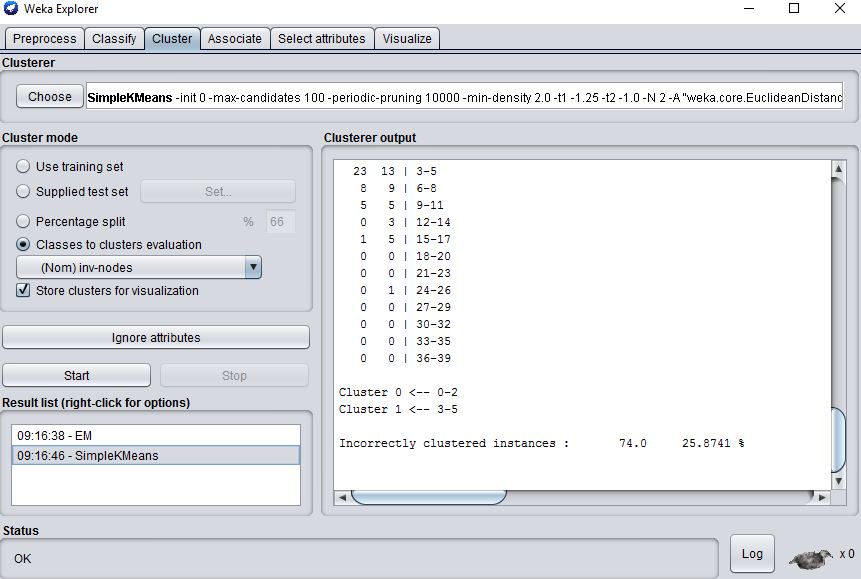


**II. Case 2. This Needs to Be a Clustering Algorithm**

**Name of Algorithm: SimpleKMeans**

**i. Output Results**

What this shows: This shows the results of the SimpleKMeans clustering algorithm with the invasive nodes (inv-node) attribute being ignored. The results how a cluster can classify with the number of clusters specified being 2. The algorithm finds the clusters based on the remaining attributes then attempts to find its classification.

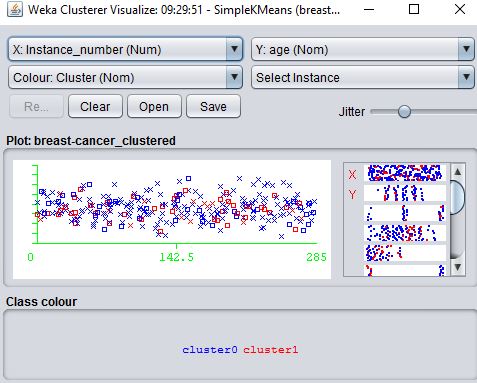


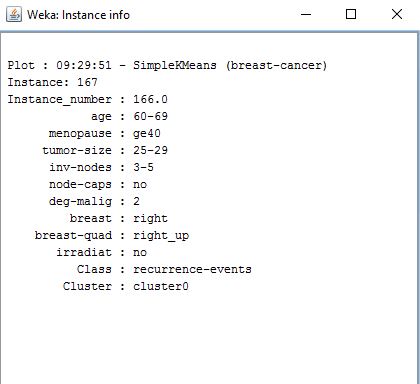
**ii. Explanation of Item**

At the bottom of the screen shot, "Incorrectly clustered instances". 27.6% of the clustered instances are not categorized in 0 – 5 as invasive. The algorithm finds 2 clusters where the user specifies how many clusters it selects, the algorithm are based on the remaining attributes if we did not know the classification this tool is good to use.

**iii. Graphical or Other Special Purpose Additional Output**

What this shows: Showing the instanced visualization along the x axis, as the y axis shows the placements of the different age ranges where the selected attribute of in this case age.



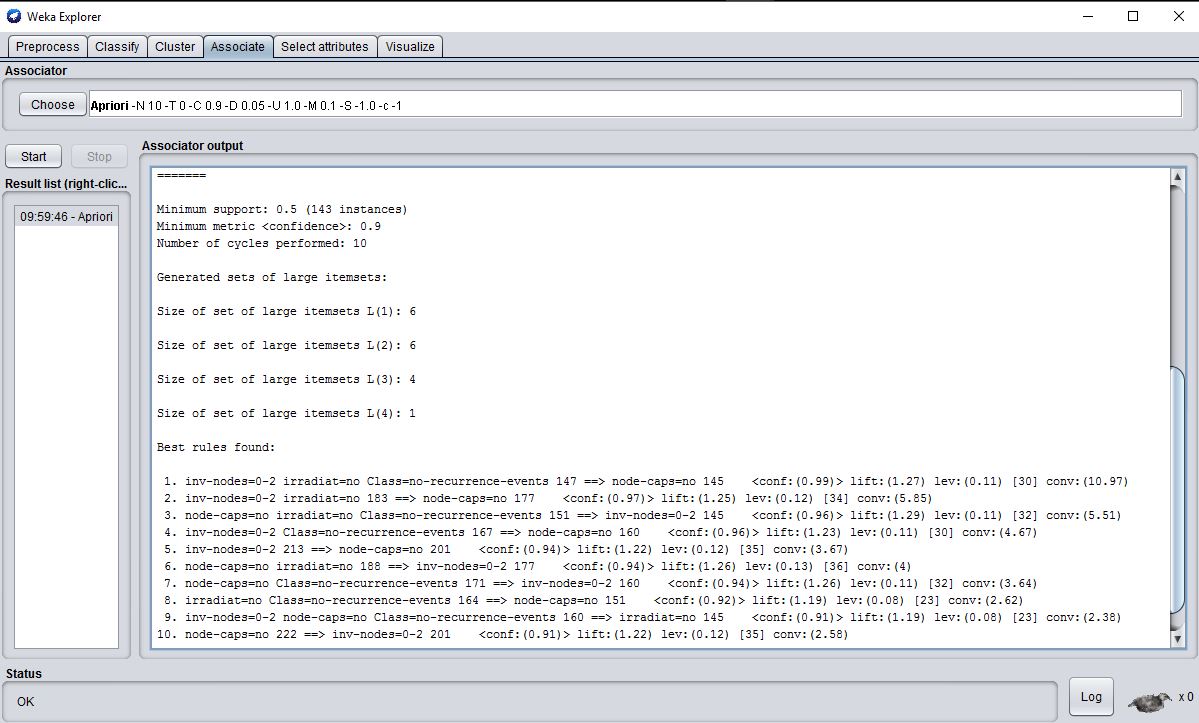


**II. Case 3. This Needs to Be an Association Mining Algorithm**

**Name of Algorithm: Apriori**

**i. Output Results**

\*\*\*What this shows: This shows the results of the Apriori association rule mining algorithm.



**ii. Explanation of Item**

**This shows that the algorithm specifies the minimum confidence level then goes about iterating to reduce the support to create rules with greater or less confidence. Based on this data we can see that the range of confidence is from .91 to, .97 which represents that the data can be redundant in this aspect. If you pay attention to rule 9 one could say it is plausible that if 2 invasive nodes are found, the nodes could not be a recurring cancerous event or a benign cancerous node.**

**iii. Graphical or Other Special Purpose Additional Output**

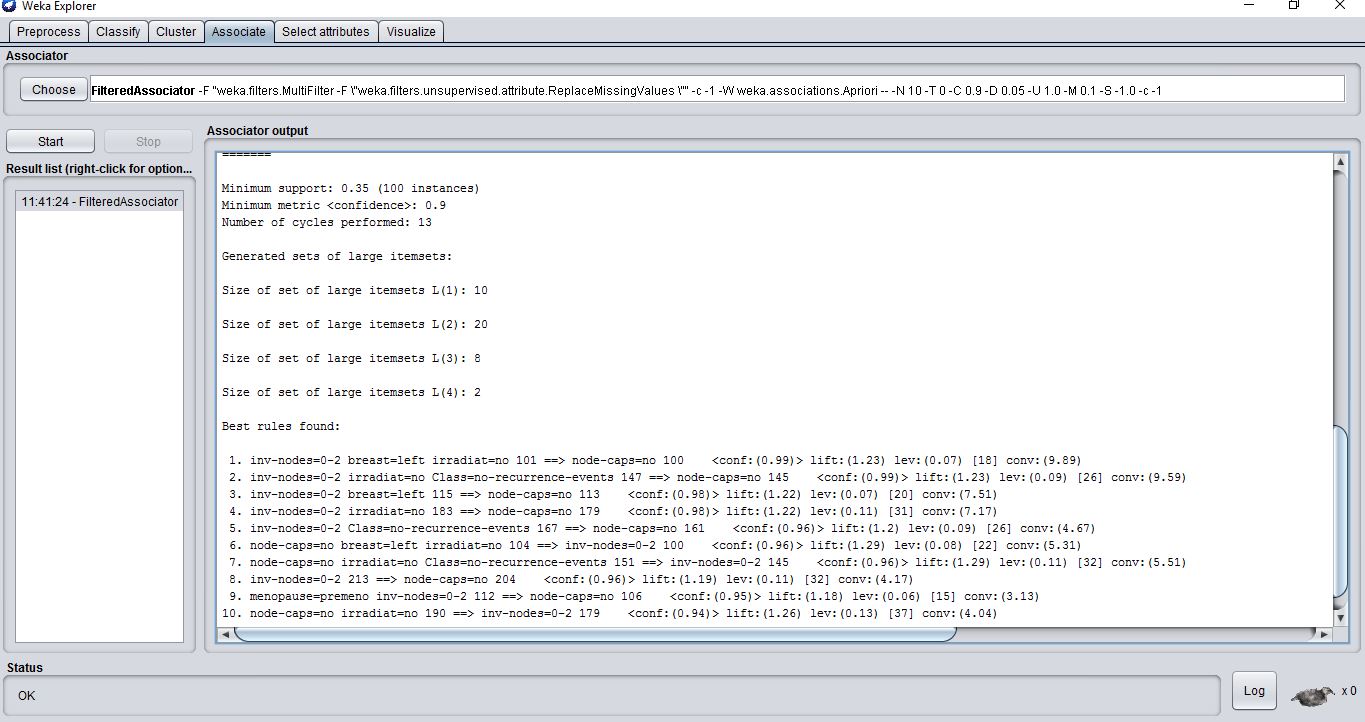
There don't appear to be any other output options for association rules. There is no standard visualization for them so nothing is included for this point.

**II. Case 4. Any Kind of Algorithm**

**Name of Algorithm: Filtered Associator**

**i. Output Results**

What this shows: This shows the results of the Filtered Associator association rule mining algorithm.



**ii. Explanation of Item**

This also shows that the algorithm specifies the arbitrary associator that will pass through a filter. Then structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. Based on this data we can see that the range of confidence is from .94 to, .99 which represents that the data can be redundant in this aspect. If you pay attention to rule 9 one could say it is plausible that if 2 invasive nodes are found, the nodes could not be a recurring cancerous event or a benign cancerous node. But in comparison to apriori we see that this took more cycles and large item sets.

**iii. Graphical or Other Special Purpose Additional Output**

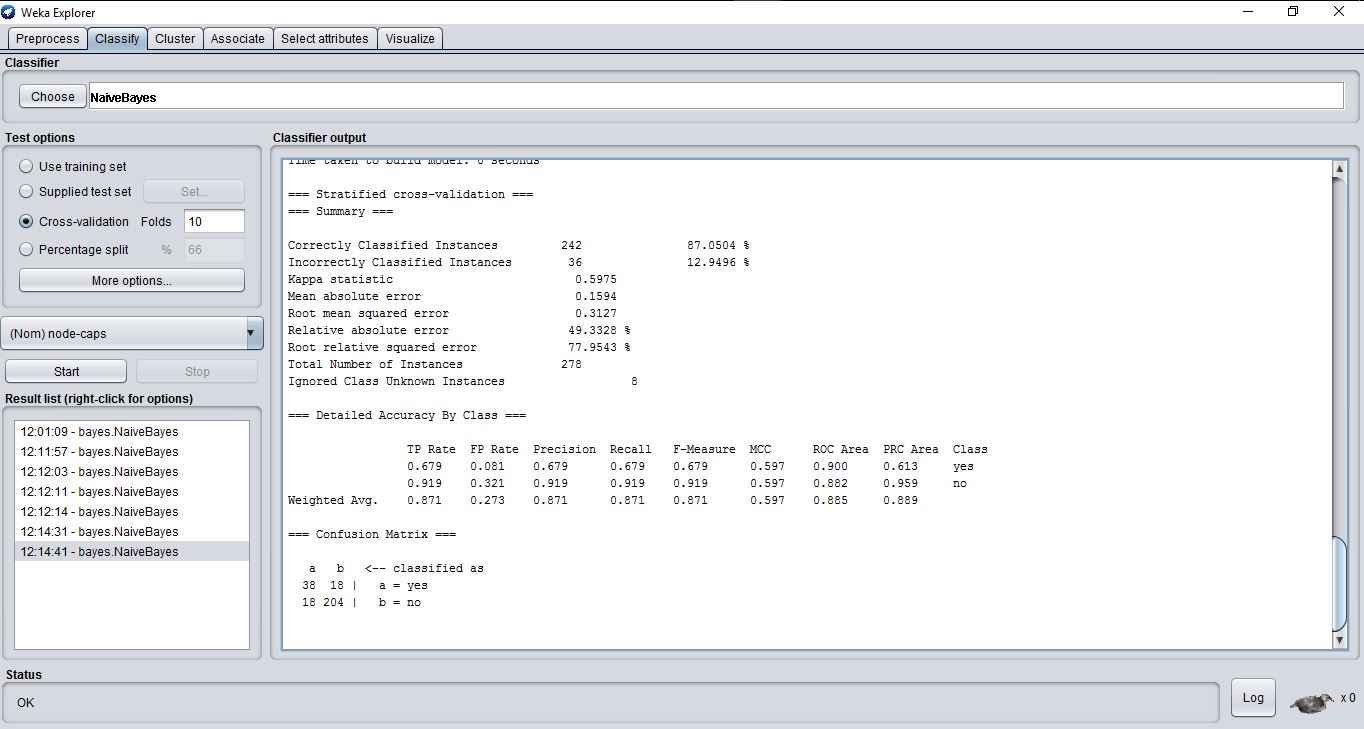
Unfortunately in WEKA there is no visualization for the association rules so there is nothing to include here.

**II. Case 5. Any Kind of Algorithm**

**Name of Algorithm: Naïve Bayes**

**i. Output Results**

\*\*\*What this shows: this screen shows the result of the Naïve Bayes algorithm. Where we have an error rate of 204/222 and a ROC area of 88.5 with a good correctly classified about 87% from the dataset;

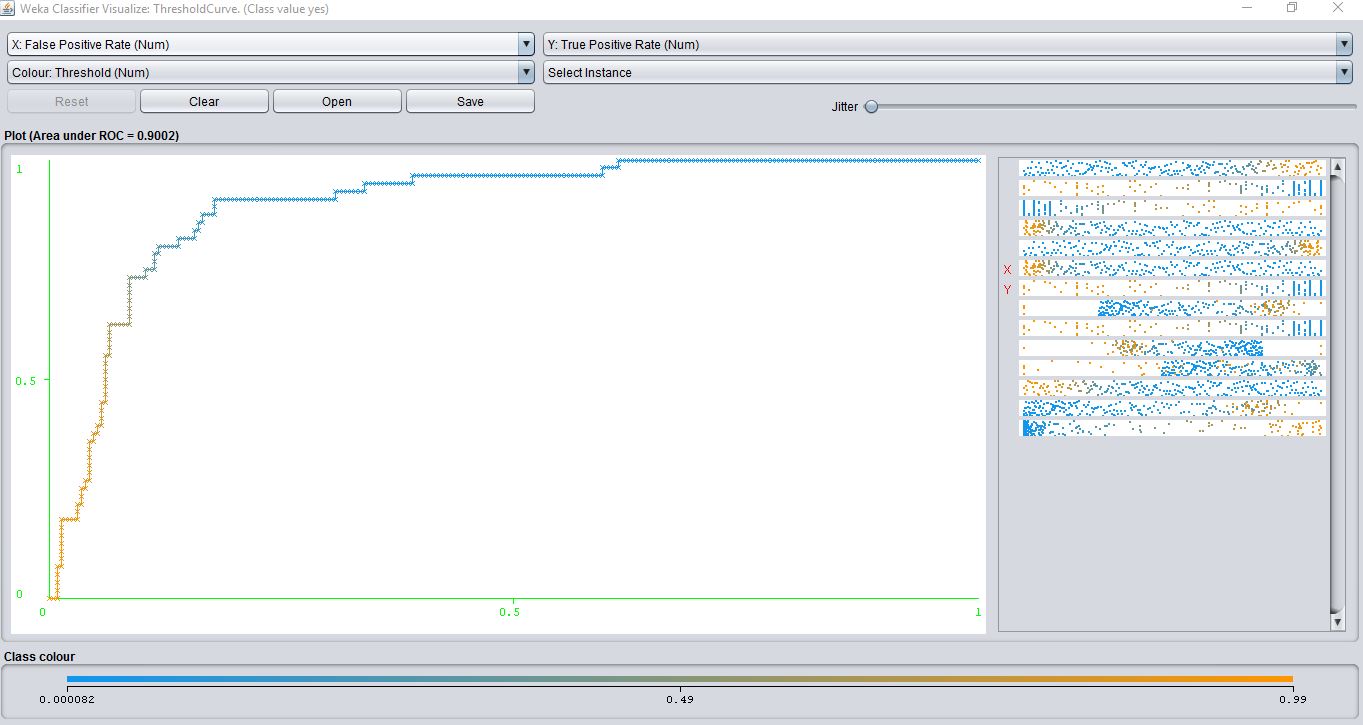


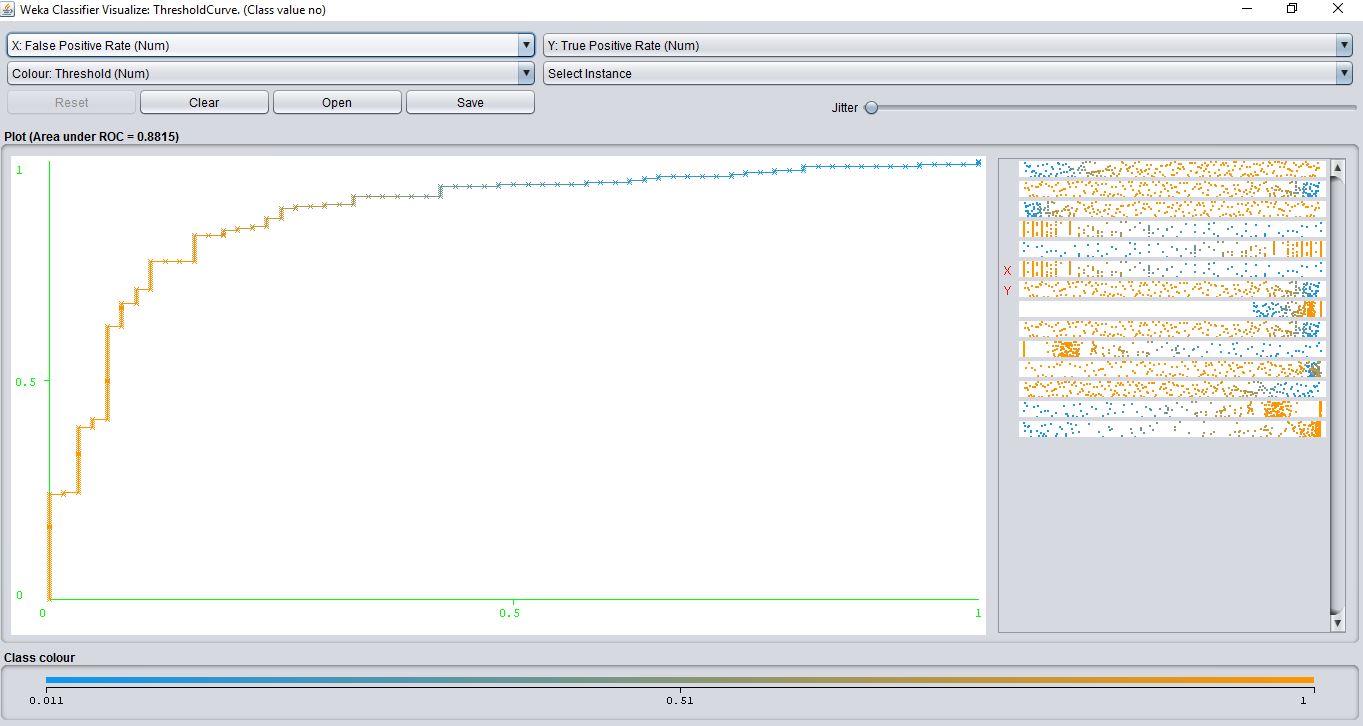
**ii. Explanation of Item**

Notice that the output includes the Mean absolute error, the Root mean squared error, the Relative absolute error, and the Root relative squared error. Even thought that they differ, the relation of the other attributes all have probabilities of the node caps determining yes or no. no being that the cancer has a possibility of spreading or where the nodes are yes nothing cancerous is there. Since we know of the instance we can see through the error rates don’t really matter if we compare the “same measure”.

**iii. Graphical or Other Special Purpose Additional Output**

**This is an ROC curve visualization where we can see on the x axis is the false positive rate and the y axis is the true positive within the data we want to be as close the t y axis as possible for a good classification which we can see on the top left a 90% based on the area under the curve. Whereas the classification of no caps for the lymph nodes being 88% which is ok.**



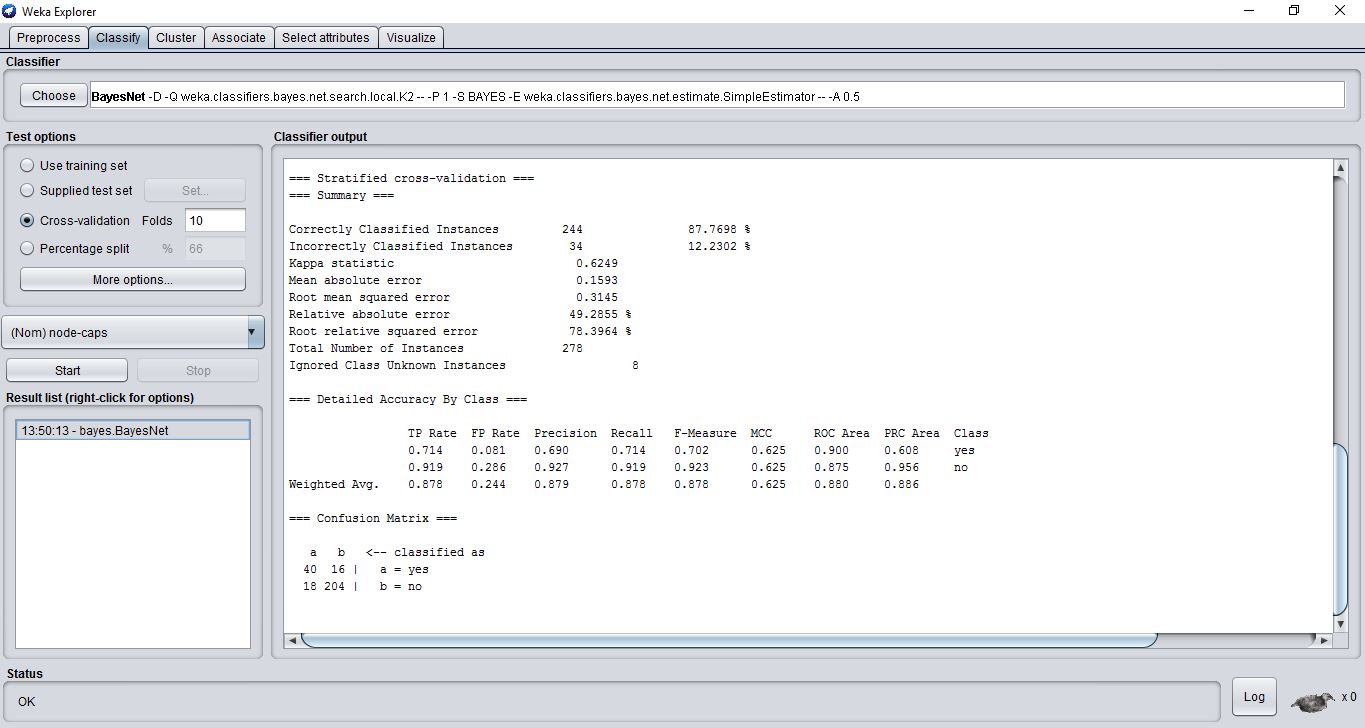


**II. Case 5. Any Kind of Algorithm**

**Name of Algorithm:** Bayes net

**i. Output Results**

What this shows: This screen shot shows the results of the, Bayes net classification algorithm.

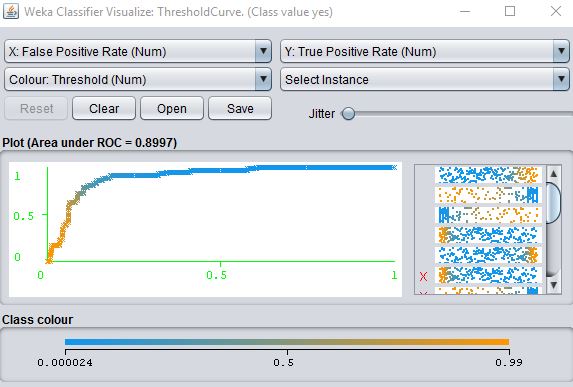
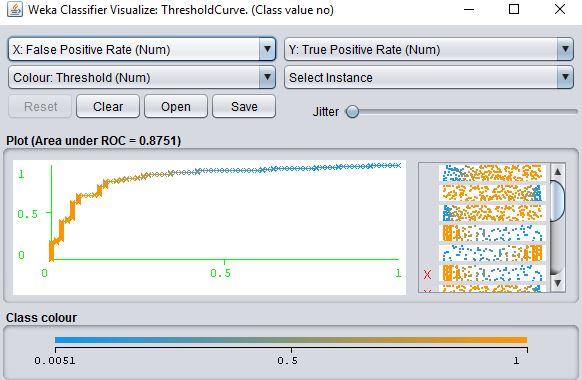


**ii. Explanation of Item**

Utilizing the probabilities of each individual node of the network of variables versus something like multinomial where it gets all possibilities and combinations on n variables. Where it creates derived versions of conditionals is how it can be debatable a more efficient as we can see by the correctness of the class instance being 87.7 versus naïve Bayes which is a slight improvement from 87%. But in the ROC curve results we see that naïve Bayes classifies better by .01 to .03%

**iii. Graphical or Other Special Purpose Additional Output**

As described we can see that our ROC curve in the previous example I thought it was interesting to see that Bayes net performed less regarding the accuracy of the ROC curve. Even thought that naïve Bayes did a better job here we know that naïve Bayes is aggressive and makes things oversimplified since in the real world variable are not truly independent hence naive Bayes assuming all features are conditionally independent is the reason why I said the comparison is debatable.

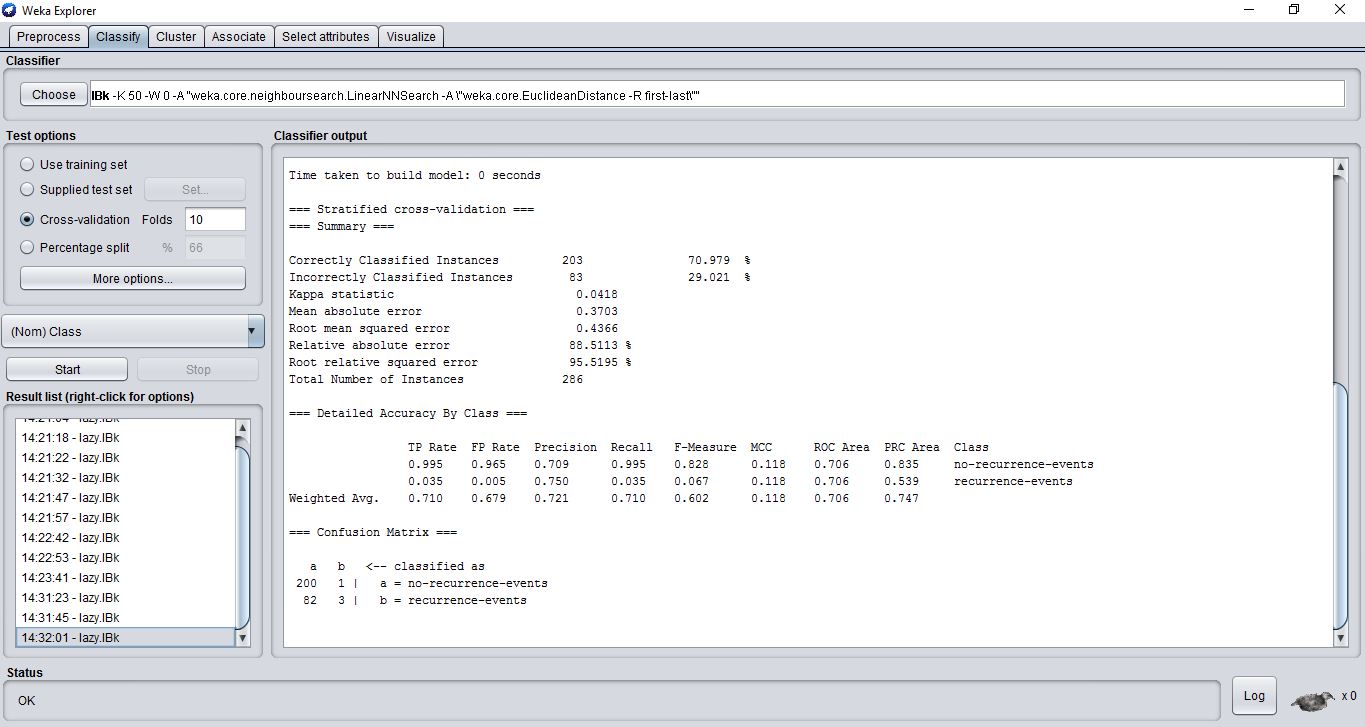
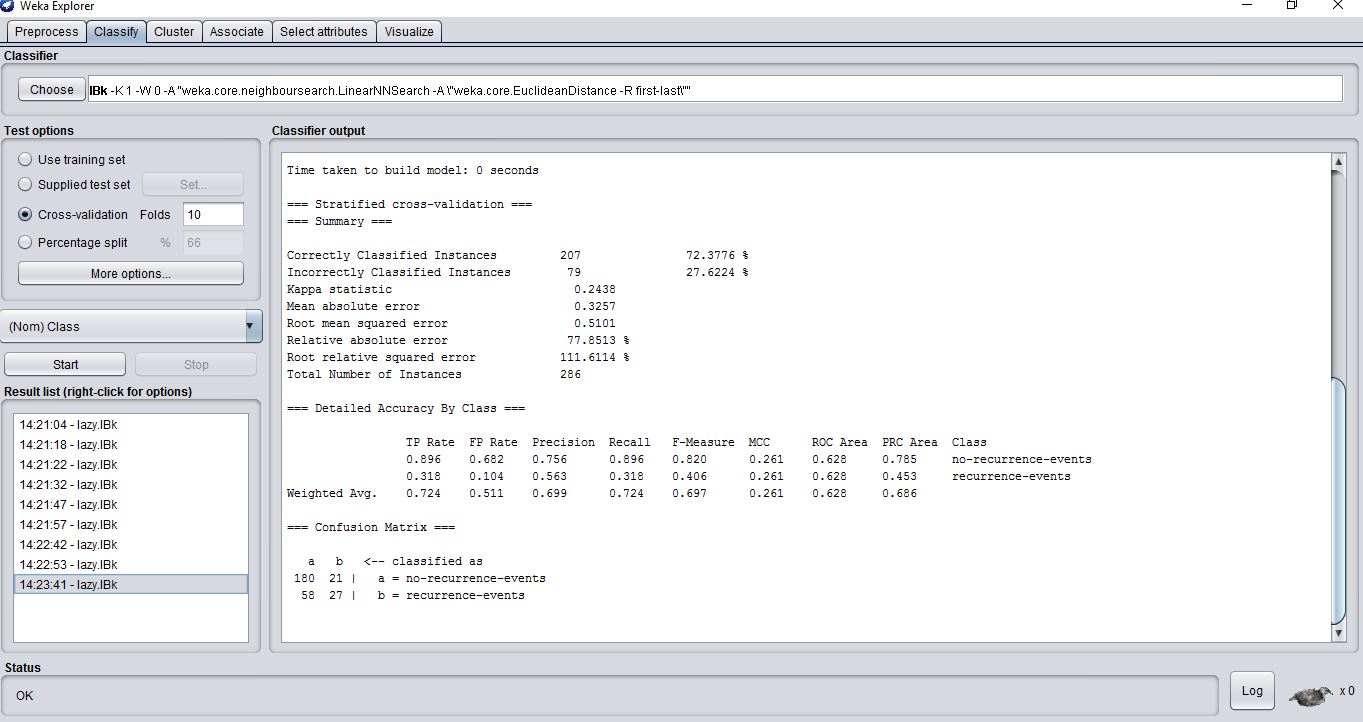


**II. Case 6. Any Kind of Algorithm**

**Name of Algorithm:** IBk (instanced based k- nearest neighbor)

**i. Output Results**

What this shows: This screen shot shows the results of the, instanced based k- nearest neighbor classification algorithm.

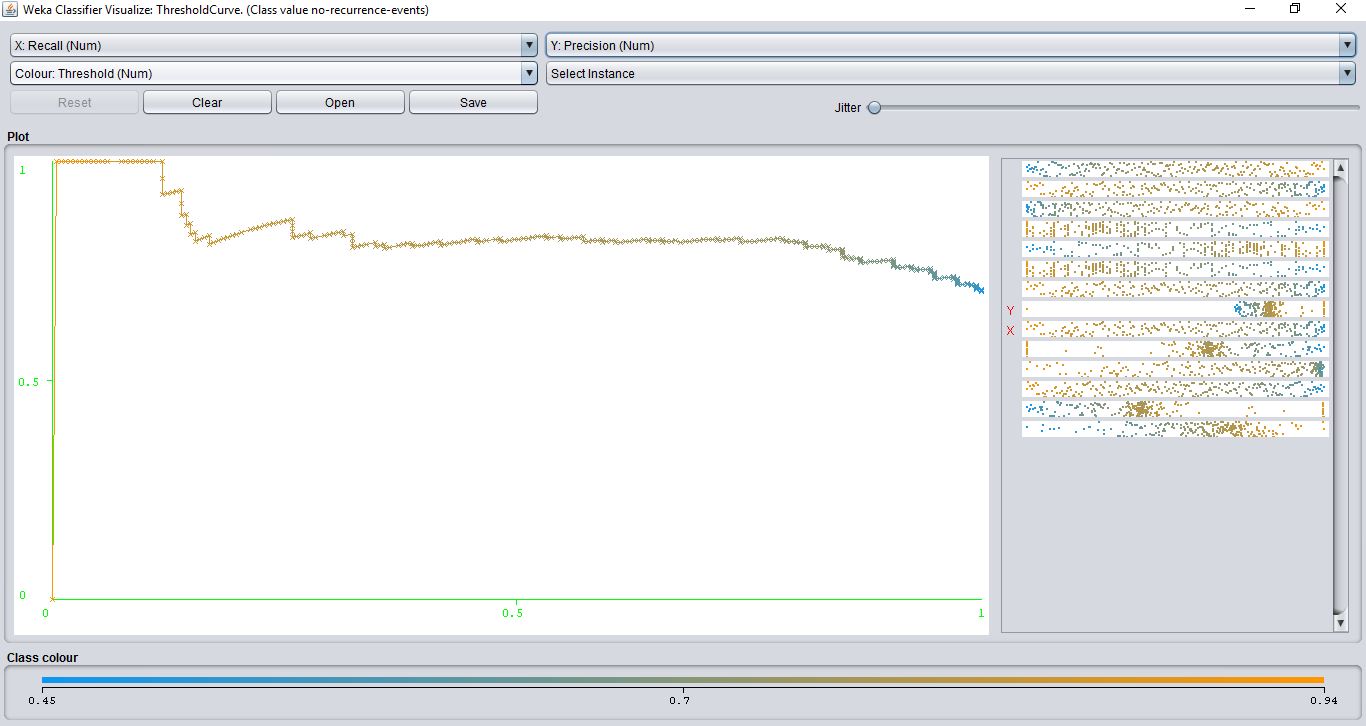


**ii. Explanation of Item**

As we can see in the data that the accuracy of the correctly classified instances is 72% if this data set were to be noisy we would see the decrease of that accuracy and to see if it was noise you would increase the k value to see if the correctly classified instance would increase. Hence in the second picture I increased the k to 50 and the correctness decreased by 2%.

**iii. Graphical or Other Special Purpose Additional Output**

The x axis being true position and the y axis is showing the precision of that in the graph we can see based on your “instanced correctness” the graph shows that it classifies ok, not too bad being that percentage is a 72%. The best classifier we have on the graph will always be (1, 1) or the top right of the graph, where in the graph we can see our best is (1, .72)

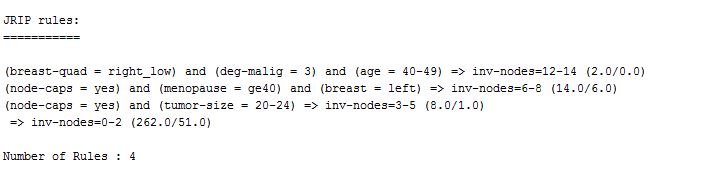
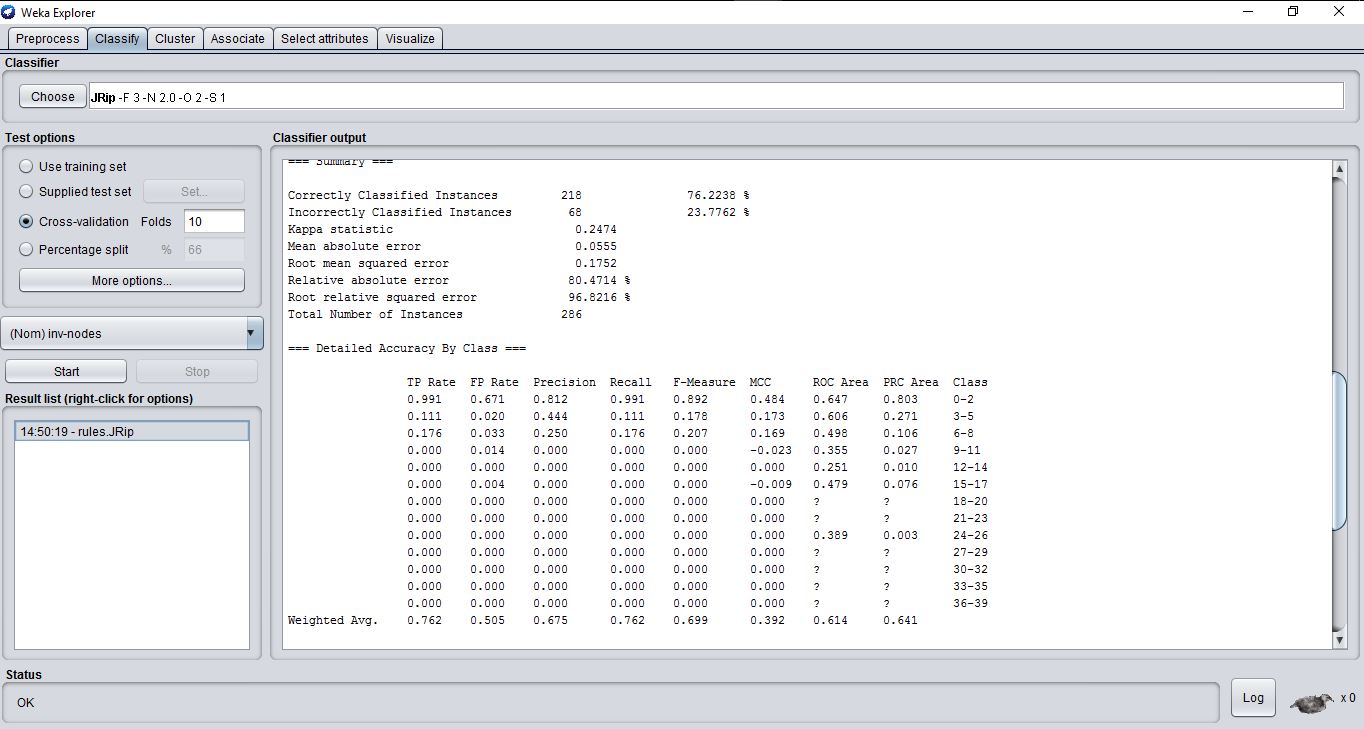


**II. Case 7. Any Kind of Algorithm**

**Name of Algorithm:** JRip

**i. Output Results**

What this shows: This screen shot shows the results of the, JRip classification algorithm.

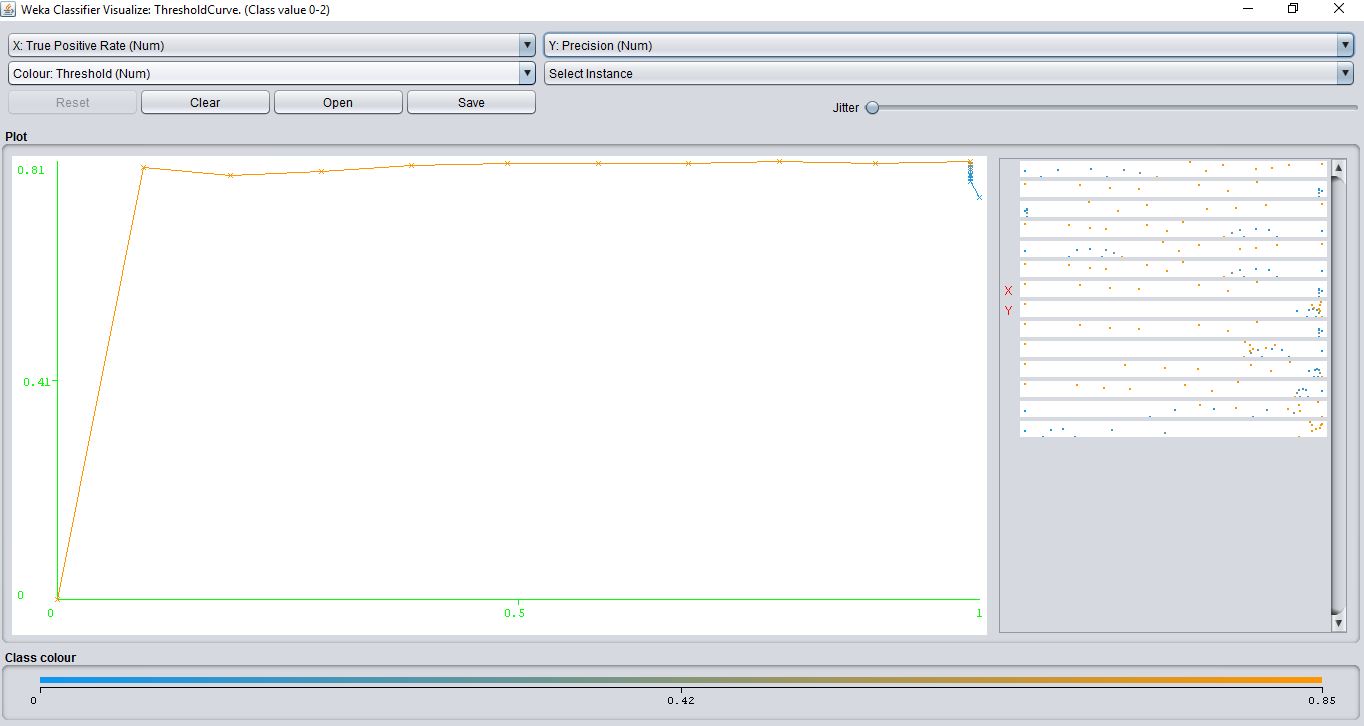


**ii. Explanation of Item**

By these rules created by JRip we can see the lesser rules give a success rate of 76% with rules being set by the other attributes then concluding to the selected attribute. By growing and pruning on the data set to reduce the number of errors following a global optimization step.

**iii. Graphical or Other Special Purpose Additional Output**

As we can see with this data set a very nice precision curve based on the 0-2 nodes being very close to (1, 1) with an 88%.

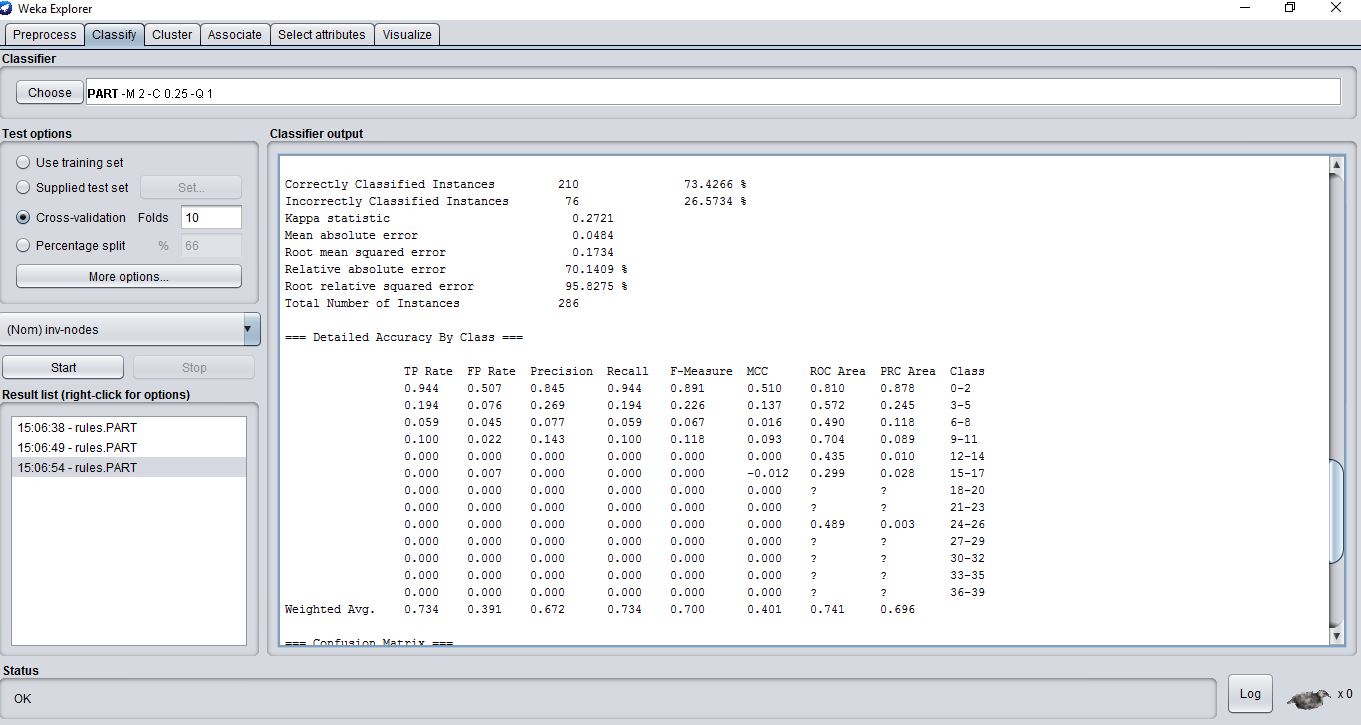


**II. Case 8. Any Kind of Algorithm**

**Name of Algorithm:** PART

**i. Output Results**

What this shows: This screen shot shows the results of the, PART classification algorithm.



node-caps = no AND

irradiat = no: 0-2 (189.6/11.0)

node-caps = no AND

age = 40-49: 0-2 (15.38/3.0)

tumor-size = 20-24: 3-5 (11.4/3.4)

age = 60-69 AND

irradiat = yes: 0-2 (14.0/5.0)

age = 30-39 AND

node-caps = yes: 6-8 (6.0/2.0)

menopause = premeno AND

tumor-size = 25-29 AND

irradiat = yes: 3-5 (4.62/1.62)

menopause = ge40 AND

age = 60-69: 6-8 (5.0/2.0)

breast-quad = left\_low AND

menopause = ge40: 6-8 (7.0/3.0)

breast-quad = left\_low AND

Class = no-recurrence-events: 0-2 (4.0/2.0)

breast-quad = left\_up AND

menopause = premeno AND

breast = right AND

deg-malig = 3: 0-2 (7.0/4.0)

breast-quad = right\_up: 3-5 (5.0/3.0)

breast-quad = left\_up AND

menopause = premeno AND

breast = right: 9-11 (4.0/1.0)

age = 40-49 AND

irradiat = yes: 12-14 (3.0/1.0)

deg-malig = 3 AND

irradiat = no: 15-17 (5.0/2.0)

deg-malig = 2: 0-2 (3.0/1.0)

: 9-11 (2.0)

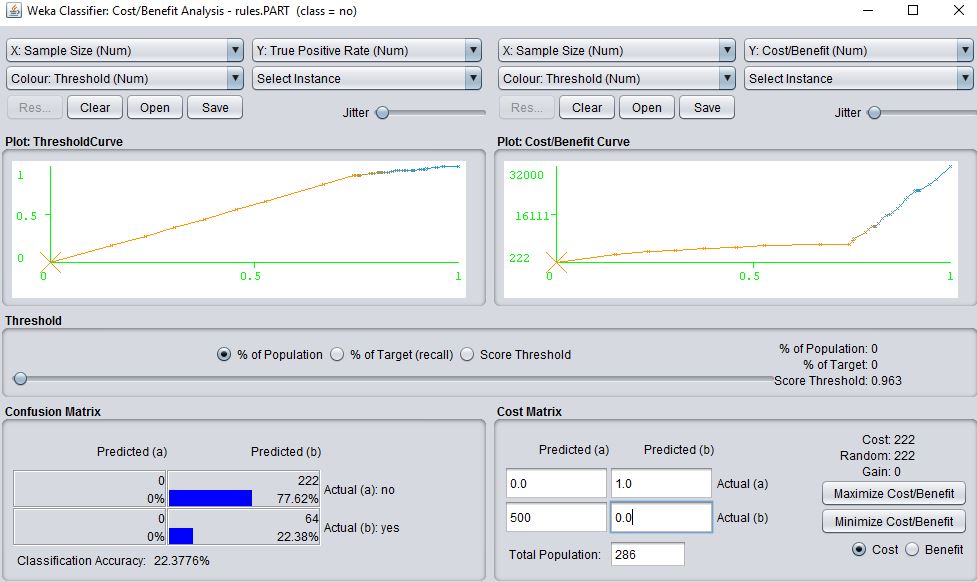
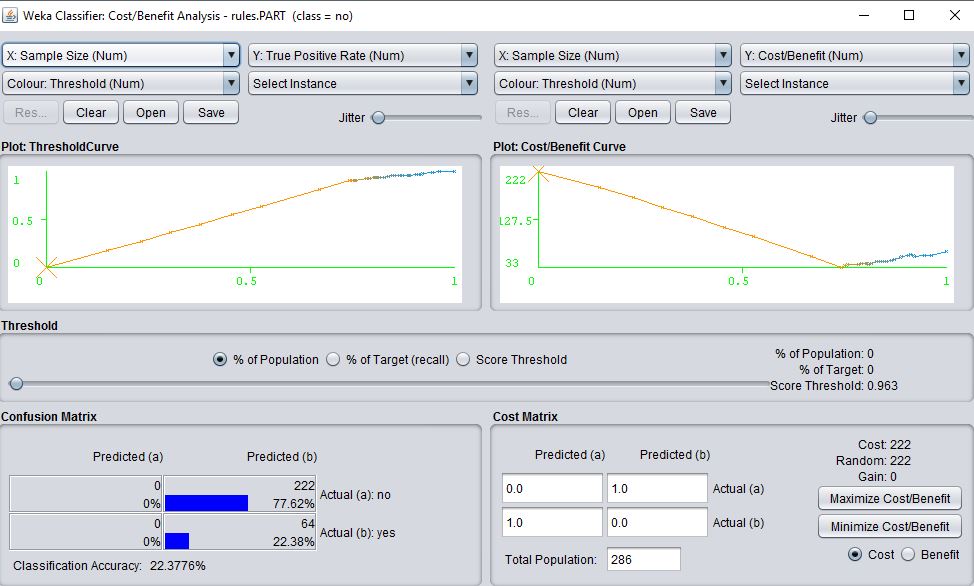
Number of Rules: 16

**ii. Explanation of Item**

These algorithms is the classic divide and conquer where it makes a rule then removes the instances it covers, then continues the process for the remaining instances. Then goes about building a tree then reads off the biggest leaf. Then discards the tree, with a success rate of 73% but test more versus the JRip algorithm;

**iii. Graphical or Other Special Purpose Additional Output**

Just like the mailer for the cost benefit, this shows the possible changes of parameters when we try to classify on both extremes ranging from 1 all the way to 500. On the left shows the costs for FP and FN where on the right it is shown that you are correctly predicting the costs based on the TP.



**III. Choosing the Best Algorithm among the Results**

Taken from Weka results: Jrip

JP sample mean = 76.2238%

JP root mean squared error = .1752

Squaring the value above:

JP mean squared error = .03069504

Taken from Weka results: PART

PT sample mean = 73.4266%

PT root mean squared error = .1734

Squaring the value above:

PT mean squared error = .03006756

This is my estimate of the standard deviation of the t statistic where the divisor is 10 because I opted for the default 10-fold cross-validation in Weka:

Estimate of paired root mean squared error (EPRMSE)

= square root ((JP mean squared error / 10) + (PT mean squared error / 10))

= .0779503688

t statistic

= (JP sample mean – PT sample mean) / EPRMSE

=35.8843715

Hence based on the value we can see that JRip wins!

**III.A. Random Babbling:**

It’s pretty cool to see that previously over the history j48 was considered the go to algorithm but as the years go by its pretty cool to see improvements as well as better algorithms. Now looking at stats a bit more closely I can never just take them at face value. Giving appreciation to people to people in the media doing data mining correctly rather than just spewing numbers for headlines.

Utilizing what I learned here the original plan was to use a gaming data set based on a Blizzard game that would show how the AI could go about formulating a goal based on different attributes the player would input and how the AI adjusts to the player level for a sense of fairness.

I still loved how this data set turned out through the research because my mom is a breast cancer survivor and it is nice to see what people do behind the scenes to make sure my mom is better ☺.

**III.B. an Application of the Paired t-test:**

Taken from Weka results:

NB sample mean = 87.054%

NB root mean squared error = .3127

Squaring the value above:

NB mean squared error = .09778129

Taken from Weka results:

BN sample mean = 87.7698%

BN root mean squared error = .3145

Squaring the value above:

BN mean squared error = .09891025

This is my estimate of the standard deviation of the t statistic where the divisor is 10 because I opted for the default 10-fold cross-validation in Weka:

Estimate of paired root mean squared error (EPRMSE)

= square root(( NB mean squared error / 10) + (BN mean squared error / 10))

= .1402467611

t statistic

= (NB sample mean – BN sample mean) / EPRMSE

=- 5.104

Based on this result we can see that the value is small so we can conclude that the alternate is better than the other algorithm.