Scenario 2: AI Classifier Development, Optimization and Selection

**1. Introduction to the task**

This task requires our company to proceed with design and development by working on a decision making system to classify the patient’s severity of whether the mass is malignant or benign. Two different types of classifiers will be engaged in classifying the data to find the optimum classification and therefore a conclusion will be reached on which classifier performed the best and why. **2. AI classifiers**

What are decision trees?

Decision trees are classifiers of supervised learning that represent a hierarchical tree structure consisting of three elements that are responsible for the anatomy of a decision tree (Jenhani, et al., 2008). The root node comprises of a decision primarily based on a test attribute. Thus, an edge or arc corresponds to a given value of the test attribute which then leads to decisions reaching a single leaf node labelled by classification which is then applied.

Representation of a Decision Tree.

Figure 1. Decision Tree Diagram

Question

Yes

No

No

Yes

Yes

No

Decision tree construction

The construction of a decision tree is based on a training set. The decision tree initially starts with an empty tree and selects the best fit test attribute for each decision node. (Jenhani, et al., 2008), contributes that the main focus by selecting attributes is by reducing the mixture of classes among training sets, making it accessible to determine the object classes.

There are different steps involved in proper construction of a decision tree namely;

1. Splitting.

Splitting involves the procedure of partitioning datasets into further subsets. These splits are generally made on a variable within a location. Each split is therefore summarized with the predictor variable used for the split, and the values for the predictor variable are split among the left and right child nodes.

1. Pruning.

Pruning involves the procedure of growing a large tree, then pruning it to its optimal size by reducing nodes that provide less information (Hastie, et al., 2001). Other methods include using a validation dataset, by dividing the sample in two sections and testing the model of the training set on the validation dataset, cross validation involving diving the dataset in 10 folds and testing the model developed from 9 folds on the 10th fold which is then repeated for 10 combinations.

Two types of pruning exist which are pre-pruning (forward pruning) and post-pruning (backward pruning).

Pre-pruning involves restriction of growth of the decision tree during the construction. Post-pruning administers the removal of parts of a decision tree after construction by trying to simplify the tree by removing nodes. Due to the removal, post-pruning discards more resources than pre-pruning but it is practically more powerful in simplification due to the access of the full decision tree for measure of exploration and impact of removal of nodes (S.E & R, 2007).

Information Gain.

Information gain is described as the entropy reduction to an attribute while splitting a node (Mitchell, 1997).

𝐺𝑎𝑖𝑛 𝑅𝑎𝑡𝑖𝑜(𝐵) = 𝐺𝑎𝑖𝑛(𝐵)/𝑆𝑝𝑙𝑖𝑡𝐼𝑛𝑓𝑜(𝐵)

Gain(B) is the reduction of information that is required which is caused by knowing (B).

Decision tree classification using J48.

Decision tree classification describes how the classification process by questioning features associated with the data. (Kingsford & Salzberg, 2008), adds that every question is comprised in a single node and each internal node points to a child node for a possible outcome.

The J48 decision tree serves as an algorithm that generates rules for the prediction of the target or class variable (Nadali, et al., 2011). J48 is an extension of the algorithm ID3 which has additional features such as computing missing values, decision tree pruning, continuous attribute value ranges (Kaur & Chhabra, 2014).

**A detailed description of decision tree parameters (individual research required) and how they might affect the performance of the classifier.**

* binarySplits.

The binary splits involve a process where the tree is grown by considering one nominal value above all other nominal values which results in a tree where 2 branches are formed from a node.

* confidenceFactor

The confidence factor is responsible for how aggressive the pruning procedure will be. If the value for the confidence factor is higher, it is assumed that the data being assessed is a good representation for all events and less pruning will occur which therefore affects the performance of the classifier.

* minNumObj

The minimum number of objects parameter determines the threshold value (Patel & Upadhyay, 2012).

**A brief description / introduction to artificial neural networks**

The term Artificial Neural Networks refers to a mathematical model that distributes architecture which consist of processing nodes with multiple connections. Each input to the neuron have their own weights, where the weights are floating numbers which are adjusted when training the network. There are many different types of Artificial Neural Networks for various types of problems, from performing pattern recognition and modelling memory which are in turn split into two different classes (Goldberg, 2001). Supervised classes tend to refer to a network that attempts to learn the relationship between data and a parameter, while unsupervised classes refer to a network that naturally finds groupings within a dataset independently. The focus for classifying the data set will be on a Multi-Layer Perceptron.

What is a Multi-Layer Perceptron?

A multilayer perceptron is an ANN architecture that is trained by using the back propagation algorithm (Panchal, et al., 2011). It comprises of several layers of nodes, where the bottom layer is an input layer from which external information is received. The output layer is where the solution to the problem is attained. In addition both the input and output layers are separated by what is called a hidden layer. All layers in an MLP contain neurons which interlink with each other from the input layer, to the hidden layer to the output layer by weights. These weights are then adjusted in accordance to mapping capabilities of the trained network.

Training a Multilayer Perceptron Network.

In order to train a multi-layer perceptron, certain instances must be looked at;

Blue (input layer) Green (hidden layer)

Red (output layer)

Figure. 2

. Multilayer Perceptron (MLP)

* The number of hidden layers to use.
* The number of neurons present to be in a hidden layer
* Validation of the neural network to test for over-fitting.

**The Use of Input layers and input nodes.**

According to (Zhang, et al., 1998), the number of input nodes is critical due to the nodes containing important information about the complexity of the non-linear pattern structure within the data.

**The Use of hidden layers and nodes.**

The hidden nodes within a hidden layer allow neural networks to capture the pattern within the data and perform the complex non-linear aligning between the input layer and the output layer. The use of one hidden layer is normally sufficient to approximate with the complexity of non-linear patterns for desired accuracy (Hornik, et al., 1989) , but due to the use of just a single hidden layer, it requires a large number of nodes which would have an effect on the training time. (Zhang, 1994), states that by working with a network containing two hidden layers, the predictions are more accurate alongside the modelling of the data structure.

**The Use of output layers.**

The number of output nodes is just one as it is directly related to the problem under study.

**The connection of nodes in a MLP (Multi-Layer Perceptron)**

Network architecture within an MLP consists of connections made between input nodes, hidden nodes and output nodes in each layer. The parameters of the network include the number of inputs per neuron, the weights, the activation function and number of layers which must be met for the neural network to learn the data well (Moussa, et al., 2006).

The Multilayer Perceptron (MLP) uses the back propagation algorithm for prediction. (Alsmadi, et al., 2009), contributes some rules in usage of the algorithm.

Rule number one suggests that due to the complexity of the relationship between input data and outcome increases, the number of hidden elements in the hidden layer must also be increased.

Rule number two suggests that additional hidden layers might be required if the process is separable into multiple stages.

Rule number three suggests that the amount of training data available sets an upper bound to the number of processing elements in hidden layers.

.

**A detailed description of artificial neural network parameters (individual research required) and how they might affect the performance of the classifier**

* Hidden Layers

The hidden layer parameter determines the number of hidden layers and hidden neurons to have in a neural network. Typically, one hidden layer is sufficient for the data set.

* Learning Rate

The learning rate parameter determines how far the weights change with response to the data set. The choice of whether to decrease or increase the learning rate greatly affects the classifiers accuracy when training. A learning rate that is too large will take longer to train data than a learning rate with a small value (Wilson & Tony, 2001).

* Momentum

(Anantwar & Shelke, 2012), contributes that the use of the momentum rate is to achieve the minimum faster if it is altered, otherwise it would take longer to attain the minimum. The momentum parameter for artificial neural networks updates weights in the equal direction, thus speeding up the training time for different applications (Istook & Martinez, 2002).

**3. The data set**

The data set used for the tests are from the UCI Machine Learning repository, which is the Mammographic Mass Data Set. The data set contains 516 benign and 445 malignant cases, which totals to 961 instances. Each of these instances comprises of 5 different attributes namely; BI-RADS, Age, Shape, Margin, Density and Severity.

Mammographic Mass Dataset Attributes

|  |  |
| --- | --- |
| Attributes | Ranges |
| BI-RADS | 1 to 5 (ordinal, non-predictive) |
| Age | Patient’s Age in years (integer) |
| Shape | Mass Shape: Round(1); Oval(2); Lobular(3); Irregular(4) (nominal) |
| Margin | Mass Margin: Circumscribed(1); Micro lobulated(2); Obscured(3); Illdefined(4); Spiculated(5) (nominal) |
| Density | Mass Density: High(1); ISO(2); Low(3); Fat-containing(4) (ordinal) |
| Severity | Benign(0) or Malignant(1) (binominal) (class) |

Missing Attributes

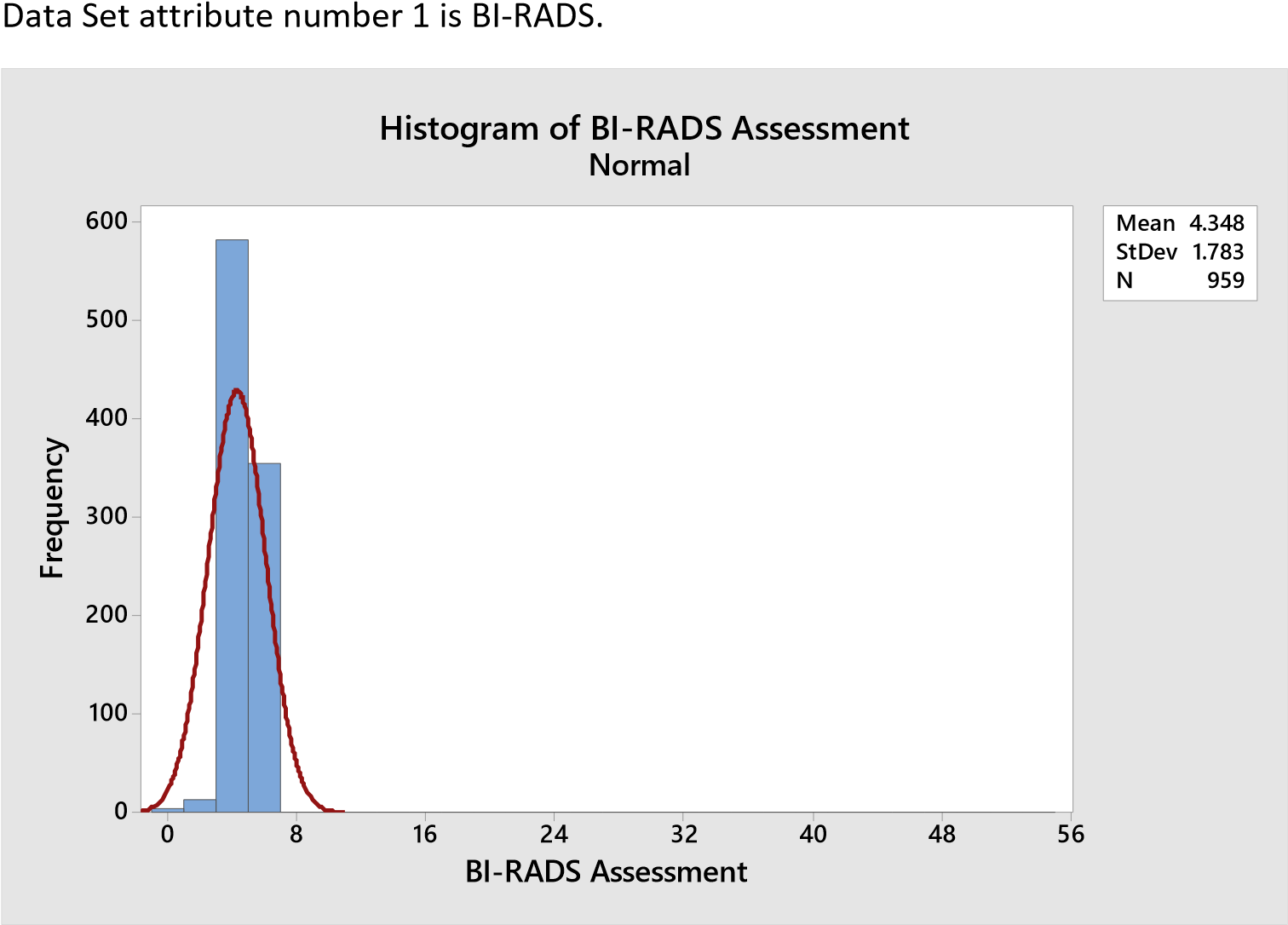
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BI-RADS | Age | Shape | Margin | Density |
| 2 | 5 | 31 | 48 | 76 |

A number of missing attributes were discovered within the Mammographic Mass Data Set.

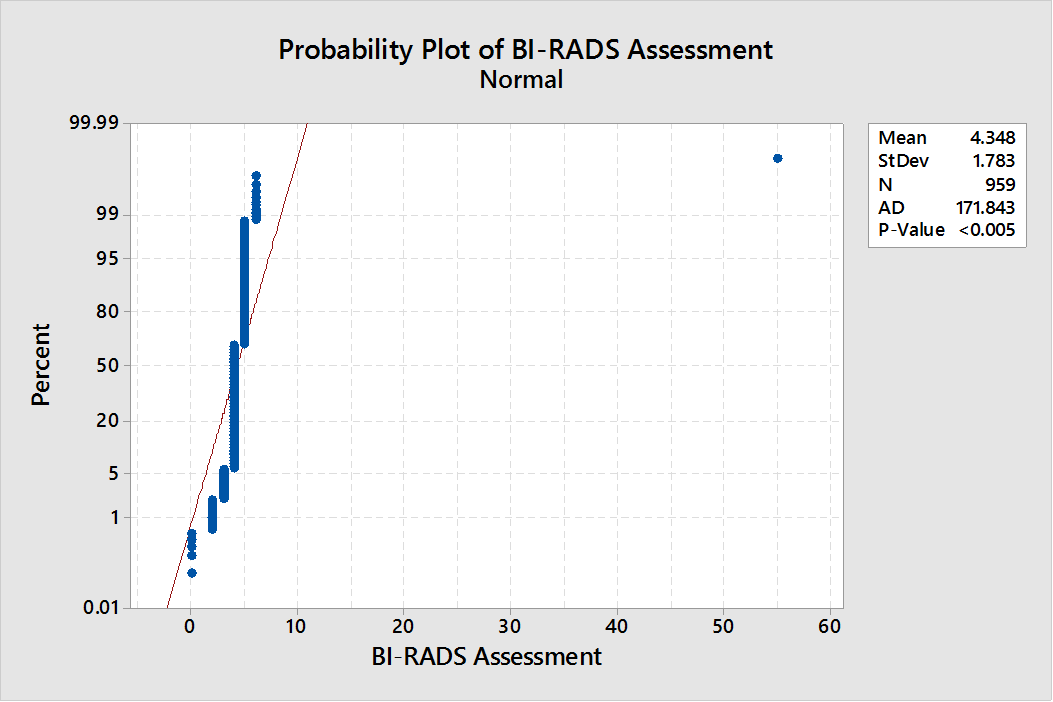
• Description of the data set attributes (e.g.):

i. Distribution.

The distribution of data set attributes will be explained by a normal distribution graph describing the correlation of the histogram and the ideal normal distribution curve, along with applying a statistical probability plot for the data set attributes.

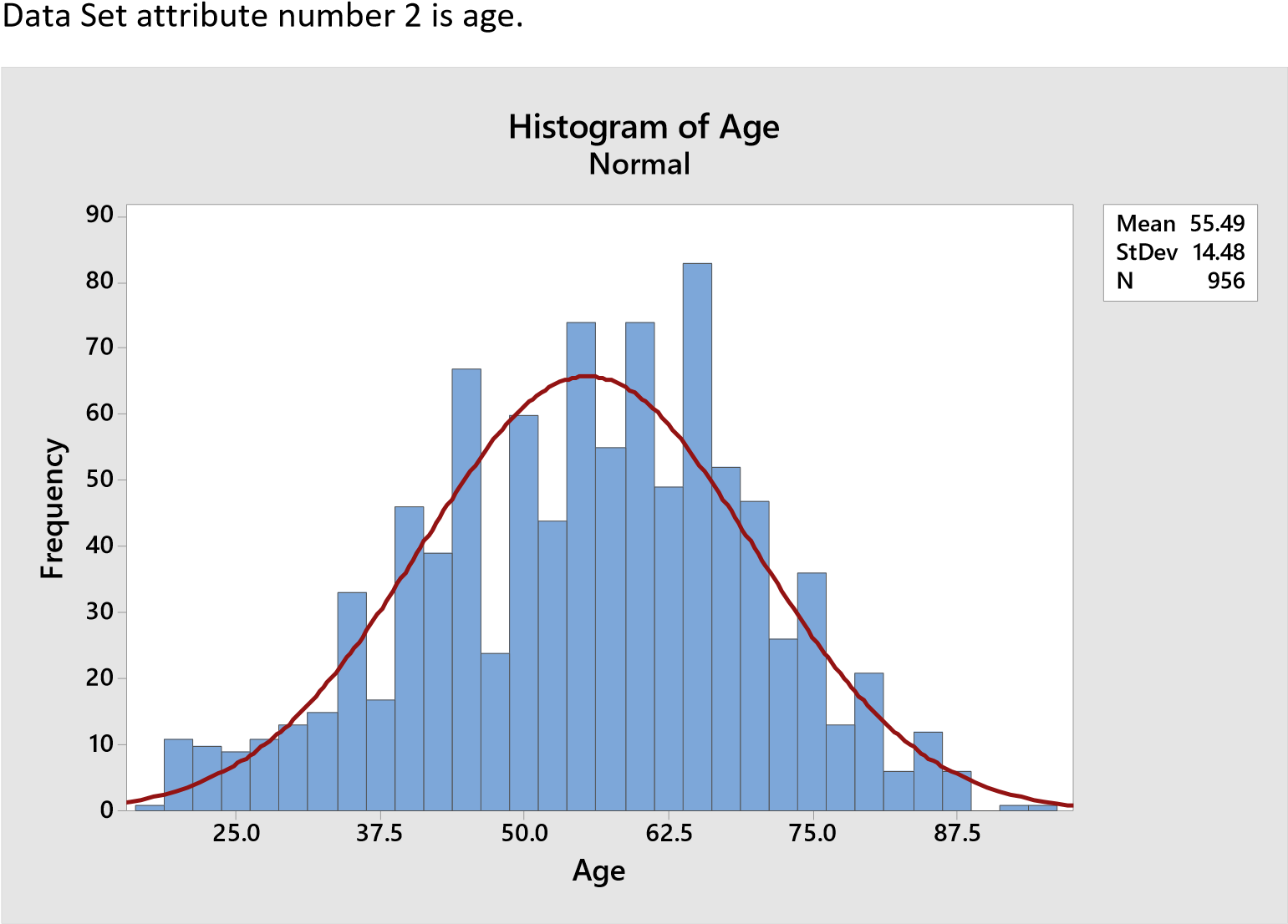


The plot resembles an ideal normal distribution, but the histogram does not resemble that.



The probability plot for the BI-RADS suggest that the points do not follow a normal distribution.

This suggests that the probability is non-normal. The P-Value is at < 0.005 which states that the data is not normally distributed.



The plot resembles an ideal normal distribution, but the histogram does not resemble that.



The probability plot for Age suggest that the points do not follow a normal distribution.

This suggests that the probability is non-normal. The P-Value is at < 0.005 which states that the data is not normally distributed.

The plot resembles an ideal normal distribution, but the histogram does not resemble that.

Data Set attribute number 3 is Shape.

**5**

**4**

**3**

**2**

**1**

**0**

**4**

**0**

**0**

**3**

**0**

**0**

**2**

**0**

**0**

**1**

**0**

**0**

**0**

**M**

**e**

**a**

**n**

**2**

**.**

**7**

**2**

**2**

**S**

**t**

**D**

**e**

**v**

**1**

**.**

**2**

**4**

**3**

**N**

**9**

**3**

**0**

**S**

**h**

**a**

**p**

**e**

**F**

**r**

**e**

**q**

**u**

**e**

**n**

**c**

**y**

**H**

**i**

**s**

**t**

**o**

**g**

**r**

**a**

**m**

**o**

**f**

**S**

**h**

**a**

**p**

**e**

**N**

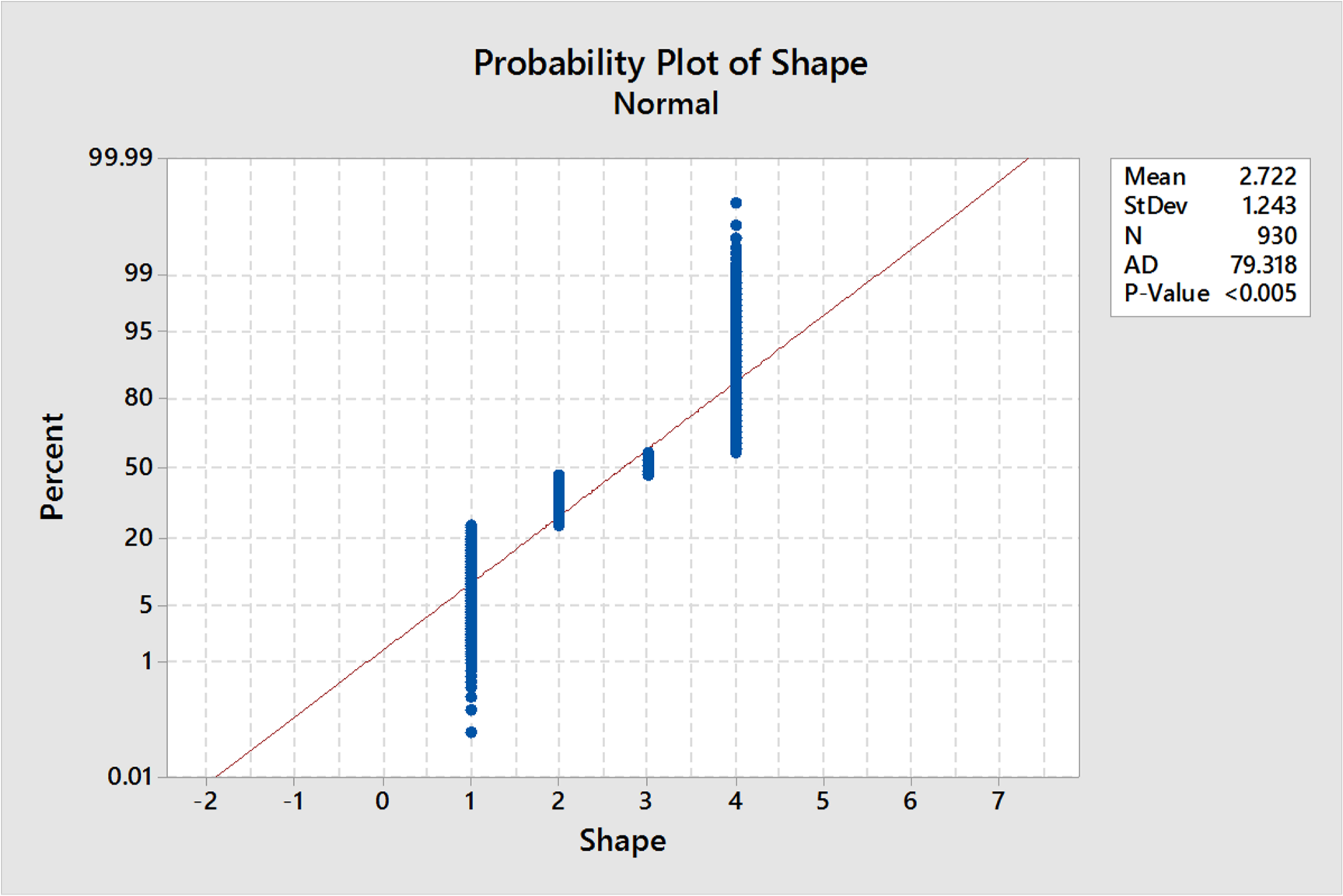
**o**

**r**

**m**

**a**

**l**



The probability plot for Shape suggest that the points do not follow a normal distribution.

Value is at < 0.005 which states that the data

The plot resembles an ideal normal distribution, but the histogram does not resemble that.

Data Set attribute number 4 is Margin.

**6**

**5**

**4**

**3**

**2**

**1**

**0**

**4**

**0**

**0**

**3**

**0**

**0**

**2**

**0**

**0**

**1**

**0**

**0**

**0**

**M**

**e**

**a**

**n**

**2**

**.**

**7**

**9**

**6**

**S**

**t**

**D**

**e**

**v**

**1**

**.**

**5**

**6**

**7**

**N**

**9**

**1**

**3**

**M**

**a**

**r**

**g**

**i**

**n**

**F**

**r**

**e**

**q**

**u**

**e**

**n**

**c**

**y**

**H**

**i**

**s**

**t**

**o**

**g**

**r**

**a**

**m**

**o**

**f**

**M**

**a**

**r**

**g**

**i**

**n**

**N**

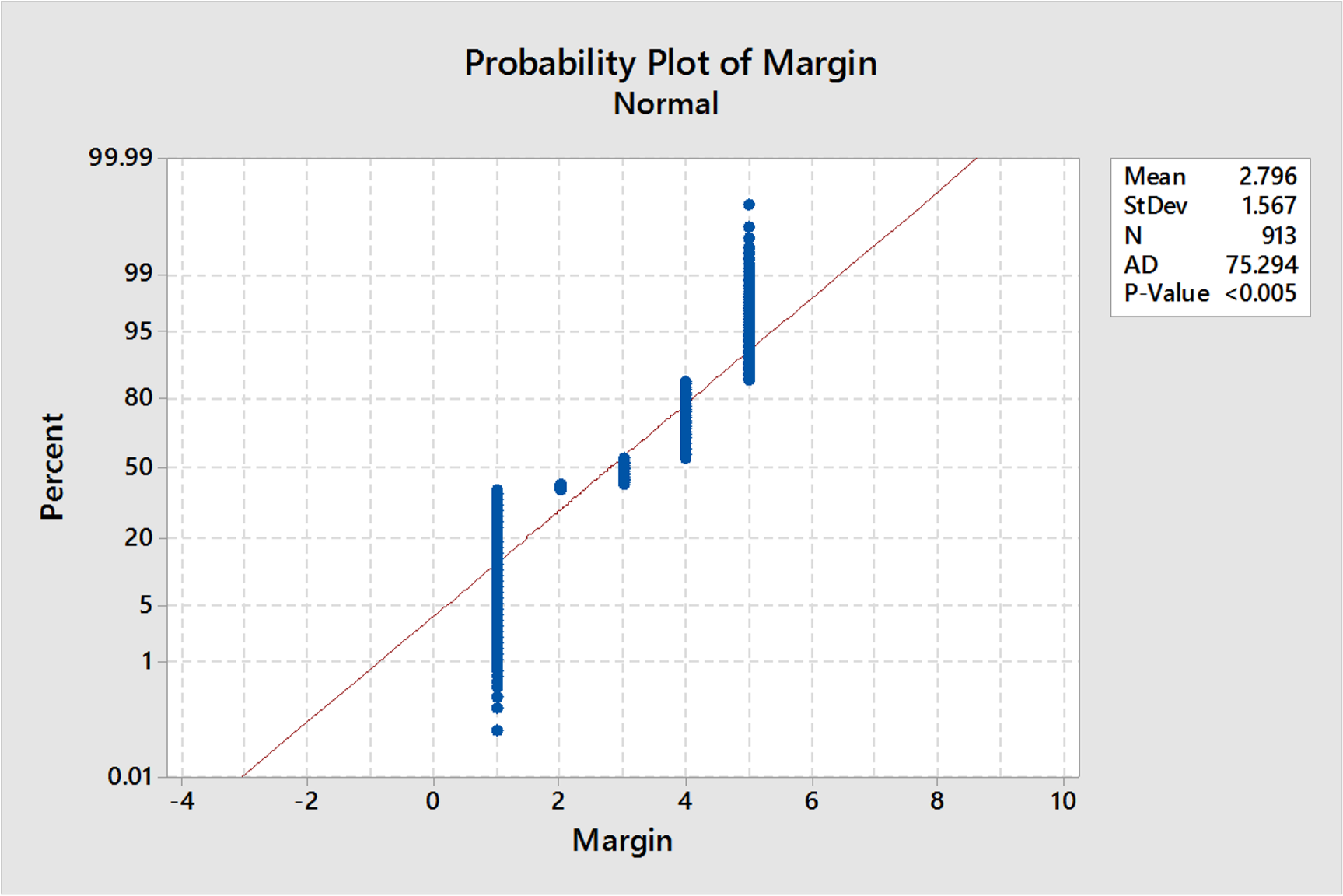
**o**

**r**

**m**

**a**

**l**



The probability plot for the Margin suggest that the points do not follow a normal distribution.

Value is at < 0.005 which states that the data

The plot resembles an ideal normal distribution, but the histogram does not resemble that.

Data Set attribute number 5 is Density.

**4**

**3**

**2**

**1**

**9**

**0**

**0**

**8**

**0**

**0**

**7**

**0**

**0**

**6**

**0**

**0**

**5**

**0**

**0**

**4**

**0**

**0**

**3**

**0**

**0**

**2**

**0**

**0**

**1**

**0**

**0**

**0**

**M**

**e**

**a**

**n**

**2**

**.**

**9**

**1**

**1**

**S**

**t**

**D**

**e**

**v**

**0**

**.**

**3**

**8**

**0**

**4**

**N**

**8**

**8**

**5**

**D**

**e**

**n**

**s**

**i**

**t**

**y**

**F**

**r**

**e**

**q**

**u**

**e**

**n**

**c**

**y**

**H**

**i**

**s**

**t**

**o**

**g**

**r**

**a**

**m**

**o**

**f**

**D**

**e**

**n**

**s**

**i**

**t**

**y**

**N**

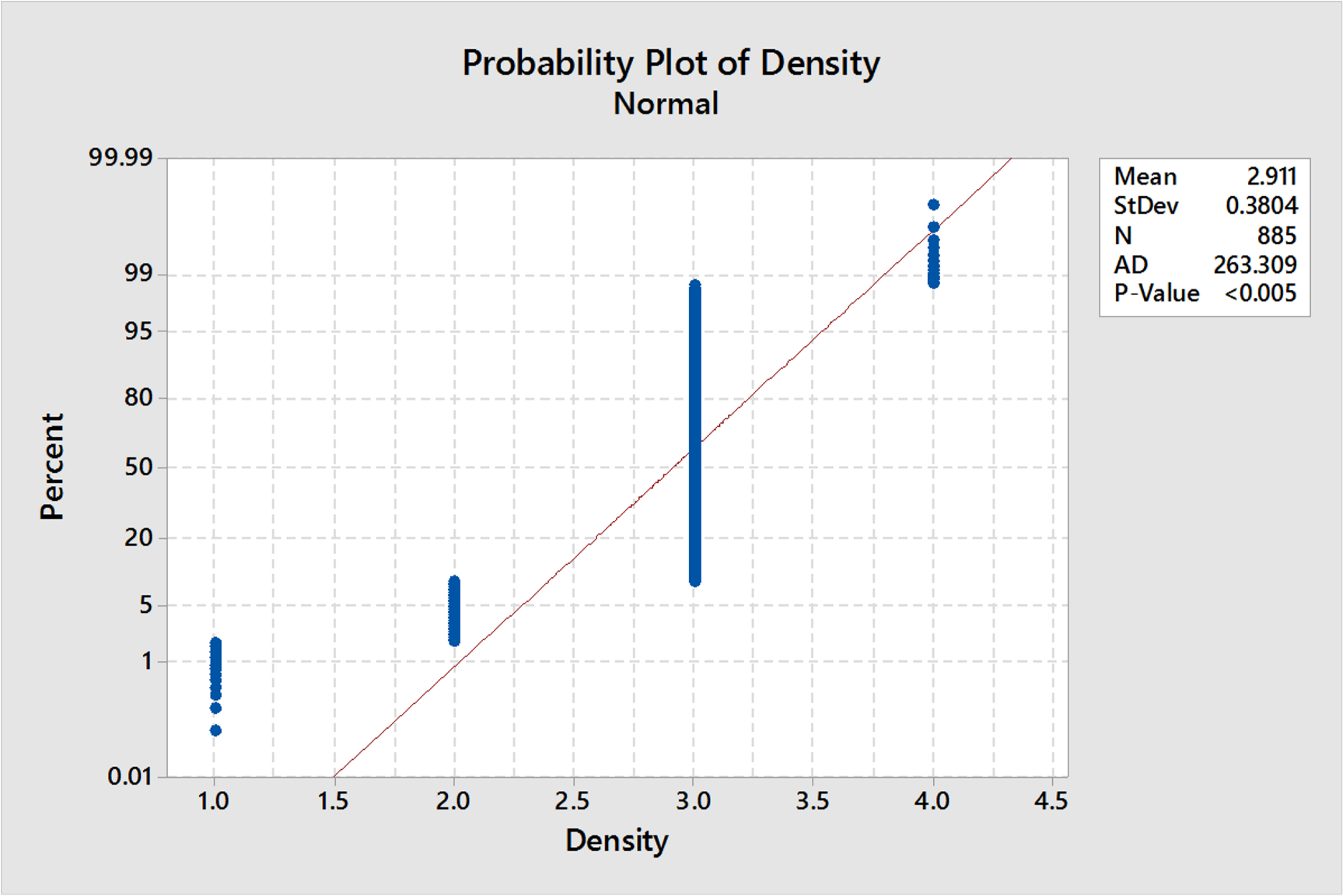
**o**

**r**

**m**

**a**

**l**



1. predictive, (individual research required)

In the mammography data set there are 4 predictive attributes. From understanding the data, the age, shape, density and margin.

The predictive value of breast biopsy leads to 70% of unnecessary breast biopsises.

1. Outliers

Outliers for BI

-

RADS.

**6**

**0**

**5**

**0**

**4**

**0**

**3**

**0**

**2**

**0**

**1**

**0**

**0**

**B**

**I**

**-**

**R**

**A**

**D**

**S**

**A**

**s**

**s**

**e**

**s**

**s**

**m**

**e**

**n**

**t**

**B**

**o**

**x**

**p**

**l**

**o**

**t**

**o**

**f**

**B**

**I**

**-**

**R**

**A**

**D**

**S**

**A**

**s**

**s**

**e**

**s**

**s**

**m**

**e**

**n**

**t**

Outliers

for Age

**1**

**0**

**0**

**9**

**0**

**8**

**0**

**7**

**0**

**6**

**0**

**5**

**0**

**4**

**0**

**3**

**0**

**2**

**0**

**1**

**0**

**A**

**g**

**e**

**B**

**o**

**x**

**p**

**l**

**o**

**t**

**o**

**f**

**A**

**g**

**e**

Outliers for Shape.

**4**

**.**

**0**

**3**

**.**

**5**

**3**

**.**

**0**

**2**

**.**

**5**

**2**

**.**

**0**

**1**

**.**

**5**

**1**

**.**

**0**

**S**

**h**

**a**

**p**

**e**

**B**

**o**

**x**

**p**

**l**

**o**

**t**

**o**

**f**

**S**

**h**

**a**

**p**

**e**

Outliers for Margin.

**5**

**4**

**3**

**2**

**1**

**M**

**a**

**r**

**g**

**i**

**n**

**B**

**o**

**x**

**p**

**l**

**o**

**t**

**o**

**f**

**M**

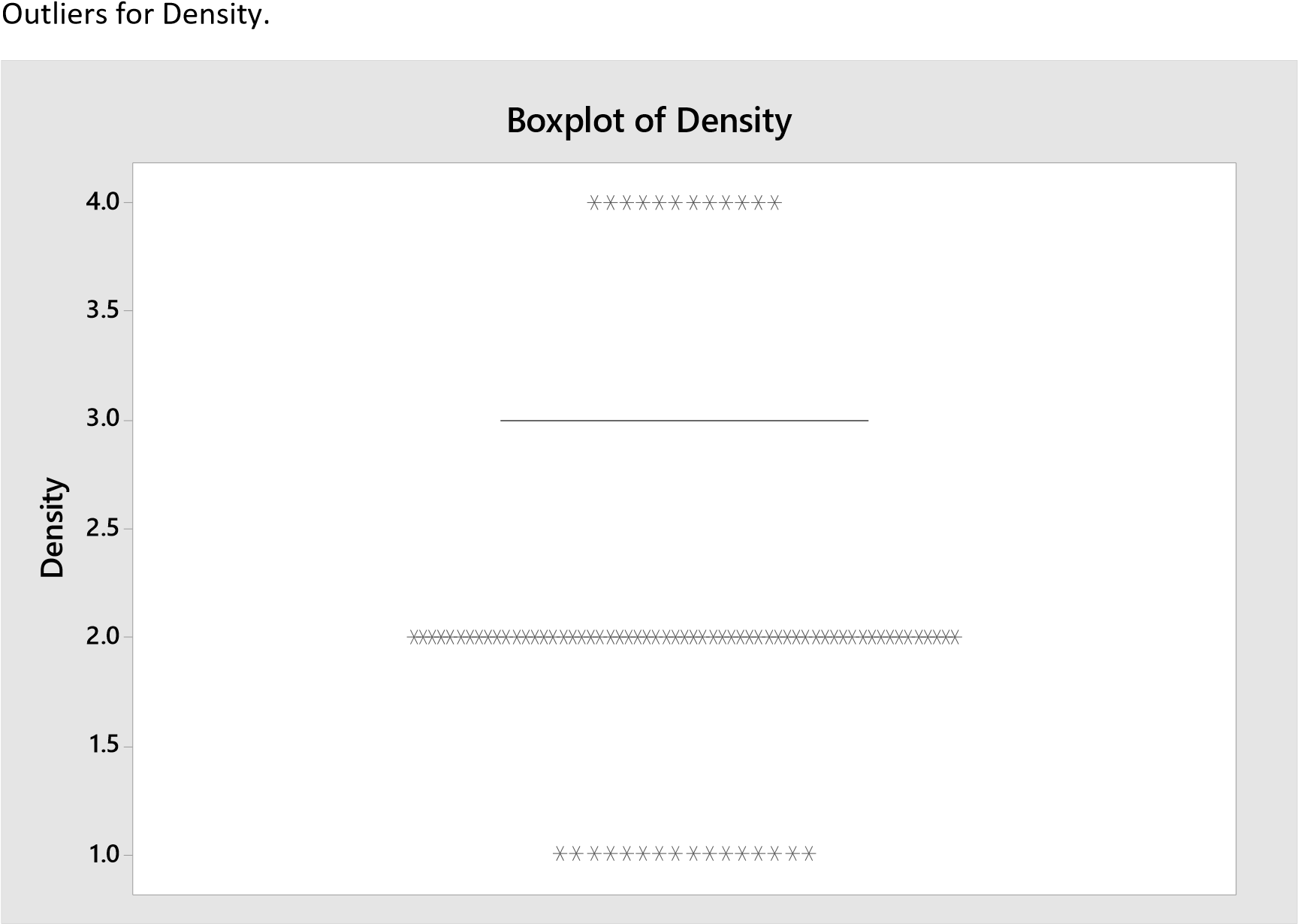
**a**

**r**

**g**

**i**

**n**



1. Attribute measurement scales e.g. nominal, ratio etc.

Attribute measurement scales consist of:

* + Nominal ( Simple Categories)

The nominal categories are Shape and Margin. The mass shape can be categorized as

Round = 1, Oval = 2, Lobular = 3, and Irregular = 4.

The mass margin can be categorized as Circumscribed = 1, Micro lobulated = 2, Obscured = 3, IllDefined = 4, Spiculated = 5.

* + Ordinal (Values that can be sorted by order)

The first ordinal attribute as described by the data set are the BI-RADS assessment ranging from 1 being definitely benign to 5 being highly suggestive of malignancy. The second ordinal attribute is described as mass density of high = 1, ISO = 2, low = 3, and fat-containing = 4.

1. **Predicting of the best classifier for the dataset**

Based on the evidence from sections 2 and 3, what are the strengths and weaknesses of a decision tree when applied to the dataset.

Strengths of decision trees (Bhargava, et al., 2013).

* + Decision trees generate rules that are easily understood by users.
  + Decision trees make up for the lack of missing attributes by either removing values if the data set tends to a non-normally distributed data or imputing values if the data set is normally distributed.
  + Decision trees work well with numeric and categorical variables.

Weaknesses of decision trees.

* + Decision trees produce large trees which reduce the understanding by users.
  + Decision trees require target values to be discrete.

Based on the evidence from sections 2 and 3, what are the strengths and weaknesses of an artificial neural network when applied to the dataset.

Strengths of Artificial Neural Networks.

* + Artificial Neural Networks can easily learn about complex non-linear relationships between the inputs and outputs of data.
  + Artificial Neural Networks can predict unseen data.

Weaknesses of Artificial Neural Networks.

* + Artificial Neural Networks take a longer time to classify the data set due to great computational procedures.
  + Artificial Neural Networks are prone to over-fitting due to a number of hidden layers (Tu, 1996).
  + Prediction: which classifier is likely to be best for the dataset (before you have done any experiments) Individual research required for this section.

The classifier which would best accompany the dataset would be the J48 decision tree due to its strengths to generate easy to understand rules and removal or imputation of missing values.

1. **Initial observations and experiments with pre-processing**

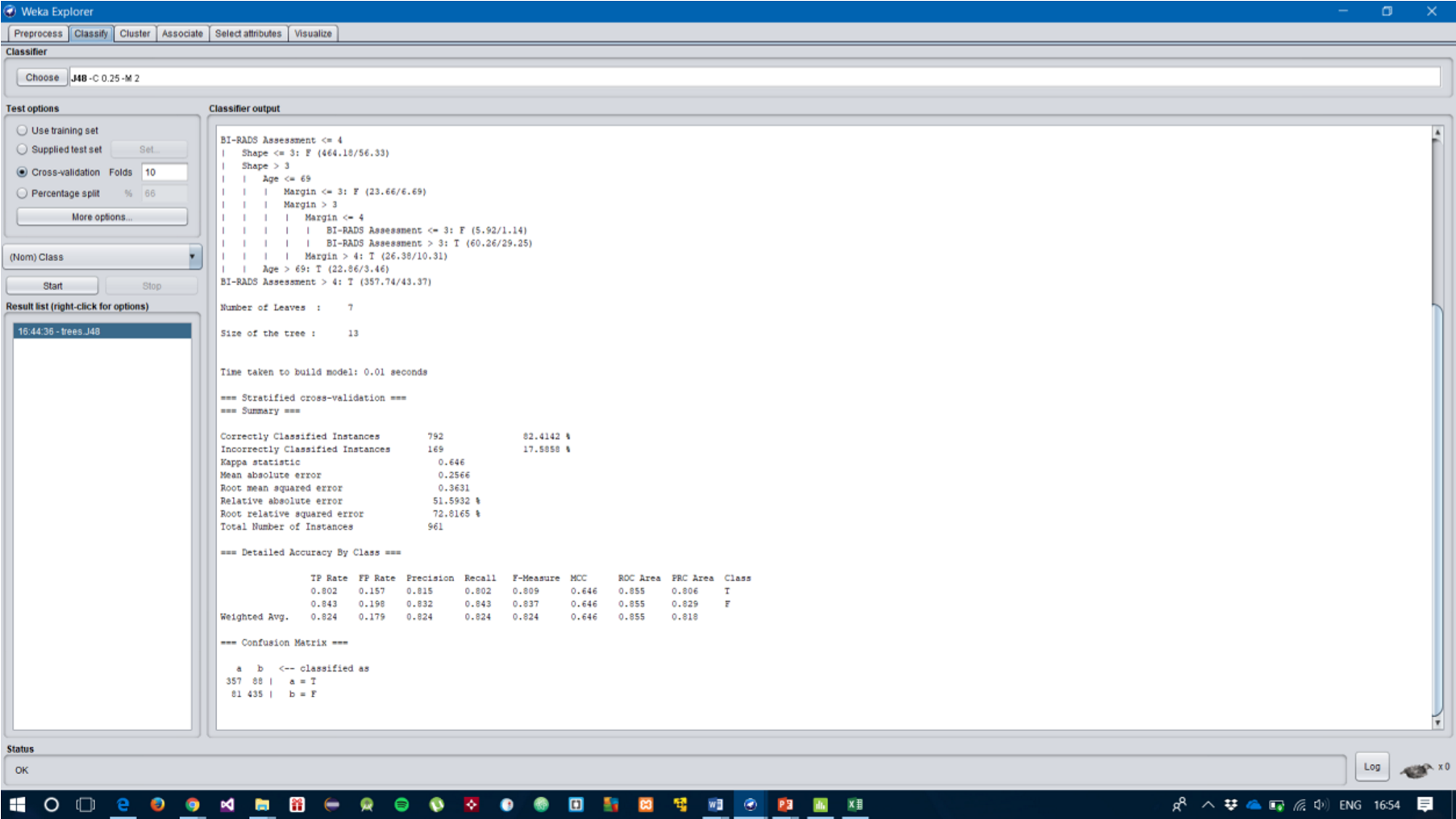
* Strategy for missing attribute values. Justified with evidence/argument

The strategy for missing attribute values becomes apparent where the data set tested for normal distribution outputs as a non-normal distribution along with the data set attributes of the nominal and ordinal scales which are not suitable for replacement. The ideal action for this data set would be to remove the missing attributes values marked with a “?”.

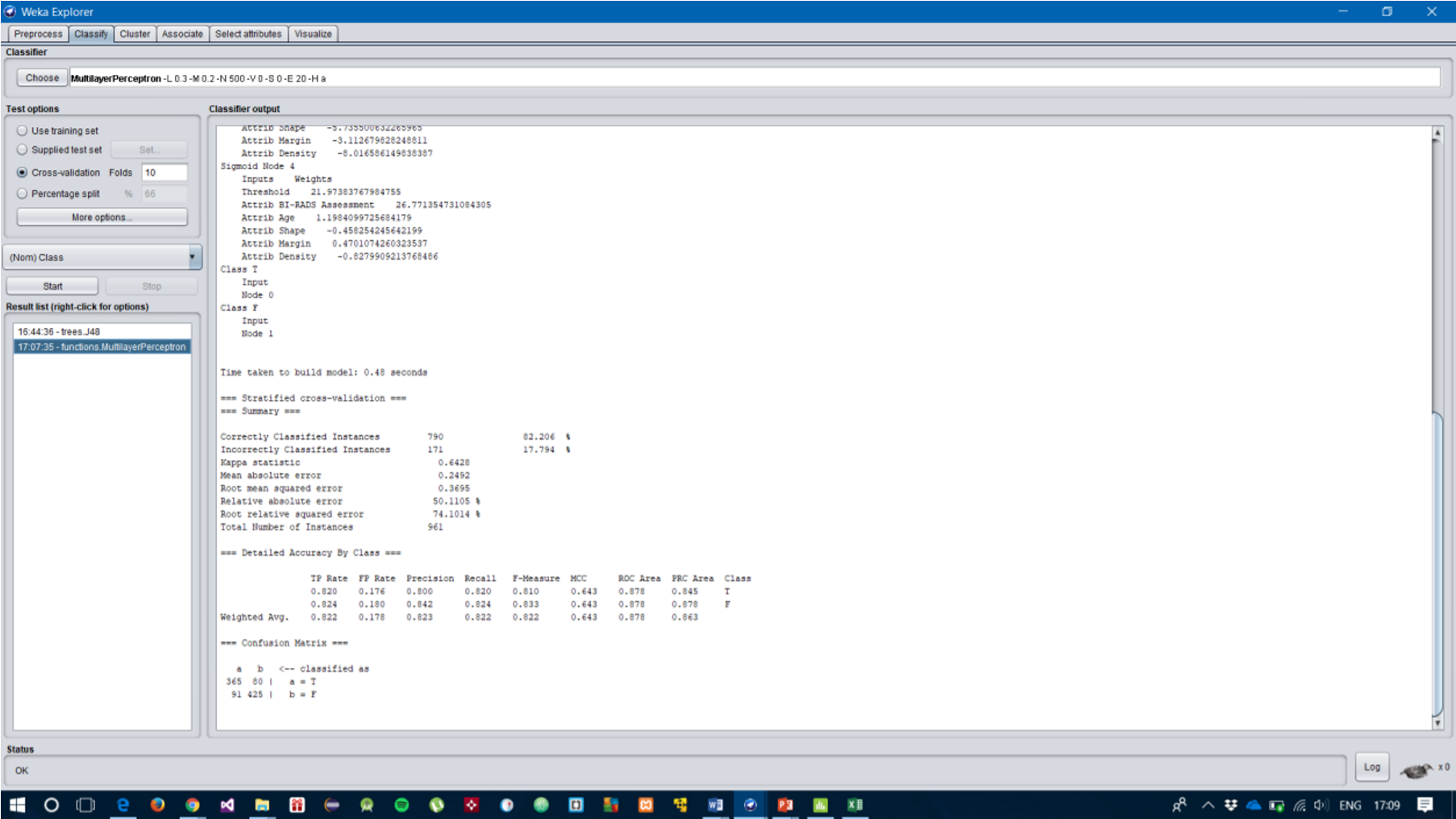
* Strategy for outliers, justified with evidence/argument

The strategy for outliers is dependent on box plots which show the outliers for different set attributes such as the BI-RADS assessment and the density. In order for the outliers to be removed, WEKA explorer will be used where the inter-quartile range function finds outliers within the data set to remove them accordingly.

* Reported results for experiments to develop / test initial strategies (Individual student judgment to be used on how far to go with each type of classifier in experiments – leave plenty of time for real experiments later)

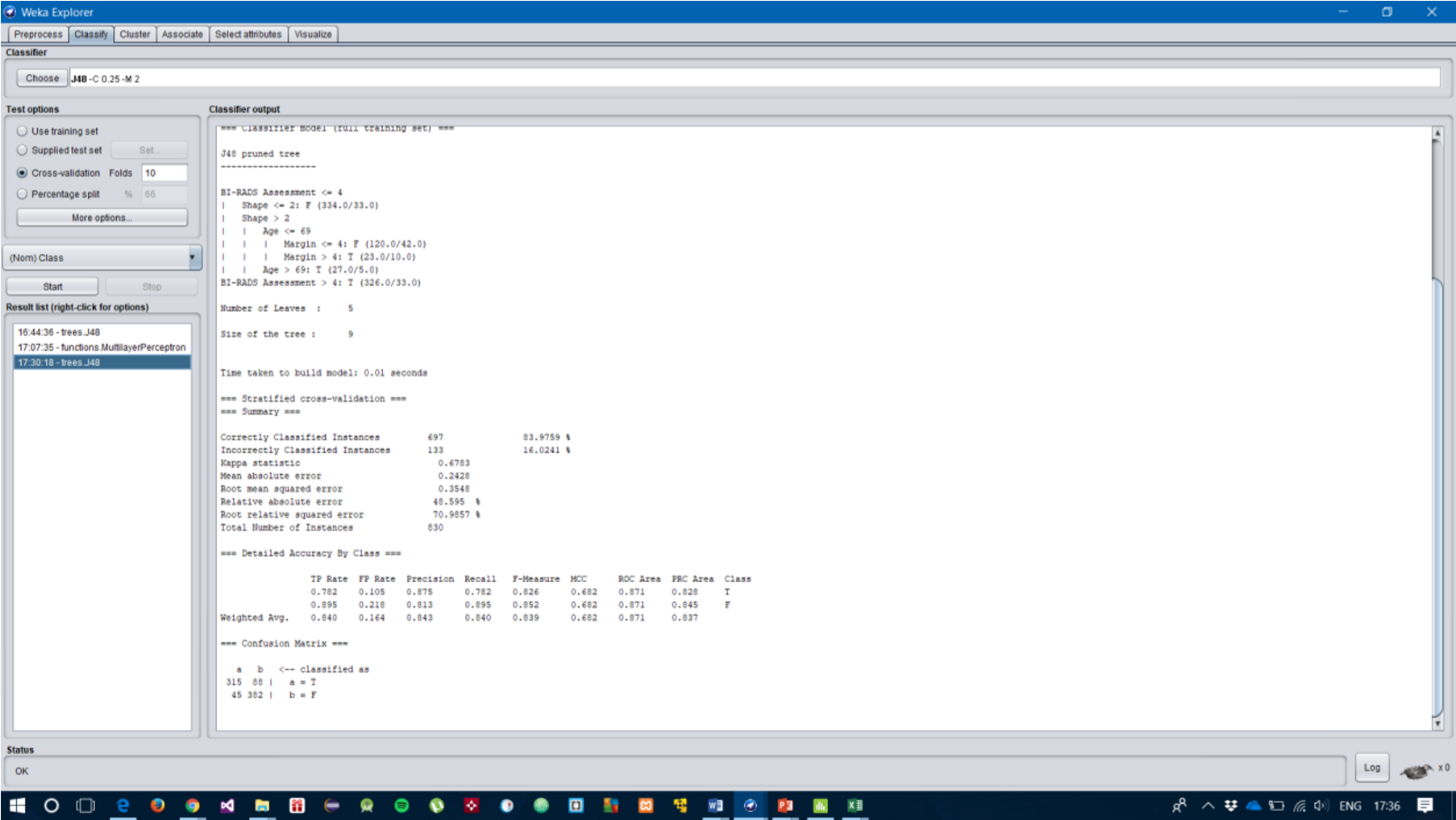


Using the WEKA explorer, the data set was classified with a J48 classifier with all default values to classify the data and the output is seen in the screen shot where the number of correctly classified instances read 82.4142% and the number of incorrectly classified instances reads 17.5858%.

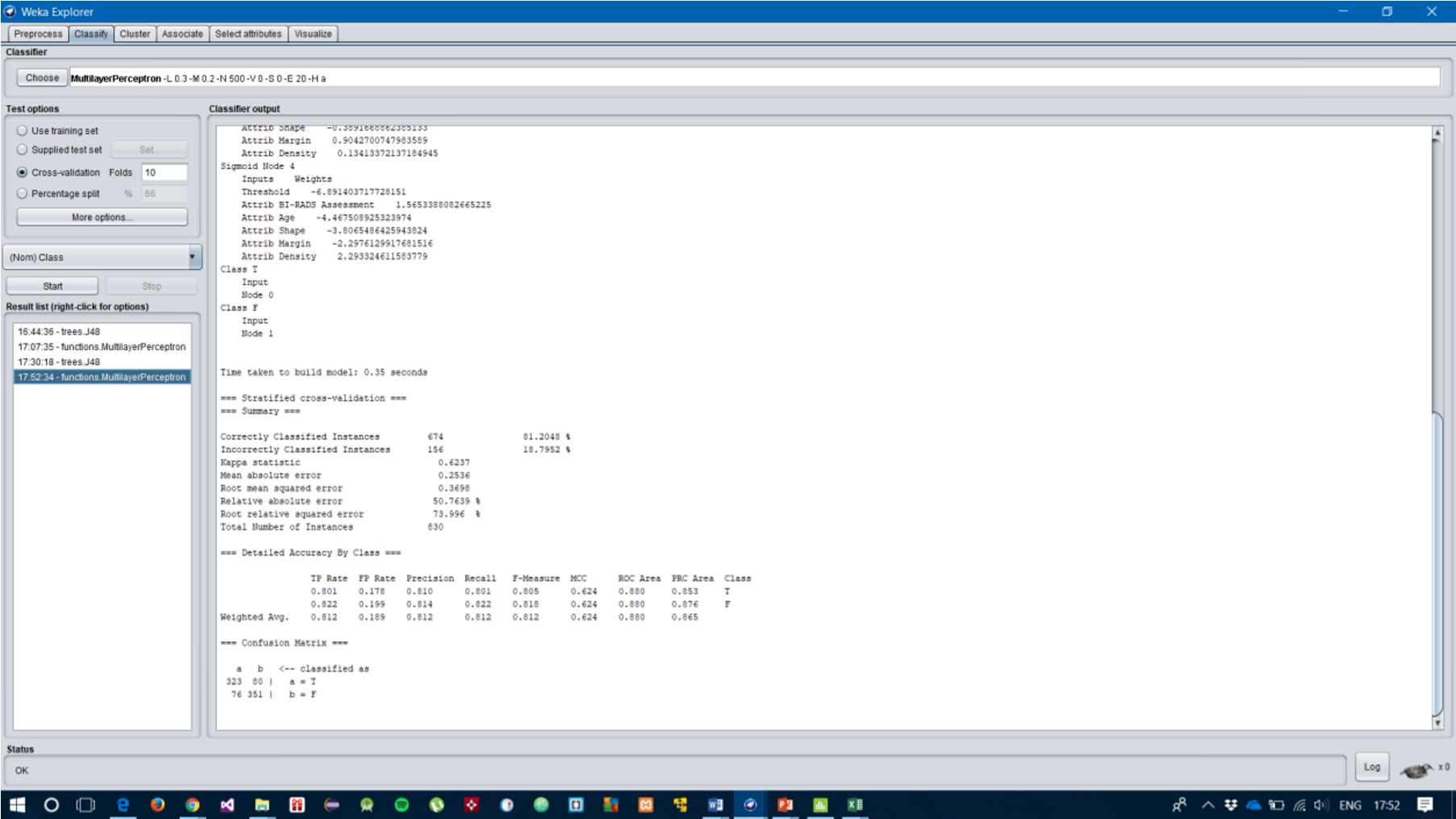


Using the WEKA explorer, the data set was classified with a Multi-Layer Perceptron classifier with all default values to classify the data and the output is seen in the screen shot where the number of correctly classified instances read 82.206% and the number of incorrectly classified instances reads 17.794%.

**Removing missing values from the dataset.**



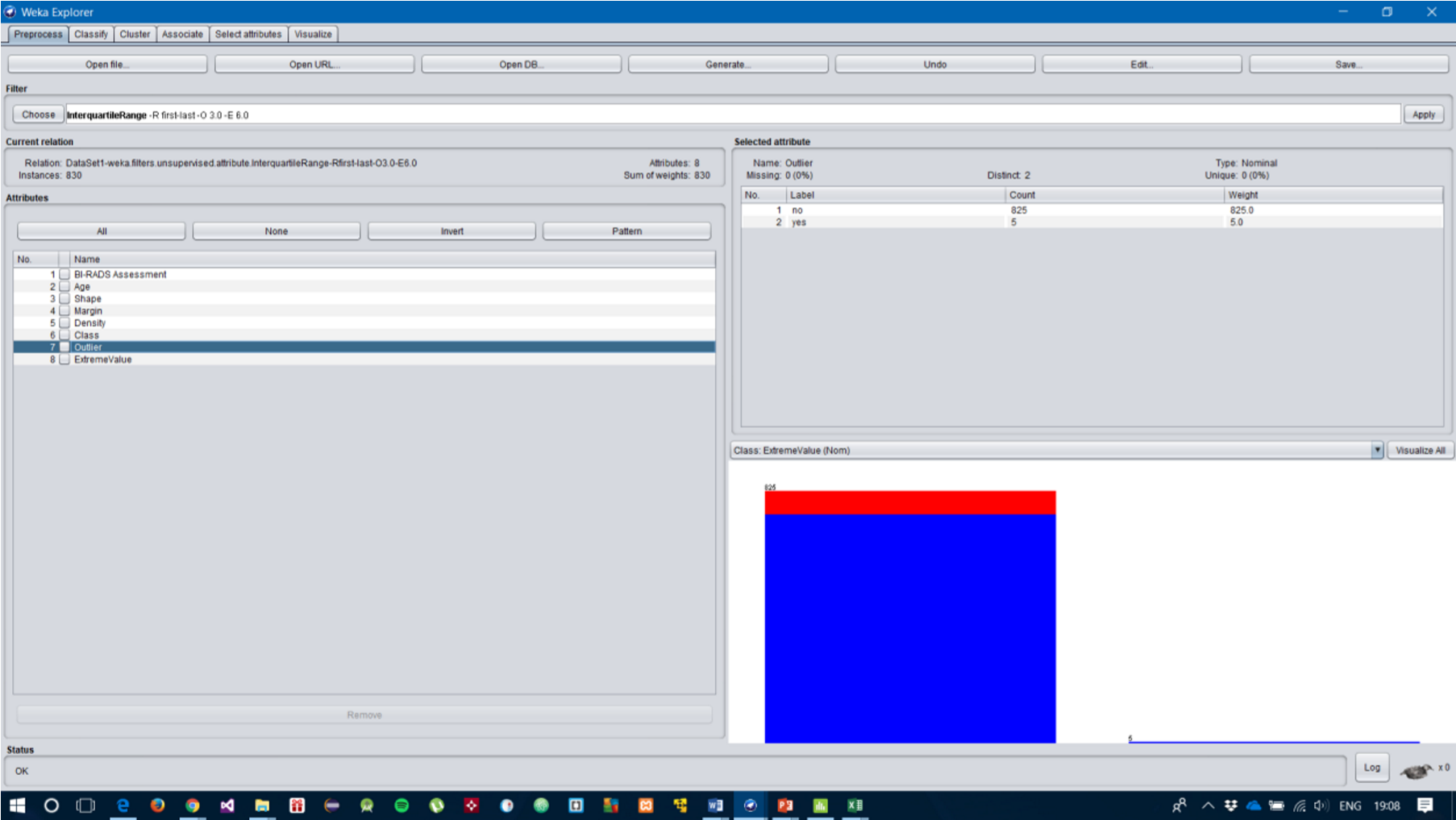
Upon classifying the Mammography dataset with using the filters and RemoveWithValues, all missing values from the dataset have been removed and re-classified with the J48 Decision Tree. The outcome was the Correctly Classified Instances have increased to 83.9759% and the Incorrectly Classified Instances have decreased to 16.0241%. The total number of instances have significantly decreased to 830 instances from 961 instances.



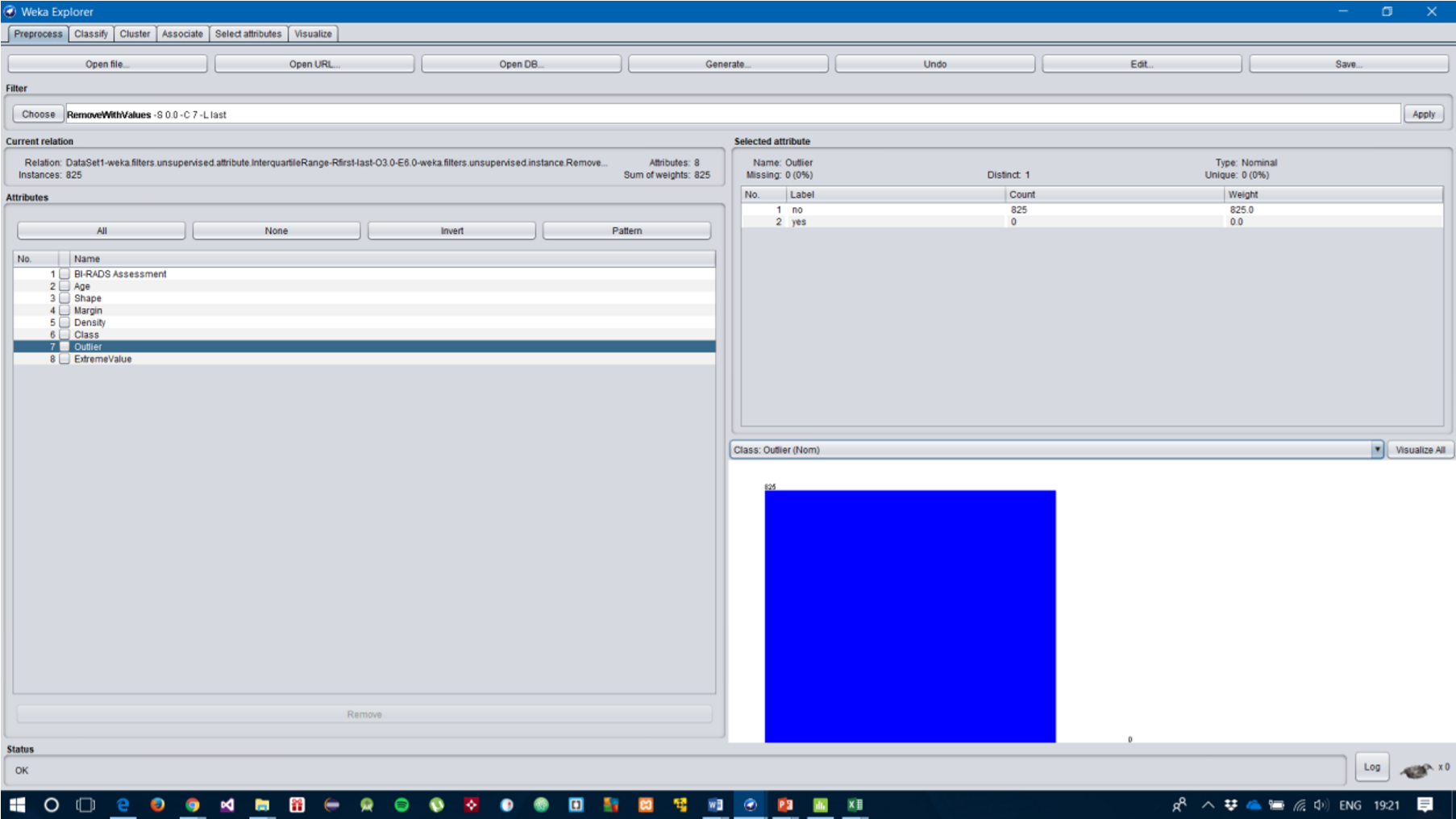
Upon classifying the Mammography dataset with using the filters and RemoveWithValues, all missing values from the dataset have been removed and re-classified with the Multi-Layer

Perceptron. The outcome was the Correctly Classified Instances have increased to 81.2048% and the Incorrectly Classified Instances have decreased to 18.7952%. The total number of instances have significantly decreased to 830 instances from 961 instances.

**Identify Outliers from data set.**



This part of scenario 2 consists of identifying outliers and removing them. First the data set is used with filters through using the inter-quartile range to identify the outliers within the data set. After using the filters, two new attributes were added to the data named: Outlier and Extreme Values. Upon identifying these values, the next step is to remove them.



By using the RemoveWithValues, we are able to remove the outliers as seen in the screenshot. The label of yes has no count of any values present as outliers.

6. Main experimental series

• The major plan for the J48 Decision experiments is to initially set a base line using the zeroR and OneR classifier algorithms and using the J48 Decision Tree to classify the first instance of the data set. The ZeroR and OneR classifier algorithms are used to look at basic properties of the dataset. The next step will be to start classifying the dataset with different parameters in search of a decision tree with the highest classification accuracy.

i. Execution of DT experiments to find DT with highest Classification Accuracy.

Initial test that was conducted with J48 classifier, ZeroR and OneR classifier. J48 DT performed well on the data set with a Classification Accuracy of 82.96%.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | ZeroR | OneR | J48 |
| Classification Accuracy | 50.33 | 82.91 | 82.96 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |  |
| Confidence Factor | 0.25 | 0.2 | 0.15 | 0.1 | 0.05 | 0.01 |
| Minimum  Number of  Objects | 2 | 2 | 2 | 2 | 2 | 2 |
| Classification Accuracy | 82.96 | 83.39 | 83.51 | 83.55 | 83.70 | 83.31 |

The first test considers altering the confidence factor parameter for J48 by using values of 0.25, 0.2, 0.15, 0.1, 0.05 and 0.01. The test results are seen in the screenshot and table above, with their respective classification accuracies. The tree with confidence factor of 0.05 reached a classification accuracy of 83.70%, stepping down from the initial confidence factor value of 0.25 which has a classification accuracy of 82.96%. The next step is to use the CA confidence factor of 0.05 and begin processing to try and get a higher classification out of it.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | | |  |  |
| Confidence Factor | 0.05 | 0.1 | 0.001 |  | 0.002 | 0.003 | 0.004 | 0.005 |
| Minimum  Number of  Objects | 2 | 2 | 2 |  | 2 | 2 | 2 | 2 |
| Classification Accuracy | 83.70 | 83.55 | 82.78 |  | 82.68 | 82.68 | 82.64 | 82.72 |

The second test involved decreasing the confidence factor pruning parameter for the DT, which decreases the classification accuracy as well. It is understood that, larger intervals of confidence factors need to be tested in order to achieve a higher classification.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | |  |  |  |
| Confidence Factor | 0.05 | 0.02 | 0.03 | 0.04 | 0.06 | 0.07 | 0.08 | 0.09 |
| Classification Accuracy | 83.70 | 83.59 | 83.70 | 83.70 | 83.70 | 83.67 | 83.62 | 83.58 |

Upon experimentation using higher pruning parameters along with the optimum saw a stable accuracy with values closest to 0.05, but the confidence factors above 0.06 saw a decrease in accuracy.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | | | |  |  |  |
| Confidence Factor | 0.05 | 0.051 | 0.052 | 0.053 | 0.054 | 0.055 | 0.049 | 0.048 | 0.047 | 0.046 |
| Classification Accuracy | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 |

The confidence intervals which are closer to the optimum was used in testing to witness any slight change in the CA. There was no significant improvement.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  |  | J48 Decision Tree | | |  |  |  |
| Confidence Factor | 0.05 | 0.03 | 0.035 | 0.036 | 0.037 | 0.038 | 0.039 | 0.0411 | 0.0412 | 0.0413 |
| Classification Accuracy | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 | 83.70 |

This last attempt at experimenting with confidence intervals within the range of 0.03 and 0.05 witnesses stagnancy within the classification accuracy of the decision tree. This suggests that the DT with the largest confidence factor is the best.

|  |  |  |
| --- | --- | --- |
| Classifier | J48 Decision Tree | |
| Confidence Factor | 0.05 | 0.25 |
| Classification Accuracy | 83.70 | 82.94 |

This experiment tested the dataset with the default value for confidence factor of 0.25 with the optimum value for confidence factor of 0.05. It shows that the DT with a lower confidence factor achieved a higher classification accuracy of 83.70%.

|  |  |  |
| --- | --- | --- |
| Classifier | J48 Decision Tree | OneR |
| Confidence Factor | 0.05 | - |
| Classification Accuracy | 83.70 | 82.91 |

The last experiment was a test to see the difference in the CA between the J48 pruned tree with a confidence factor of 0.05 and the OneR rule. The J48 decision tree performed better than the oneR rule.

**Experiments with MNO Pruning.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |
| Minimum  Number of  Objects | 2 | 4 | 6 | 8 | 10 |
| Classification Accuracy | 82.96 | 83.03 | 83.06 | 83.26 | 83.30 |

The first experiment involved using the MNO pruning parameters which achieved an increase in CA due to the increase of minimum number of objects. This suggests that the higher the MNO parameters, the best the classifier performance.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | |  |  |  |
| Minimum number of objects | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| Classification Accuracy | 83.30 | 83.50 | 82.54 | 83.94 | 83.31 | 82.54 | 82.47 | 82.46 |

The second experimentation was carried out further by increasing the MNO pruning values to find a DT with higher classification accuracy. In doing so, tree number 4, with 40 MNOs, gained the highest classification of 83.94%. Increasing the MNO, significantly decreased the classification accuracy of the DT.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |
| Minimum  Number of  Objects | 40 | 42 | 44 | 46 | 48 |
| Classification Accuracy | 83.94 | 84.02 | 83.74 | 83.16 | 83.20 |

The third experiment involved pruning the decision tree with MNO values closer to the optimum tree that achieved a classification accuracy of 83.94%. The decision tree with a MNO parameter of 42 gained a better classification accuracy of 84.02% making it the best DT so far.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |
| Minimum  Number of  Objects | 42 | 41 |  | 43 | 45 |
| Classification Accuracy | 84.02 | 83.99 |  | 83.99 | 83.39 |

The last experiment using MNO pruning saw altered values closer to the optimum decrease significantly. The best pruned tree with highest CA was 84.02%.

ii. Execution of DT experiments to find the most highly pruned DT that does not have a significantly lower Classification Accuracy than the tree with best CA.

**Using Confidence Interval Pruning.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | | |  |  |
| Confidence Factor | 0.05 | 0.005 | 5.0E-4 |  | 4.0E-4 | 3.0E-4 | 2.0E-4 | 1.0E-4 |
| Classification Accuracy | 83.70 | 82.72 | 82.86 |  | 82.87 | 82.87 | 82.87 | 82.87 |

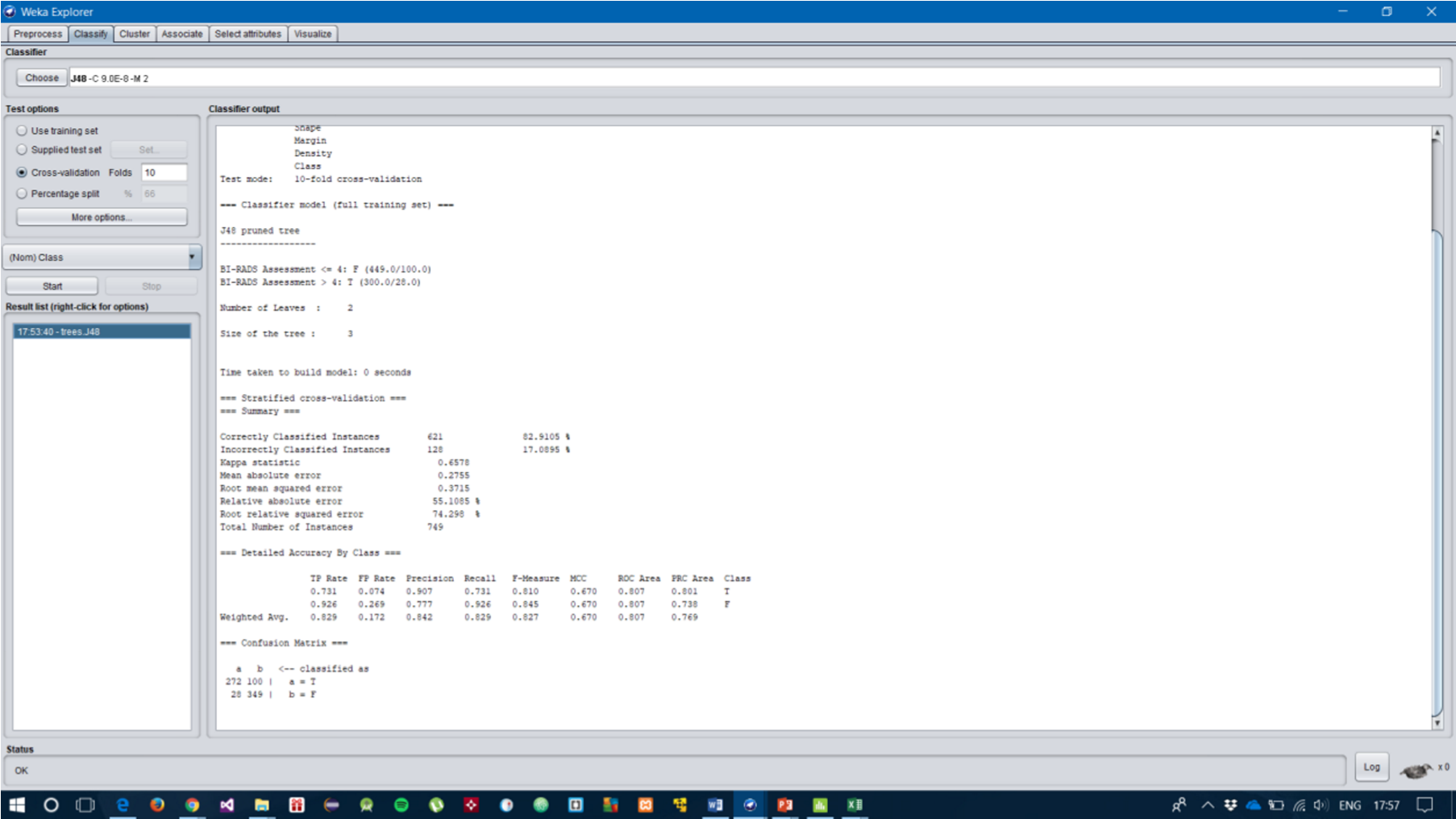
The first experiment while using the optimum confidence factor with lower pruning values saw a decrease in CA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J4 | 8 Decision Tree | |  |  |
| Confidence Factor | 0.05 | 0.1 | 5.0E-4 | 1.0E-4 | 2.0E-4 | 3.0E-4 | 4.0E-4 |
| Classification Accuracy | 83.70 | 83.55 | 82.86 | 82.87 | 82.87 | 82.87 | 82.87 |

Using even lower values decreased the performance of the DT and in turn, the optimum value of 0.05 will be used along with pruning parameters closest to it.

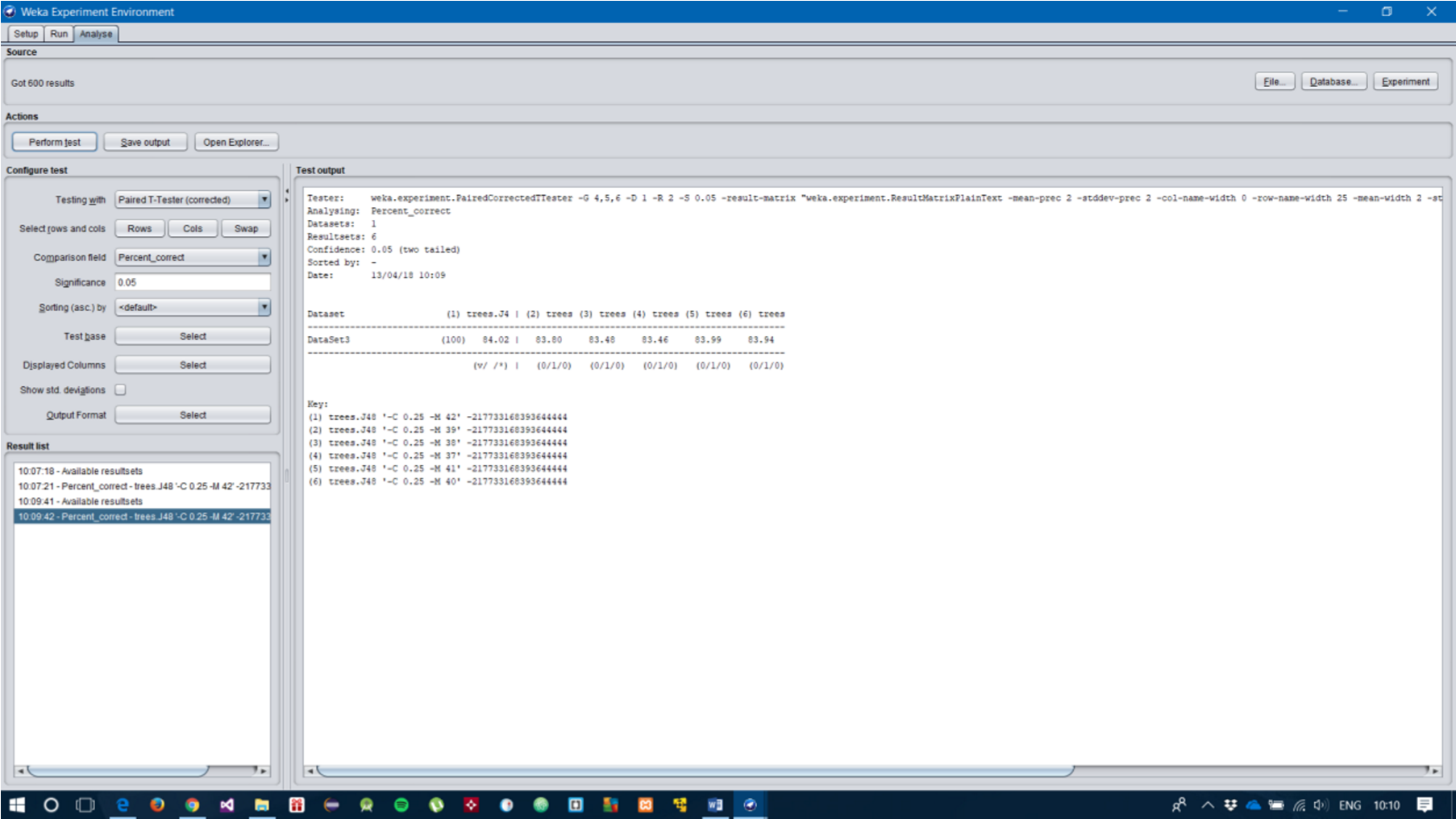
|  |  |  |  |
| --- | --- | --- | --- |
| Classifier |  | J48 Decision Tree |  |
| Confidence Factor | 0.05 | 5.0E-8 | 8.0E-8 |
| Classification Accuracy | 83.70 | 82.87 | 82.87 |

The pruning has been pushed even lower, where the DT with the confidence interval of 8.0E-8 achieved an 82.87 CA.



This last test concluded that the most highly pruned tree that does not have a significantly lower classification accuracy achieved a correctly classified instance of 82.9105%.

**MNO pruning to find DT that is not significantly lower in classification than DT with best CA.**



Experimenting with the minimum number of objects parameter, the value of 41 MNO was applied to the classification achieved 83.99%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |  |
| Minimum  Number of  Objects | 42 | 41 | 35 | 30 | 25 | 20 |
| Classification Accuracy | 84.02 | 83.99 | 83.03 | 82.54 | 83.06 | 83.50 |

Using less number of minimum objects saw the CA decrease and therefore will be pruned even more to find the most highly pruned DT.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | J48 Decision Tree | |  |  |  |
| Minimum  Number of  Objects | 42 | 41 | 20 | 10 | 5 | 38 | 37 | 36 |
| Classification Accuracy | 84.02 | 83.99 | 83.50 | 83.30 | 83.10 | 83.48 | 83.46 | 83.40 |

The DT which was used in conjunction with a rather small MNO parameter, achieved a CA of

83.10%, making it the most highly pruned tree so far. Further experiments will be conducted to find the most highly pruned DT.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | J48 Decision Tree | |  |  |
| Minimum  Number of  Objects | 42 | 41 | 20 | 21 | 1 | 2 |
| Classification Accuracy | 84.02 | 83.99 | 83.50 | 83.42 | 82.98 | 82.96 |

The DT which was pruned the most using an MNO parameter of 1, saw a drop in its CA of 82.98%. This concludes that the most highly pruned tree which does not have a significant lower classification than the best CA, was tree number 5.

**Plan for artificial neural network experiments, with justification**



The initial plan for the classification of the dataset with use of artificial neural networks was first testing how well the Multi-Layer Perceptron accuracy in terms of classification when compared to a j48, OneR and ZeroR rule. Evidently, the MLP achieved a high CA, which therefore drives a plan where comparing the use of single layer neurons and multiple layer of neurons that have different effects on the classifier.

i. Execution of artificial neural network experiments to find artificial neural network with highest classification accuracy using a single layer of neurons.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-layer Perceptron | | |  |
| Hidden layers/neurons | 2 | 4 |  | 8 | 12 | 18 |
| Learning Rate | 0.3 | 0.3 |  | 0.3 | 0.3 | 0.3 |
| Momentum | 0.2 | 0.2 |  | 0.2 | 0.2 | 0.2 |
| Classification Accuracy | 83.36 | 83.03 |  | 82.95 | 82.79 | 82.74 |

This first experiment uses a single layer of neurons within a single hidden layer with incrementing values to find the most appropriate number of neurons in a single layer with default values for learning rate and momentum which decreased the performance of the MLP classifier.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-layer Perceptron | | |  |  |
| Hidden layers/neurons | 3 | 6 |  | 9 | 12 | 15 | 19 |
| Learning Rate | 0.3 | 0.3 |  | 0.3 | 0.3 | 0.3 | 0.3 |
| Momentum | 0.2 | 0.2 |  | 0.2 | 0.2 | 0.2 | 0.2 |
| Classification Accuracy | 83.35 | 83.16 | | 82.84 | 82.79 | 82.84 | 82.84 |

The second experiment applied another varied set of the number of neurons within a single hidden layer which also decreased the performance of the classifier.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-layer Perceptron | | |  |
| Hidden layers/neurons | 1 | 2 |  | 3 | 4 | 5 |
| Learning Rate | 0.3 | 0.3 |  | 0.3 | 0.3 | 0.3 |
| Momentum | 0.2 | 0.2 |  | 0.2 | 0.2 | 0.2 |
| Classification Accuracy | 84.03 | 83.36 |  | 83.35 | 83.03 | 82.87 |

This experiment involved keeping the number of neurons in a single layer to single digits, which increased the CA where the learning rate and momentum for a single neuron in a single hidden layer were default. The CA is 84.03%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-layer Perceptron | | |  |
| Hidden layers/neurons | 2 | 4 |  | 6 | 8 | 10 |
| Learning Rate | 0.1 | 0.1 |  | 0.1 | 0.1 | 0.1 |
| Momentum | 0.2 | 0.2 |  | 0.2 | 0.2 | 0.2 |
| Classification Accuracy | 83.76 | 83.55 |  | 82.59 | 83.51 | 83.43 |

The third experiment used a varied set of number of neurons in a single layer, with a drop in the learning rate see it achieve a higher CA of 83.76%. This suggests that altering the learning rate parameter increases the classifiers performance.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier |  | Multi-Layer Perceptron |  |
| Hidden layers/neurons | 1 | 1 | 1 |
| Learning Rate | 0.1 | 0.2 | 0.3 |
| Momentum | 0.2 | 0.1 | 0.2 |
| Classification Accuracy | 84.51 | 84.14 | 84.03 |

The learning rate and momentum have different values for each along with the default values of Learning rate: 0.3, and Momentum: 0.2. The learning rate of 0.1 and momentum of 0.2 achieved a CA of 84.51%.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier |  | Multi-Layer Perceptron | |
| Hidden layers/neurons | 1 |  | 1 |
| Learning Rate | 0.1 |  | 0.3 |
| Momentum | 0.2 |  | 0.2 |
| Classification Accuracy | 84.51 |  | 84.03 |

This last experiment concluded the comparison of default values and decreased values of both parameters of learning rate and momentum. The default values achieved an 84.03% while the decreased values achieved an 84.51% CA. This shows the optimum Neural Network for an MLP. ii. Execution of artificial neural network experiments to find artificial neural network with highest classification accuracy using a multiple layers of neurons (Individual student judgment to be used on how many layers to be used).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | Multi-Layer Perceptron | | |  |  |
| Hidden layers/neurons | 1,3 | 1,5 | 1,7 | 1,9 | 2,3 | 2,5 | 2,7 | 2,9 |
| Classification Accuracy | 83.76 | 83.91 | 84.02 | 83.92 | 82.87 | 82.94 | 83.23 | 83.15 |

The first experiment resulted in using multiple neurons within two hidden layers with incrementing values for the number of neurons. Each run achieved different CA with the highest achieving an 84.02% accuracy using two hidden layers with a single neuron in the first hidden layer and 7 neurons in the second hidden layer. This experiment also used default values of learning rate and momentum.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-Layer Perceptron | |  |  |
| Hidden layers/neurons | 1,1 | 2,1 | 3,1 | 4,1 | 5,1 | 6,1 |
| Classification Accuracy | 83.75 | 82.92 | 83.04 | 83.11 | 82.82 | 82.94 |

This experiment that was used in finding a CA using a single layer of neurons, used multiple hidden layers with multiple neurons. This achieved a CA of 83.75% while using more number of hidden layers saw a decrease in performance. This experiment also used default values of learning rate and momentum.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier |  |  |  | Multi-Layer Perceptron | | |  |  |
| Hidden layers/neurons | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 | 1,7 | 1,8 |
| Classification Accuracy | 84.03 | 84.57 | 84.51 | 84.46 | 84.34 | 84.22 | 84.23 | 83.23 |

The second experiment witnessed an increase in CA with a single neuron in the first hidden layer and 2 neurons in the second hidden layer with the learning rate and momentum at 0.1 and 0.2. This alteration saw an increase CA of 84.57%.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Multi-Layer Perceptron | | |  |  |  |  |  |  |  |
| Hidden  Layers/  Neurons | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 | 1,7 | 1,8 | 1,9 | 1,10 |
| Classification Accuracy | 84.08 | 84.53 | 84.53 | 84.49 | 84.38 | 84.27 | 84.29 | 84.24 | 84.30 | 84.30 |

This experiment used multiple neurons per hidden layer. In addition, the learning rate and momentum parameters were changed from the default values in order to test for higher classification accuracy. The parameters that achieved a higher CA of 84.08%, was the learning rate of

0.1, momentum 0.1.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier |  | Multi-Layer Perceptron |  |
| Hidden layers/neurons | 2,4,6 | 3,6,9 | 4,8,12 |
| Classification Accuracy | 82.87 | 83.24 | 83.15 |

This experiment used three hidden layers with multiple neurons for each. Upon experimentation, using more number of neurons within the hidden layers saw a decrease in CA. This experiment also used default values of learning rate and momentum.

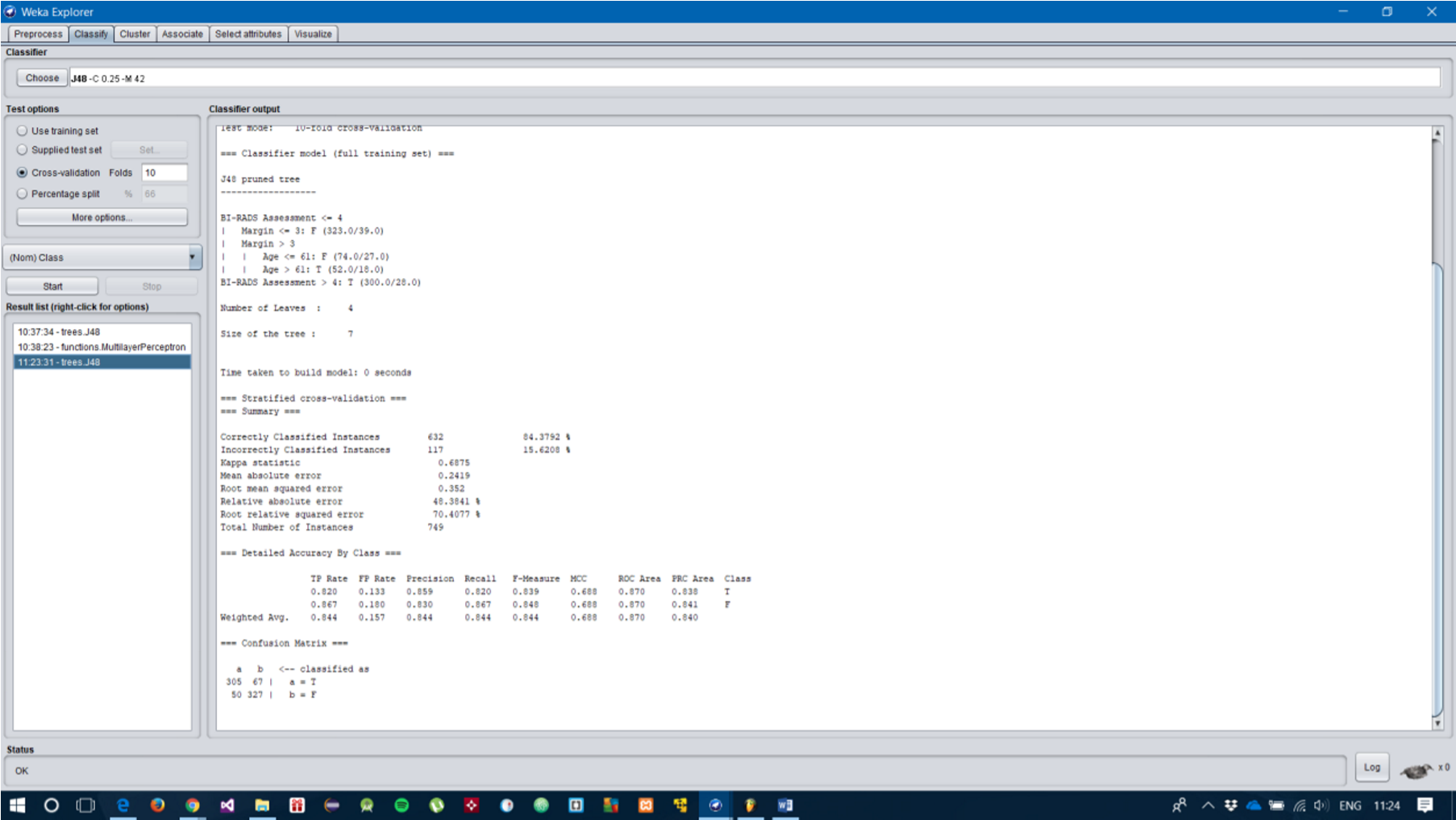
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier |  |  | Multi-layer Perceptron | |
| Hidden layers/neurons | 3,6,9 |  |  | 5,10,15 |
| Classification Accuracy | 83.24 |  |  | 83.14 |

This experiment used the default value for the learning rate and momentum with three hidden layers and multiple neurons for each. No significant improvement which suggest to keep the number of hidden layers to 2. This experiment also used default values of learning rate and momentum.

8. Conclusions

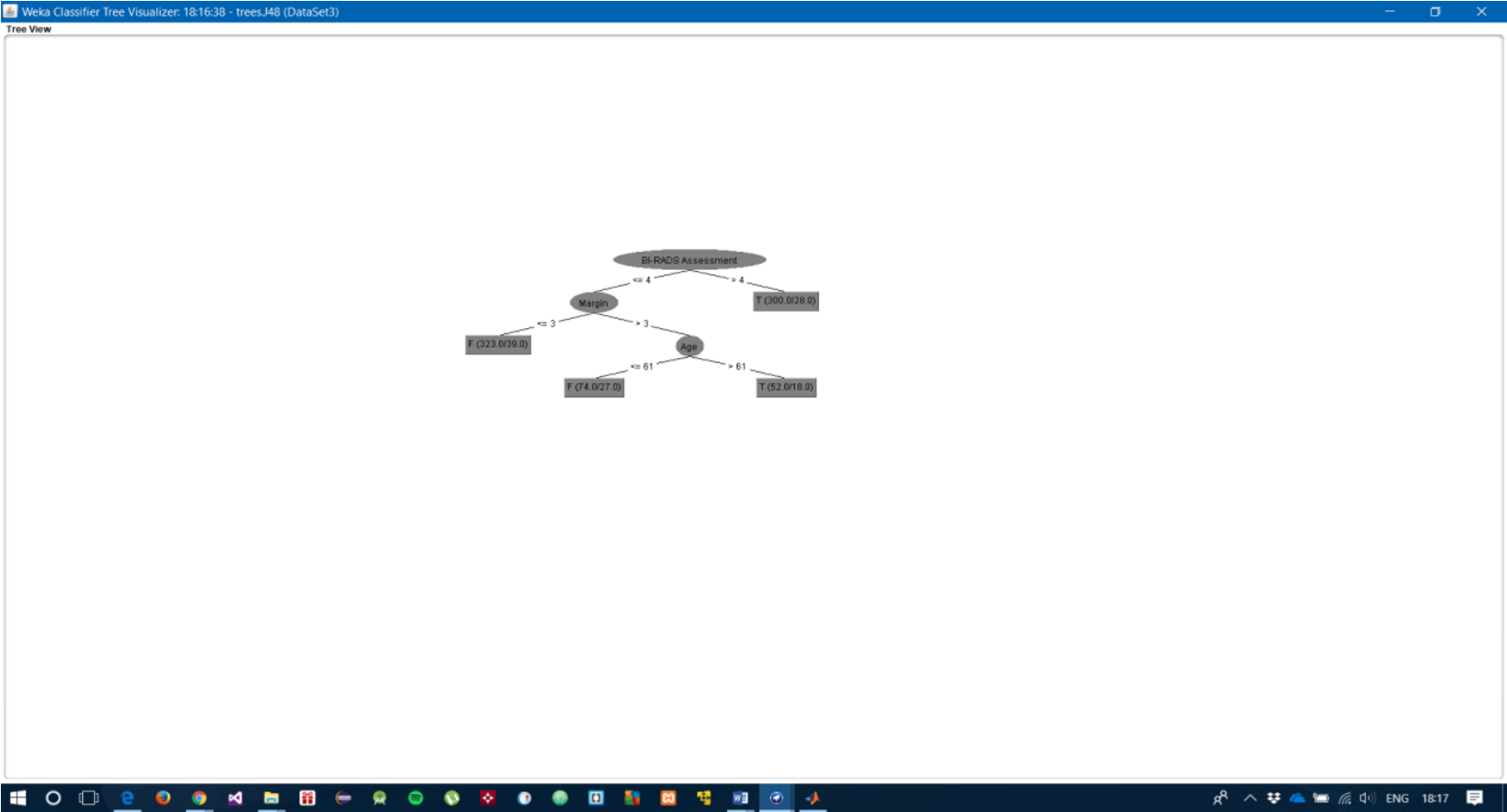
* A description of the best (pruned) decision tree for the dataset (including number of nodes, leaves, pruning parameters etc.).

The best pruned decision tree that was applied to the dataset used the MNO parameter to classify the dataset. Upon testing, the MNO parameter was set at 42 minimum number of objects that achieved a CA of 84.02%.

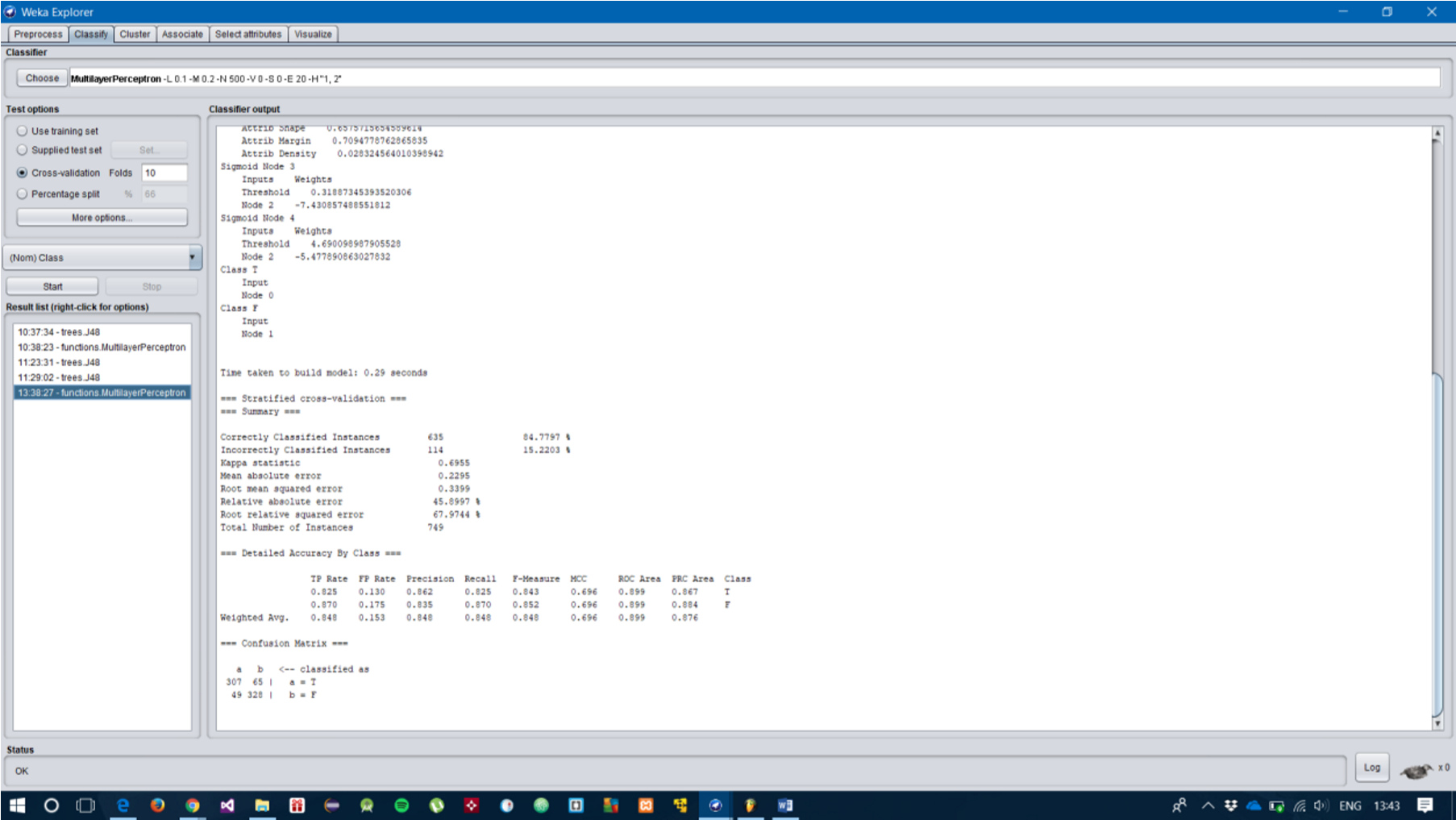


When testing the dataset in WEKA explorer with MNO pruning parameter of 42 MNOs, the correctly classified instances peaked at 84.3792%. There are 4 leaves in the decision tree, the number of nodes equal to 3 and the size of the tree is 7.

* A diagram of the best pruned decision tree for the dataset.

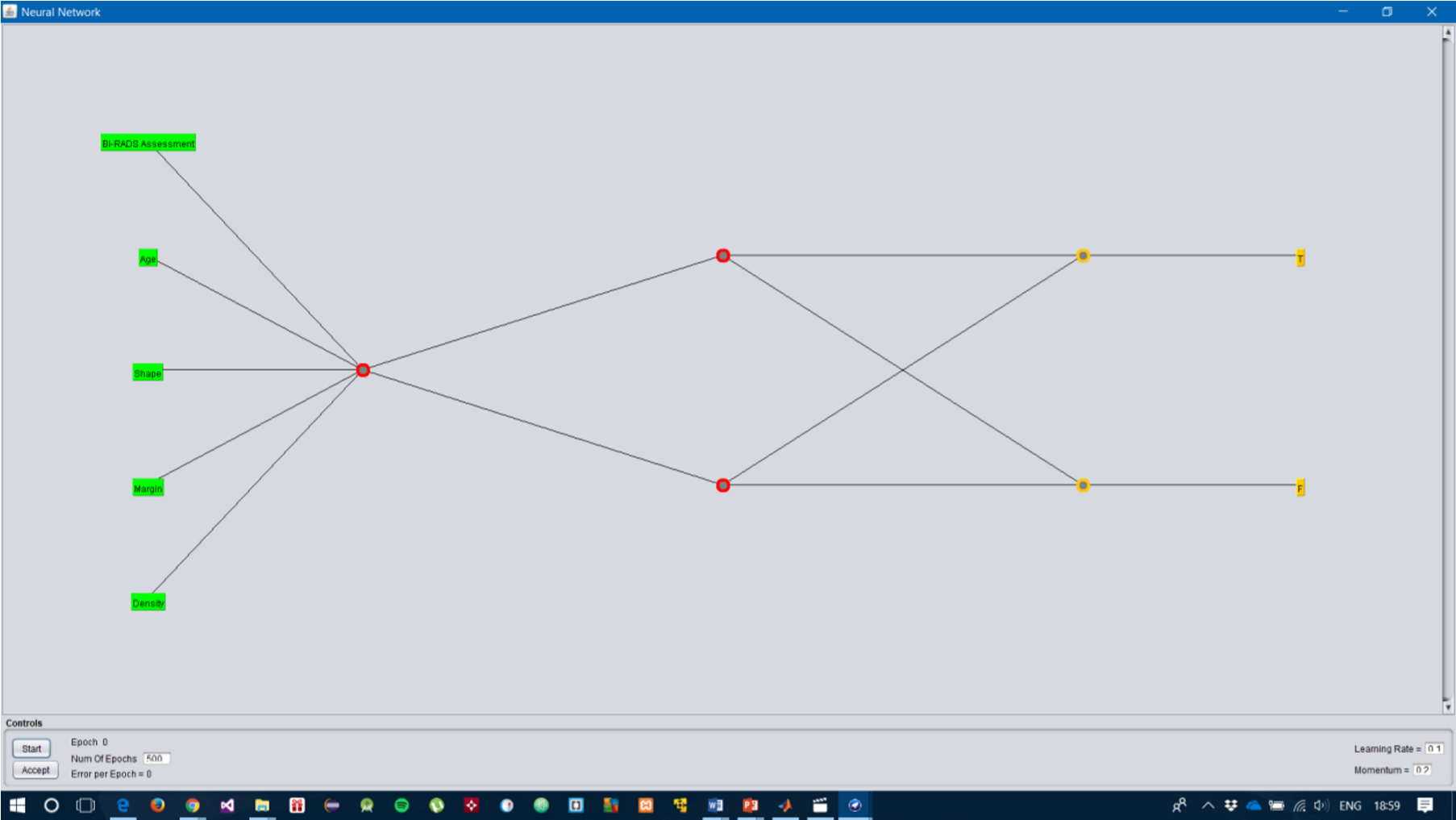


* A description of the best artificial neural network for the dataset (numbers of neurons, layers etc.)



The best ANN for training the dataset was using a single hidden layer with 2 neurons and a selected parameter value for the learning rate of 0.1 and momentum of 0.2.

* A diagram of the best artificial neural network for the dataset.



* A summary table of best classifiers

|  |  |  |
| --- | --- | --- |
| **Name of Classifier** | **J48/C4.5** | **Multi-layer Perceptron** |
| Correctly Classified Instances | 83.4792 | 84.7797 |
| Incorrectly Classified Instances | 15.6208 | 15.2203 |
| Time to build model | 0 | 0.3 seconds |
| Kappa Statistic | 0.6875 | 0.6955 |
| Mean Absolute Error | 0.2419 | 0.2295 |
| Root Mean Squared Error | 0.352 | 0.3399 |
| Relative Absolute Error | 48.3841 | 45.8997 |
| Root Relative Squared Error | 70.4077 | 67.9744 |

* A statement of which performed best and therefore which you would recommend to the client to solve the task

Based on evidence from experimenting with the Mammography Dataset, the classifier that performed the best under various parameters was the Multi-Layer Perceptron.

The Multi-Layer Perceptron is recommended for the client to solve the task.

* Short analysis or discussion of results Results of J48.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **ROC Area** |
| Malignant | 0.820 | 0.133 | 0.859 | 0.820 | 0.870 |
| Benign | 0.867 | 0.180 | 0.830 | 0.867 | 0.870 |
| Weighted Avg. | 0.844 | 0.157 | 0.844 | 0.688 | 0.870 |

Results of Multi-Layer Perceptron.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **ROC Area** |
| Malignant | 0.825 | 0.130 | 0.862 | 0.825 | 0.899 |
| Benign | 0.870 | 0.175 | 0.835 | 0.870 | 0.899 |
| Weighted Avg. | 0.848 | 0.153 | 0.848 | 0.848 | 0.899 |

Confusion Matrix for J48.

|  |  |  |
| --- | --- | --- |
|  | A True (Actual) | B False (Actual) |
| A True (Test) | 305 (TP) | 67 (FP) |
| B False (Test) | 60 (FN) | 327 (TN) |

Confusion Matrix for Multi-Layer Perceptron.

|  |  |  |
| --- | --- | --- |
|  | A True (Actual) | B False (Actual) |
| A True (Test) | 307 (TP) | 65 (FP) |
| B False (Test) | 49 (FN) | 328 (TN) |

References

Alsmadi, M. K. S., Omar, K. B. & Noah, S. A., 2009. Back Propagation Algorithm: The Best Algorithm Among the Multi-layer Perceptron Algorithm. *International Journal of Computer Science and Network Security,* 9(4), pp. 378-381.

Anantwar, S. G. & Shelke, R. R., 2012. Simplified Approach of ANN: Strengths and Weakness. *International Journal of Engineering and Innovative Technology,* 1(4), pp. 73-76.

Bhargava, N., Sharma, G., Bhargava, R. & Mathuria, M., 2013. Decsion Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering,* 3(6), pp. 1114-1119.

Goldberg, D. E., 2001. *Genetic algorithm in search, optimization, and machine learning.* s.l.:Pearson Education.

Hastie, T., Tibshirani, R. & Fridman, J., 2001. *The Elements of Statistical Learning.* 2nd ed. California: Springer.

Hornik, K., Stinchcombe, M. & White, H., 1989. Multilayer Feedforward Networks and Universal Approximators. *Nerual Networks,* Volume 2, pp. 359-366.

Istook, E. & Martinez, T., 2002. Improved Backpropagation learning in neural networks with windowed momentum. *Internation Journal of Neural Systems,* 12(3&4), pp. 303-318.

Jenhani, I., Amor, N. B. & Elouedi, Z., 2008. Decision trees as possibilistic classifiers. *Internationa Journal of Approximate Reasoning,* 48(3), pp. 784-807.

Kaur, G. & Chhabra, A., 2014. Improved J48 Classification Algorith for the Prediction of Diabetes. *International Journal of Computer Applications,* 98(22), pp. 13-17.

Kingsford, C. & Salzberg, S. L., 2008. What are decision trees. *Nature Biotechnology,* 26(9), pp. 10111013.

Mitchell, T. M., 1997. *Machine Learning.* s.l.:McGraw-Hill Science.

Moussa, M., Ariebi, S. & Nichols, K., 2006. *Arithmetic precision for implementing BP networks on FPGA.* 1 ed. Ontario : Springer.

Nadali, A., Kahhky, E. & Nosratabadi, H., 2011. *Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system.* s.l., s.n.

Panchal, G., Ganatra, A., Kosta, Y. P. & Panchal, D., 2011. Behaviour Analysis of Multilayer Perceptrons with Multiple Hidden Neurons and Hidden Layers. *International Journal of Computer Theory and Engineering,* 3(2), pp. 332-337.

Patel, N. & Upadhyay, S., 2012. Study of Various Decision Pruning Methods with their Empirical Comparision in Weka.. *International Journal of Computer Applications,* 60(12), pp. 20-25.

S.E, H. & R, O., 2007. *Improving decision tree pruning through automatic programming.* Olso, Proceedings of the Norwegian Conference on Informatics.

Tu, J. V., 1996. Advantages and Disadvantages of using Artificial Neural Networks versus Logsitic Regression for predicting Medical Outcomes. 49(11), pp. 1225-1231.

Wilson, D. R. & Tony, R. M., 2001. *The Need for Small Learning Rates on Large Problems.* s.l., s.n.

Zhang, G., Patuwo, B. E. & Hu, M. Y., 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* Volume 14, pp. 35-62.

Zhang, X., 1994. Time series analysis and prediction by neural networks. *Optimization Methods and Software,* Volume 4, pp. 151-170.

10. Appendices

• Any other supporting material (e.g. tables of extra experiments you performed which are not used directly in the main report)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Multi-Layer Perceptron | | |  |  |  |  |  |
| Hidden  Layers/  Neurons | 1,1 | 1,6 | 1,7 | 1,8 | 2,1 | 2,6 | 2,7 | 2,8 |
| Classification Accuracy | 83.75 | 83.95 | 84.02 | 83.98 | 82.92 | 83.15 | 83.23 | 83.26 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Multi-Layer Perceptron | | |  |  |  |  |  |
| Hidden  Layers/  Neurons | 2,1 | 2,2 | 2,3 | 2,4 | 2,5 | 2,6 | 2,7 | 2,8 |
| Classification Accuracy | 83.00 | 83.36 | 83.30 | 83.24 | 83.35 | 83.47 | 83.72 | 83.59 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Multi-Layer Perceptron | | |  |  |  |  |  |  |  |
| Hidden  Layers/  Neurons | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 | 1,7 | 1,8 | 1,9 | 1,10 |
| Classification Accuracy | 83.75 | 82.92 | 83.04 | 83.11 | 82.82 | 82.94 | 82.70 | 82.75 | 82.98 | 82.59 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Multi-Layer Perceptron | | |  |  |  |  |  |
| Hidden  Layers/  Neurons | 1,2 | 1,4 | 1,6 | 1,8 | 2,2 | 2,4 | 2,6 | 2,8 |
| Classification Accuracy | 83.70 | 83.86 | 83.95 | 83.98 | 82.92 | 82.98 | 83.15 | 83.26 |

This experiment involved using two hidden layers and multiple neurons per layer, which evidently

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classification | Multi-Layer Perceptron | |  |  |  |
| Hidden  Layers/  Neurons | 1,1 | 1,1 | 1,1 | 1,1 | 1,1 |
| Learning Rate | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| Momentum | 0.2 | 0.3 | 0.5 | 0.4 | 0.6 |
| Classification Accuracy | 84.03 | 83.86 | 83.40 | 83.50 | 83.47 |

This experiment uses two hidden layer with a two layer of multiple neurons where two parameters of learning rate and momentum have different values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classification |  |  |  |  |  |  |  |  |  |  |
| Hidden  Layers/  Neurons | 2,5 | 3,10 | 4,15 | 5,20 | 6,25 | 7,30 | 8,35 | 9,40 | 10,45 | 11,50 |
| Classification Accuracy | 83.34 | 83.47 | 83.31 | 83.54 | 83.83 | 83.59 | 83.54 | 83.71 | 83.67 | 83.62 |

This experiment used two hidden layers with multiple neurons in each. When testing the dataset, the time took to classify the dataset took too long and the outcome was decreased.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Multi-layer Perceptron | |  |  |
| Hidden layers/neurons | 2 | 4 | 8 | 12 |
| Learning Rate | 0.4 | 0.6 | 0.2 | 0.8 |
| Momentum | 0.3 | 0.4 | 0.3 | 0.8 |
| Classification Accuracy | 83.38 | 82.86 | 82.99 | 82.70 |

The last experiment for classification, saw a decrease in CA with altered parameters for the number of hidden layers, the learning rate and the momentum.