Task 2

SUMMARISE

Section 1. Introduction

In this paper secondary student dataset evaluated in which the researcher wants to tell how the grades of the students get varied due to the other factors like Mjob, Pjob, Absence, internet, romance, finance, free time etc. In this the highest accuracy for models changed as the different situations were created so basically comparison of different model results mathematics and the Portuguese score. The factors on which different results were there are A- the final grade(G3) of all the student s were removed B – The second period(G2) score are also removed C – the first grade(G1) score is also removed so the evaluation is only done on the basis of other factors .Further more data was seen in 3 types Discrete data, 5 level data and Linearly.

i) binary classification (pass/fail);
ii) classification with five levels (from I very good or excellent to V - insufficient); and
iii) regression, with a numeric output that ranges between zero (0%) and twenty (100%).

In the preprocessing the data was normalized and then standardized so the mean and the SD valve will become 1 and 0 the models tested were Naïve – Bayes, Neural Network, SVM, Decision Tree and Random Forest for Binary and 5 level classification PCC was seen and for linear regression RMSE was seen. The changes in the model result are on the basis of other factors involved and the background and interests.

for the binary classification

Section 2: Naïve Bayesian Classifier

Following all the Factors as A, B & C the NVB showed the best result for Mathematics. For Portuguese data the model could not show the highest accuracy as it had the least accuracy.

Section 3: Neural Network

After the neurons also this model was not a great model for maths data as it obtained an accuracy of only 88.3.In the Portuguese results for G2 & G1 the model had accuracy of 87 % approx

Section 4: Support Vector Machine

As when different factors were involved the value of SVM also decreased as it was not coming as the best model for the dataset. For Portuguese data the SVM was not considered the best model

Section 5: Decision Tree

Decision Tree is considered as a better model than SVM but then also its value was varying around 90.1(not the best model). It has the highest accuracy of 93.3 % for Portuguese dataset

Section 6: Random Forest

In this model randomly data is picked and the evaluation is done on that to remove the biasness as this was not considered as the best as with the accuracy of 91.2 so there was no such biasness in the data. In the Portuguese data it could not show any Difference from maths data.

for the 5 – level classification

Naïve – Bayes shows the highest result with 78.5% for mathematics data and Decision tree with 76.1% as the highest where the data was divided into 5 parts.

for the linear regression While comparing all model and all the factors the best results were obtained for **Random Forest** and Root mean square error was seen for regression with the accuracy of around 1.75-1.32 value. As other were showing higher values.

Section 7: Experimental Evaluation

The accuracy of the model was decreasing as other factors were added as B and C. The best results can be seen when G3 was only removed and as we can say when there were no grades in the C Factor the result were not good at all because no grade were included in that. For the binary classification the data was sectioned into two parts as >10 pass otherwise fail so it showed different results and for the 5 level the data was sampled into parts from which NVB and DT were showing a good result and at last for the regression the data we could see in the error was showing the best result with Random Forest.

Sections 8 & 9: Related Work and Conclusion

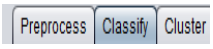
Related work applied different types of model which also have different results for different inputs which can conclude that for this dataset there can't be any particular model which we can consider as the best fit but we can say that there are so many other factors that can be considered as on the basis of factors which may lead to changes in the student results. We can also consider other subjects and the portioning we are doing can be considered with more variables and more diversity

EVALUATE

Experiment Plan

To replicate the experiments 2 datasets from the UCI repository were used, one with Mathematics data and grades and the other with Portuguese data and grade. While the paper also tested the algorithm on a large synthetic dataset, the focus in this experiment is on the results from the publicly available datasets. As the three types are not seen the experiment only binary classification and the linear regression is done.

Experiments

Following the methodology as delineated in the paper, the experiments were conducted in two phases: Binary Classification and Linear regression in which it was sectioned into 3 parts A, B & C as G3 removed, G2 also removed and finally G1was also removed. As the experiment was conducted in Weka the preprocessing was done in was as the Mjob was classified further into its instances then it was filtered with normalization and then Standardization was done to make both the dataset of mat and portu skewed as the mean and the SD was changed to 0 and 1 there was no class for this processes. G3 was made as the Class and then data was filtered in unsupervised filtering from numerical to nominal then supervised discrete filter was applied which made the data into parts to verify it classifier was added as Logistic Regression which partitioned it into the best way. Now finally G3 is removed and then we go to classify section 

where the models were applied one by one and there result was calculated .then again the process was repeated when G2 was removed then same steps for G1.

After this we undo all the steps and normalize and Standardize again and the in the Filter Linear Regression was added so the data was now classified as Linear regression and then in the classify section models were applied and evaluated to ROC value

Similar steps were done for the Portuguese data and then they are compared.

Analysis

Comparing the results of the experiments in WEKA with those of the original research paper, the following items are of note:

- In weka there is no coding and it is a tool which applies the model and the process which we tell it to do.
- In the sampling process t was not letting us give the sample it was taking a random sample which gave us a different result from the paper which we are replicating.
- When numeric or nominal data was involved some models we can't apply as we can see in the linear regression we can see NVB data and similarly in the binary Regression when no grades were involved.

- It gave very different results from the R-minor As the research paper has done the analysis in R, because of sampling and because of the preprocessing as I added some steps and excluded some.
- As seen in the table below are the results of the experiment conducted which are little bit of similar to those in the research paper but not exact.

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	75.18	74.68	78 .73	74.17	76.15	83.3	75.5	83.4	81.0	81.1
B	74.17	76.96	77.72	78.22	76.70	61.5	56.9	61.0	58.1	58.8
C	na	76.55	75.96	65.02	64.72	66.8	60.9	50.0	65.7	64.8

Table1: BINARY CLASSIFICATION

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	na	0.98	0.52	0.59	0.64	na	0.98	0.52	0.56	0.58
B	na	1.45	0.96	1.06	0.90	na	1.69	0.85	0.90	0.82
C	na	1.67	0.98	1.01	0.94	na	1.84	0.98	1.02	0.97

Table2: LINEAR REGRESSION

Discussion

There are several reasons why the results diverge from those in the original research paper, including that WEKA uses a different version of the C4.5mdecision tree algorithm and the fact that appear to be each and every step was not mentioned in the paper which can be replicated for preprocessing which is a important part to receive some results whilst small, they could certainly affect the experiment results. Further some models can't be tested for some data as Weka does not allows doing that which could also have altered some results.

On the whole the Results in the paper are different from the experiment conducted but the process was same and Max Replication is done as the Results comparing the model is may not be the same but the best model for different situations is same and can be said in the paper including other factors may alter the results but the grade will eventually matter for the result and predicting final grades.

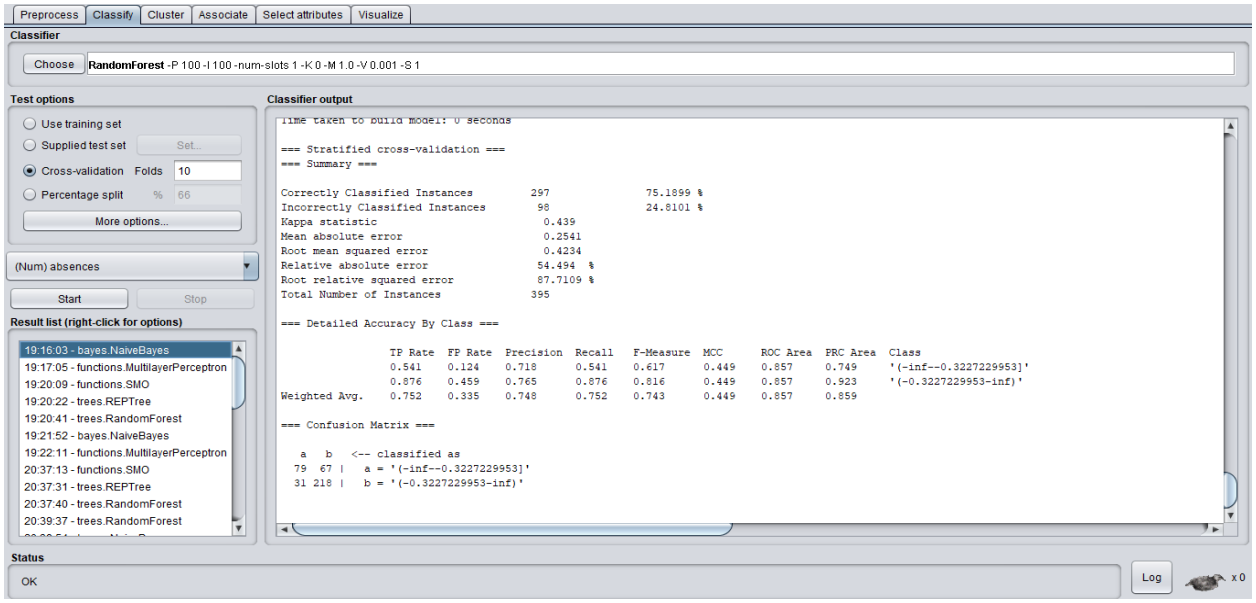
Appendix

Mathematics Results

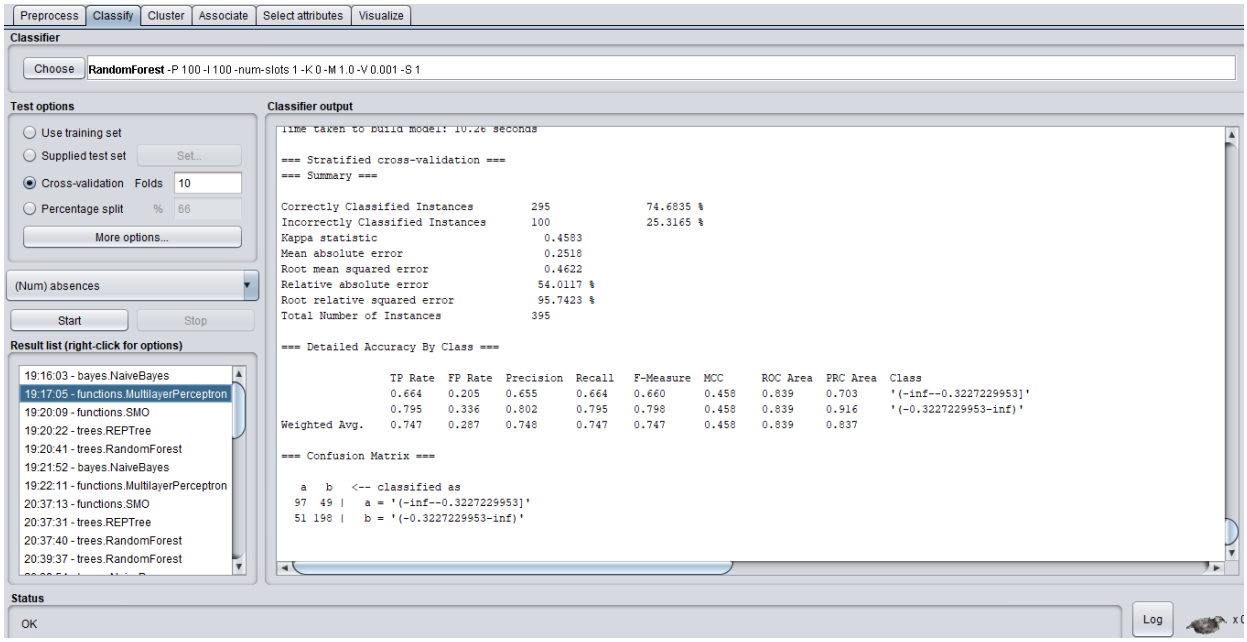
Discrete DATA (binary classification)

A- Grade was removed

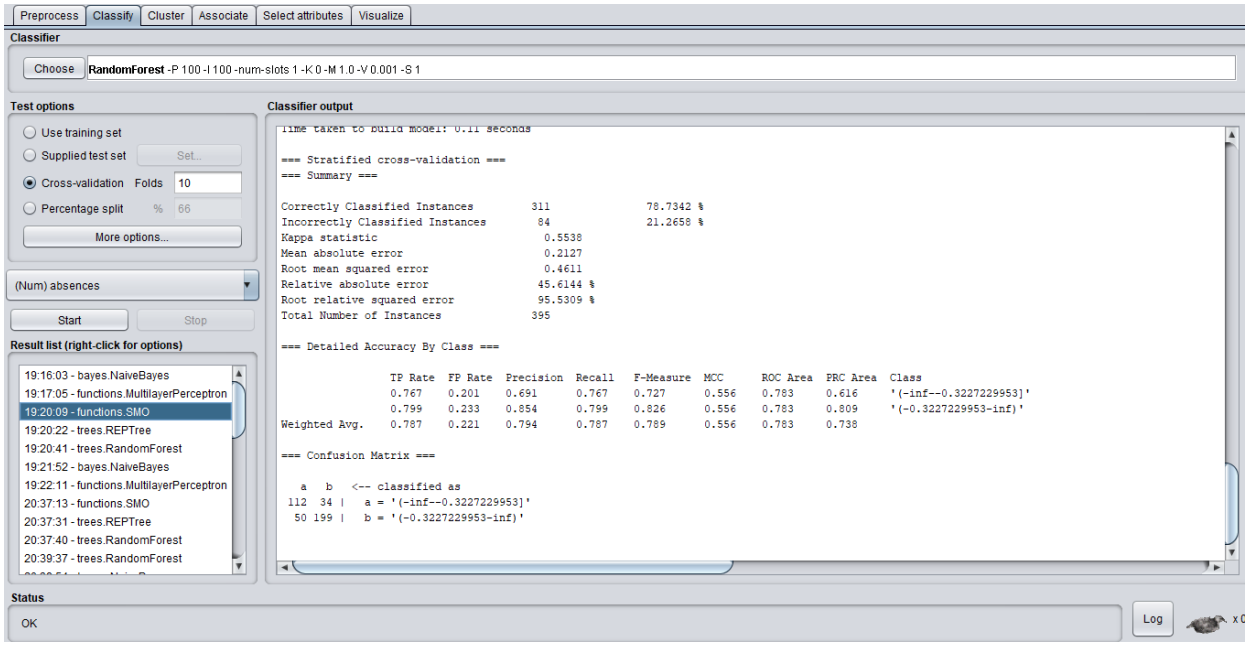
Naive-bayes



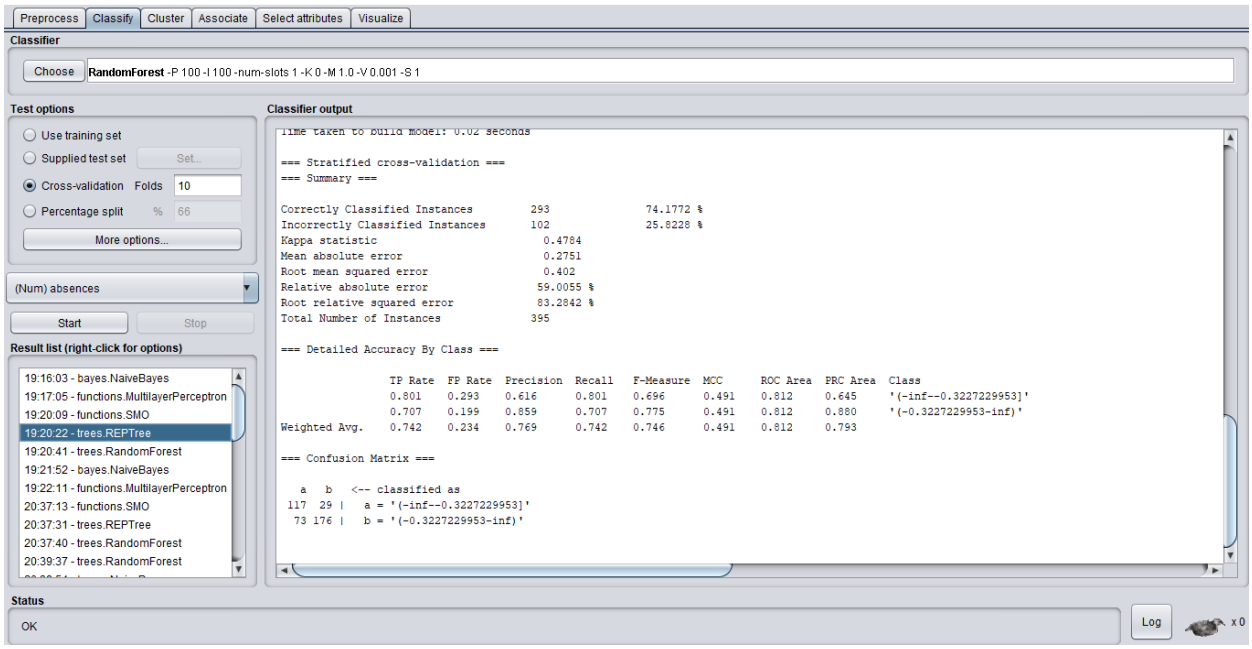
Neural Network

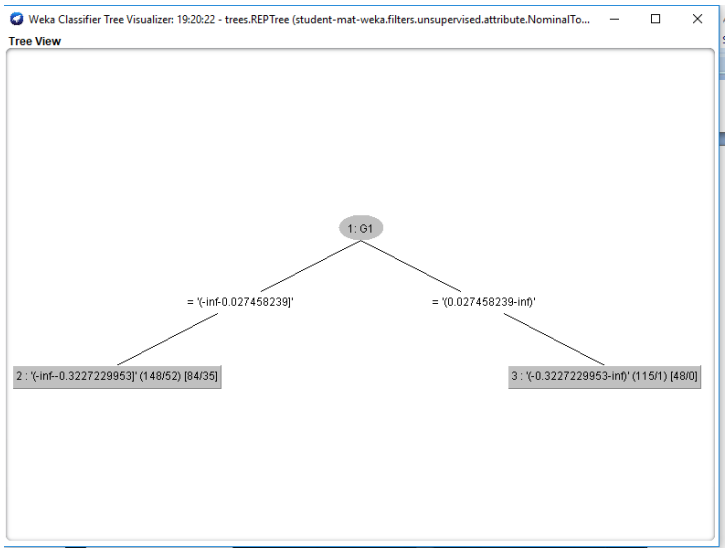


SVM

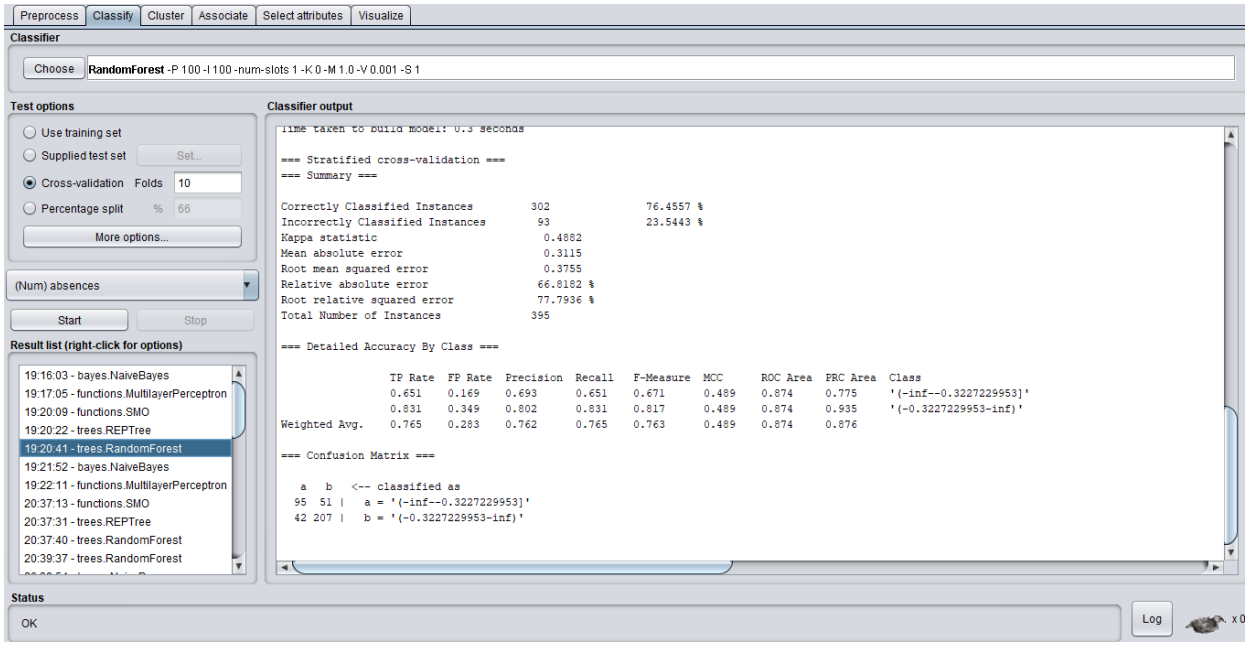


Decision Tree



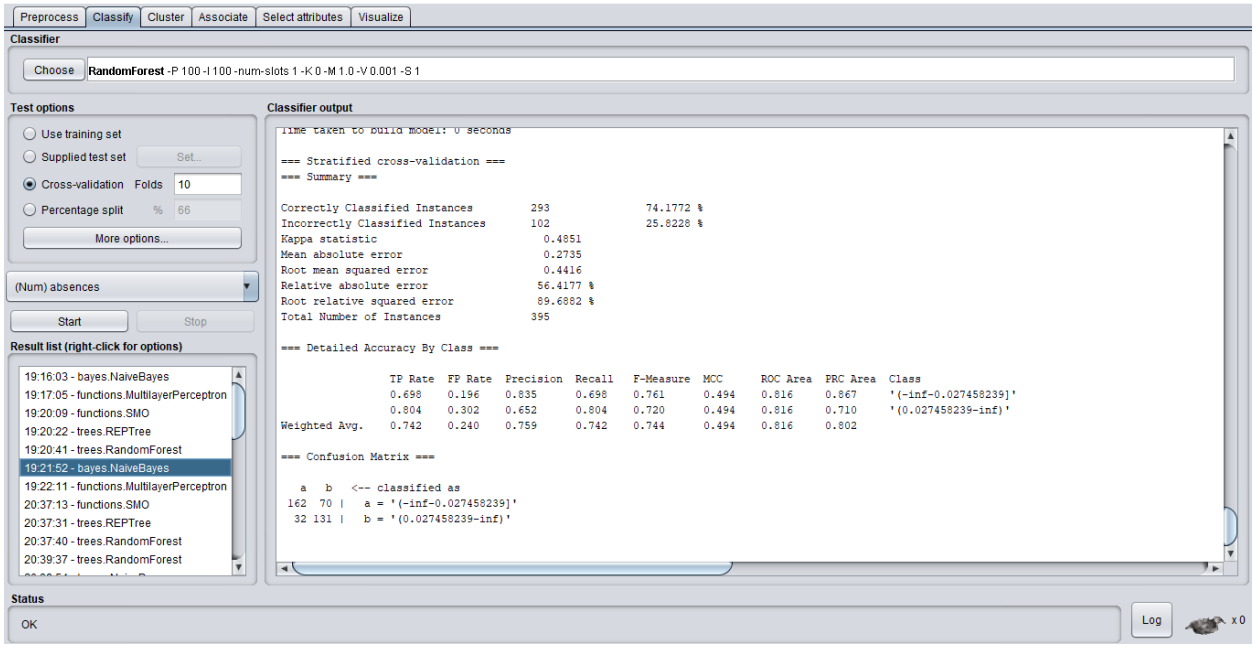


Random Forest

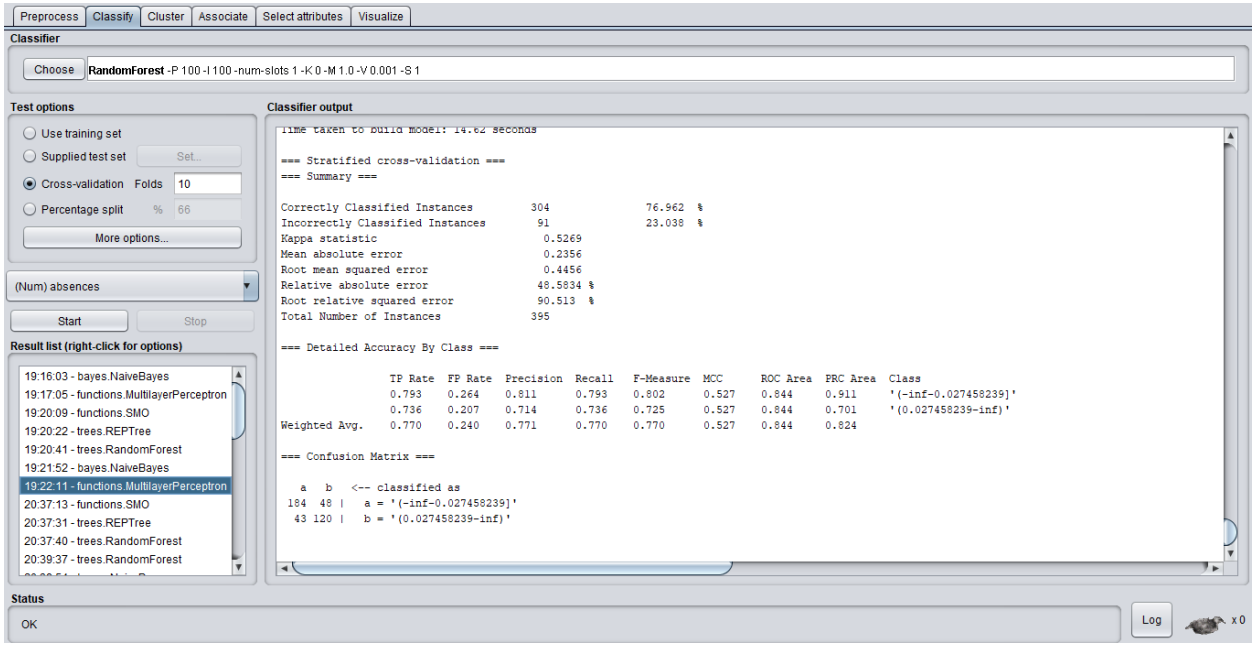


B - G2 also removed

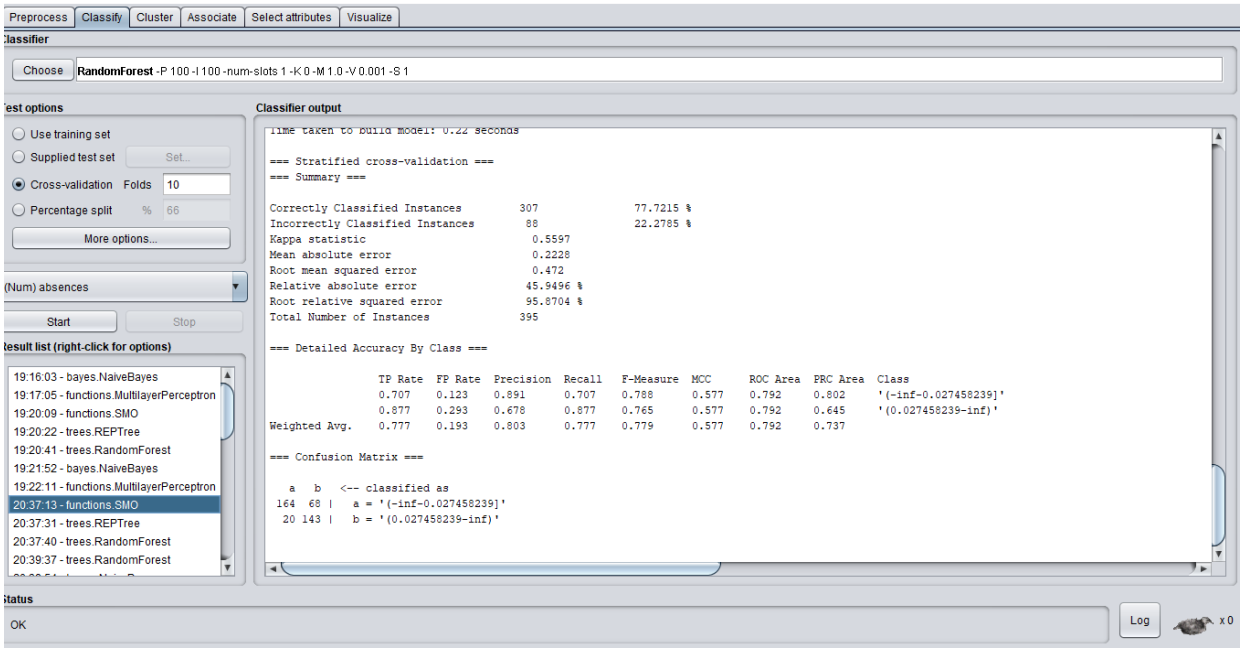
Naive Bayes



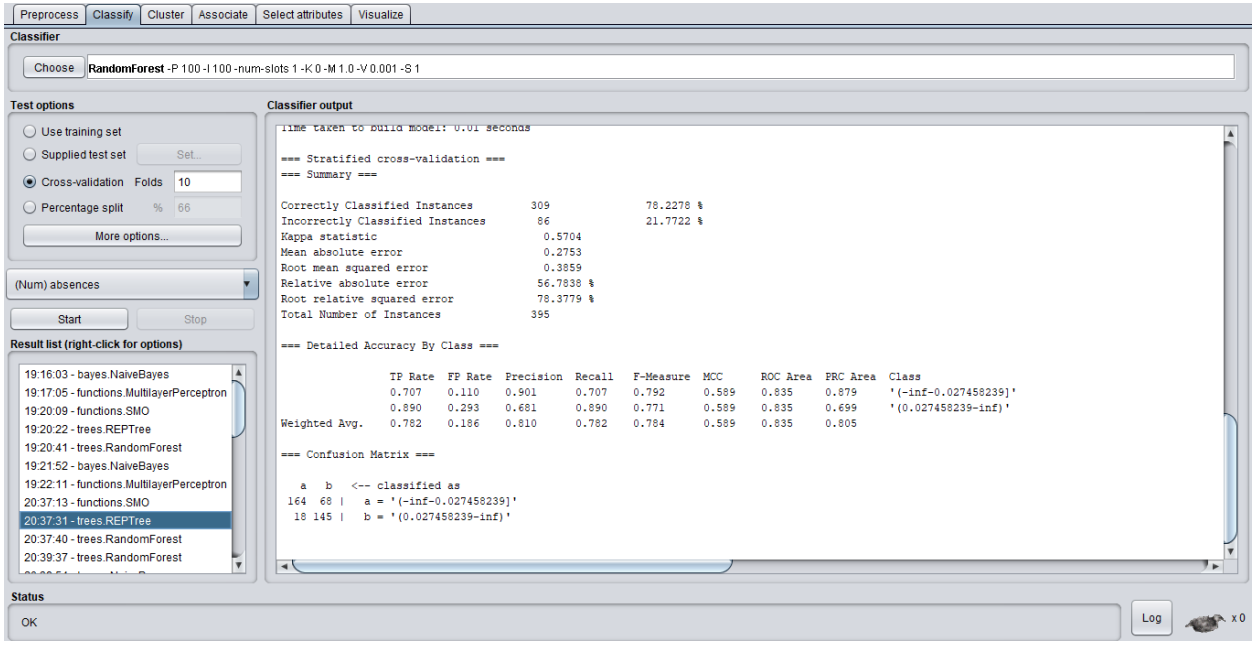
Neural network

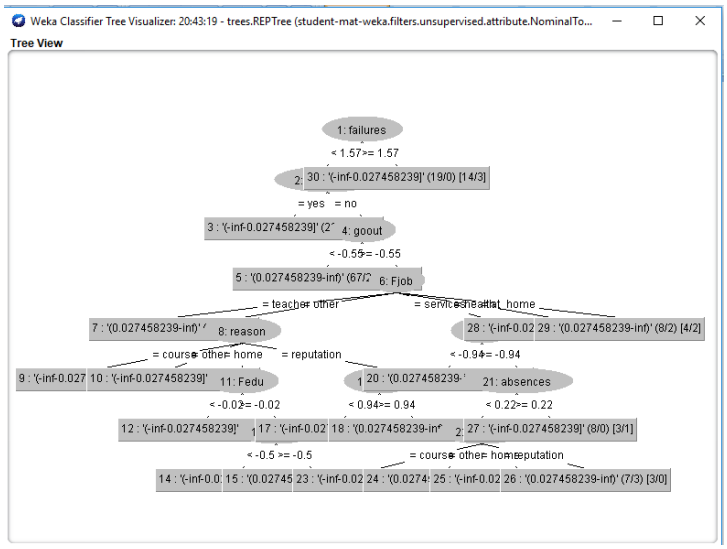


SVM

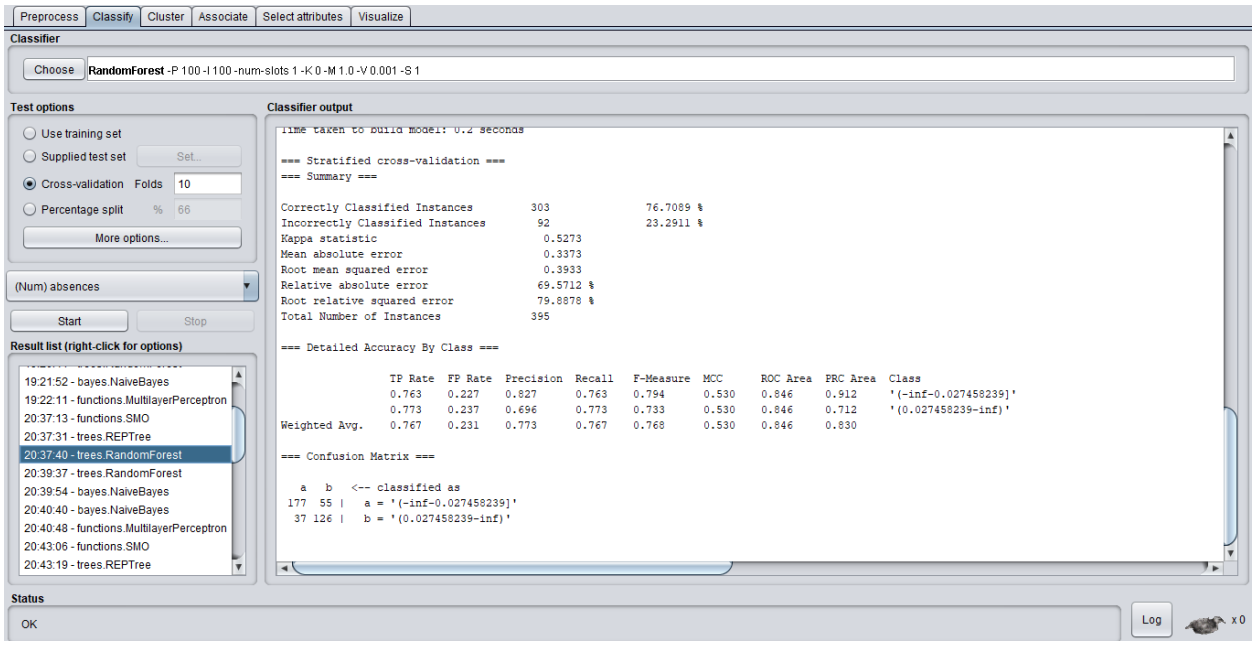


Decision tree



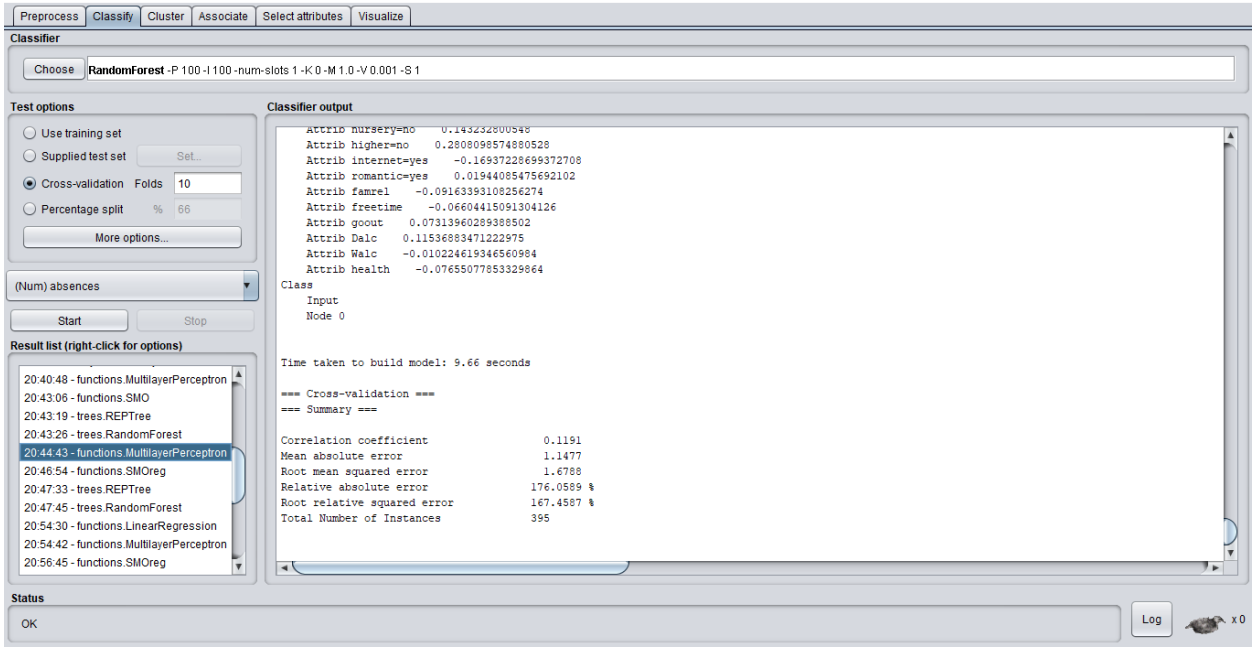


Random Forest

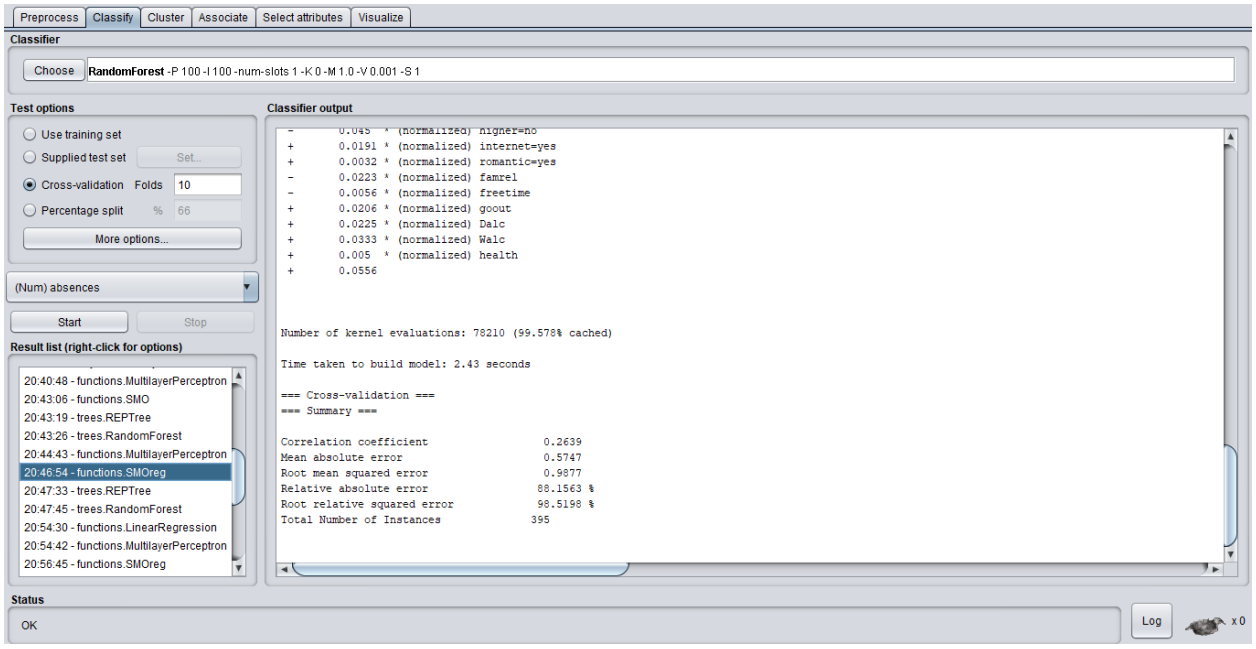


C – G1 also removed

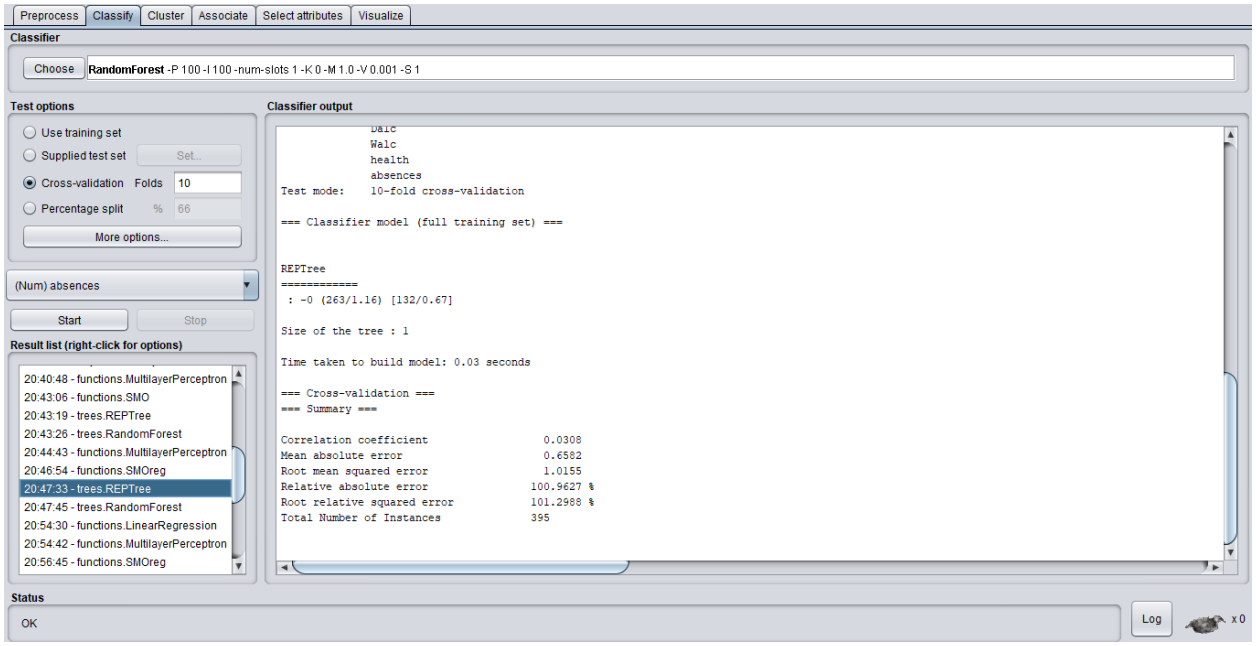
Neural network

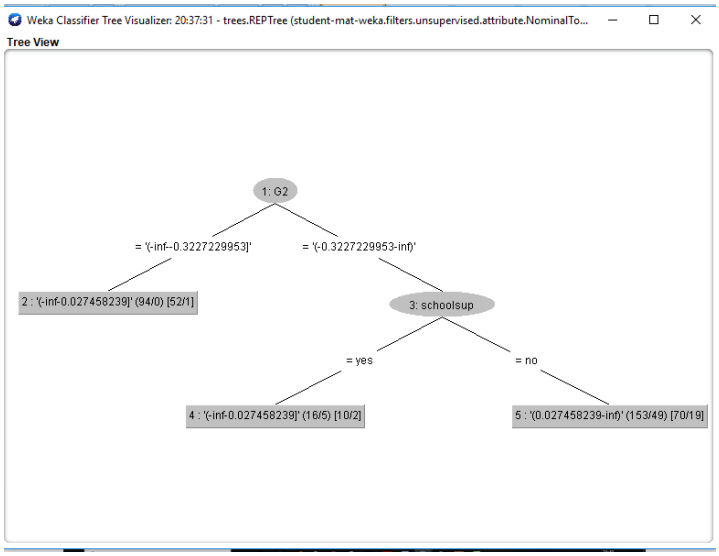


SVM

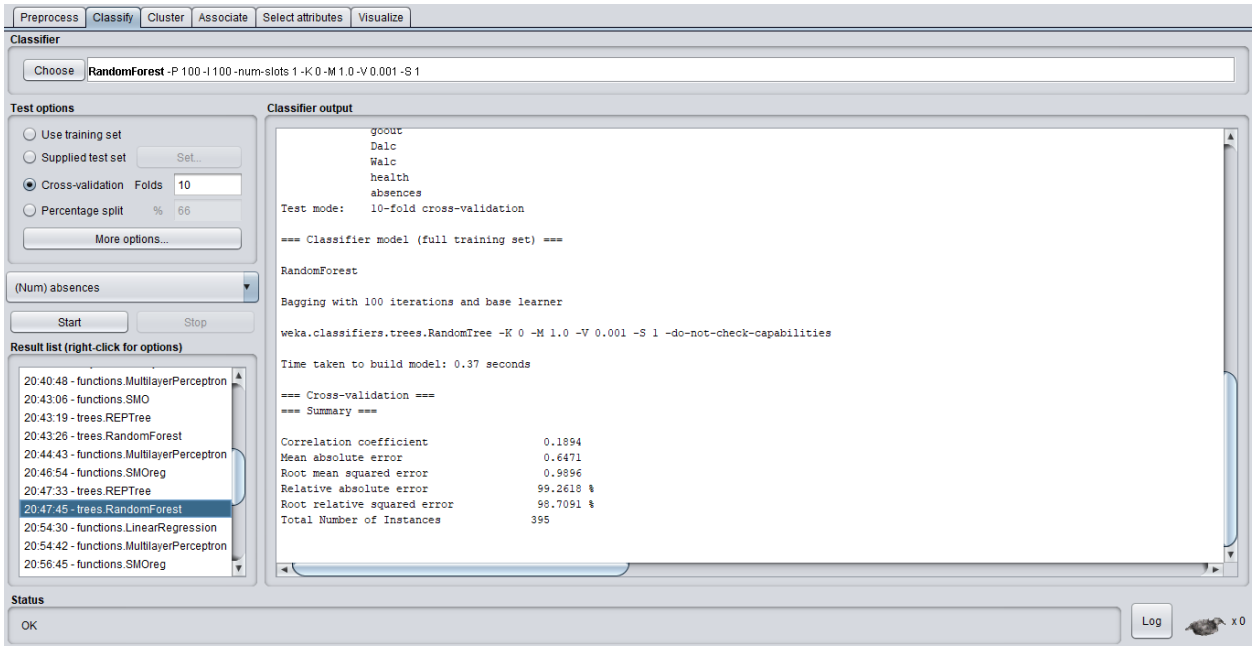


Decision tree





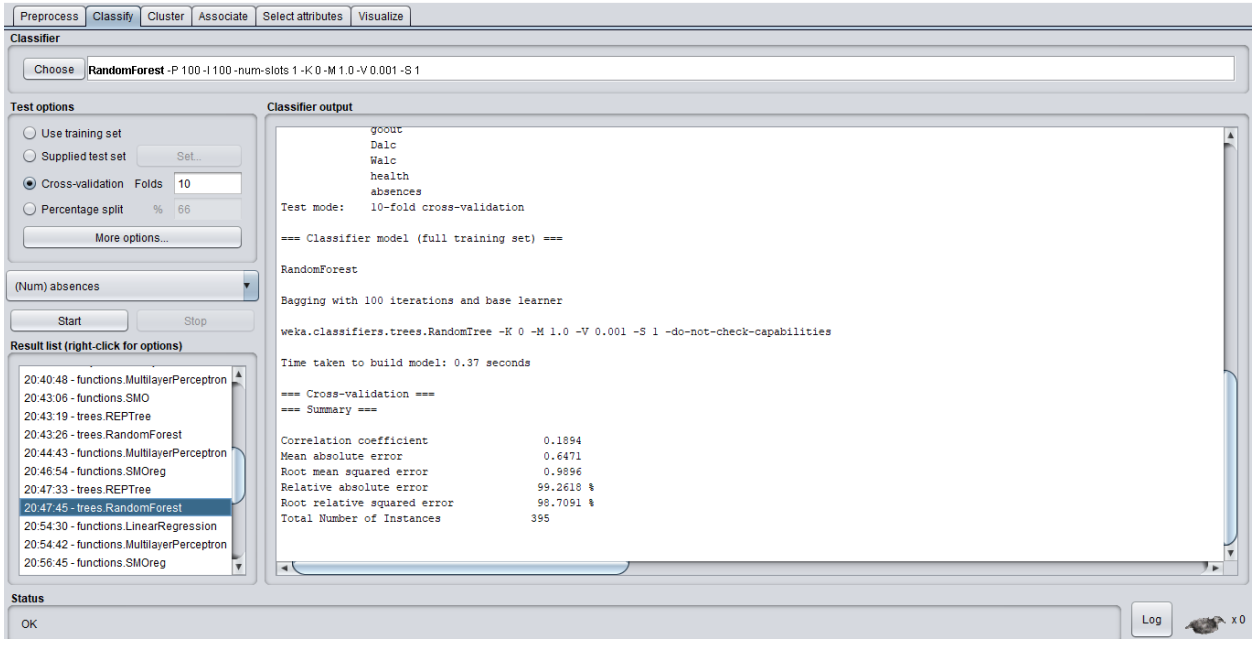
Random Forest



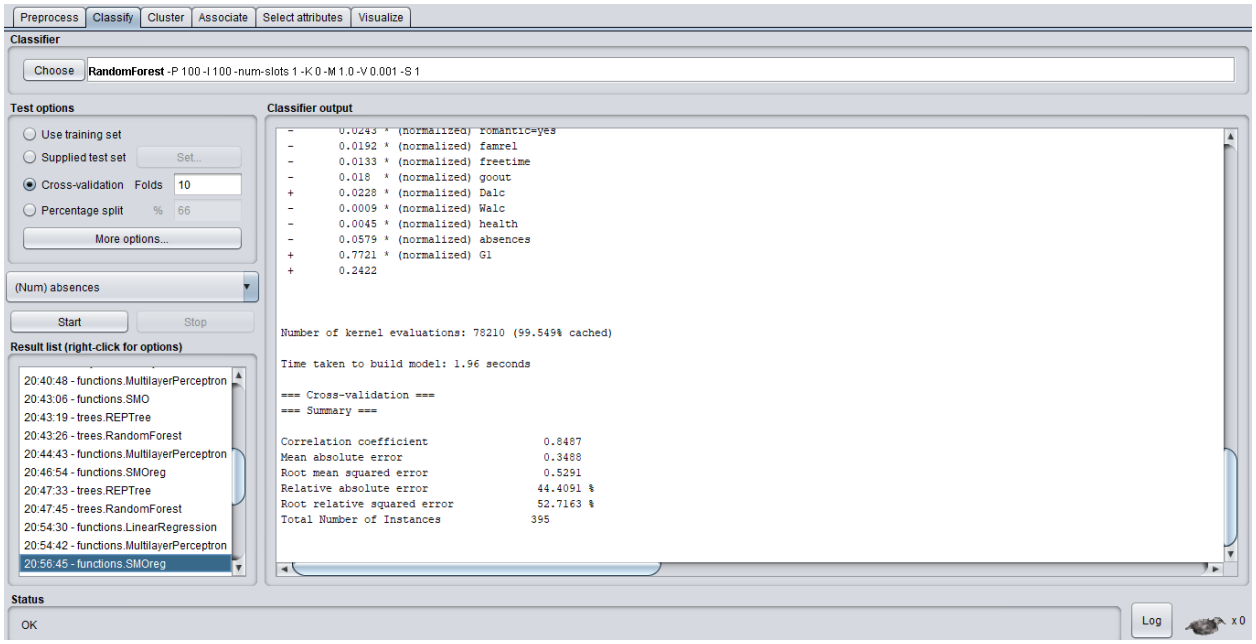
Linear regression

A- Final Grade was removed

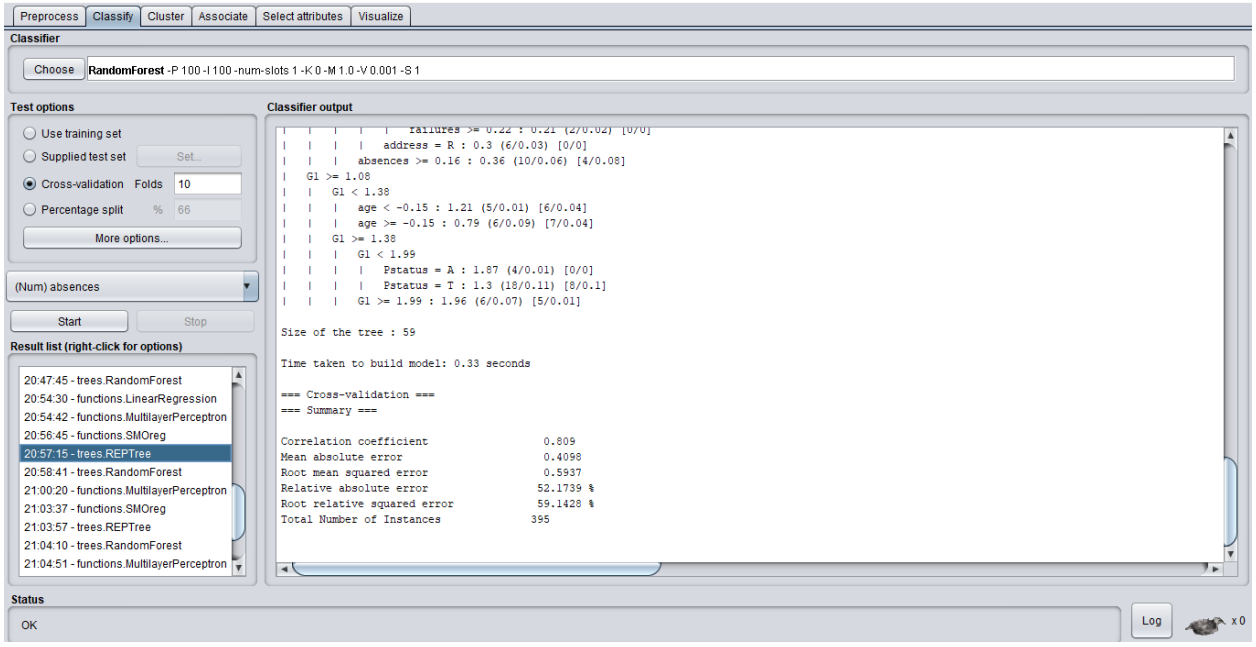
Neural network



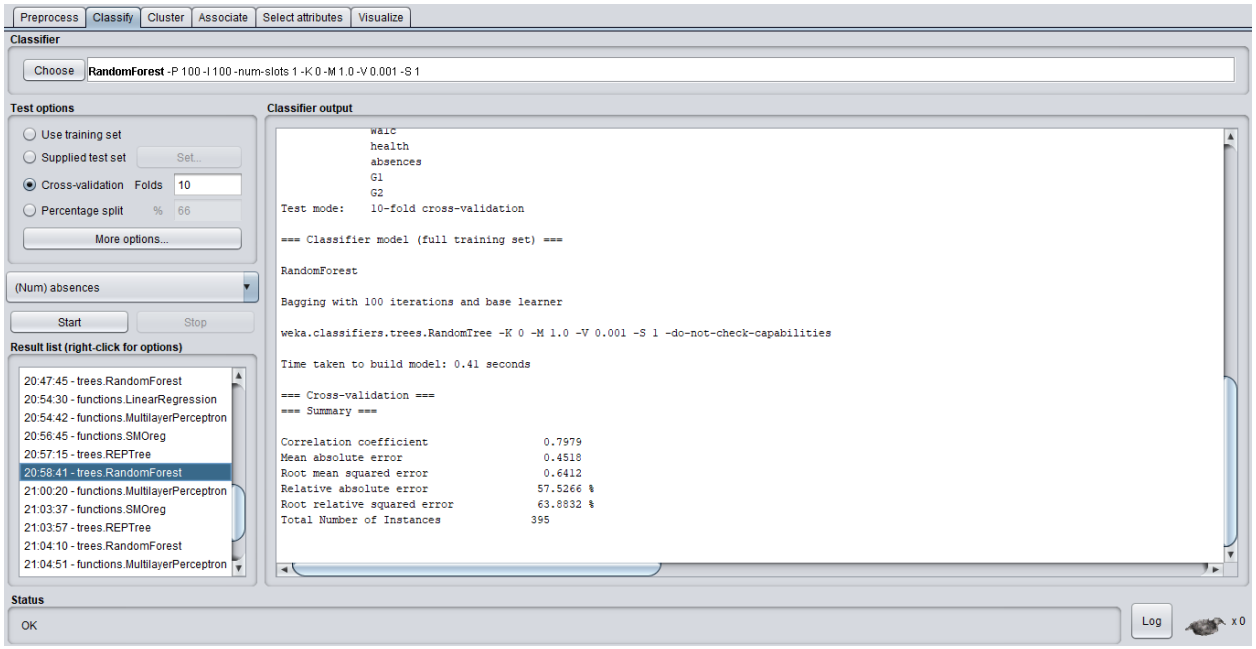
SVM



Decision Tree

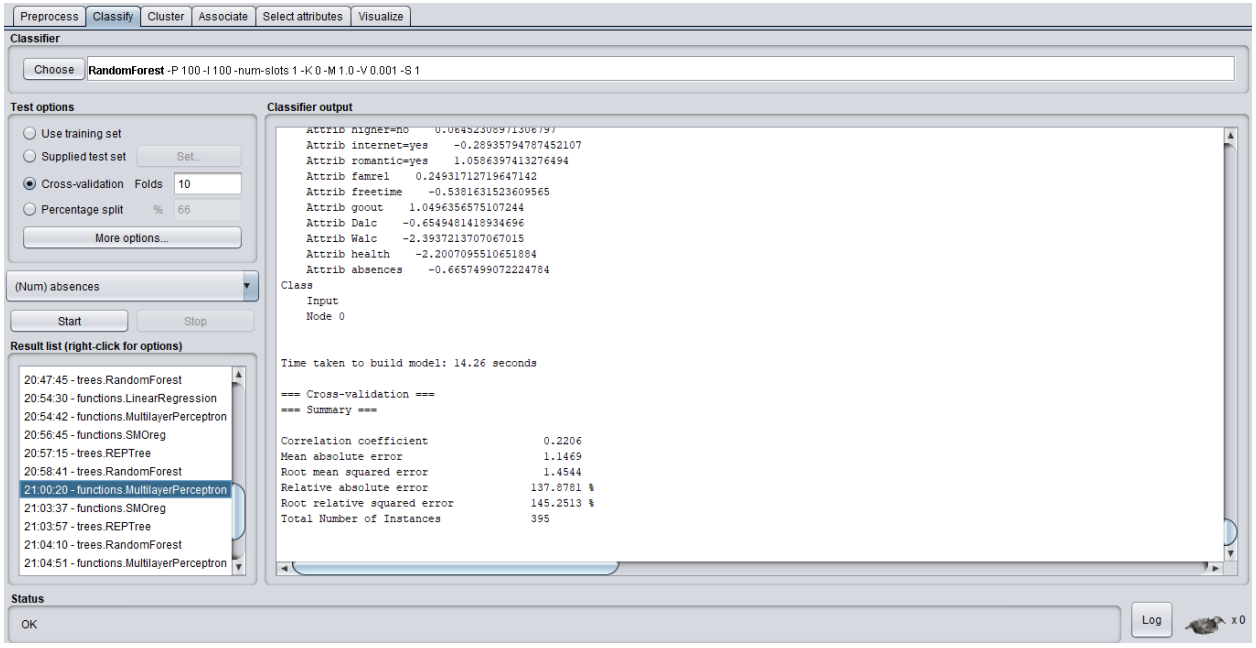


Random forest

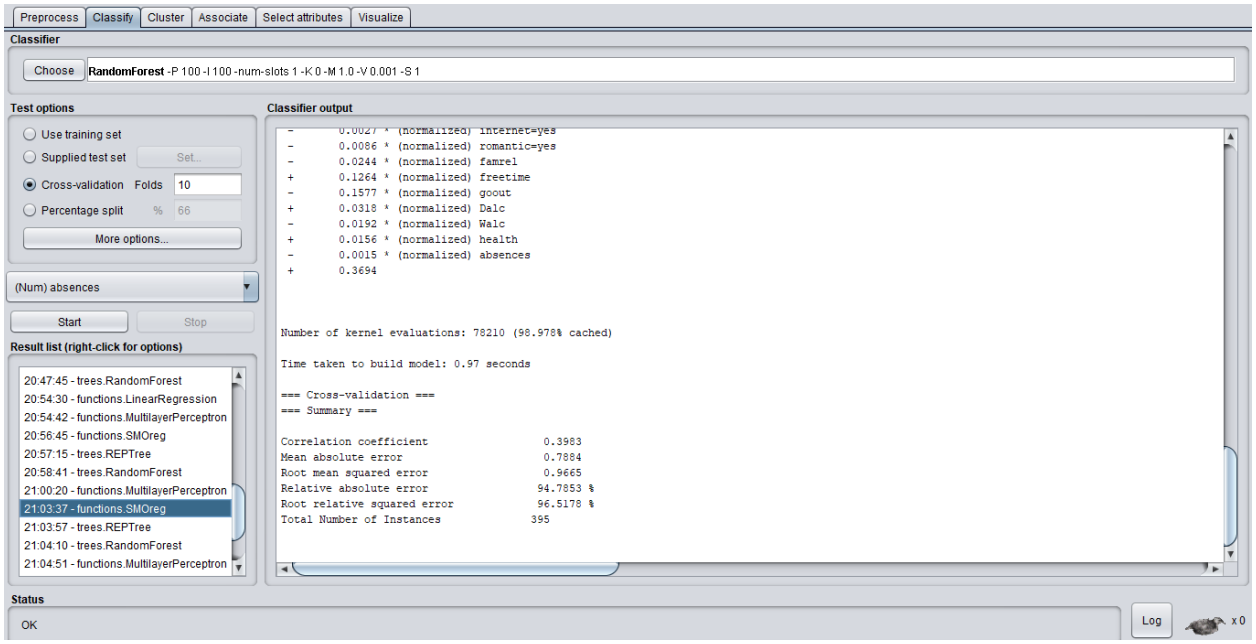


B – G2 also removed

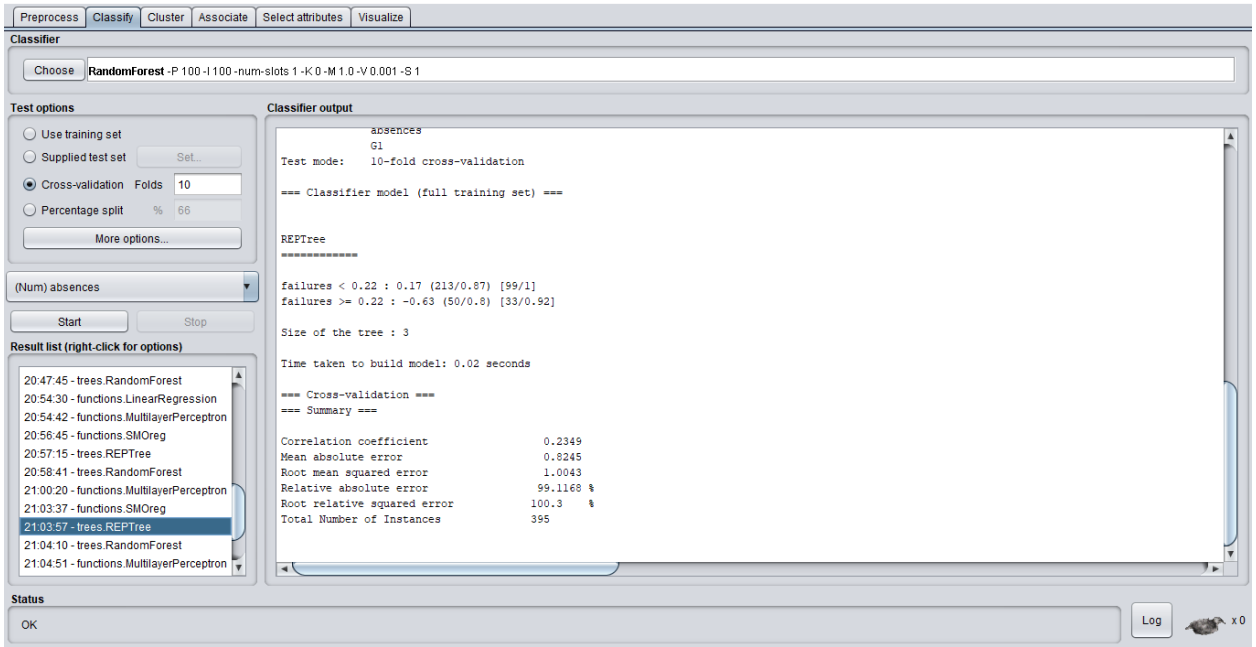
neural network

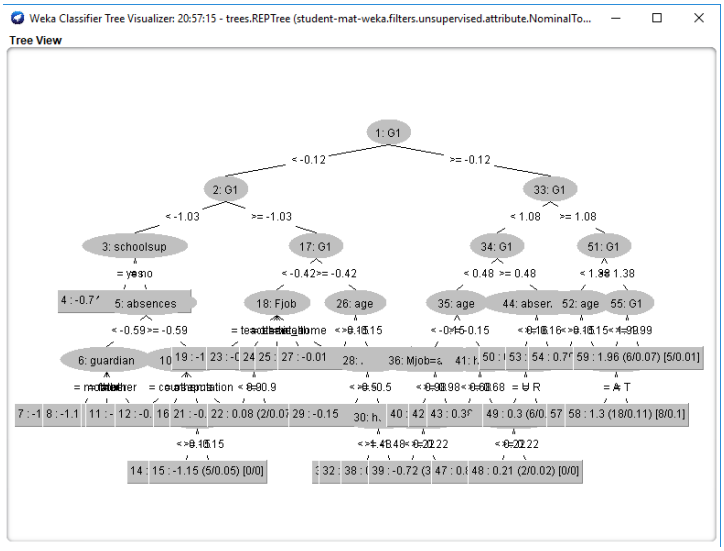


SVM

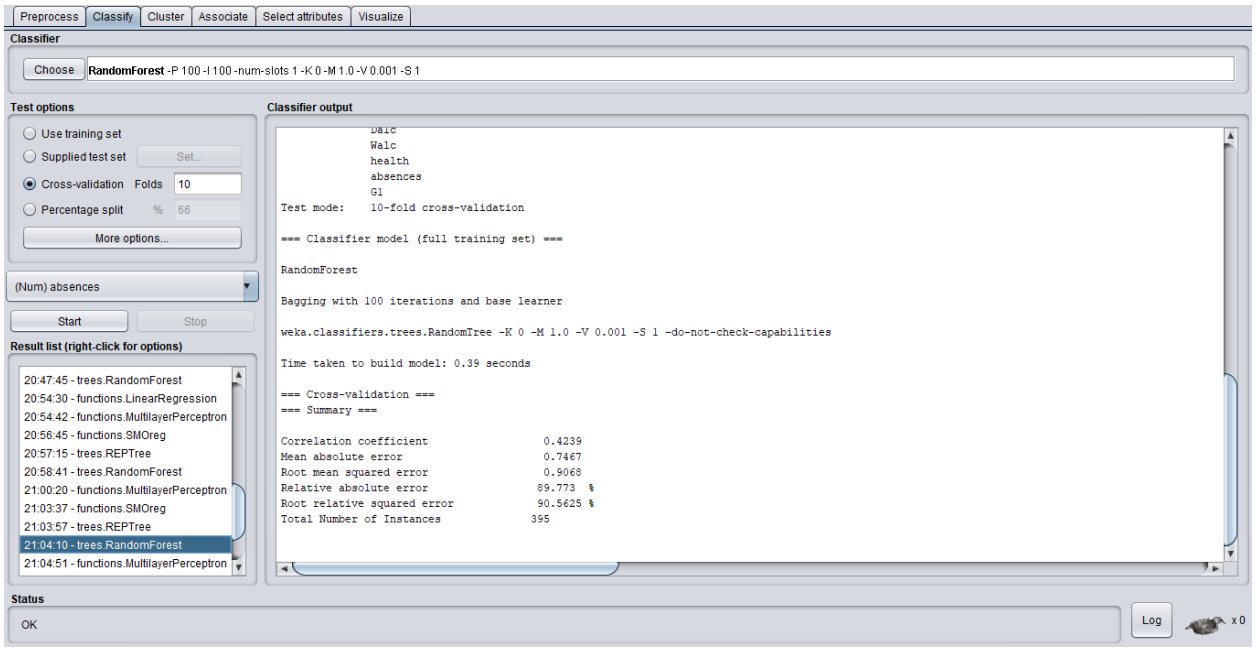


Decision tree



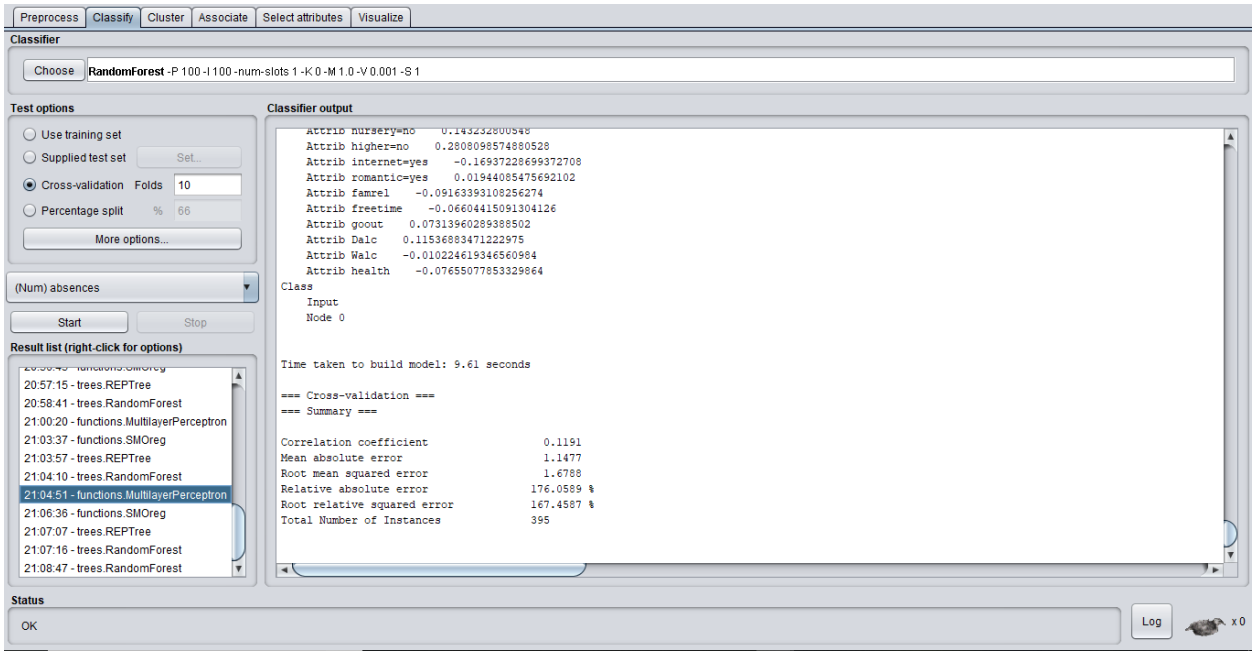


Random Forest

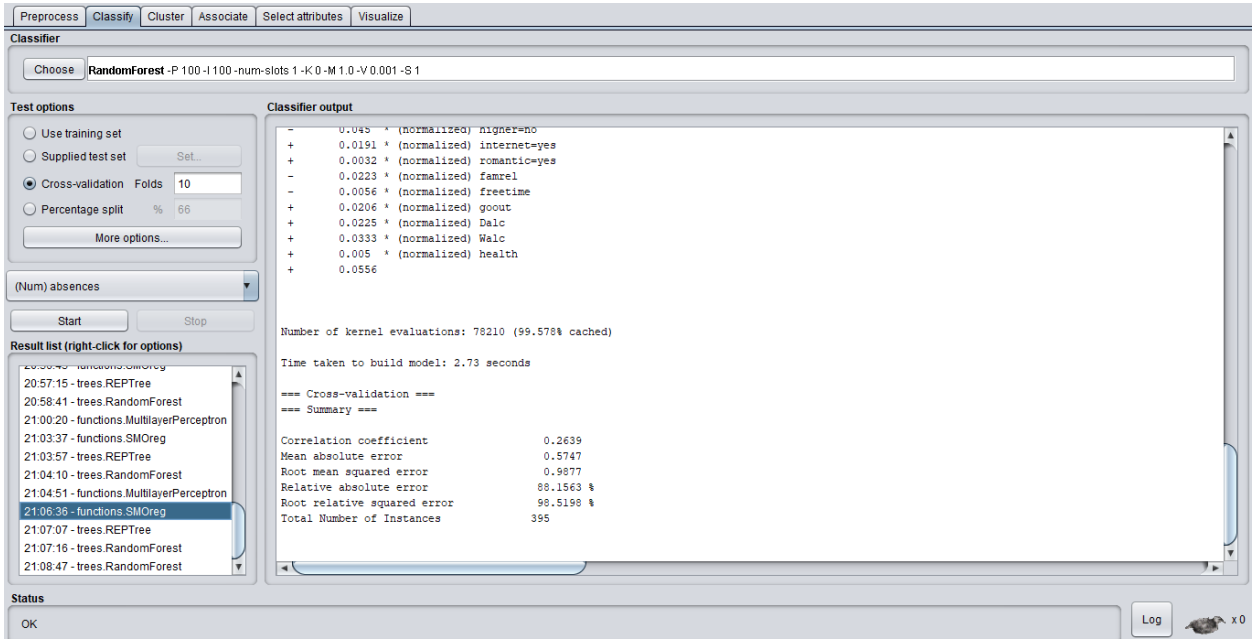


C – G1 also removed

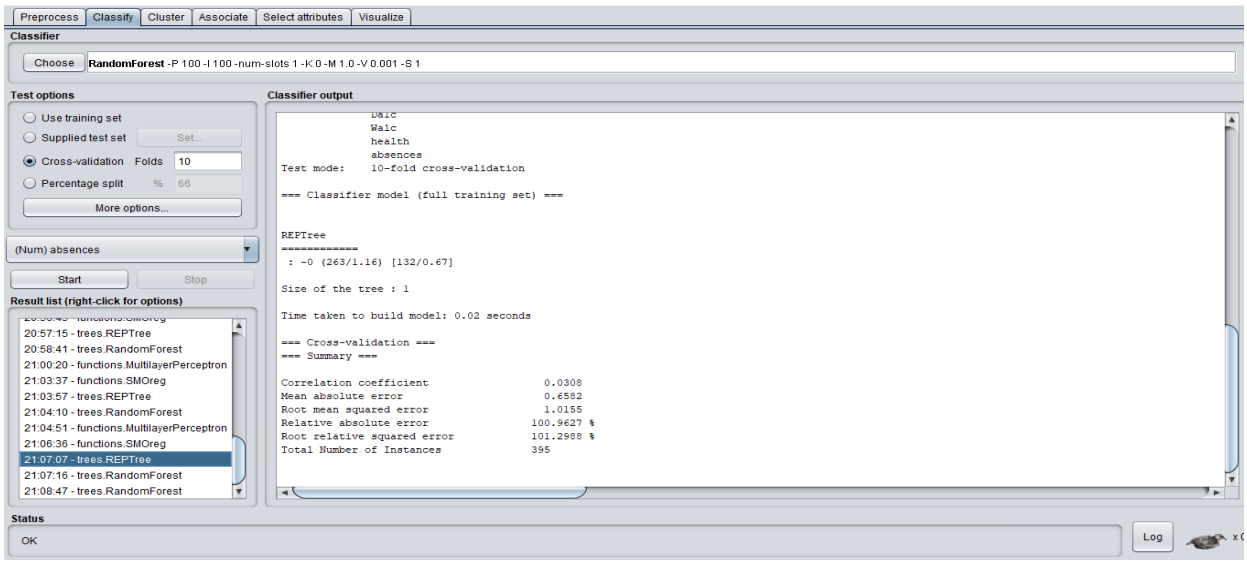
Neural network

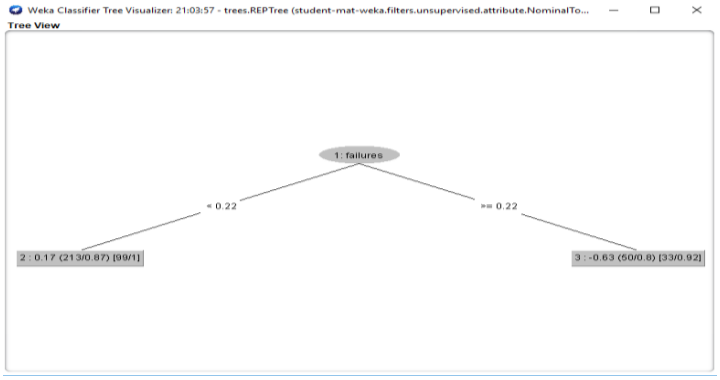


SVM

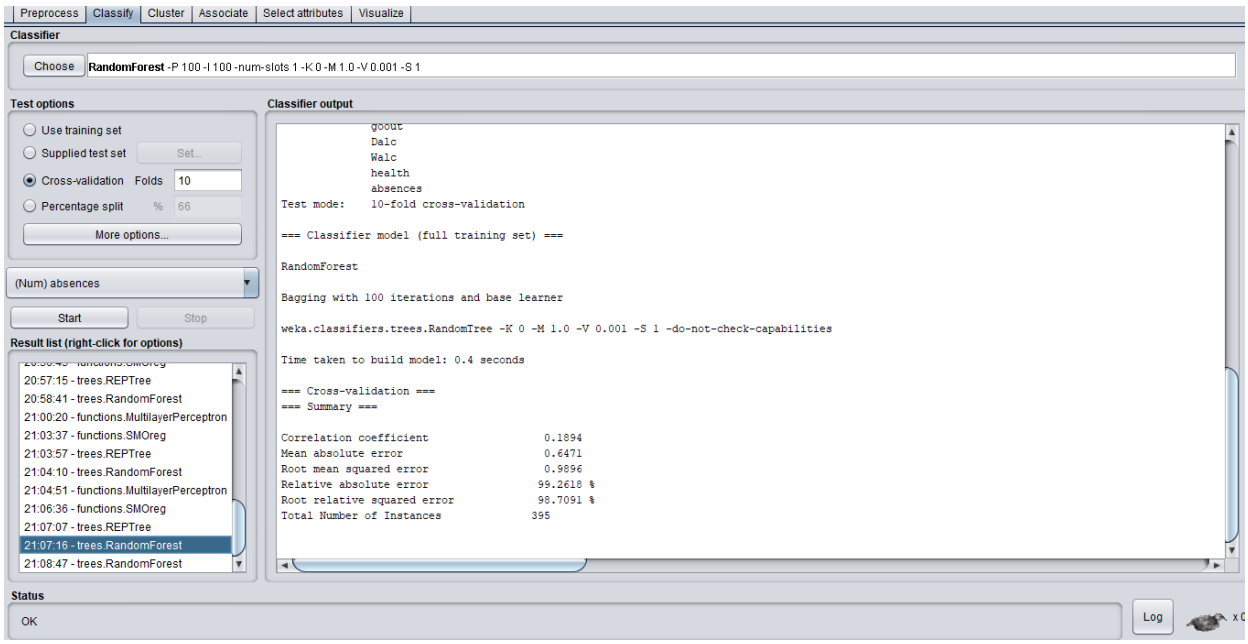


Decision Tree





Random Forest

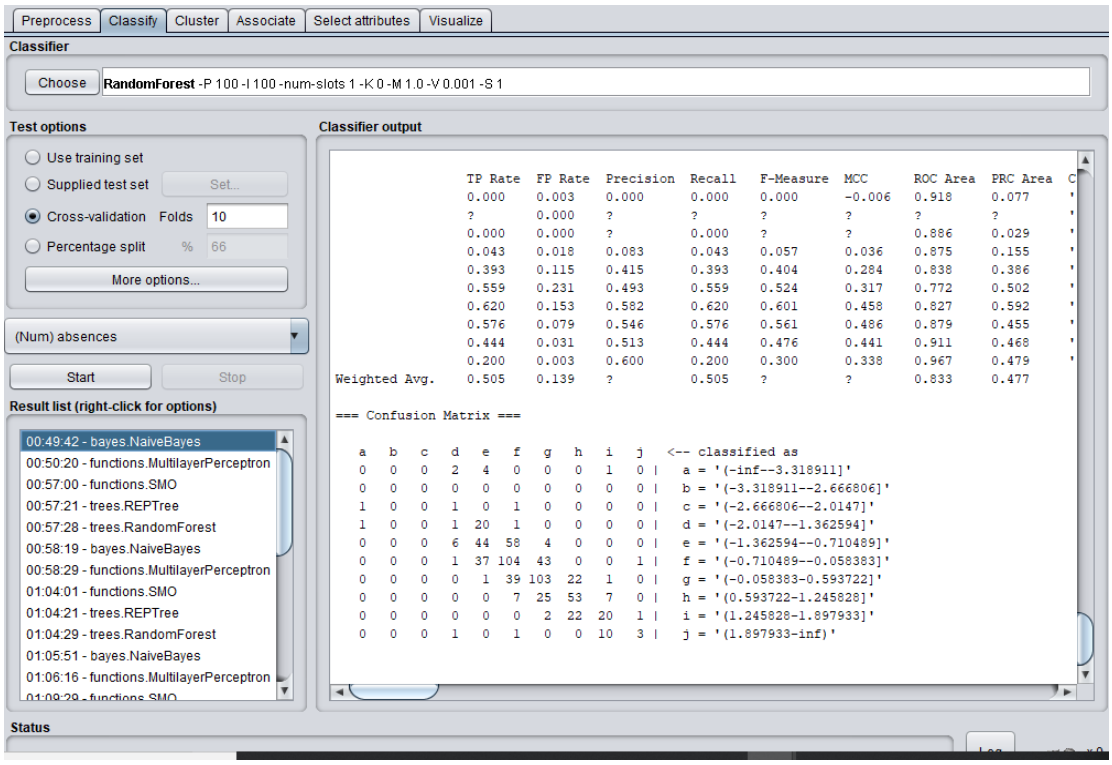


PROTUGESE RESULTS

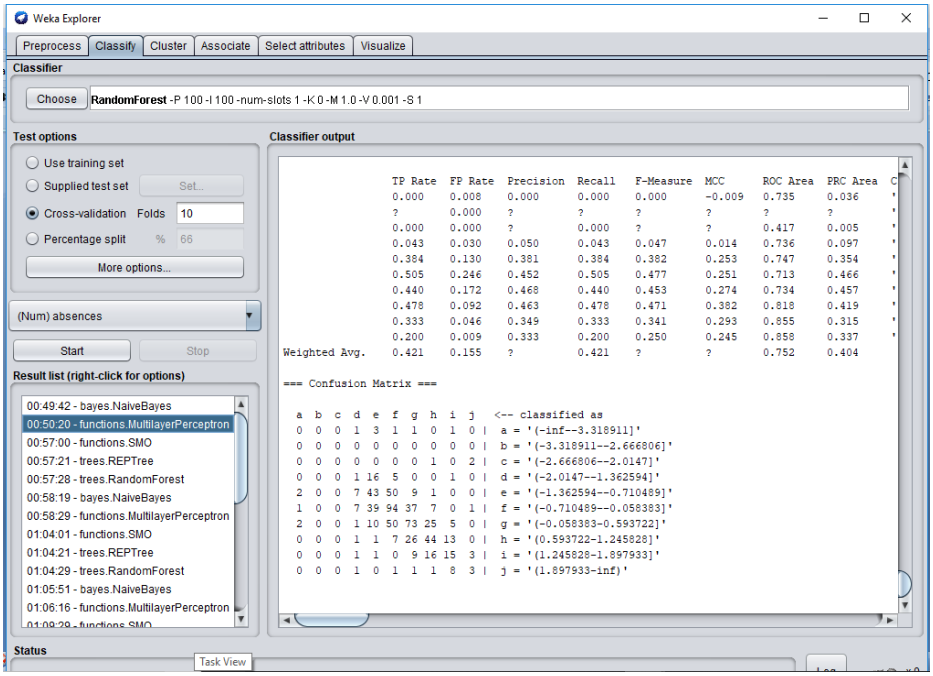
BINARY CLASSIFICATION

A – Final grade removed

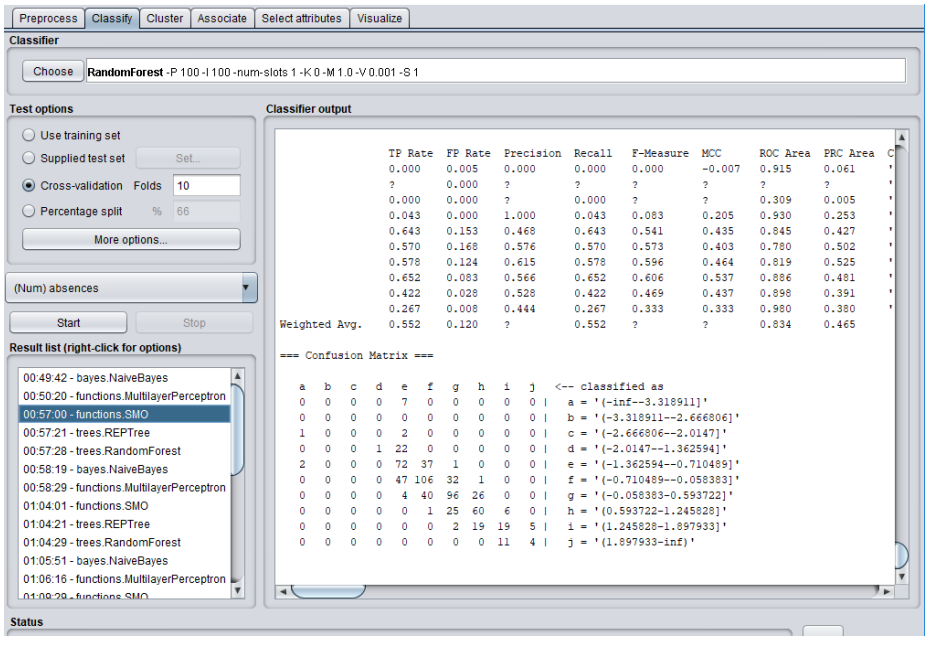
Naive bayes



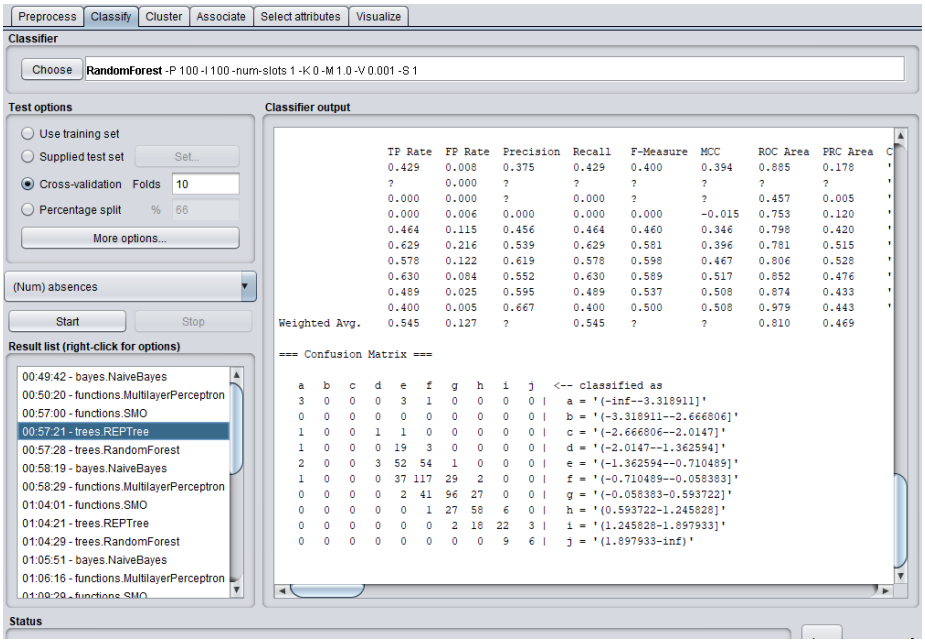
Neural network

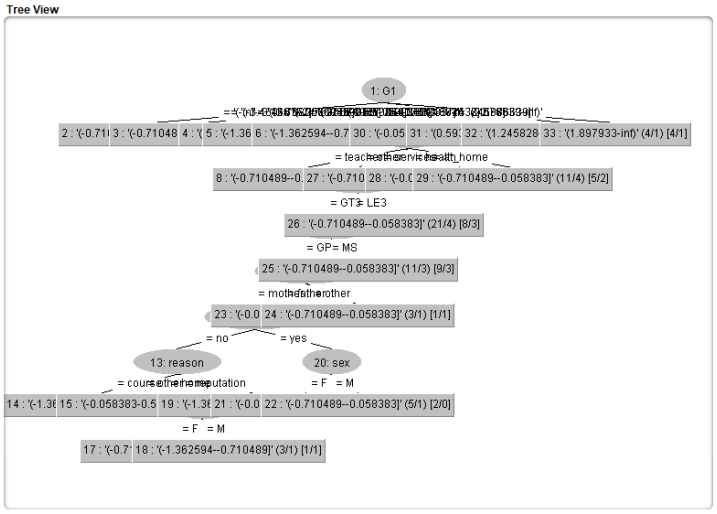


SVM

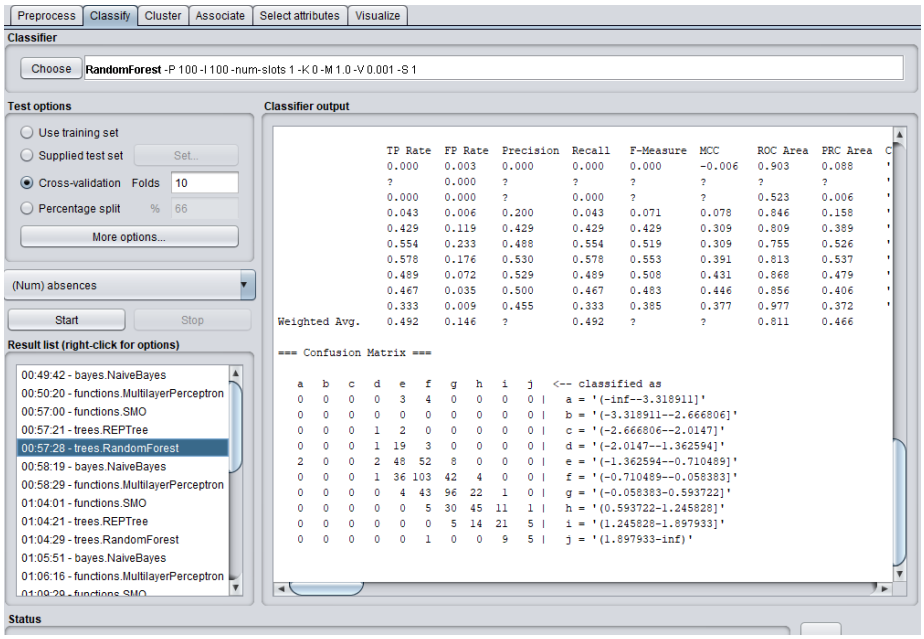


Decision Tree



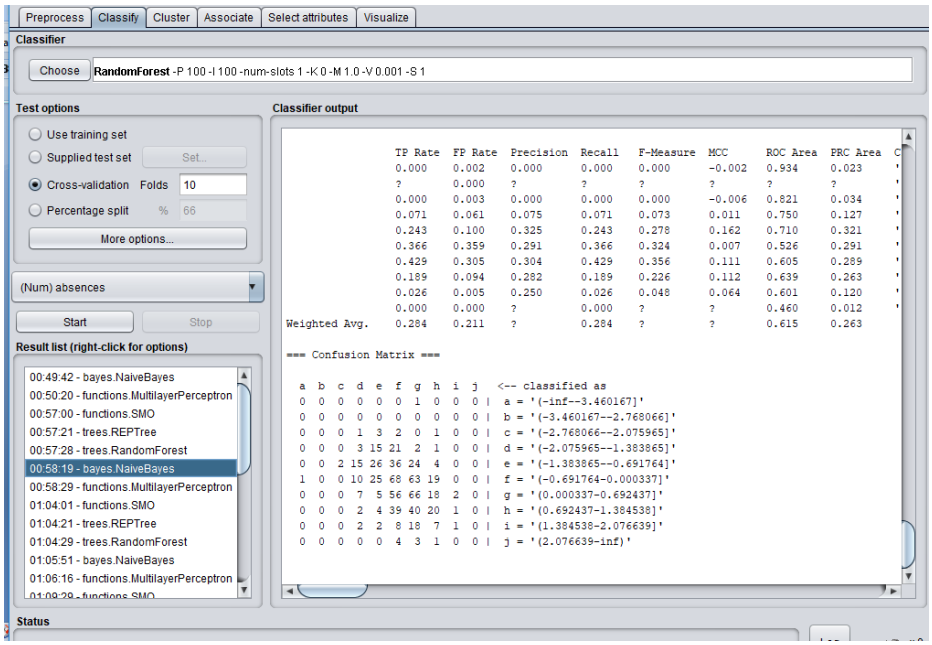


Random forest

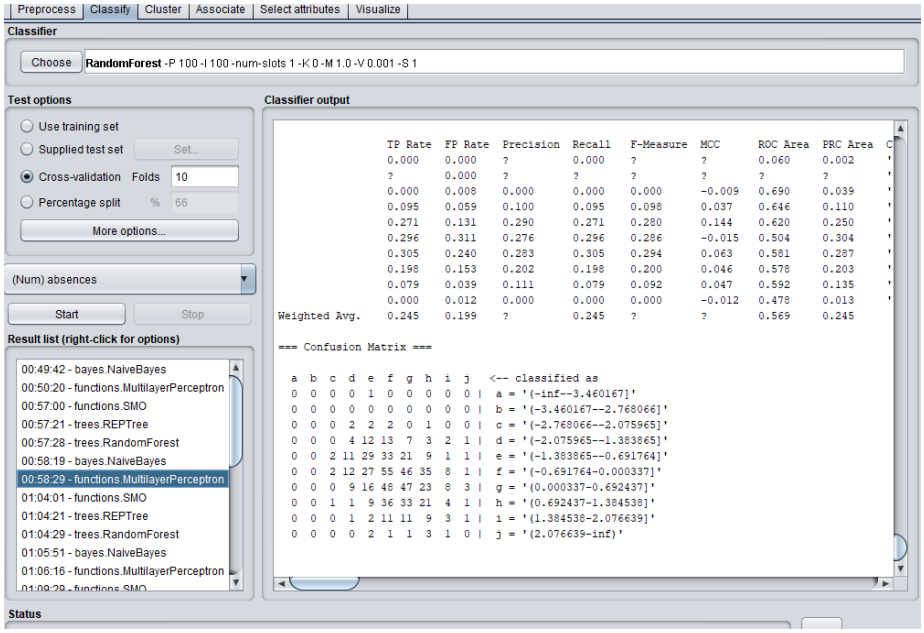


B – G2 also removed

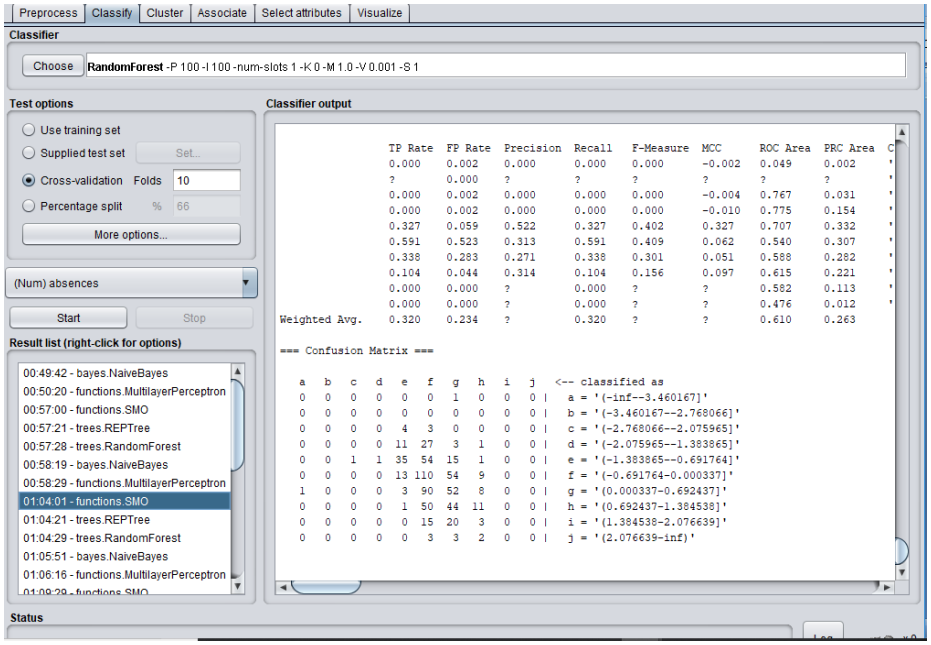
<u>Naïve- Bayes</u>



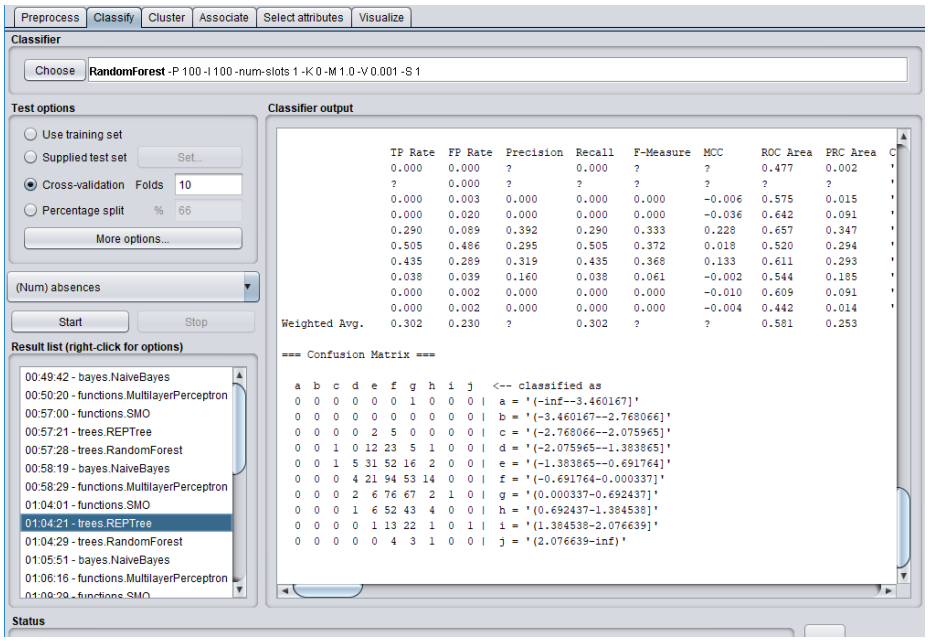
Neural Network

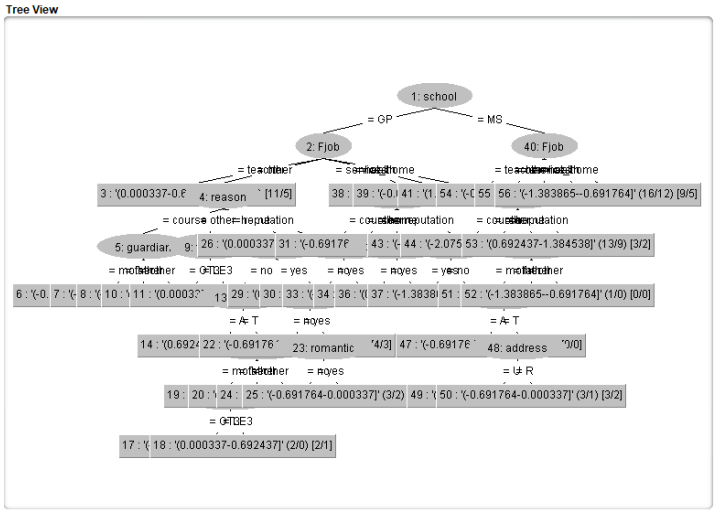


SVM

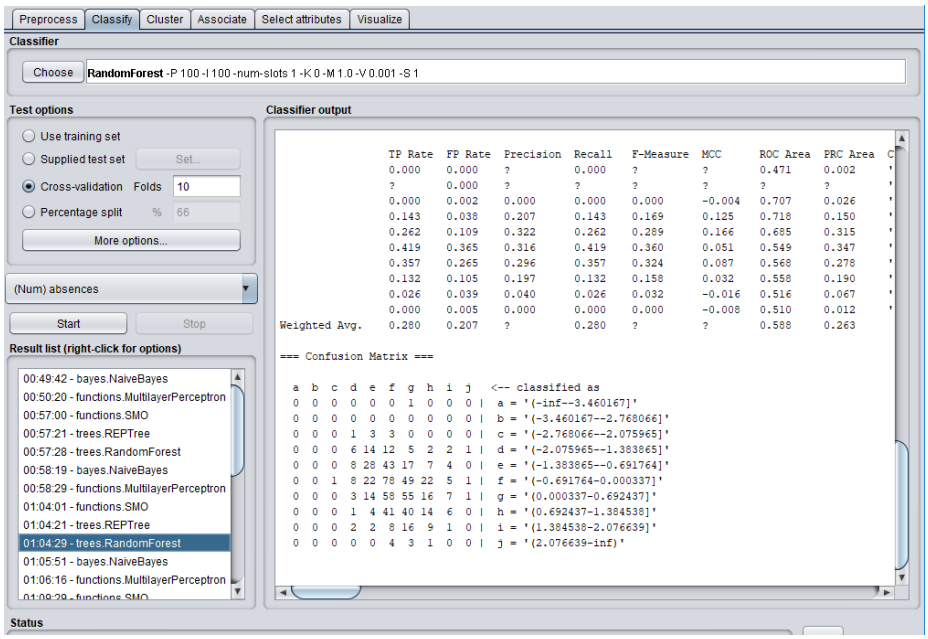


Decision Tree



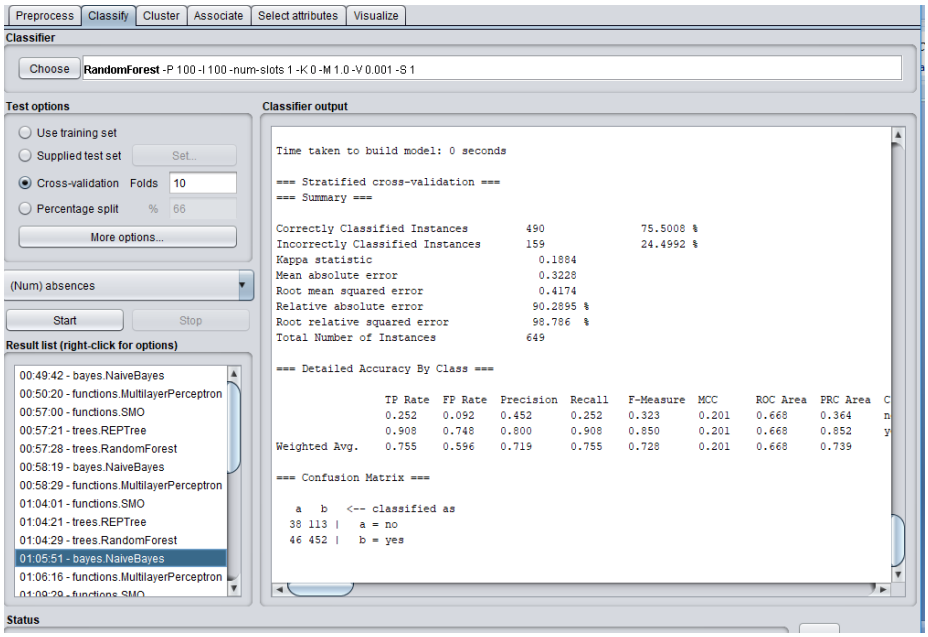


Random forest

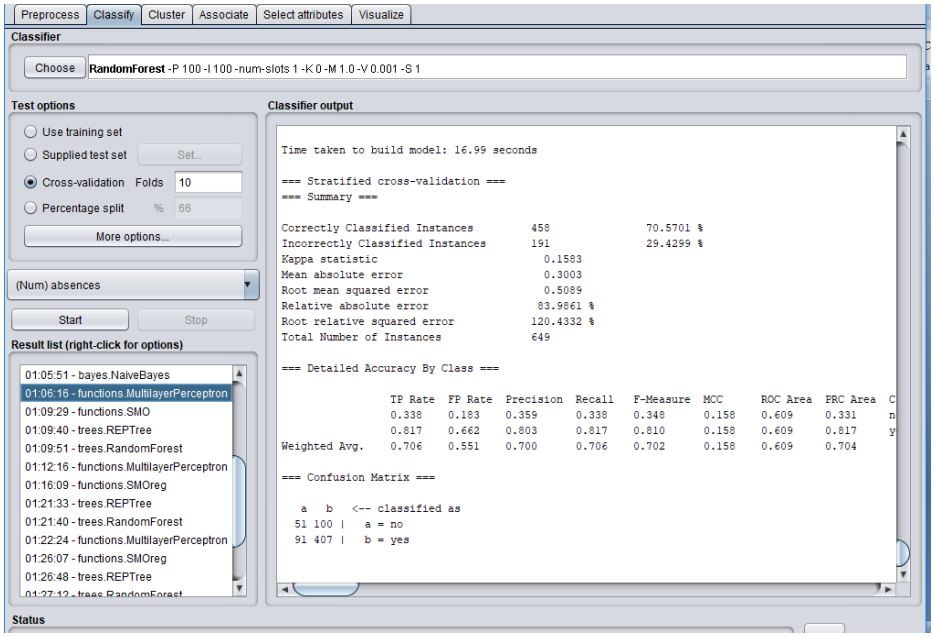


C – G1 is also removed

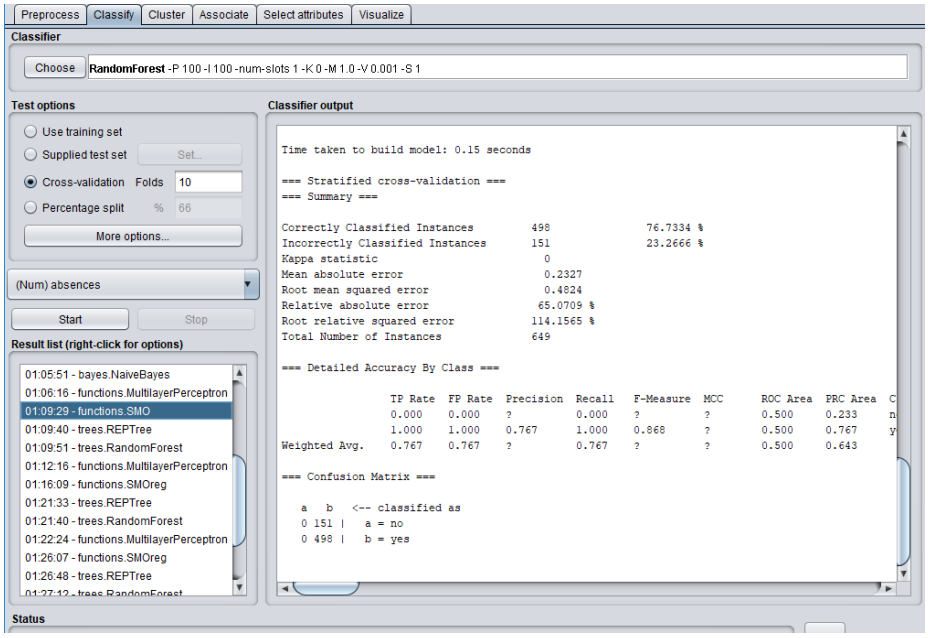
Naïve - Bayes



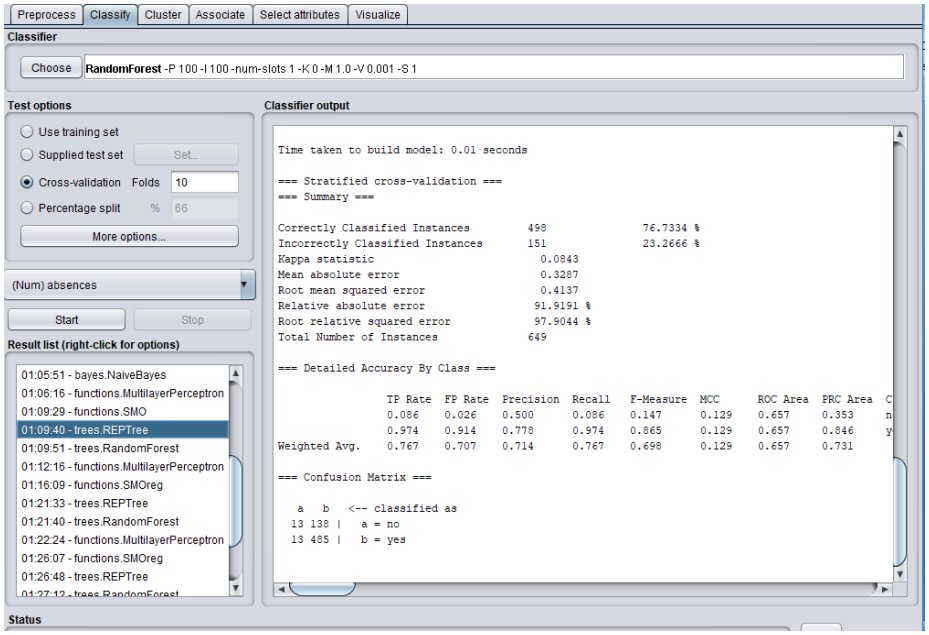
Neural Network

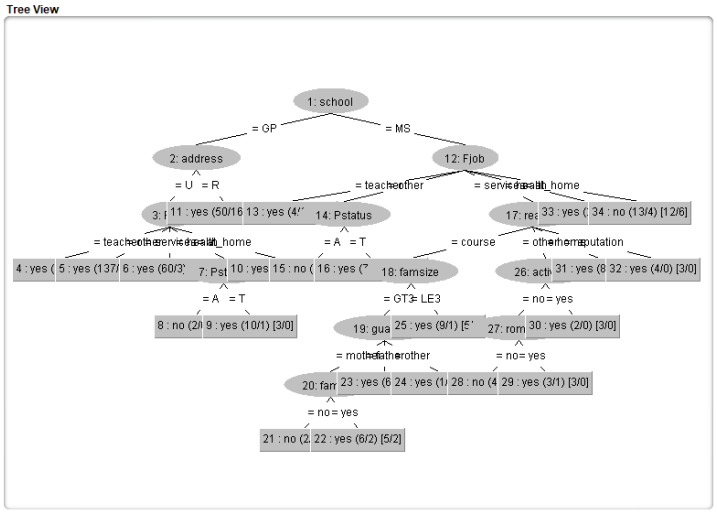


SVM

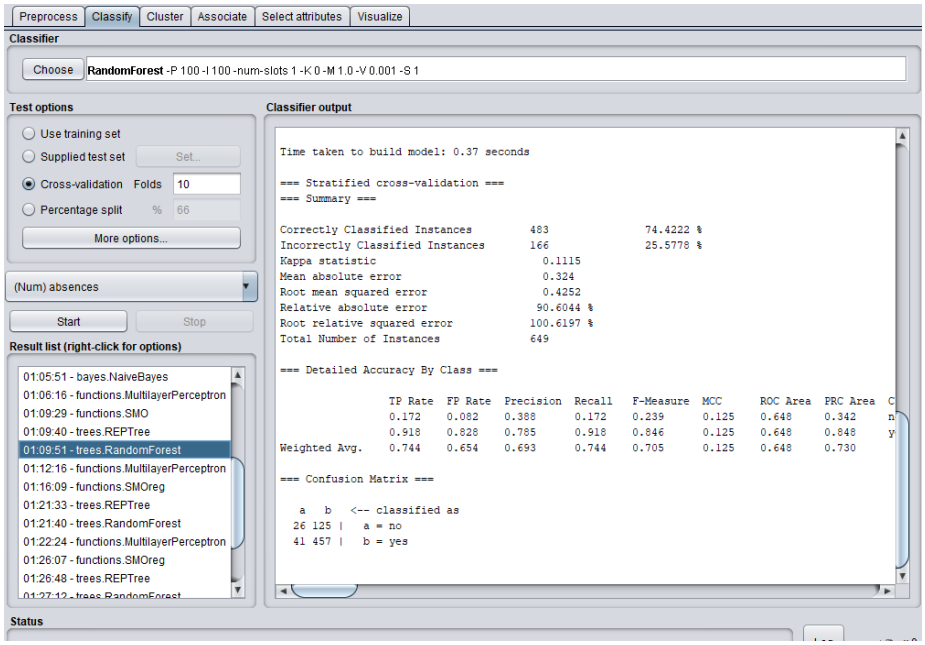


Decision Tree





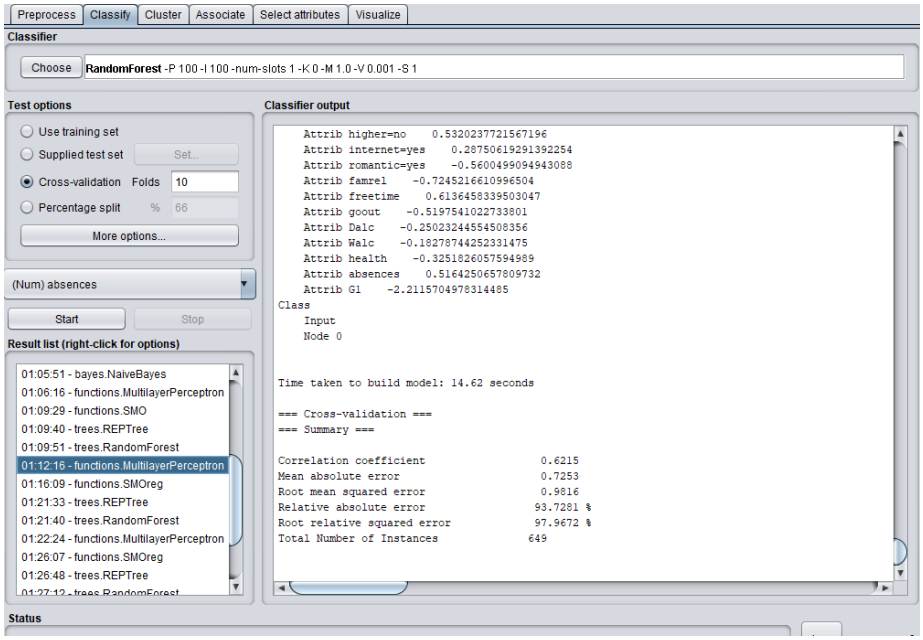
Random Forest



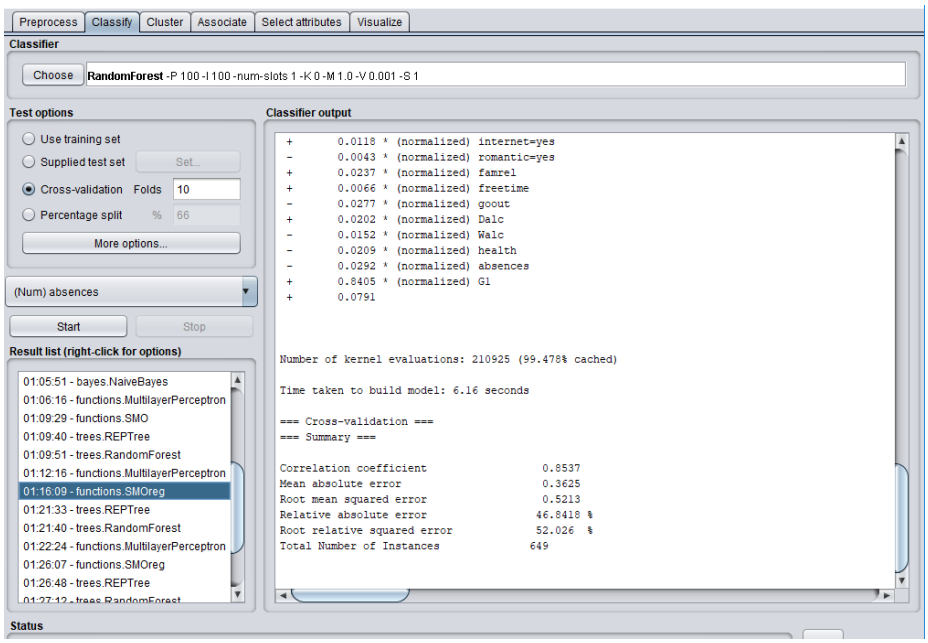
Linear regression

A - Final grade removed

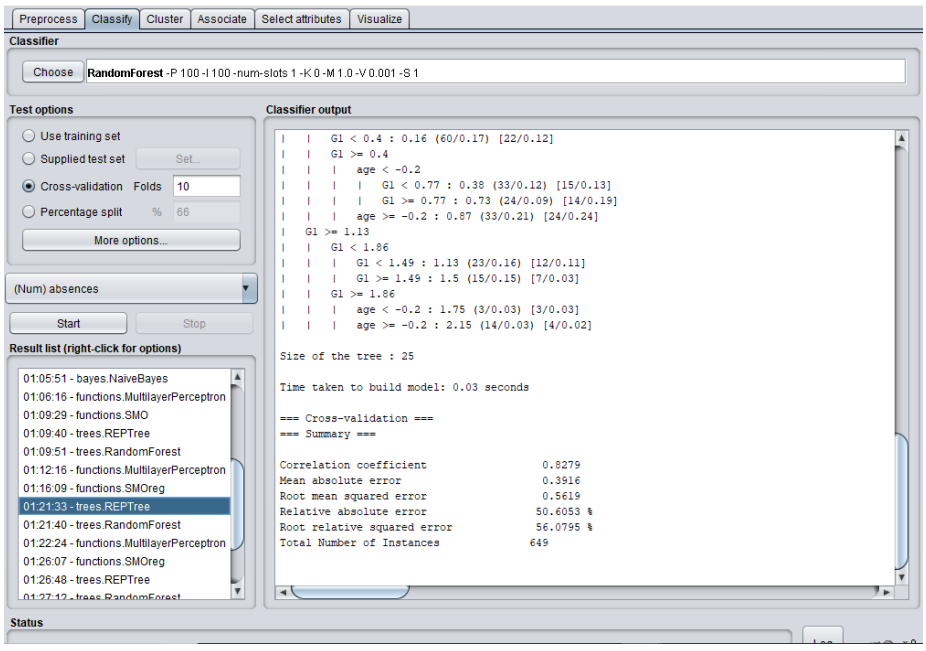
Neural Network

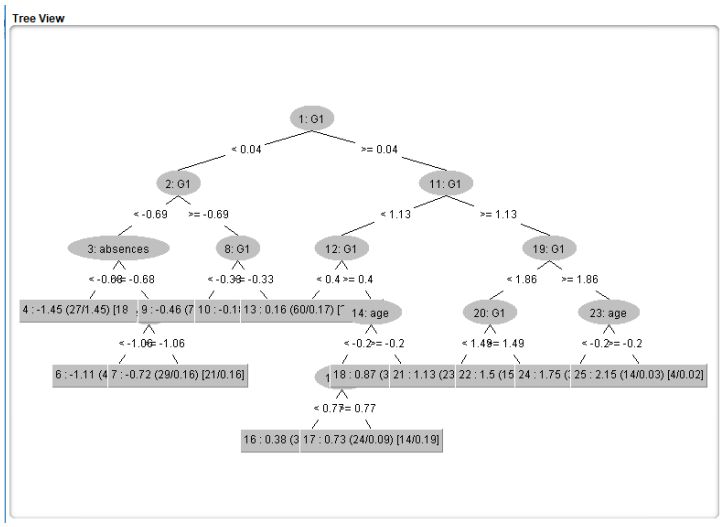


SVM

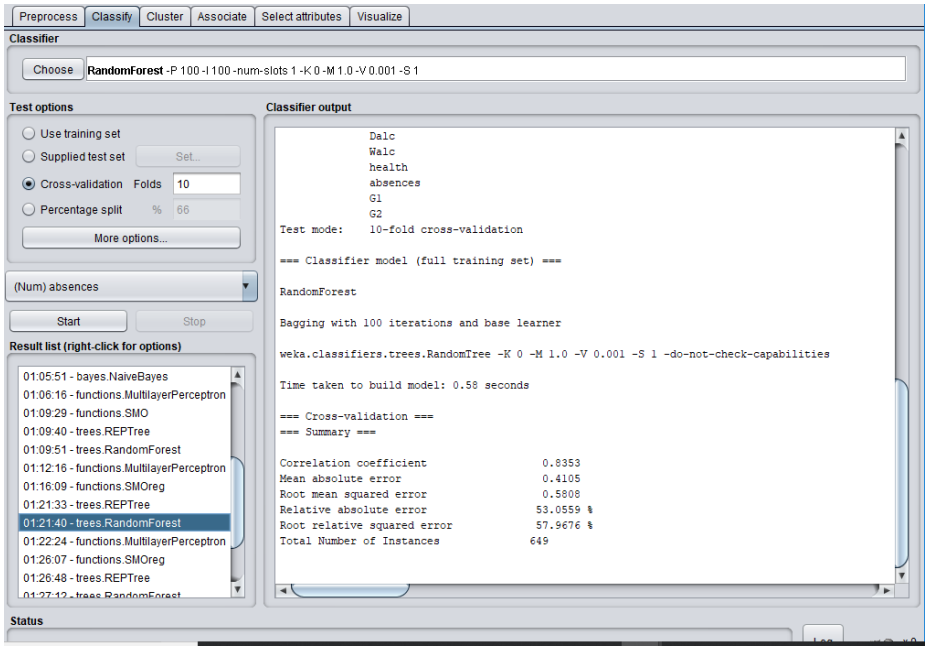


Decision Tree



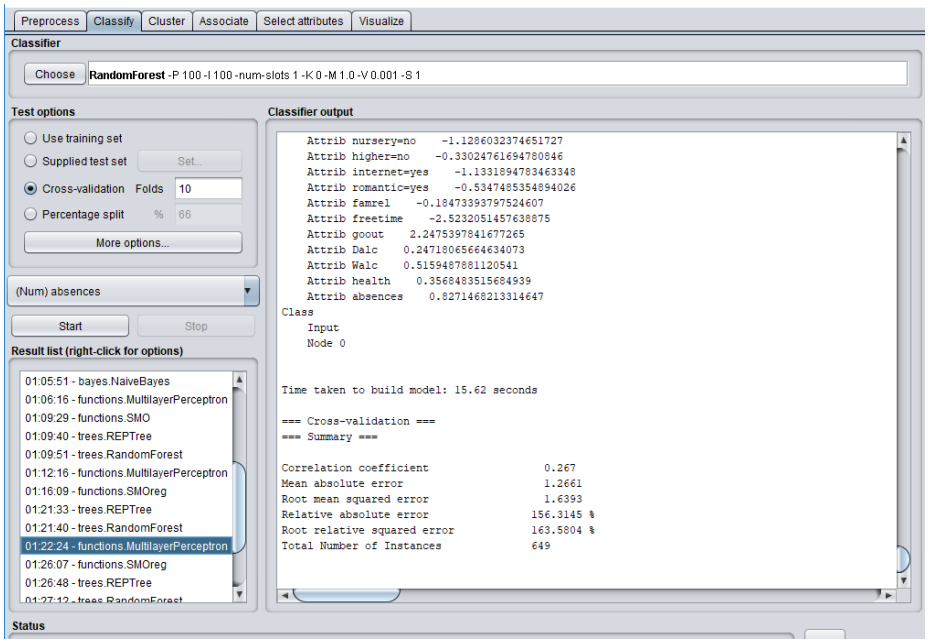


Random Forest

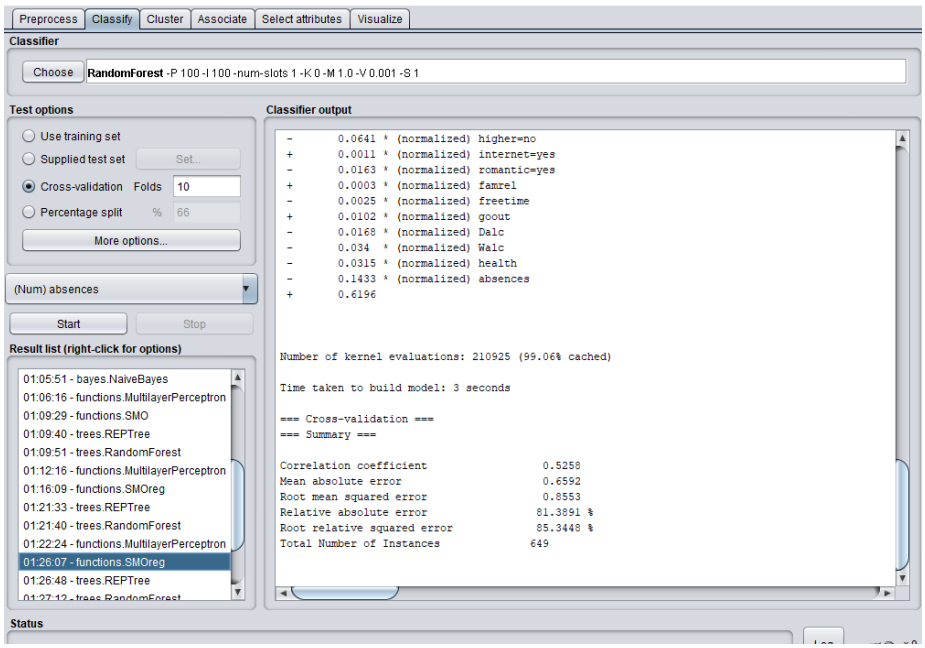


B – G2 also removed

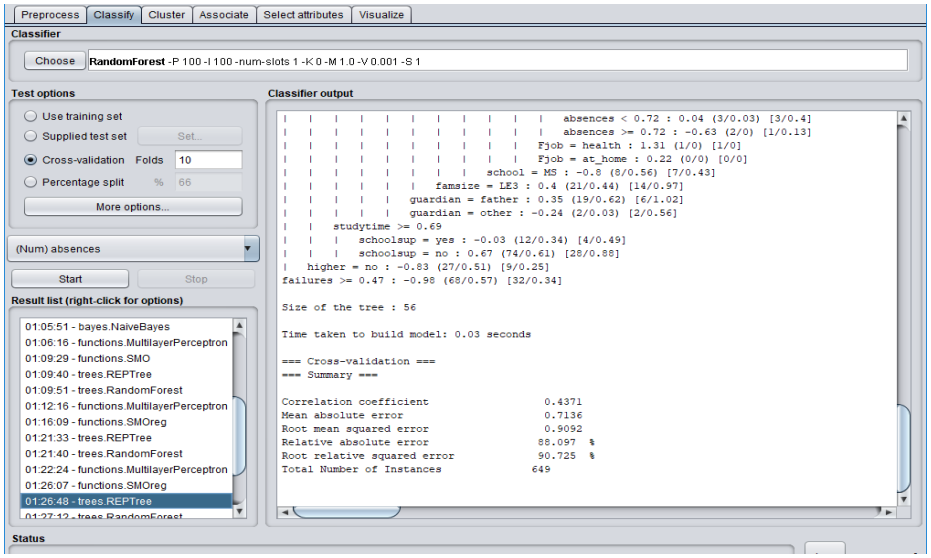
Neural Network

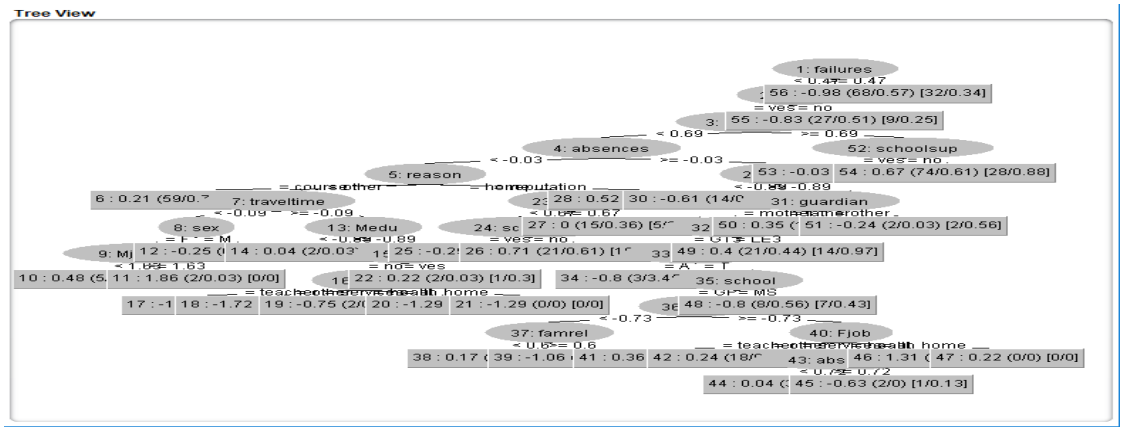

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Num) absences
Start
Stop
Result list (right-click for options)
01:05:51 - bayes.NaiveBayes
01:06:16 - functions.MultilayerPerceptron
01:09:29 - functions.SMO
01:09:40 - trees.REPTree
01:09:51 - trees.RandomForest
01:12:16 - functions.MultilayerPerceptron
01:16:09 - functions.SMOreg
01:21:33 - trees.REPTree
01:21:40 - trees.RandomForest
01:22:24 - functions.MultilayerPerceptron
01:26:07 - functions.SMOreg
01:26:48 - trees.REPTree
Classifier output
Attrib nursery=no -1.1286032374651727
Attrib higher=no -0.33024761694780846
Attrib internet=yes -1.1331894783463348
Attrib romantic=yes -0.5347485354894026
Attrib famrel -0.18473393797524607
Attrib freetime -2.5232051457638875
Attrib goout 2.2475397841677265
Attrib Dalc 0.24718065664634073
Attrib Walc 0.5159487881120541
Attrib health 0.3568483515684939
Attrib absences 0.8271468213314647
Class
Input
Node 0
Time taken to build model: 15.62 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.267
Mean absolute error 1.2661
Root mean squared error 1.6393
Relative absolute error 156.3145 %
Root relative squared error 163.5804 %
Total Number of Instances 649
Status

SVM

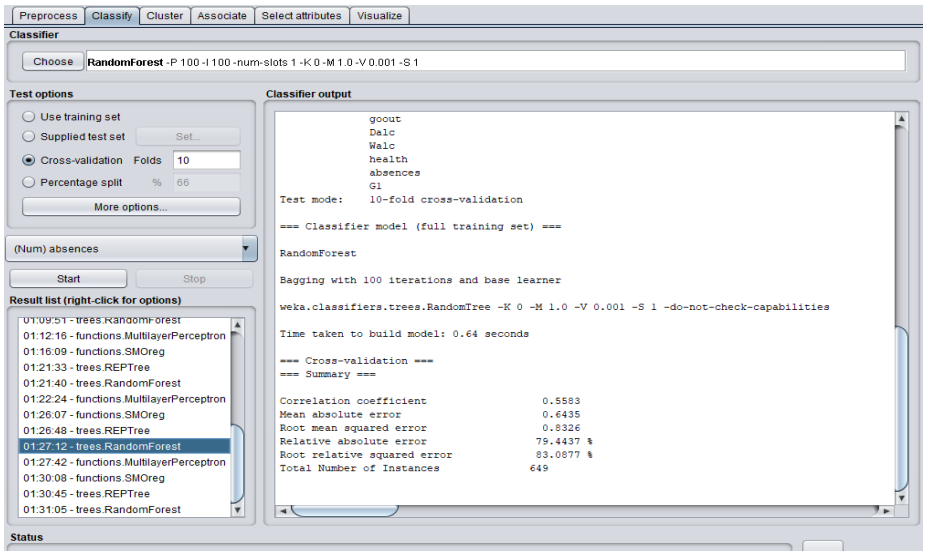

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Num) absences
Start
Stop
Result list (right-click for options)
01:05:51 - bayes.NaiveBayes
01:06:16 - functions.MultilayerPerceptron
01:09:29 - functions.SMO
01:09:40 - trees.REPTree
01:09:51 - trees.RandomForest
01:12:16 - functions.MultilayerPerceptron
01:16:09 - functions.SMOreg
01:21:33 - trees.REPTree
01:21:40 - trees.RandomForest
01:22:24 - functions.MultilayerPerceptron
01:26:07 - functions.SMOreg
01:26:48 - trees.REPTree
Classifier output
- 0.0641 * (normalized) higher=no
+ 0.0011 * (normalized) internet=yes
- 0.0163 * (normalized) romantic=yes
+ 0.0003 * (normalized) famrel
- 0.0025 * (normalized) freetime
+ 0.0102 * (normalized) goout
- 0.0168 * (normalized) Dalc
- 0.034 * (normalized) Walc
- 0.0315 * (normalized) health
- 0.1433 * (normalized) absences
+ 0.6196
Number of kernel evaluations: 210925 (99.06% cached)
Time taken to build model: 3 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.5258
Mean absolute error 0.6592
Root mean squared error 0.8553
Relative absolute error 81.3891 %
Root relative squared error 85.3448 %
Total Number of Instances 649
Status

Decision Tree



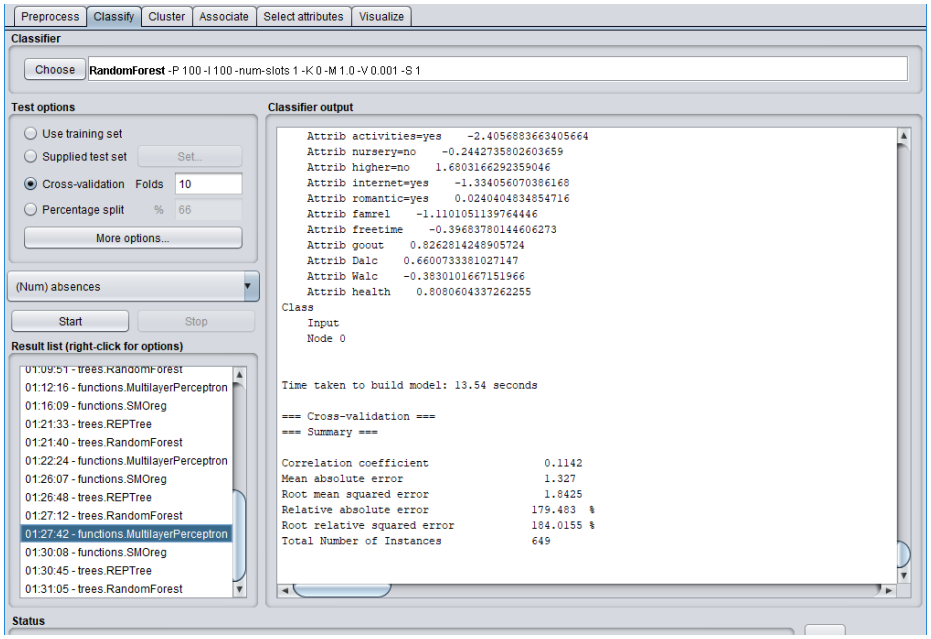


Random Forest

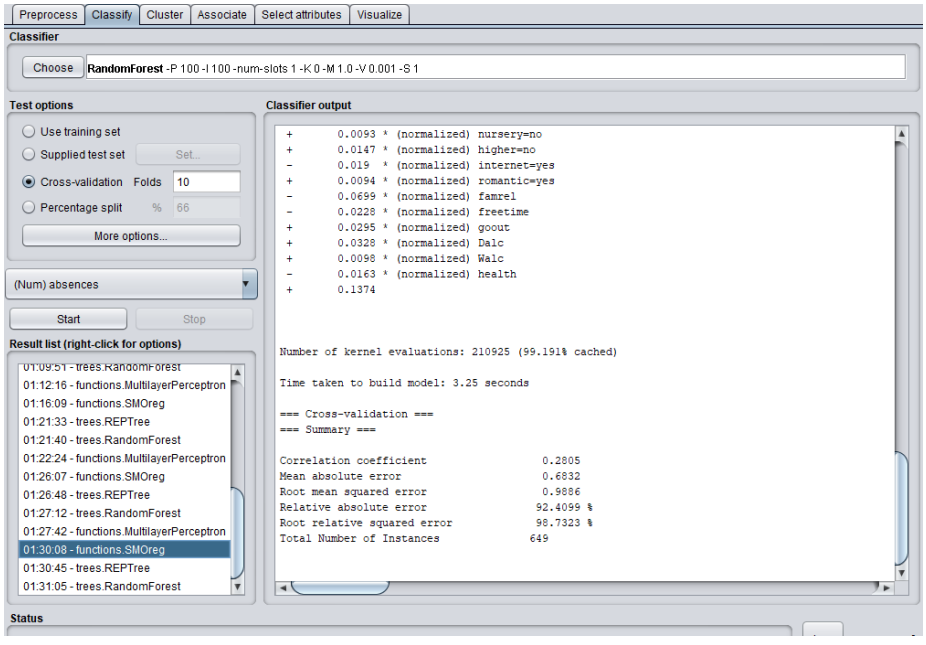


C – G1 also removed

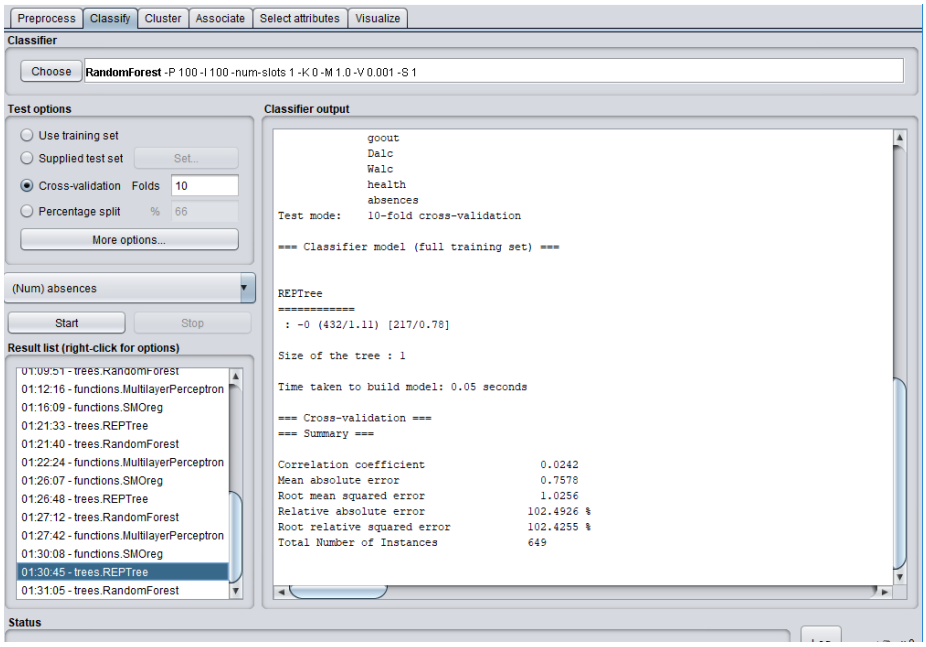
Neural Network



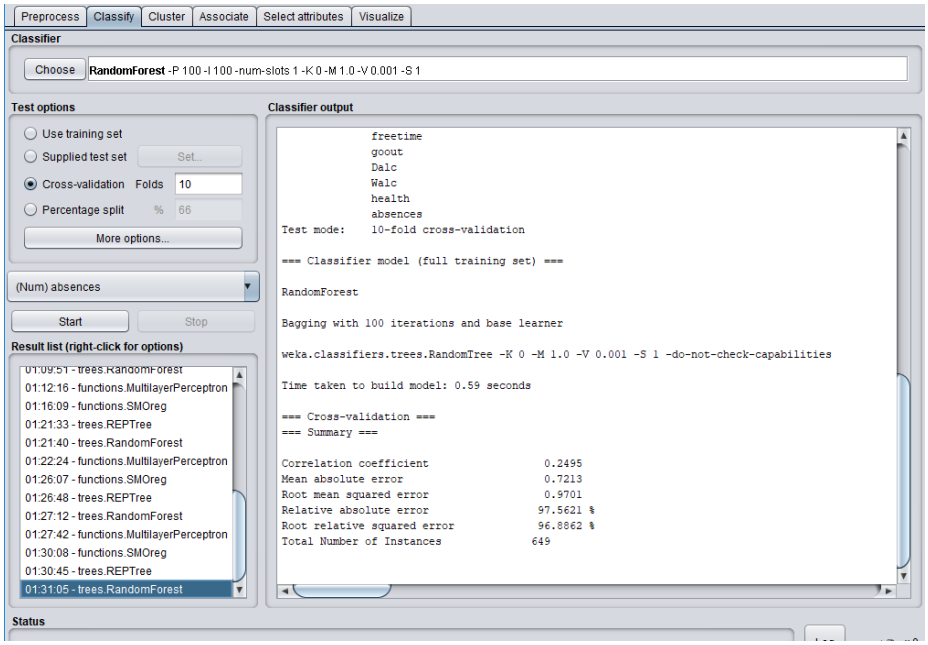
SVM



Decision Tree



Random Forest



Graph representing the attributes and their pre-processed form