**Data Mining III**

**CSE 40977**

[**Assignment I**](https://ucsdextension.blackboard.com/webapps/assignment/uploadAssignment?content_id=_1937780_1&course_id=_4953458_1&assign_group_id=&mode=view)

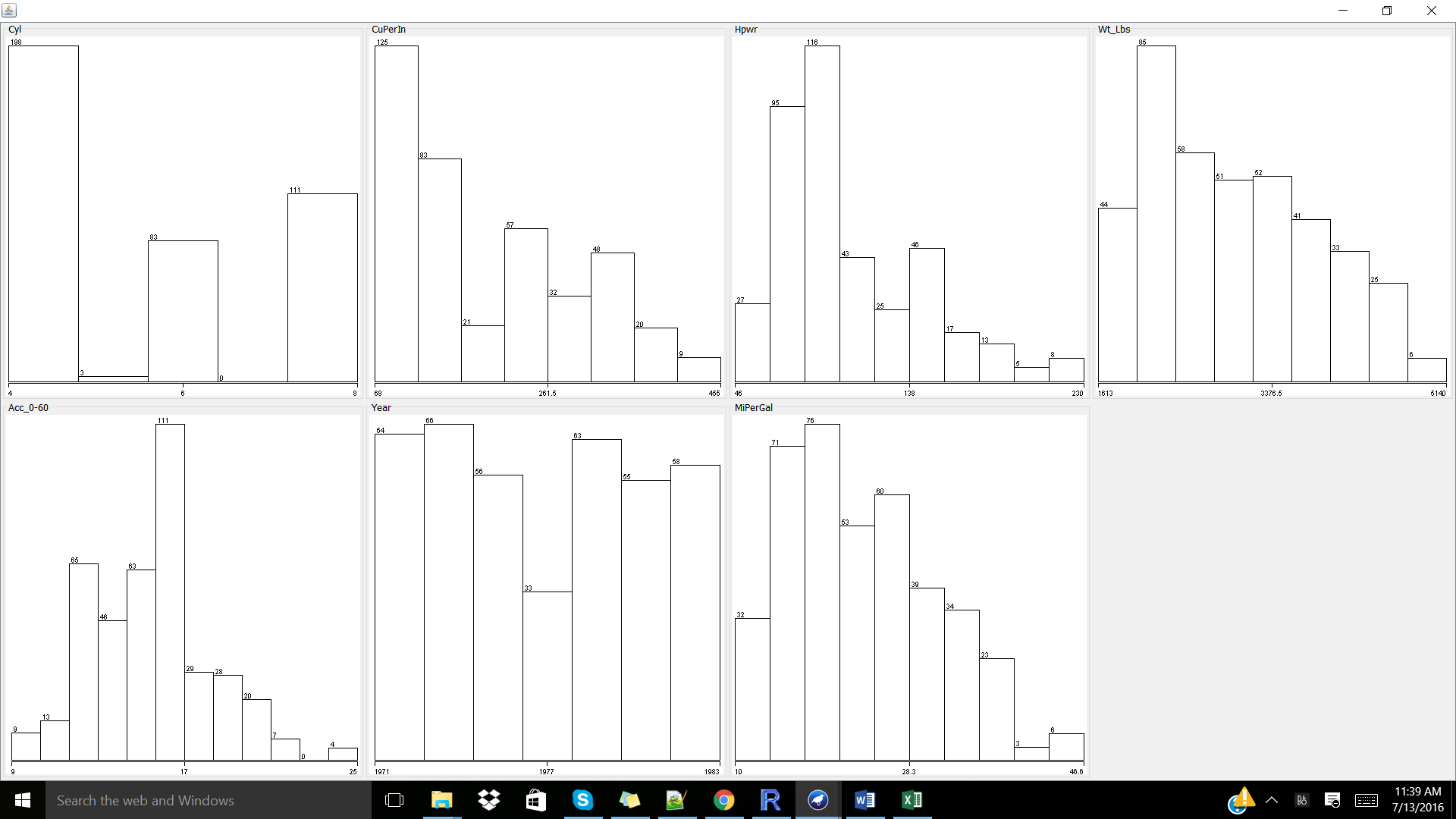
1 - Open CARS1.arff that we saved during Lesson 2. Using the following guidelines, perform additional modeling of the CARS1 data:  
  
(i) using other methods such as SVM, SMO, Decision Stump etc. (choose 4 additional methods)  
  
(ii) Select the top 3 performing methods out of the models we built in the Lesson and the models you built in (i).  
  
(iii) Build additional models by performing parameter tuning (such as useSmoothed and minNumberInstances in the Model Tree method) on your top three models. Perform 6 runs with the changed parameters and report your best score (for example, two different runs for each of the top three models)..

**Workflow**

**(1).** Following Instructions from Lesson 2: [Hands-On Case Studies, Step by Step Data Preparation and Modeling of Real-World Data Part I](https://ucsdextension.blackboard.com/bbcswebdav/pid-1937774-dt-content-rid-6752306_1/xid-6752306_1):

The CARS1 dataset was narrowed down from 10 attributes to 3, removing attributes Brand, Model, and Origin from the attribute list when predicting the class attribute, MiPerGal. Furthermore, non-real world values of 3 for Cyl were removed, missing values of MiPerGal were removed, other filtering was completed during pre-processing, and all of the attributes are of type numeric, including the class attribute, MiPerGal.

The “Visualize All” option demonstrates the instance distributions for the CARS1 dataset.



As completed following instructor direction, the M5P or Model Tree classifier was used to create a model to fit the class attribute, MiPerGal. The following model was created with a correlation coefficient of 0.9331 and 10 linear model leaves:

=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:

(using smoothed linear models)

CuPerIn <= 190.5 : LM1 (217/40.719%)

CuPerIn > 190.5 :

| Hpwr <= 141 :

| | CuPerIn <= 241 : LM2 (44/17.514%)

| | CuPerIn > 241 :

| | | Year <= 1978.5 : LM3 (41/20.043%)

| | | Year > 1978.5 : LM4 (15/28.906%)

| Hpwr > 141 :

| | Wt\_Lbs <= 4361.5 :

| | | Year <= 1977.5 :

| | | | Wt\_Lbs <= 3682.5 : LM5 (9/14.809%)

| | | | Wt\_Lbs > 3682.5 : LM6 (31/11.381%)

| | | Year > 1977.5 :

| | | | Wt\_Lbs <= 3997 : LM7 (4/8.682%)

| | | | Wt\_Lbs > 3997 : LM8 (7/4.756%)

| | Wt\_Lbs > 4361.5 :

| | | Year <= 1974.5 : LM9 (22/8.254%)

| | | Year > 1974.5 : LM10 (7/11.491%)

LM num: 1

MiPerGal =

-0.0219 \* Cyl

- 0.0438 \* CuPerIn

- 0.0582 \* Hpwr

- 0.0053 \* Wt\_Lbs

+ 0.911 \* Year

- 1750.8314

LM num: 2

MiPerGal =

-0.0261 \* Cyl

- 0.0163 \* CuPerIn

- 0.0019 \* Hpwr

- 0.0031 \* Wt\_Lbs

+ 0.3404 \* Year

- 638.9453

LM num: 3

MiPerGal =

-0.0261 \* Cyl

- 0.0242 \* CuPerIn

- 0.01 \* Hpwr

- 0.0023 \* Wt\_Lbs

- 0.2648 \* Acc\_0-60

+ 0.2856 \* Year

- 526.4988

LM num: 4

MiPerGal =

-0.0261 \* Cyl

+ 0.011 \* CuPerIn

- 0.0169 \* Hpwr

- 0.0052 \* Wt\_Lbs

+ 1.4864 \* Year

- 2905.0233

LM num: 5

MiPerGal =

-0.0261 \* Cyl

- 0.0017 \* CuPerIn

- 0.0143 \* Hpwr

- 0.0014 \* Wt\_Lbs

- 0.0578 \* Acc\_0-60

+ 0.0007 \* Year

+ 22.552

LM num: 6

MiPerGal =

-0.0261 \* Cyl

- 0.0009 \* CuPerIn

- 0.0056 \* Hpwr

- 0.0014 \* Wt\_Lbs

- 0.0578 \* Acc\_0-60

+ 0.1733 \* Year

- 319.7441

LM num: 7

MiPerGal =

-0.0261 \* Cyl

- 0.0056 \* Hpwr

- 0.0023 \* Wt\_Lbs

- 0.0578 \* Acc\_0-60

+ 0.2464 \* Year

- 460.0164

LM num: 8

MiPerGal =

-0.0261 \* Cyl

- 0.0056 \* Hpwr

- 0.002 \* Wt\_Lbs

- 0.0578 \* Acc\_0-60

+ 0.2464 \* Year

- 461.4961

LM num: 9

MiPerGal =

-0.0261 \* Cyl

- 0.0204 \* Hpwr

- 0.0024 \* Wt\_Lbs

- 0.4546 \* Acc\_0-60

+ 0.3107 \* Year

- 579.9211

LM num: 10

MiPerGal =

-0.0261 \* Cyl

- 0.0041 \* Hpwr

- 0.0022 \* Wt\_Lbs

- 0.2926 \* Acc\_0-60

+ 0.3885 \* Year

- 738.4909

Number of Rules : 10

Time taken to build model: 0.09 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9331

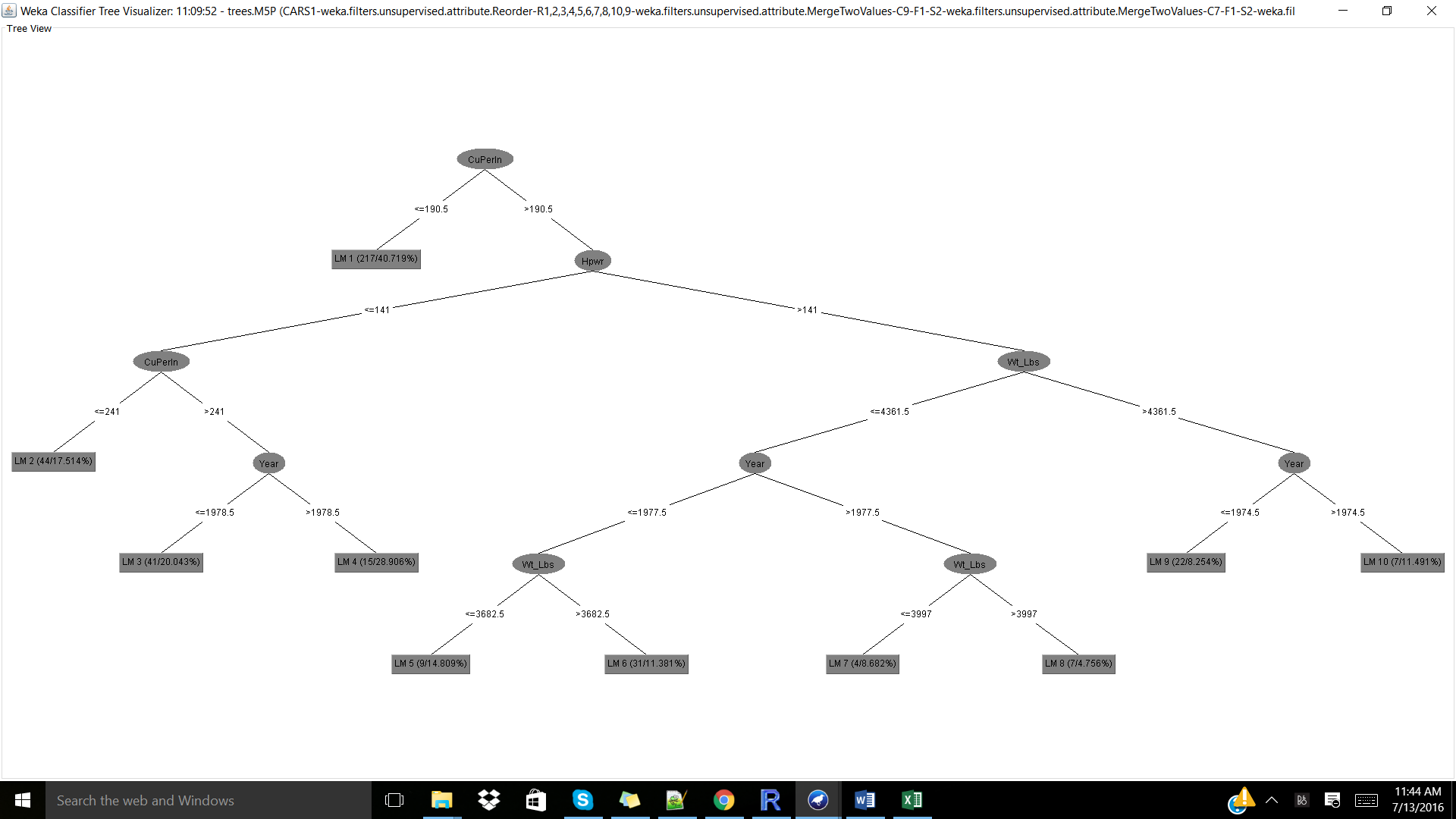
Mean absolute error 2.0088

Root mean squared error 2.801

Relative absolute error 30.5358 %

Root relative squared error 35.8942 %

Total Number of Instances 397



**(i)**

Furthermore, 4 additional classifiers were implemented on the CARS dataset. Note: all classifiers were run using Cross-validation Folds 10.

**Classifier.functions.SMOreg:**

=== Run information ===

Scheme:weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

SMOreg

weights (not support vectors):

- 0.0492 \* (normalized) Cyl

- 0.0185 \* (normalized) CuPerIn

- 0.0003 \* (normalized) Hpwr

- 0.5381 \* (normalized) Wt\_Lbs

- 0.0793 \* (normalized) Acc\_0-60

+ 0.191 \* (normalized) Year

+ 0.531

Number of kernel evaluations: 79003 (93.383% cached)

Time taken to build model: 0.14 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.8962

Mean absolute error 2.5546

Root mean squared error 3.5159

Relative absolute error 38.8325 %

Root relative squared error 45.0552 %

Total Number of Instances 397

**Classifier.trees.DecisionStump:**

=== Run information ===

Scheme:weka.classifiers.trees.DecisionStump

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Decision Stump

Classifications

Cyl <= 5.5 : 29.27661691542288

Cyl > 5.5 : 17.221649484536073

Cyl is missing : 16.5

Time taken to build model: 0.02 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.7506

Mean absolute error 3.9553

Root mean squared error 5.1493

Relative absolute error 60.1237 %

Root relative squared error 65.9872 %

Total Number of Instances 397

**Classifier.functions.MultiLayerPerceptron:**

=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Linear Node 0

Inputs Weights

Threshold 0.7338230160677247

Node 1 -1.0490847386186852

Node 2 -1.192466466221603

Node 3 -1.0750729557322694

Sigmoid Node 1

Inputs Weights

Threshold -2.3709367417922005

Attrib Cyl 0.5869382221436913

Attrib CuPerIn -0.5338810277776553

Attrib Hpwr 0.6166310499024139

Attrib Wt\_Lbs 0.6273561798093562

Attrib Acc\_0-60 0.1699677694457307

Attrib Year -0.24636824586124895

Sigmoid Node 2

Inputs Weights

Threshold 3.283637269097142

Attrib Cyl -0.3533600556624747

Attrib CuPerIn 1.3194101137803353

Attrib Hpwr 1.1295008989276456

Attrib Wt\_Lbs 2.340340055792048

Attrib Acc\_0-60 -0.15073863143399452

Attrib Year -1.3030662697111357

Sigmoid Node 3

Inputs Weights

Threshold -2.3139550911045985

Attrib Cyl 0.33359473083320645

Attrib CuPerIn -0.04362327372448996

Attrib Hpwr 0.1972504595637276

Attrib Wt\_Lbs 0.7749845598709538

Attrib Acc\_0-60 0.31649992187475295

Attrib Year -0.37375962138858015

Class

Input

Node 0

Time taken to build model: 0.16 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9032

Mean absolute error 2.5442

Root mean squared error 3.3694

Relative absolute error 38.6747 %

Root relative squared error 43.1782 %

Total Number of Instances 397

**Classifier.functions.LinearRegression:**

=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

MiPerGal =

-0.3388 \* Cyl +

-0.0061 \* Wt\_Lbs +

0.7251 \* Year +

-1389.9404

Time taken to build model: 0 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9009

Mean absolute error 2.5918

Root mean squared error 3.379

Relative absolute error 39.3974 %

Root relative squared error 43.3016 %

Total Number of Instances 397

Model Results Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Correlation Coefficient | Mean Absolute error | Comments |
| M5P | 0.9331 | 2.0088 | Strong classifier |
| SMOreg | 0.8962 | 2.5546 | Weaker classifier |
| DecisionStump | 0.7506 | 3.9553 | Weaker classifier |
| MultiLayerPerceptron | 0.9032 | 2.5442 | Strong classifier |
| LinearRegression | 0.9009 | 2.5918 | Strong classifier |

**(ii)**

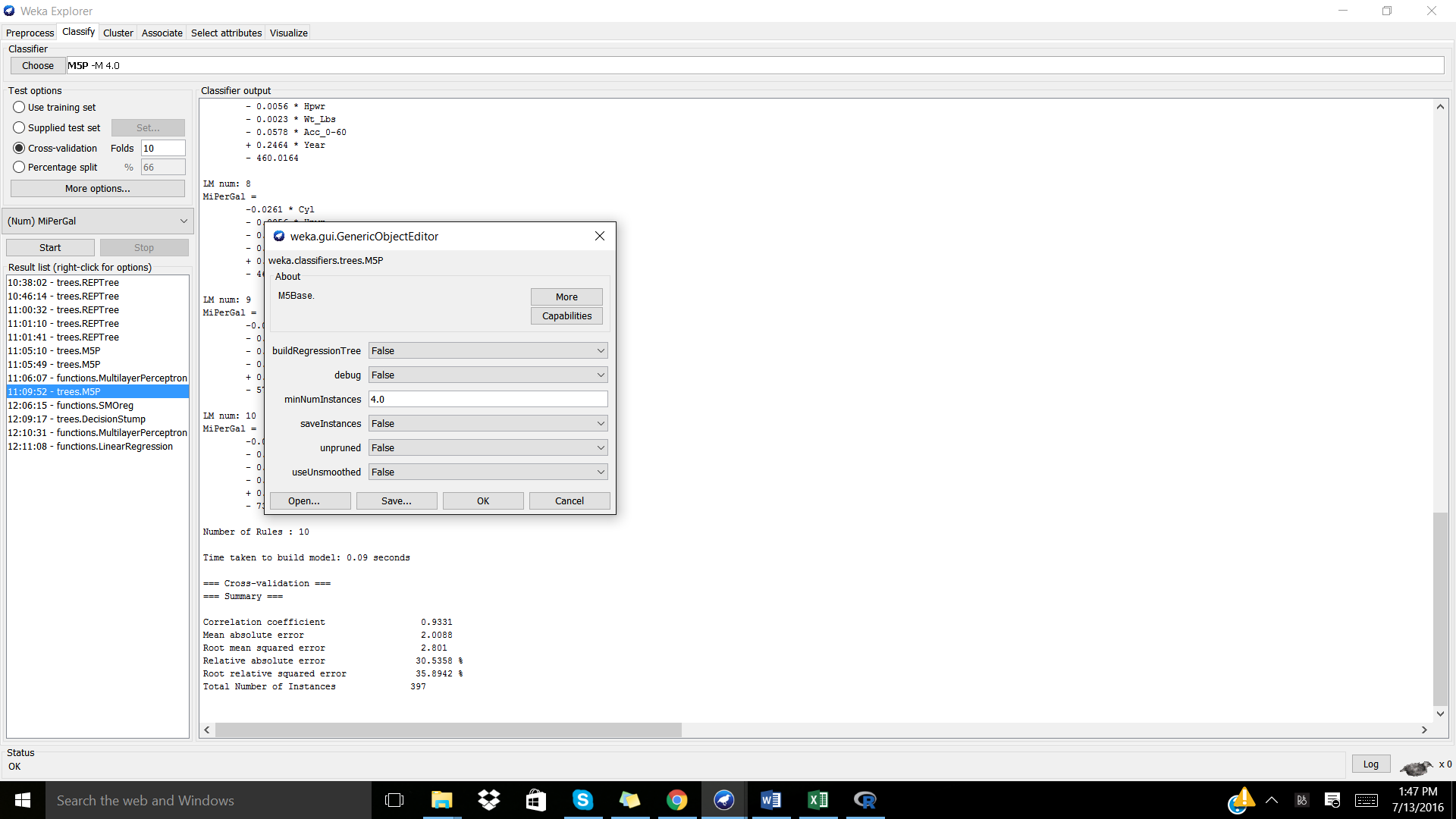
The M5P, MultiLayerPerceptron, and LinearRegression classifier models were found to be strong classifiers (Correlation coefficient > 0.90) and were thus considered for further analysis.

**(iii)**

Parameter tuning (ie. useSmoothed, minNumberInstances, etc.) was implemented on the top three models:

**M5P**

**At default**



=== Summary ===

Correlation coefficient 0.9331

Mean absolute error 2.0088

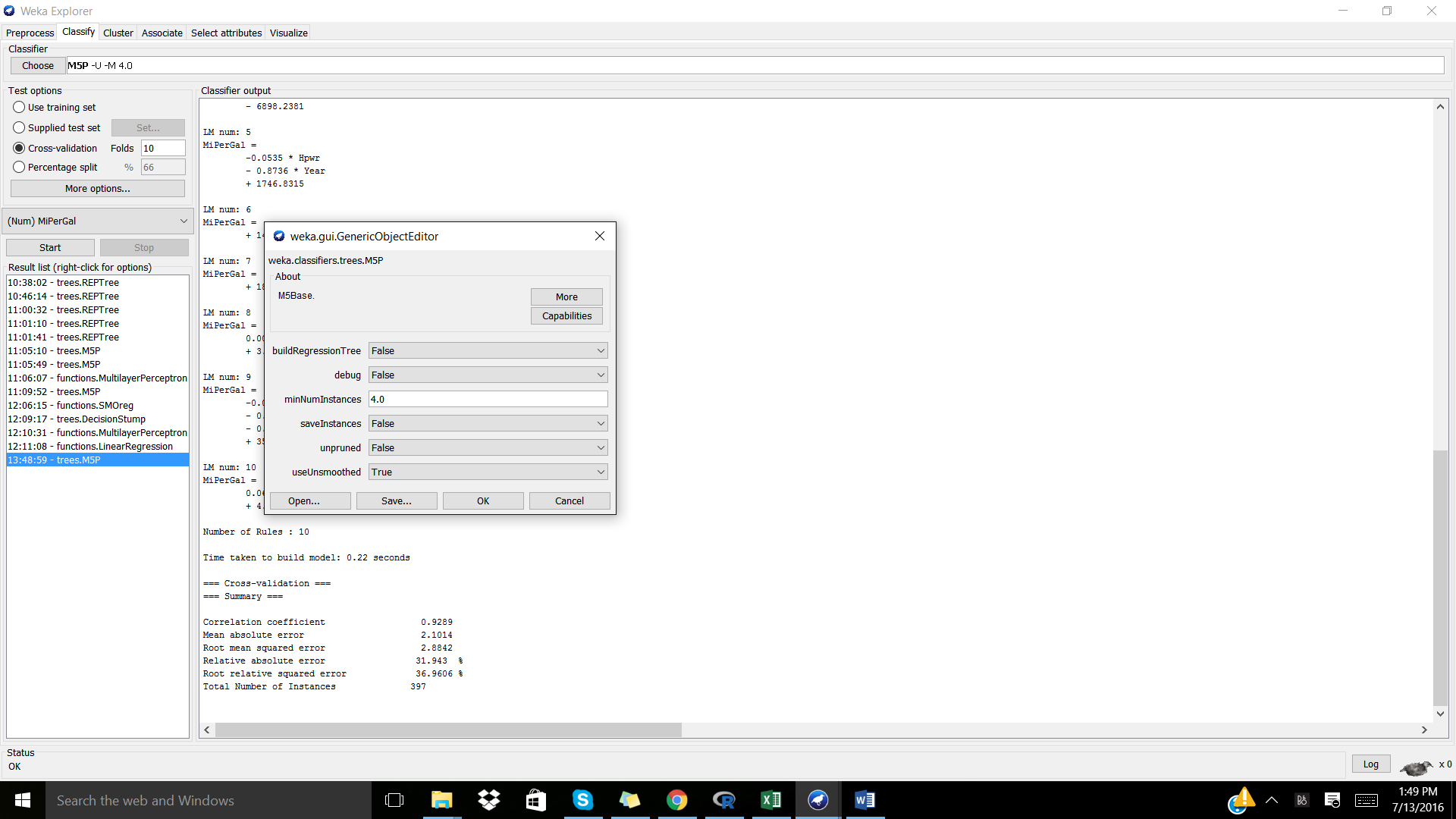
Root mean squared error 2.801

Relative absolute error 30.5358 %

Root relative squared error 35.8942 %

Total Number of Instances 397

**useUnsmoothed: True**



=== Summary ===

Correlation coefficient 0.9289

Mean absolute error 2.1014

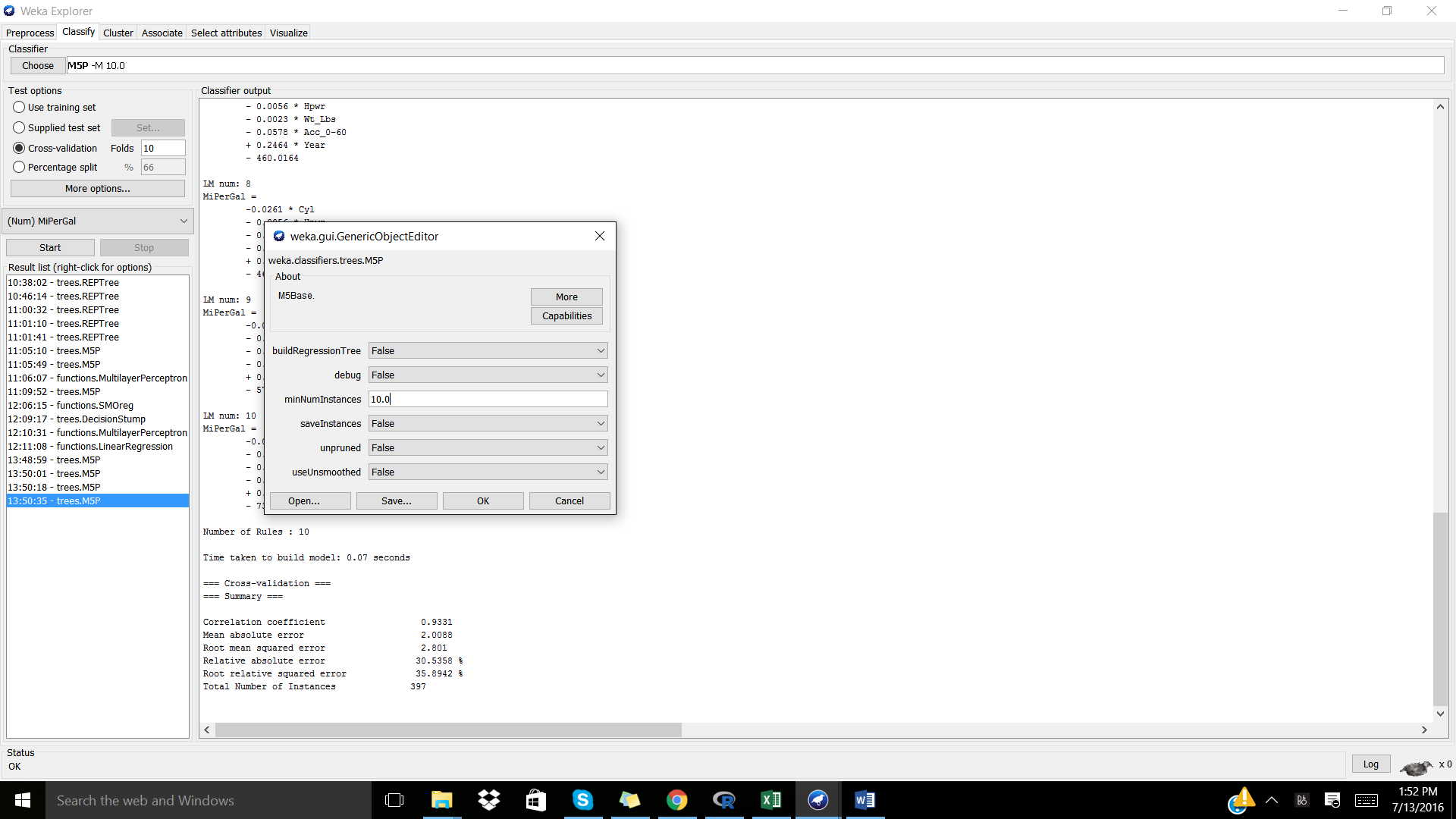
Root mean squared error 2.8842

Relative absolute error 31.943 %

Root relative squared error 36.9606 %

Total Number of Instances 397

No effect on the correlation coefficient was seen when decreasing the minNumInstances, but the correlation coefficient varied when **minNumInstance = 10.0:**



=== Summary ===

Correlation coefficient 0.9343

Mean absolute error 1.9765

Root mean squared error 2.7765

Relative absolute error 30.0439 %

Root relative squared error 35.5799 %

Total Number of Instances 397

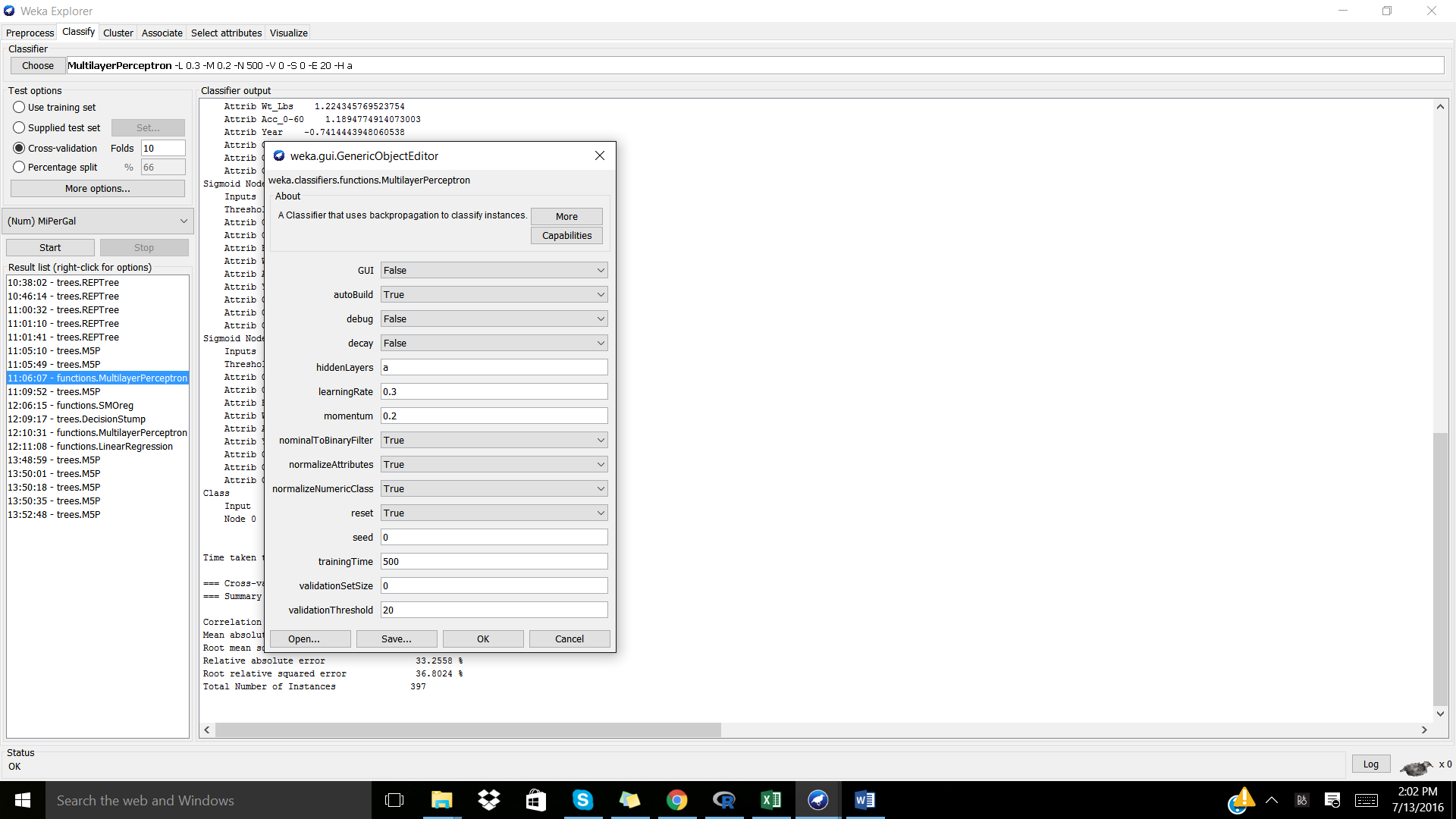
Performance matrix when parameter tuning to M5P classifier

|  |  |  |  |
| --- | --- | --- | --- |
| minNumInstances | useUnsmoothed | Correlation coefficient | Mean absolute error |
| 4.0 | False | 0.9331 | 2.0088 |
| 4.0 | True | 0.9289 | 2.1014 |
| 10.0 | False | 0.9343 | 1.9765 |

The highest correlation coefficient for the M5P classifier was achieved when minNumInstances = 10.0 and useUnsmoothed was set to False. Although the correlation coefficients were comparable, this M5P classifier had the lowest mean absolute error as well.

**MultiLayerPerceptron**

**At default**



=== Summary ===

Correlation coefficient 0.9032

Mean absolute error 2.5442

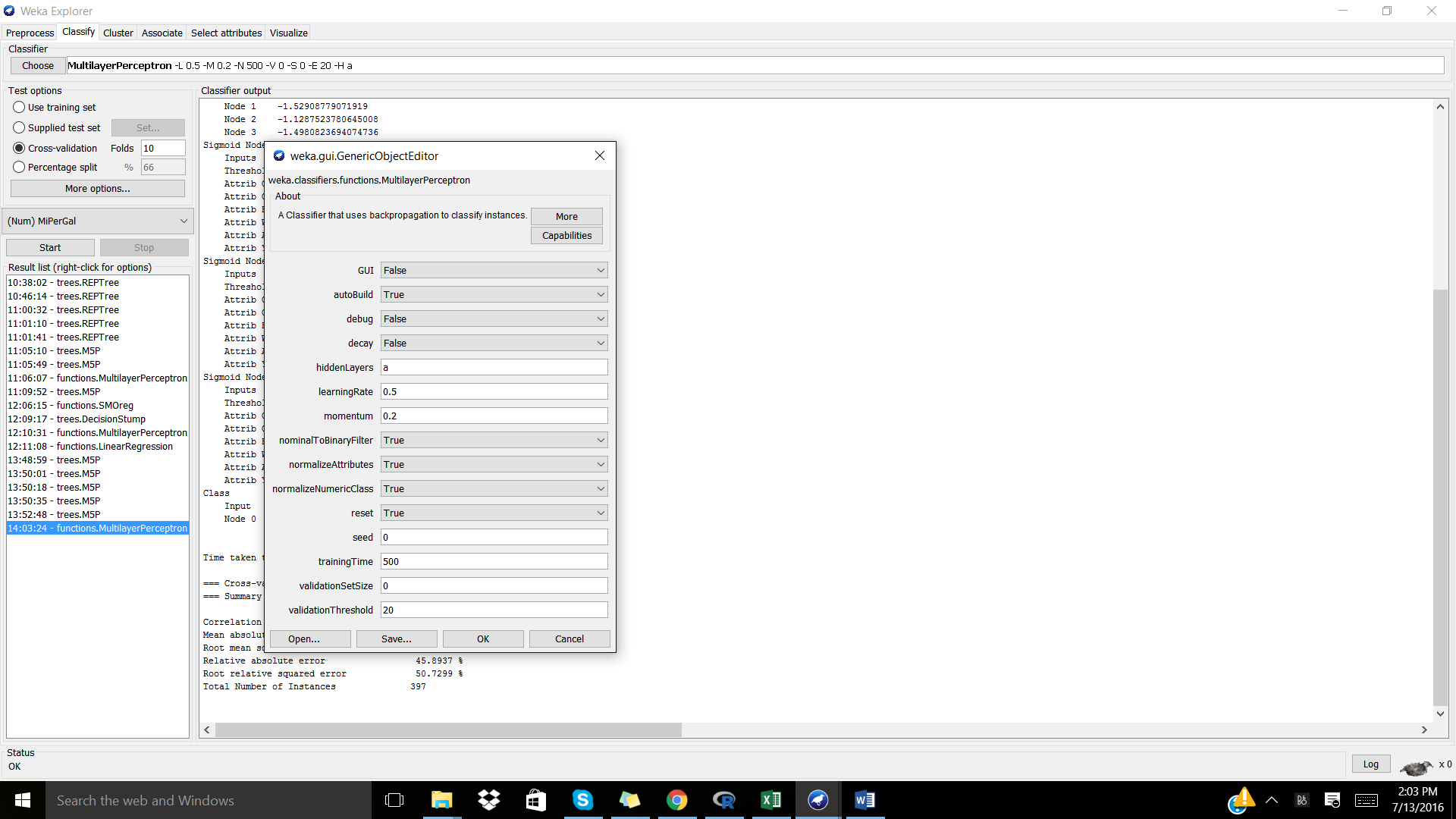
Root mean squared error 3.3694

Relative absolute error 38.6747 %

Root relative squared error 43.1782 %

Total Number of Instances 397

**Increasing learningRate from 0.3 to 0.5:**



=== Summary ===

Correlation coefficient 0.868

Mean absolute error 3.0191

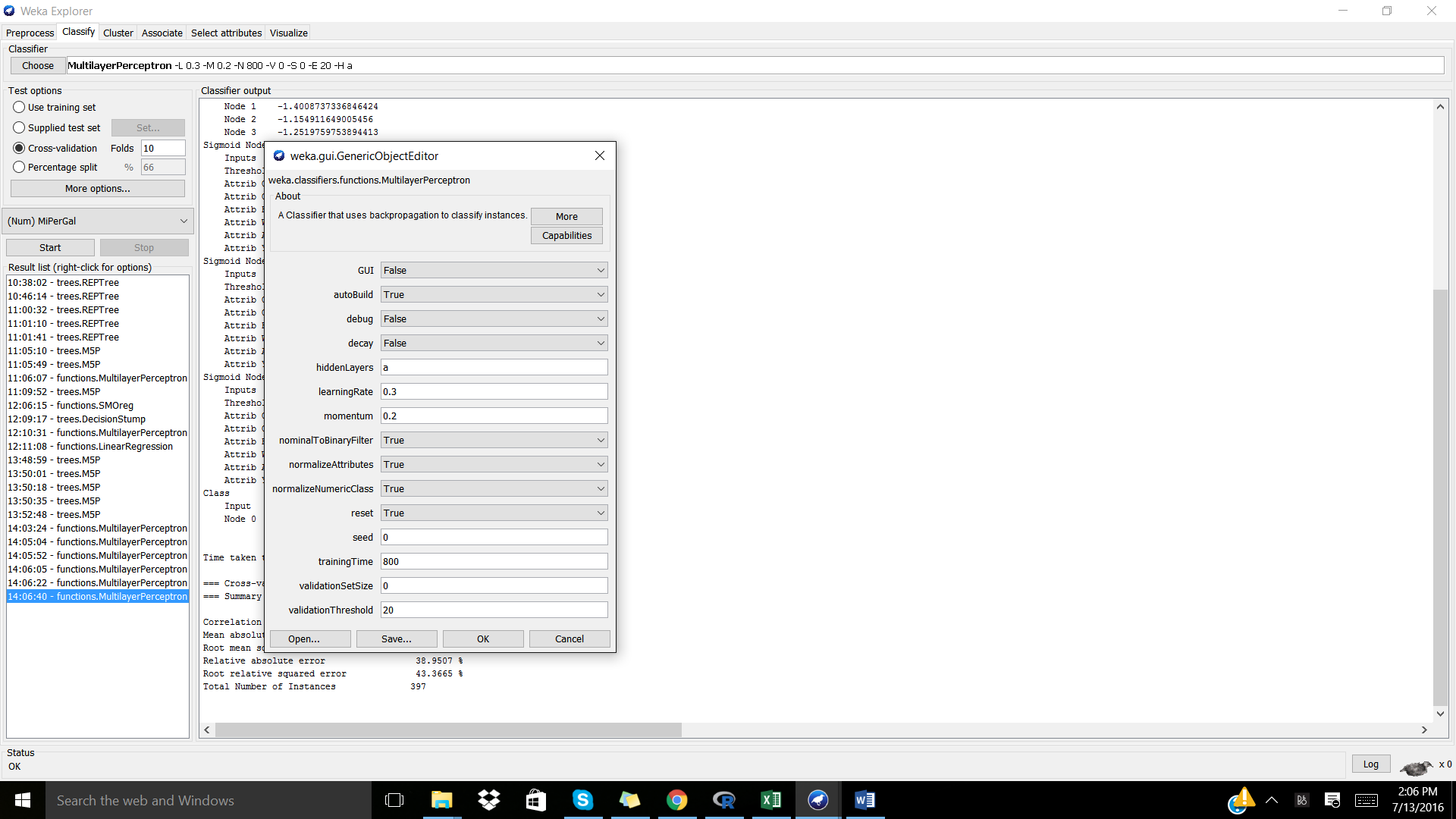
Root mean squared error 3.9587

Relative absolute error 45.8937 %

Root relative squared error 50.7299 %

Total Number of Instances 397

**Increasing trainingTime from 500 to 800:**



=== Summary ===

Correlation coefficient 0.9023

Mean absolute error 2.5624

Root mean squared error 3.3841

Relative absolute error 38.9507 %

Root relative squared error 43.3665 %

Total Number of Instances 397

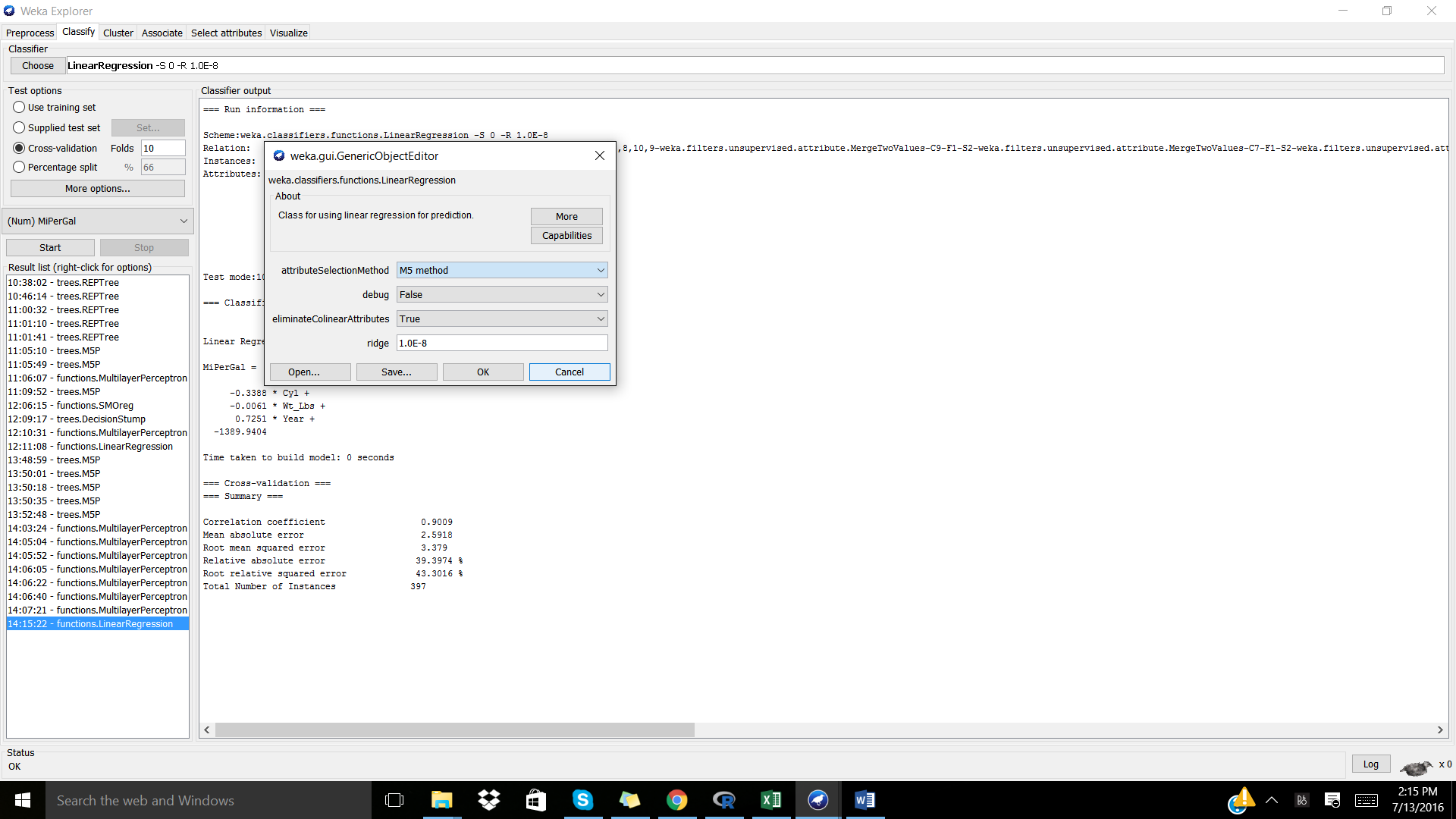
Performance matrix when parameter tuning to MultiLayerPerceptron classifier

|  |  |  |  |
| --- | --- | --- | --- |
| learningRate | trainingTime | Correlation coefficient | Mean absolute error |
| 0.3 | 500 | 0.9032 | 2.5442 |
| 0.5 | 500 | 0.868 | 3.0191 |
| 0.3 | 800 | 0.9023 | 2.5624 |

Evaluation of the correlation coefficient following learningRate and trainingTime tuning showed that the best MultiLayerPerceptron classifier was achieved using the default settings of learningRate = 0.3 and trainingTime = 500.

**LinearRegression**

**At default**



=== Summary ===

Correlation coefficient 0.9009

Mean absolute error 2.5918

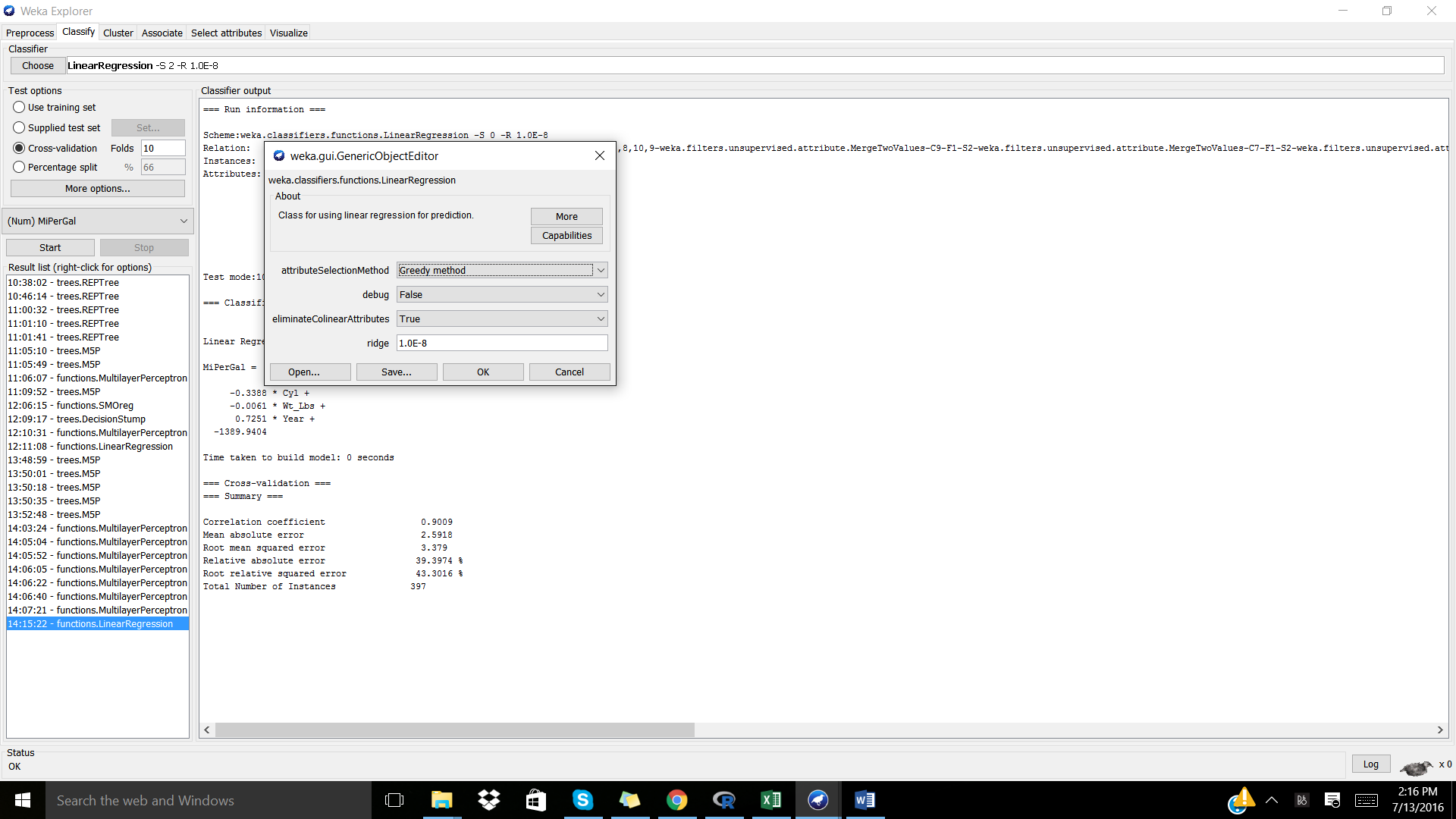
Root mean squared error 3.379

Relative absolute error 39.3974 %

Root relative squared error 43.3016 %

Total Number of Instances 397

**attributeSelectionMethod: Greedy Method**



=== Summary ===

Correlation coefficient 0.9009

Mean absolute error 2.5918

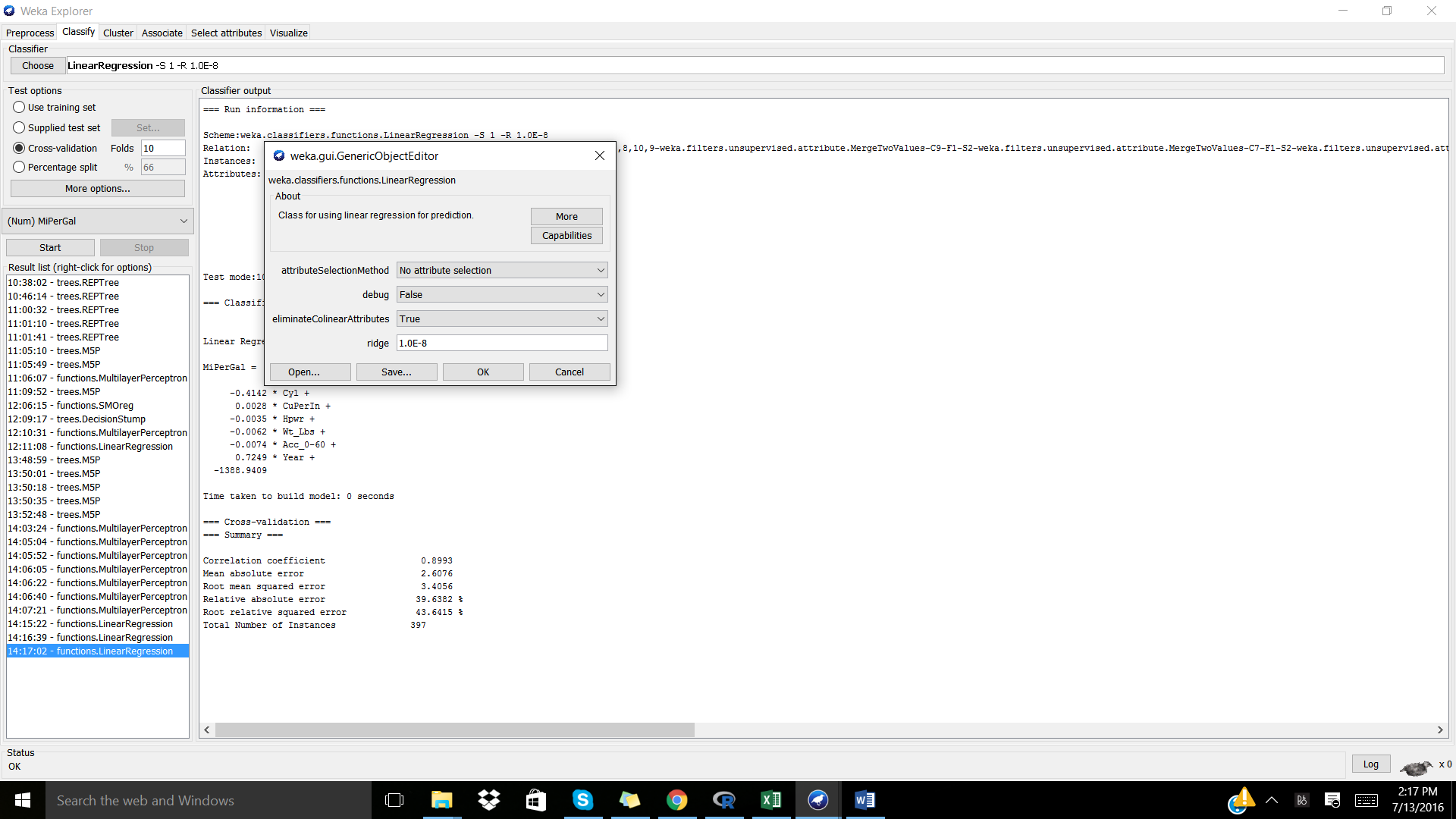
Root mean squared error 3.379

Relative absolute error 39.3974 %

Root relative squared error 43.3016 %

Total Number of Instances 397

**attributeSelectionMethod: No attribute selection**



=== Summary ===

Correlation coefficient 0.8993

Mean absolute error 2.6076

Root mean squared error 3.4056

Relative absolute error 39.6382 %

Root relative squared error 43.6415 %

Total Number of Instances 397

Performance matrix when parameter tuning to LinearRegression classifier

|  |  |  |
| --- | --- | --- |
| attributeSelectionMethod | Correlation coefficient | Mean absolute error |
| M5 method | 0.9009 | 2.5918 |
| Greedy method | 0.9009 | 2.5918 |
| No attribute selection | 0.8993 | 2.6076 |

Varying the attributeSelectionMethod from M5 to Greedy did not affect the performance of the model, thus both modifying the LinearRegression classifier to use M5 or Greedy method selection produces strong models.

**Conclusion**

Following comparisons of all of the models produced above, the M5P classifier when minNumInstances = 10.0 and useUnsmoothed was set to False was with best classifier with a Correlation coefficient = 0.9343 and a Mean absolute error = 1.9765.

**Full Model - M5P: minNumInstances = 10.0 and useUnsmoothed: False**

=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 10.0

Relation: CARS1-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,10,9-weka.filters.unsupervised.attribute.MergeTwoValues-C9-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F1-S2-weka.filters.unsupervised.attribute.MergeTwoValues-C7-F27-S29-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-L15,26-H-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,8,7-weka.filters.unsupervised.attribute.Remove-R7

Instances: 397

Attributes: 7

Cyl

CuPerIn

Hpwr

Wt\_Lbs

Acc\_0-60

Year

MiPerGal

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:

(using smoothed linear models)

CuPerIn <= 190.5 : LM1 (217/40.719%)

CuPerIn > 190.5 : LM2 (180/26.89%)

LM num: 1

MiPerGal =

-0.0438 \* CuPerIn

- 0.0582 \* Hpwr

- 0.0054 \* Wt\_Lbs

+ 0.9119 \* Year

- 1752.8

LM num: 2

MiPerGal =

-0.0149 \* Hpwr

- 0.0034 \* Wt\_Lbs

+ 0.4218 \* Year

- 801.8689

Number of Rules : 2

Time taken to build model: 4.69 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9343

Mean absolute error 1.9765

Root mean squared error 2.7765

Relative absolute error 30.0439 %

Root relative squared error 35.5799 %

Total Number of Instances 397

