# ASSIGNMENT 1B

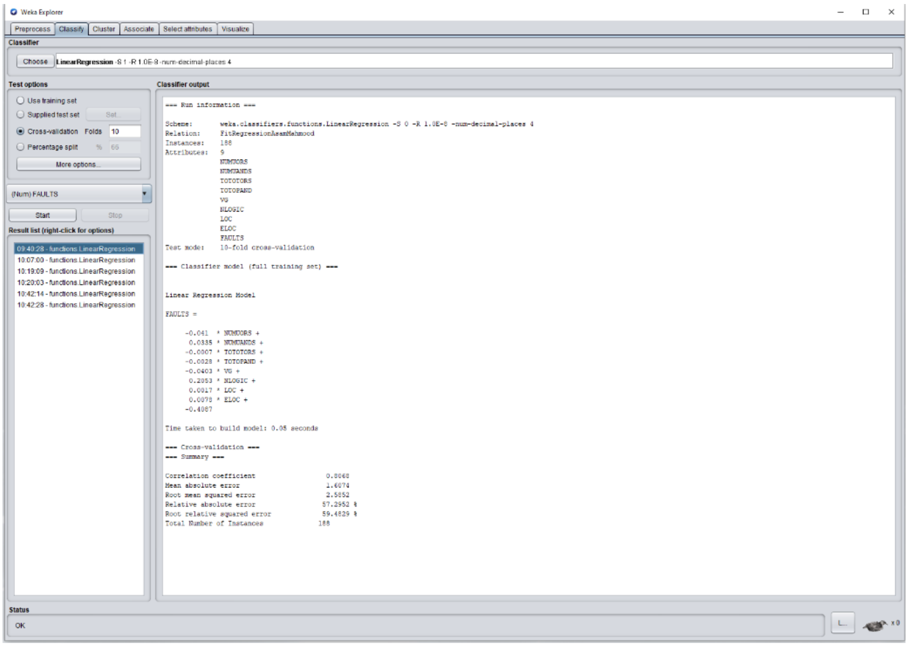
Asam Mahmood

Florida Atlantic University

PROF. TAGHI M KOSHGOFTAR

Course CAP 6673 Data Mining and Machine Learning

***Figure 1 M5 method***



Choose **LinearRegression** -S 1 -R 1.0E-8 -num-decimal-places 4

Test options:

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

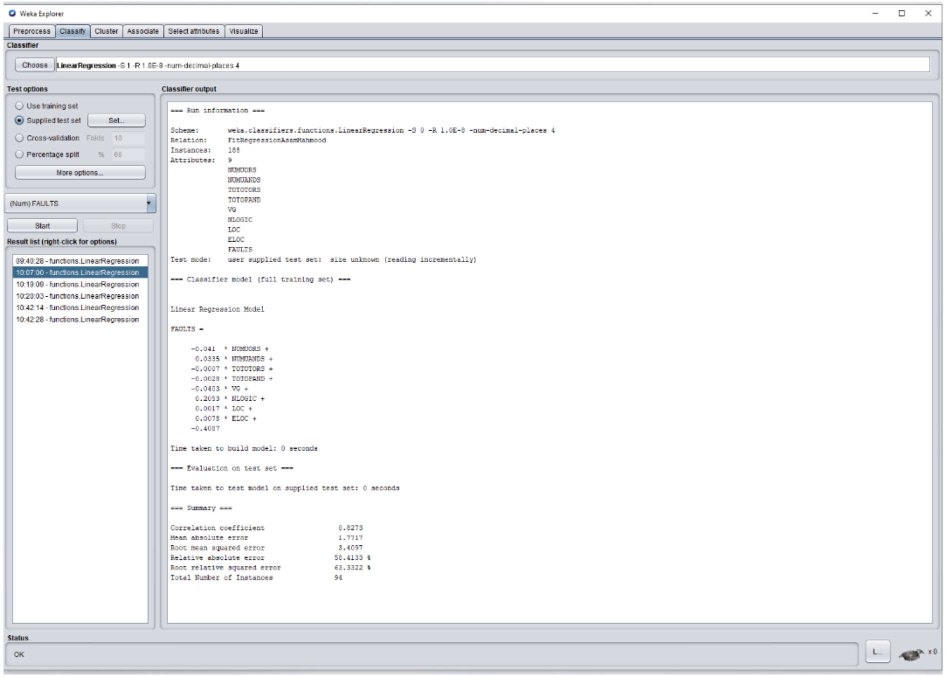
(Num) FAULTS

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    10-fold cross-validation

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


Linear Regression Model

FAULTS =

     -0.041  * NUMUORS +
      0.0335 * NUMUANDS +
     -0.0007 * TOTOTORS +
     -0.0028 * TOTOPAND +
     -0.0403 * VG +
      0.2053 * NLOGIC +
      0.0017 * LOC +
      0.0078 * ELOC +
     -0.4087

Time taken to build model: 0.05 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8068
Mean absolute error                      1.6074
Root mean squared error                  2.5852
Relative absolute error                 57.2952 %
Root relative squared error             59.4829 %
Total Number of Instances              188
```

Status: OK

***Figure 2 M5 method test***



Choose **LinearRegression** -S 1 -R 1.0E-8 -num-decimal-places 4

Test options:

○ Use training set

● Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

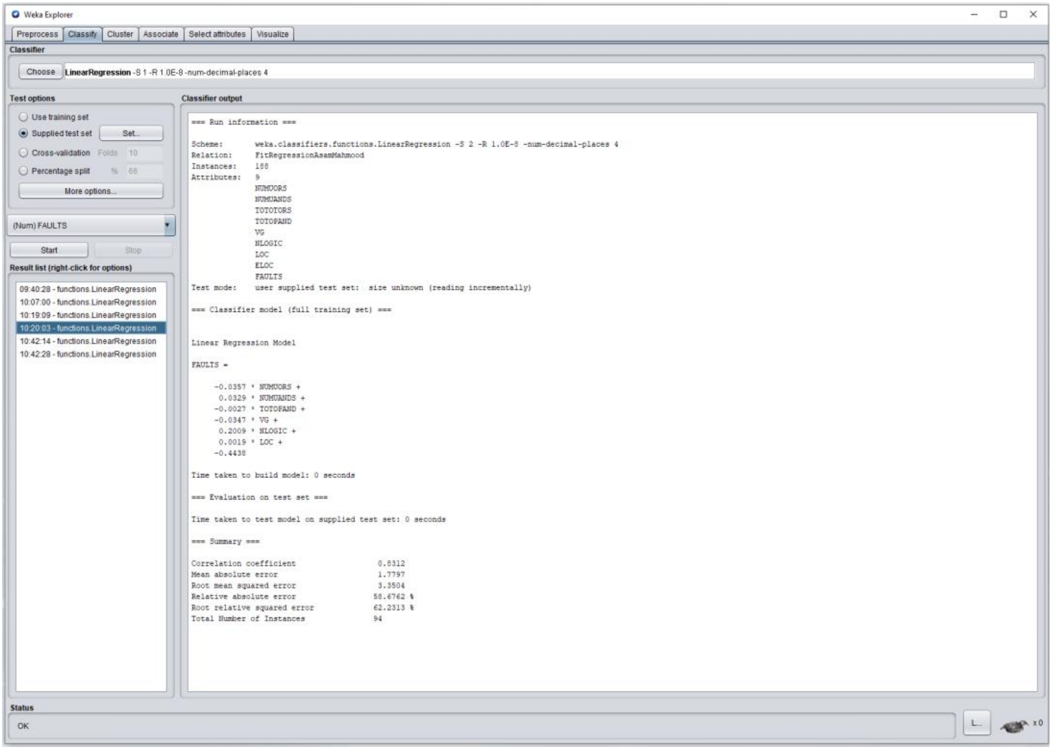
(Num) FAULTS

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    user supplied test set:  size unknown (reading incrementally)

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


Linear Regression Model

FAULTS =

     -0.041  * NUMUORS +
      0.0335 * NUMUANDS +
     -0.0007 * TOTOTORS +
     -0.0028 * TOTOPAND +
     -0.0403 * VG +
      0.2053 * NLOGIC +
      0.0017 * LOC +
      0.0078 * ELOC +
     -0.4087

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient                  0.8273
Mean absolute error                      1.7717
Root mean squared error                  3.4097
Relative absolute error                 58.4133 %
Root relative squared error             63.3322 %
Total Number of Instances               94
```

Status: OK

## ***Figure 3 Greedy Test***

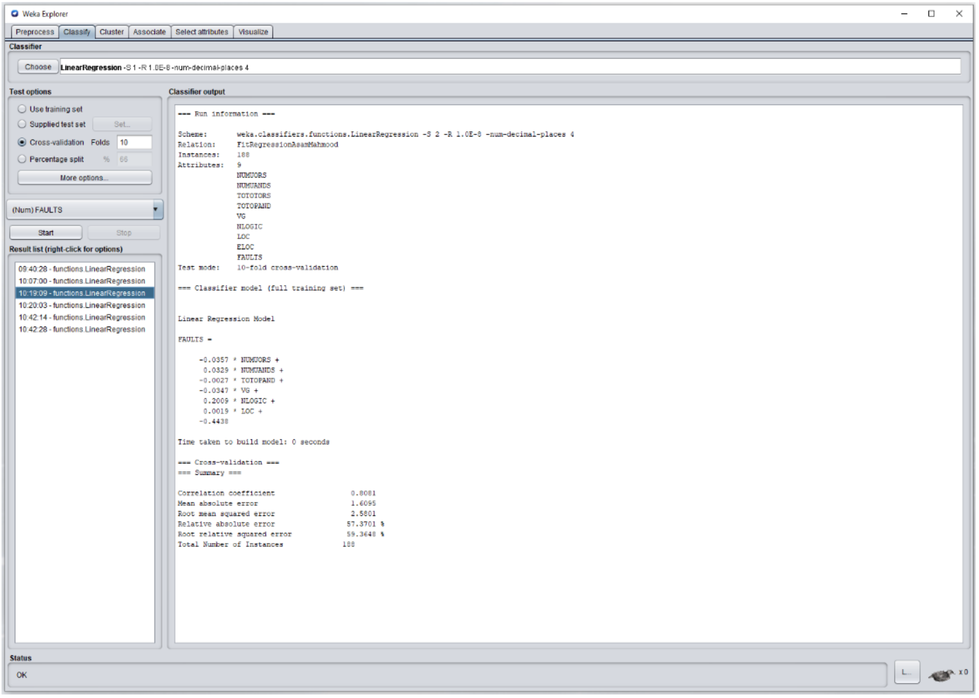


LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    user supplied test set:  size unknown (reading incrementally)

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


Linear Regression Model

FAULTS =

     -0.0357 * NUMUORS +
      0.0329 * NUMUANDS +
     -0.0027 * TOTOPAND +
     -0.0347 * VG +
      0.2009 * NLOGIC +
      0.0019 * LOC +
     -0.4438

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient                  0.8312
Mean absolute error                      1.7797
Root mean squared error                  3.3504
Relative absolute error                 58.6762 %
Root relative squared error             62.2313 %
Total Number of Instances               94
```

## ***Figure 4 Greedy 10-fold***

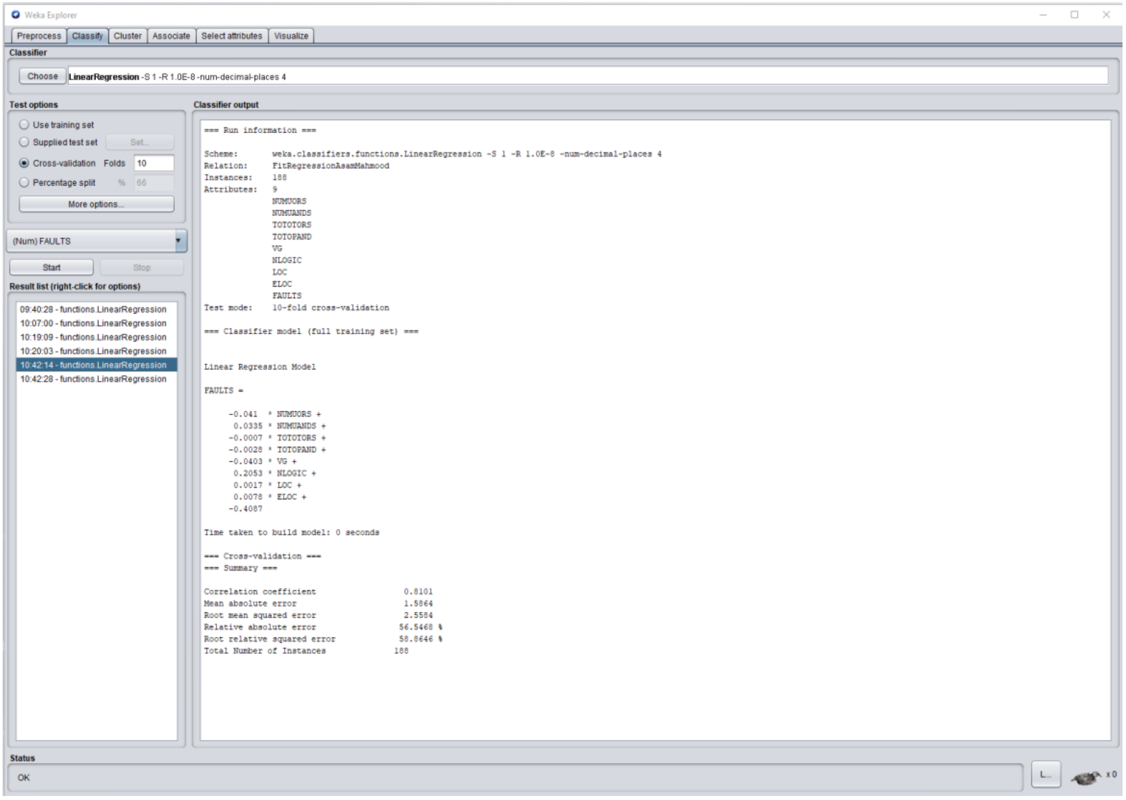


LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    10-fold cross-validation

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


Linear Regression Model

FAULTS =

     -0.0357 * NUMUORS +
      0.0329 * NUMUANDS +
     -0.0027 * TOTOPAND +
     -0.0347 * VG +
      0.2009 * NLOGIC +
      0.0019 * LOC +
     -0.4438

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8081
Mean absolute error                      1.6095
Root mean squared error                  2.5801
Relative absolute error                 57.3701 %
Root relative squared error             59.3648 %
Total Number of Instances              188
```

***<u>Figure 5 No attribute 10 fold</u>***



**LinearRegression** -S 1 -R 1.0E-8 -num-decimal-places 4

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

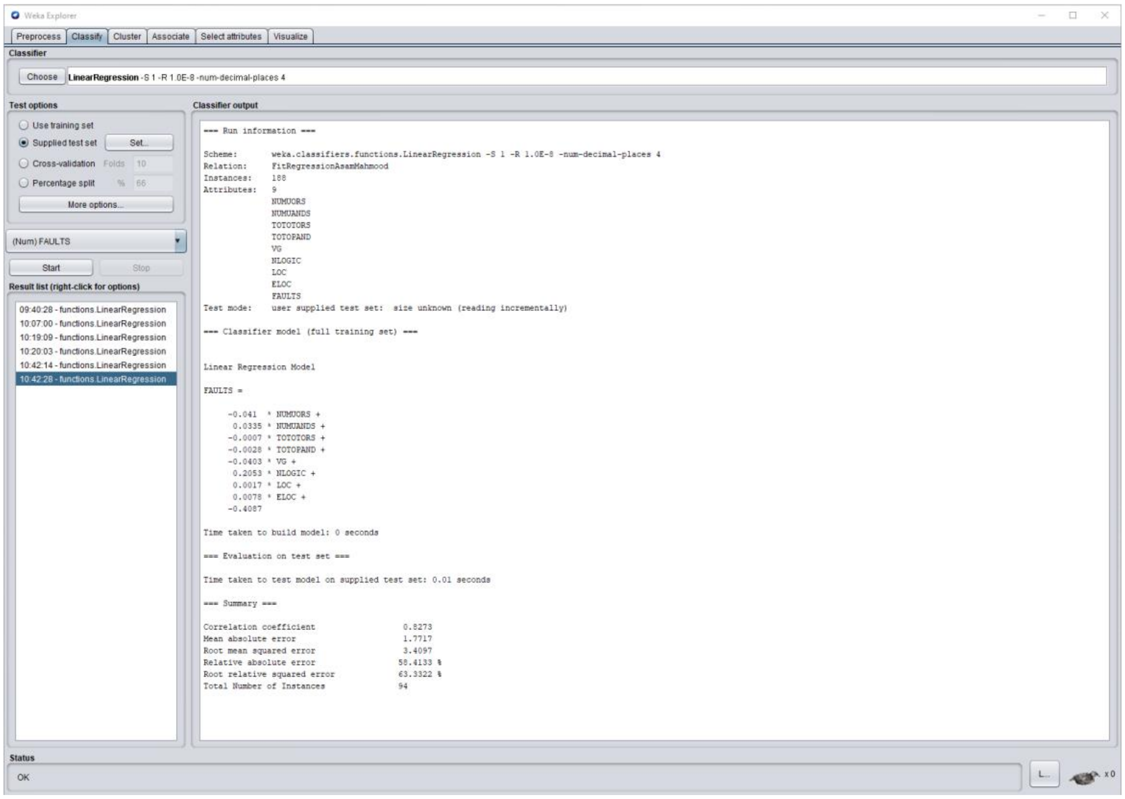
(Num) FAULTS

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    10-fold cross-validation

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


Linear Regression Model

FAULTS =

     -0.041  * NUMUORS +
      0.0335 * NUMUANDS +
     -0.0007 * TOTOTORS +
     -0.0028 * TOTOPAND +
     -0.0403 * VG +
      0.2053 * NLOGIC +
      0.0017 * LOC +
      0.0078 * ELOC +
     -0.4087

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8101
Mean absolute error                      1.5864
Root mean squared error                  2.5584
Relative absolute error                 56.5468 %
Root relative squared error             58.8646 %
Total Number of Instances              188
```

OK

***<u>Figure 6 No Attributed Selected Test</u>***



**LinearRegression** -S 1 -R 1.0E-8 -num-decimal-places 4

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

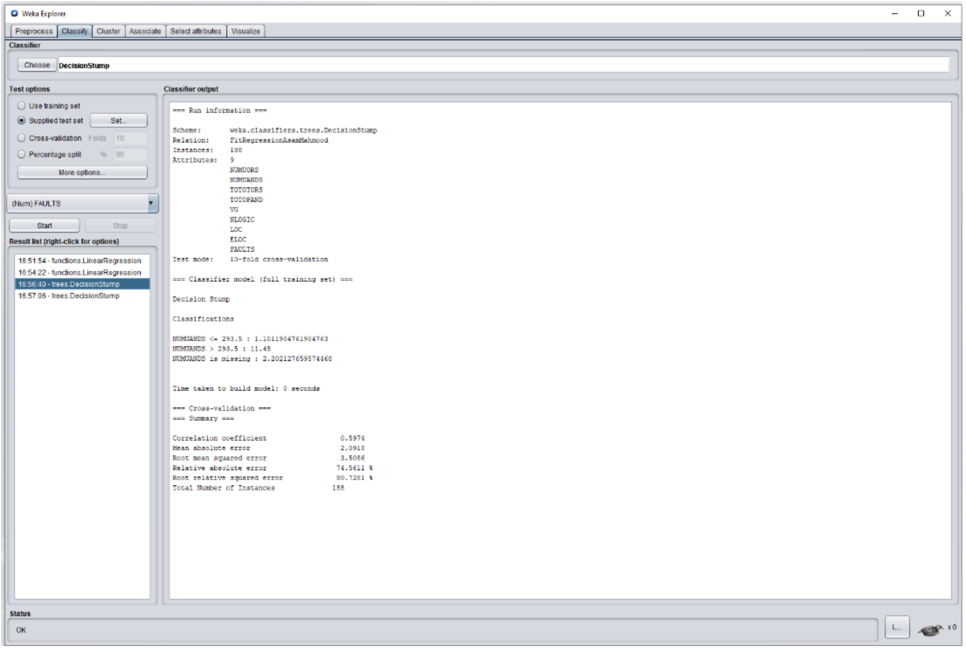
(Num) FAULTS

```
09:40:28 - functions.LinearRegression
10:07:00 - functions.LinearRegression
10:19:09 - functions.LinearRegression
10:20:03 - functions.LinearRegression
10:42:14 - functions.LinearRegression
10:42:28 - functions.LinearRegression
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    user supplied test set:  size unknown (reading incrementally)

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


Linear Regression Model

FAULTS =

     -0.041  * NUMUORS +
      0.0335 * NUMUANDS +
     -0.0007 * TOTOTORS +
     -0.0028 * TOTOPAND +
     -0.0403 * VG +
      0.2053 * NLOGIC +
      0.0017 * LOC +
      0.0078 * ELOC +
     -0.4087

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correlation coefficient                  0.8273
Mean absolute error                      1.7717
Root mean squared error                  3.4097
Relative absolute error                 58.4133 %
Root relative squared error             63.3322 %
Total Number of Instances               94
```

OK

## ***Figure 7 Decision Stump***



Classifier: DecisionStump

Test options
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) FAULTS

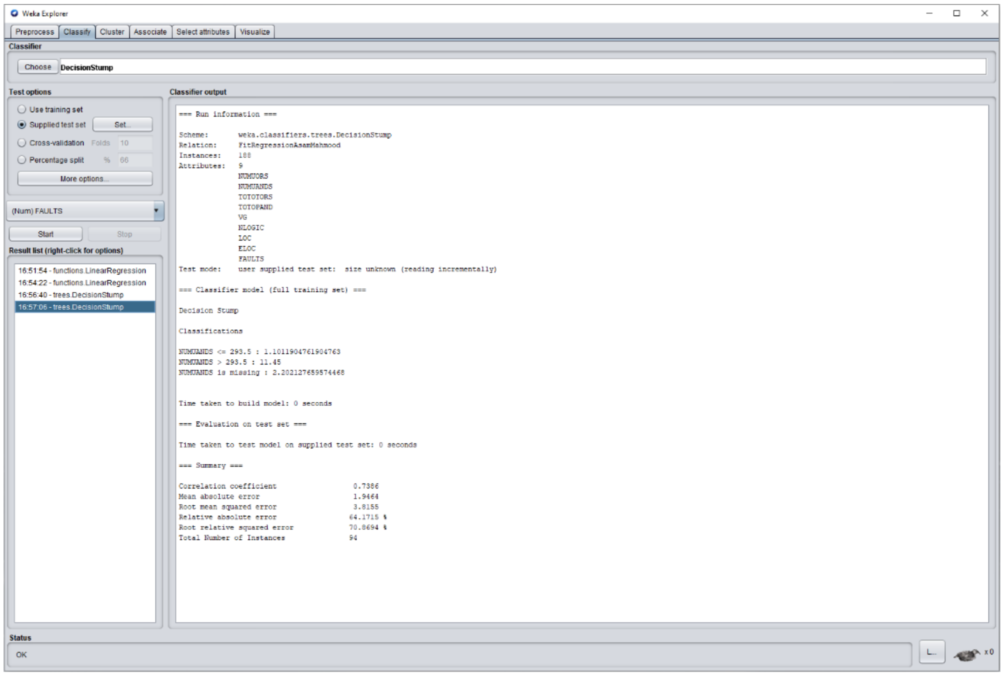
Result list (right-click for options)
16:51:54 - functions.LinearRegression
16:54:22 - functions.LinearRegression
16:56:40 - trees.DecisionStump
16:57:06 - trees.DecisionStump

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.DecisionStump
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    10-fold cross-validation

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

Decision Stump

Classifications

NUMUANDS <= 293.5 : 1.1011904761904763
NUMUANDS > 293.5 : 11.45
NUMUANDS is missing : 2.202127659574468


Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5974
Mean absolute error                      2.0918
Root mean squared error                  3.5086
Relative absolute error                 74.5611 %
Root relative squared error             80.7281 %
Total Number of Instances              188
```

Status
OK

## ***Figure 8 Decision Stump Test***



Classifier: DecisionStump

Test options
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) FAULTS

Result list (right-click for options)
16:51:54 - functions.LinearRegression
16:54:22 - functions.LinearRegression
16:56:40 - trees.DecisionStump
16:57:06 - trees.DecisionStump

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.DecisionStump
Relation:     FitRegressionAsamMahmood
Instances:    188
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPAND
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS
Test mode:    user supplied test set:  size unknown (reading incrementally)

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

Decision Stump

Classifications

NUMUANDS <= 293.5 : 1.1011904761904763
NUMUANDS > 293.5 : 11.45
NUMUANDS is missing : 2.202127659574468


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient                  0.7386
Mean absolute error                      1.9464
Root mean squared error                  3.8155
Relative absolute error                 64.1715 %
Root relative squared error             70.8694 %
Total Number of Instances               94
```

Status
OK

| *FIT* | *Greedy* | *M5* | *No Attribute* | *Decision Stump* |
|---|---|---|---|---|
| *Correlation coefficient* | 0.8081 | 0.8068 | 0.8101 | 0.5974 |
| *Mean absolute error* | 1.6095 | 1.6074 | 1.5864 | 2.0918 |
| *Root mean squared error* | 2.5801 | 2.5852 | 2.5584 | 3.5086 |
| *Relative absolute error* | 57.3701 % | 57.2952 % | 56.5468 % | 74.5611 % |
| *Root relative sqrd error* | 59.3648 % | 59.4829 % | 58.8646 % | 80.7281 % |
| *Instances* | 188 | 188 | 188 | 188 |
| *Linear Regression Model* | FAULTS =<br>-0.0357 * NUMUORS +<br>0.0329 * NUMUANDS +<br>-0.0027 * TOTOPAND +<br>-0.0347 * VG +<br>0.2009 * NLOGIC +<br>0.0019 * LOC +<br>-0.4438 | FAULTS =<br>-0.041 * NUMUORS +<br>0.0335 * NUMUANDS +<br>-0.0007 * TOTOTORS +<br>-0.0028 * TOTOPAND +<br>-0.0403 * VG +<br>0.2053 * NLOGIC +<br>0.0017 * LOC +<br>0.0078 * ELOC +<br>-0.4087 | Faults =<br>-0.041 * NUMUORS +<br>0.0335 * NUMUANDS +<br>-0.0007 * TOTOTORS +<br>-0.0028 * TOTOPAND +<br>-0.0403 * VG +<br>0.2053 * NLOGIC +<br>0.0017 * LOC +<br>0.0078 * ELOC +<br>-0.4087 | Classifications<br><br>NUMUANDS <= 293.5 :<br>1.1011904761904763<br>NUMUANDS > 293.5 : 11.45<br>NUMUANDS is missing :<br>2.202127659574468 |

| *Test* | *Greedy - Test* | *M5 - Test* | *No Attribute - Test* | *Decision Stump* |
|---|---|---|---|---|
| *Correlation coefficient* | 0.8312 | 0.8273 | 0.8273 | 0.7386 |
| *Mean absolute error* | 1.7797 | 1.7717 | 1.7717 | 1.9464 |
| *Root mean squared error* | 3.3504 | 3.4097 | 3.4097 | 3.8155 |
| *Relative absolute error* | 58.6762 % | 58.4133 % | 58.4133 % | 64.1715 % |
| *Root relative sqrd error* | 62.2313 % | 63.3322 % | 63.3322 % | 70.8694 % |
| *Instances* | 94 | 94 | 94 | 94 |
| *Linear Regression Model* | FAULTS =<br>-0.0357 * NUMUORS +<br>0.0329 * NUMUANDS +<br>-0.0027 * TOTOPAND +<br>-0.0347 * VG +<br>0.2009 * NLOGIC +<br>0.0019 * LOC +<br>-0.4438 | FAULTS =<br>-0.041 * NUMUORS +<br>0.0335 * NUMUANDS +<br>-0.0007 * TOTOTORS +<br>-0.0028 * TOTOPAND +<br>-0.0403 * VG +<br>0.2053 * NLOGIC +<br>0.0017 * LOC +<br>0.0078 * ELOC +<br>-0.4087 | Faults =<br>-0.041 * NUMUORS +<br>0.0335 * NUMUANDS +<br>-0.0007 * TOTOTORS +<br>-0.0028 * TOTOPAND +<br>-0.0403 * VG +<br>0.2053 * NLOGIC +<br>0.0017 * LOC +<br>0.0078 * ELOC +<br>-0.4087 | Classifications<br><br>NUMUANDS <= 293.5 :<br>1.1011904761904763<br>NUMUANDS > 293.5 : 11.45<br>NUMUANDS is missing :<br>2.202127659574468 |

Analysis:
The first table "Fit" has all performance metrics for the 4 models. It has the linear model output as well as the classification model output in the last row. We used 10-fold cross per instructions as well.

The second table "Test" has all performance metrics using the model trained from the FIT data set given to us from the professor and we used that for data prediction on test data set. Also, can be seen in screenshots above of weka outputs.

- The first column labeled Greedy is using The Greedy attribute selection which you can see uses 6 attributes from the data to predict the fault. Initially the correlation coefficient predicted value of linear regression model with "no-attribute" based on 10-fold cross did better than Greedy. Until after using test data set for evaluation you can see Greedy method was the best performer out of all of them. I figure it would be good to note that it omitted ELOC and TOTOTORS from linear aggression model output. Also, the root relative squared was smaller than the other methods.

- The second column with Label M5 has identical output with "no attribute" after test evaluation. But initially performed better than decision stump and performed lesser than "no attribute" by .0033 (.8101-.8068) and lesser than "greedy" by .0013 (.8081-.8068). Which are not huge performance differences at all. After test evaluation "M5" and "no-attribute" had identical performance metrics. I validated that its truly the same because they have the same weights generated.M5 performed less than Greedy and more than decision stump. Also 8 of the attributes were selected in this method.

- The third column "no attribute" was initial best performer before test runs. It also had a lower MAE than the rest also. After test evaluation it performed identically with M5 in performance metrics and less then Greedy(.8273 -"No-Attribute" < .8312-"Greedy") and better than decision stump(.7386 - Decision-stump < .8273 -"No-Attribute"). When I mean better the correlation coefficient was less than Greedy and better than decision stump. Also 8 of the attributes were selected in this method as well can be seen on both tables.

- The Fourth Column Decision Stump did worse than all other linear regression models. It only used one attribute NUMUANDS. I was able to translate the classifier model for decision stump after some testing and validation. It is saying if NUMUADS <= 293.5 than fault will be 1.1011904761904763. If NUMUADS > 293.5 than predicted fault will be 2.202127659574468. If it is missing than predicted value will 2.202127659574468. But by using one value it explains why it has a higher MAE. I also validated that before test runs and after it was the worst performer correlation Coefficients both were the worst compared to rest of linear regression runs

Inconclusion Greedy performed the best(by small difference) and the worst performance was the Decision stump after test evaluation. Linear regression is the best for simple numeric modeling. The metrics helped me answer how valid is the model we built with different types of regression runs. The metrics helped me also understand how to improve the model accordingly with testing if needed. Also, which one is more accurate when comparing errors and coefficient value with unseen test data. I also gained a better understanding how to validate with test data, compared metrics and learned how to translate performance metrics to determine how to properly train models.