Assignment - 5

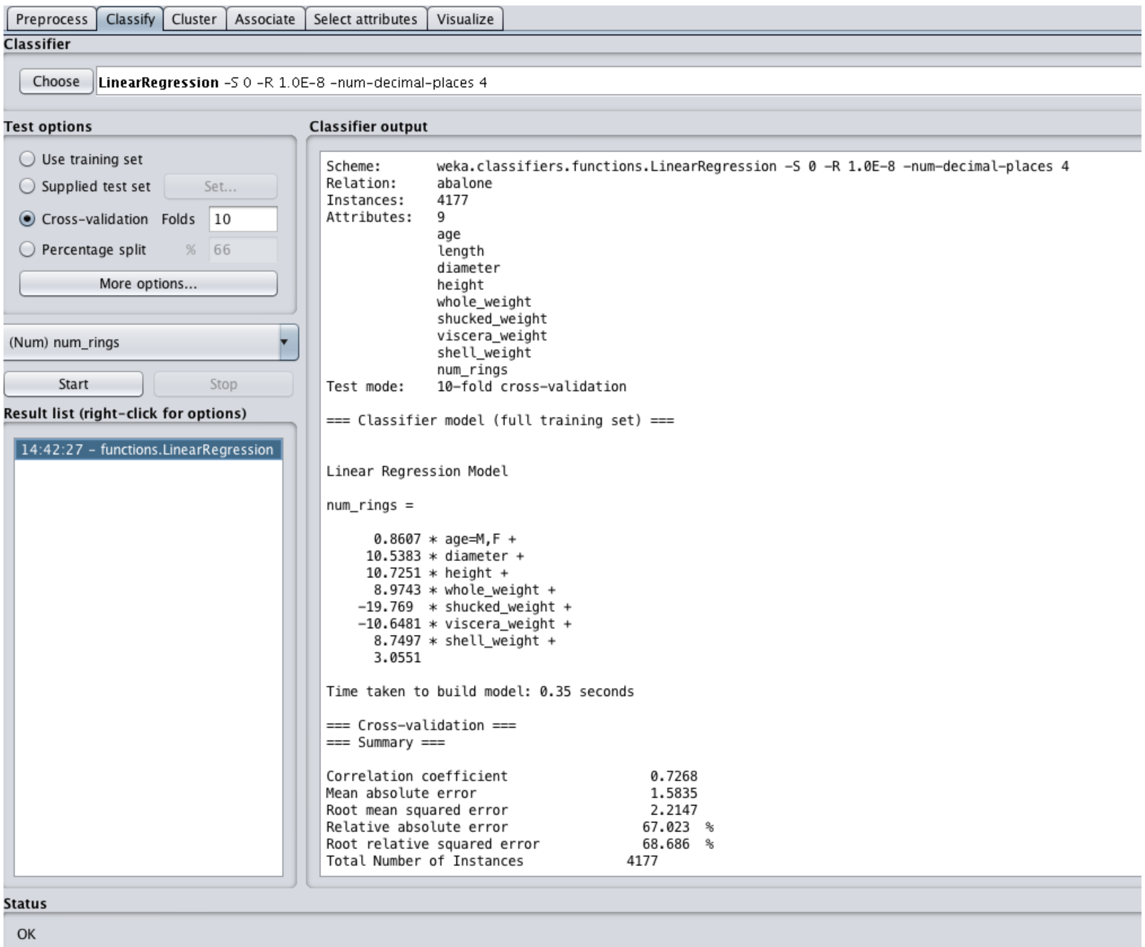
1. **WEKA** (Linear regression is done on the given complete data set.)

Steps:

- shells.arff file is selected. preprocess->open file->shells.arff
- Select Linear regression on all the attributes. classify—>choose—>classifiers—>functions
- —>Linear regression. Cross validation is taken as 10 folds, which is default value.

Mean Absolute Error :
The mean absolute error is the average over verification sample of absolute values of the differences between the selected attribute and its respective corresponding attribute.

Here the Mean absolute Error is 1.5835



LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

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

(Num) num_rings

14:42:27 - functions.LinearRegression

```
Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     abalone
Instances:    4177
Attributes:   9
              age
              length
              diameter
              height
              whole_weight
              shucked_weight
              viscera_weight
              shell_weight
              num_rings
Test mode:    10-fold cross-validation

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


Linear Regression Model

num_rings =

      0.8607 * age=M,F +
     10.5383 * diameter +
     10.7251 * height +
      8.9743 * whole_weight +
    -19.769  * shucked_weight +
    -10.6481 * viscera_weight +
      8.7497 * shell_weight +
      3.0551

Time taken to build model: 0.35 seconds

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

Correlation coefficient                  0.7268
Mean absolute error                      1.5835
Root mean squared error                  2.2147
Relative absolute error                 67.023  %
Root relative squared error             68.686  %
Total Number of Instances             4177
```

OK

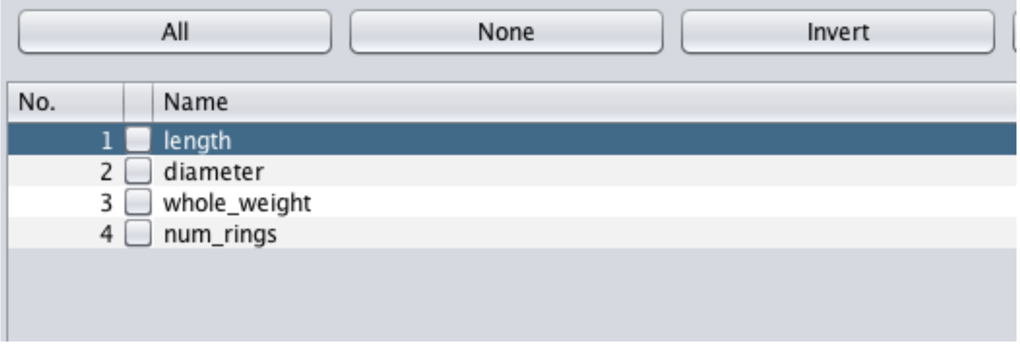
Equation :

num_rings =
(0.86070.8607 * age=M,F) + (10.5383 * diameter) +(10.7251 * height)+ (8.9743 * whole_weight)+ (-19.769 * shucked_weight) + (-10.6481 * viscera_weight) + (8.7497 * shell_weight) + (3.0551)

Here the equation is in the form of y= (c1*x1 + c2 *x2 + ….) where y=num_rings and c1, c2 ,…. are the respective coefficients of the attributes which are x1(age) , x2(diameter) and so on.

Finding the equation using length, diameter, whole_weight, num_rings :

The other parameters are removed from the attribute list after loading the file.

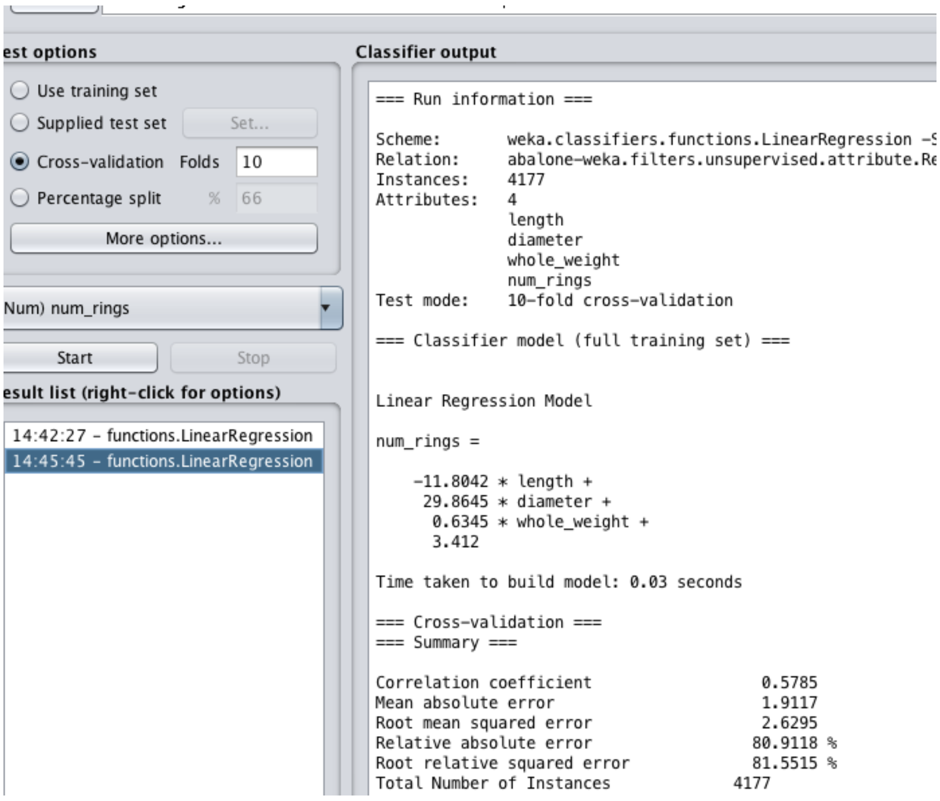


No.		Name
1	□	length
2	□	diameter
3	□	whole_weight
4	□	num_rings

Linear regression is run again.

The equation now becomes

num_rings = (-11.8042 * length) + (29.8645 * diameter) + (0.6345 * whole_weight) + 3.412



est options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

Num) num_rings

Start Stop

esult list (right-click for options)

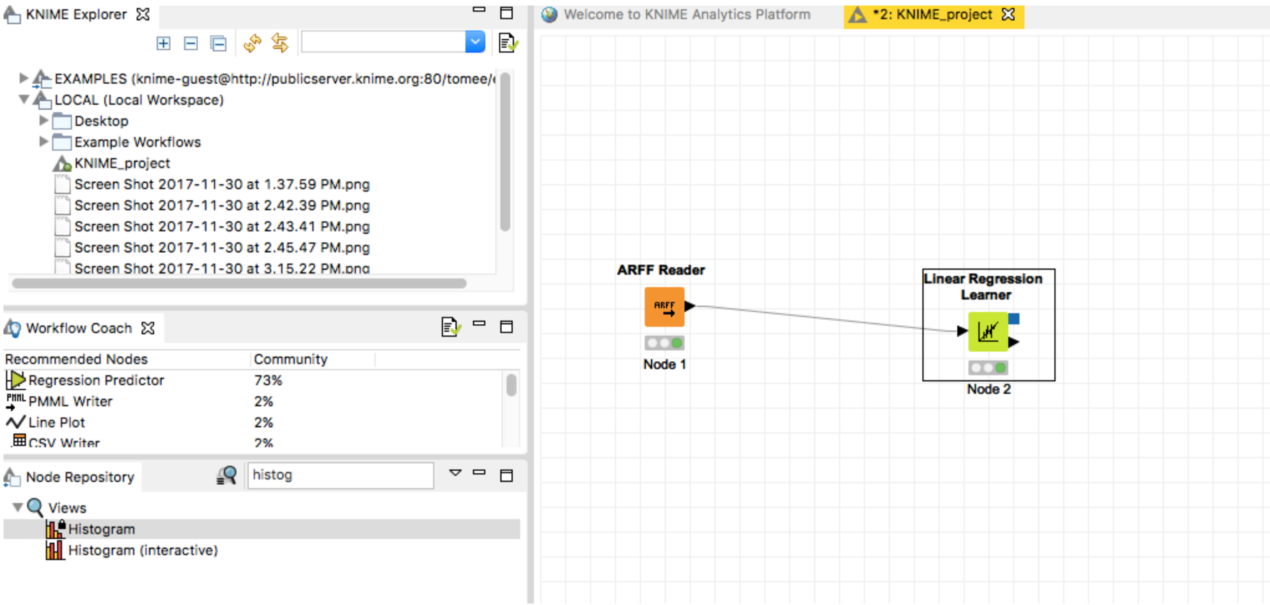
14:42:27 - functions.LinearRegression
14:45:45 - functions.LinearRegression

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S
Relation:     abalone-weka.filters.unsupervised.attribute.Re
Instances:    4177
Attributes:   4
              length
              diameter
              whole_weight
              num_rings
Test mode:    10-fold cross-validation

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


Linear Regression Model

num_rings =

    -11.8042 * length +
     29.8645 * diameter +
      0.6345 * whole_weight +
      3.412

Time taken to build model: 0.03 seconds

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

Correlation coefficient                  0.5785
Mean absolute error                      1.9117
Root mean squared error                  2.6295
Relative absolute error                 80.9118 %
Root relative squared error             81.5515 %
Total Number of Instances             4177
```

2. **KNIME**

Linear regression is performed in Knime using all the attributes.
Arff reader node and linear regression nodes are used and executed connecting the output of arff reader node to the input of linear regression node.

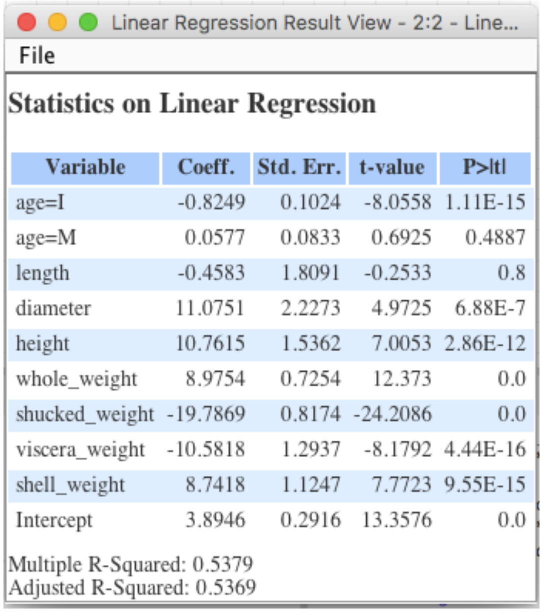


KNIME Explorer
EXAMPLES (knime-guest@http://publicserver.knime.org:80/tomee/
LOCAL (Local Workspace)
Desktop
Example Workflows
KNIME_project
Screen Shot 2017-11-30 at 1.37.59 PM.png
Screen Shot 2017-11-30 at 2.42.39 PM.png
Screen Shot 2017-11-30 at 2.43.41 PM.png
Screen Shot 2017-11-30 at 2.45.47 PM.png
Screen Shot 2017-11-30 at 3.15.22 PM.png
Workflow Coach
Recommended Nodes
Community
Regression Predictor 73%
PMML Writer 2%
Line Plot 2%
CSV Writer 2%
Node Repository
histog
Views
Histogram
Histogram (interactive)
Welcome to KNIME Analytics Platform
*2: KNIME_project
ARFF Reader
Node 1
Linear Regression Learner
Node 2

The equation is

num_rings =
(-0.8249* age=I) + (0.0577* age=M) + (-0.4583 * length) +(11.0751 * diameter) +(10.7615* height)+(8.9754 * whole_weight)+(-19.7869 * shucked_weight) + (-10.5818 * viscera_weight) + (8.7418* shell_weight) + (3.8946)

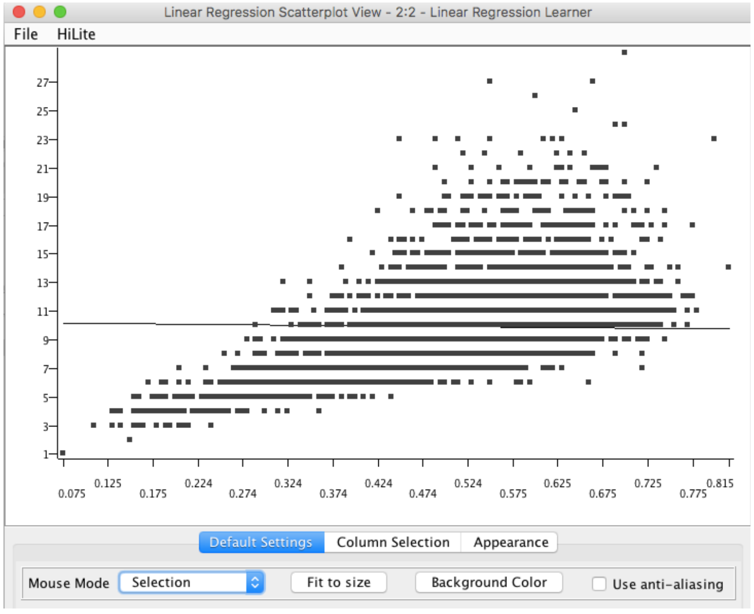
The parameters that have similar coefficients i, e differ by 0.5 almost when compared to weka's co-efficients are height, whole_weight, shucked_weight, viscera_weight, shell_weight.



Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
age=I	-0.8249	0.1024	-8.0558	1.11E-15
age=M	0.0577	0.0833	0.6925	0.4887
length	-0.4583	1.8091	-0.2533	0.8
diameter	11.0751	2.2273	4.9725	6.88E-7
height	10.7615	1.5362	7.0053	2.86E-12
whole_weight	8.9754	0.7254	12.373	0.0
shucked_weight	-19.7869	0.8174	-24.2086	0.0
viscera_weight	-10.5818	1.2937	-8.1792	4.44E-16
shell_weight	8.7418	1.1247	7.7723	9.55E-15
Intercept	3.8946	0.2916	13.3576	0.0

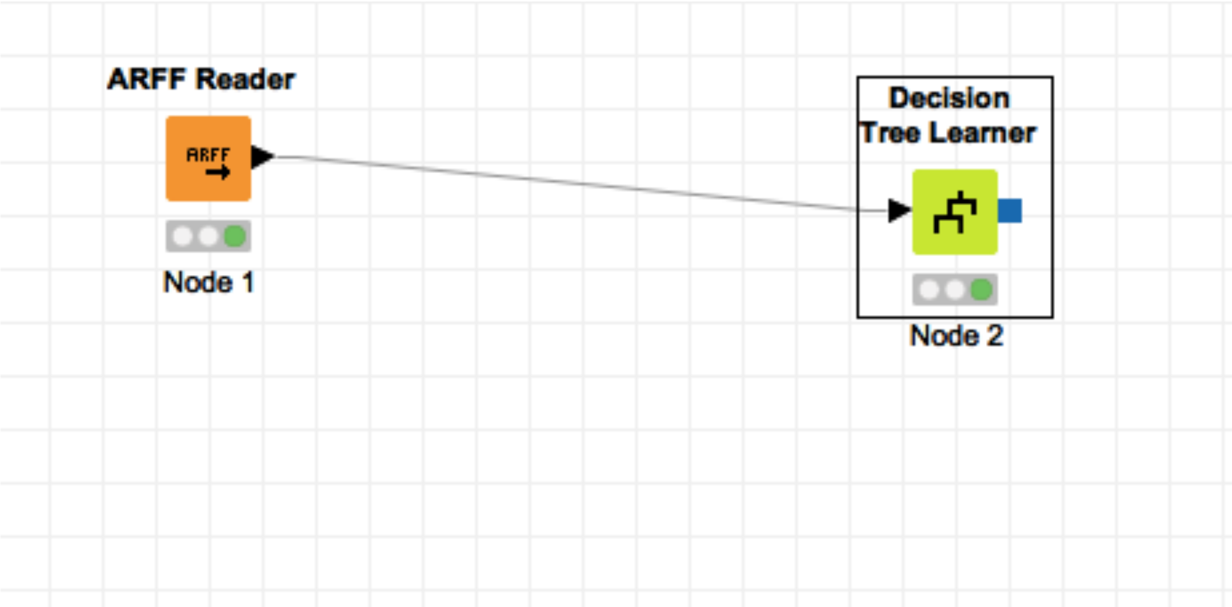
Multiple R-Squared: 0.5379
Adjusted R-Squared: 0.5369

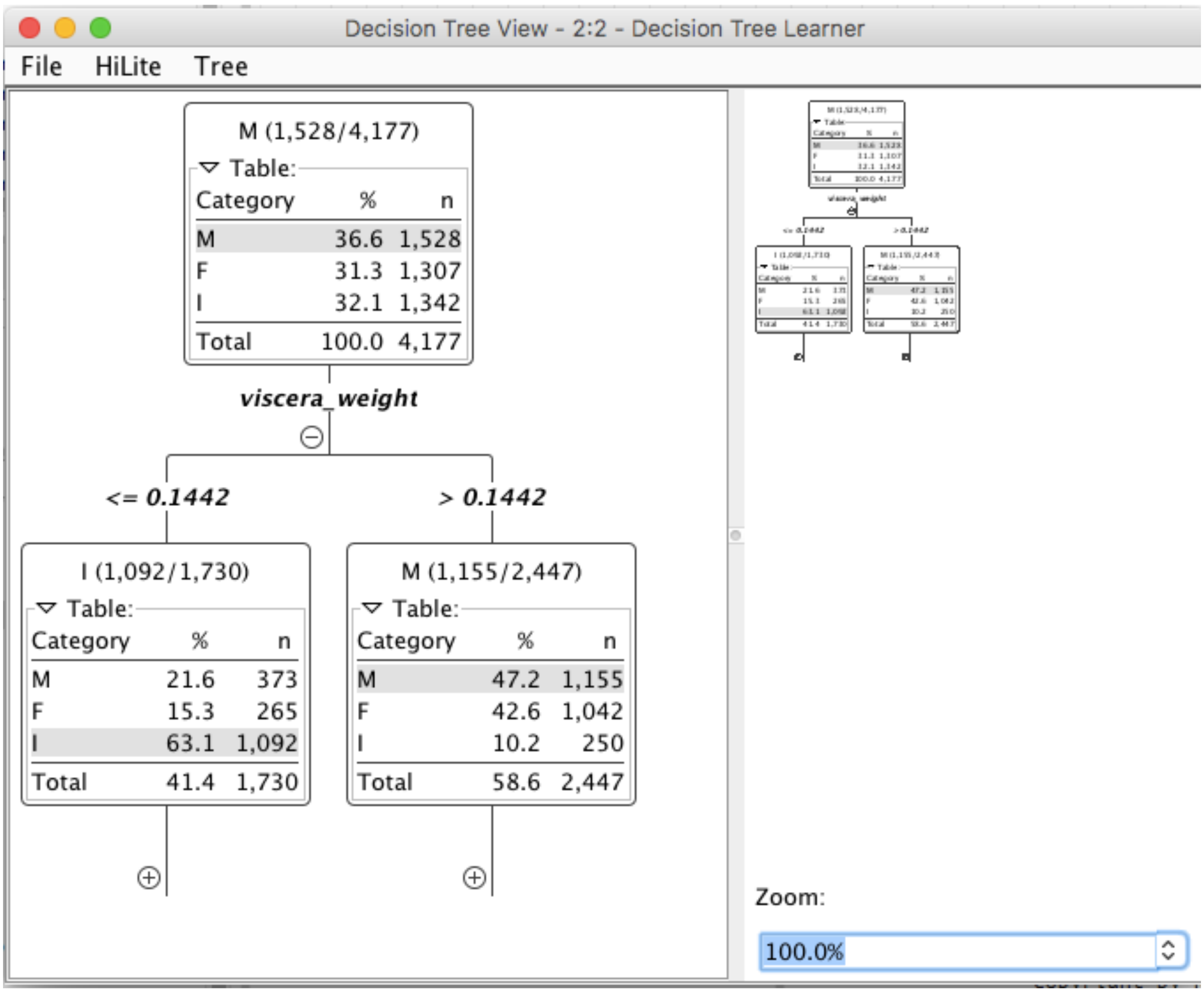


Decision Tree Learner predictor :

Decision tree learner node is connected to the arff reader.

The output of the decision tree Learner is viewed by Decision tree learner node right click—> View : Decision Tree view.





M (1,528/4,177)

Table:

Category	%	n
M	36.6	1,528
F	31.3	1,307
I	32.1	1,342
Total	100.0	4,177

I (1,092/1,730)

Table:

Category	%	n
M	21.6	373
F	15.3	265
I	63.1	1,092
Total	41.4	1,730

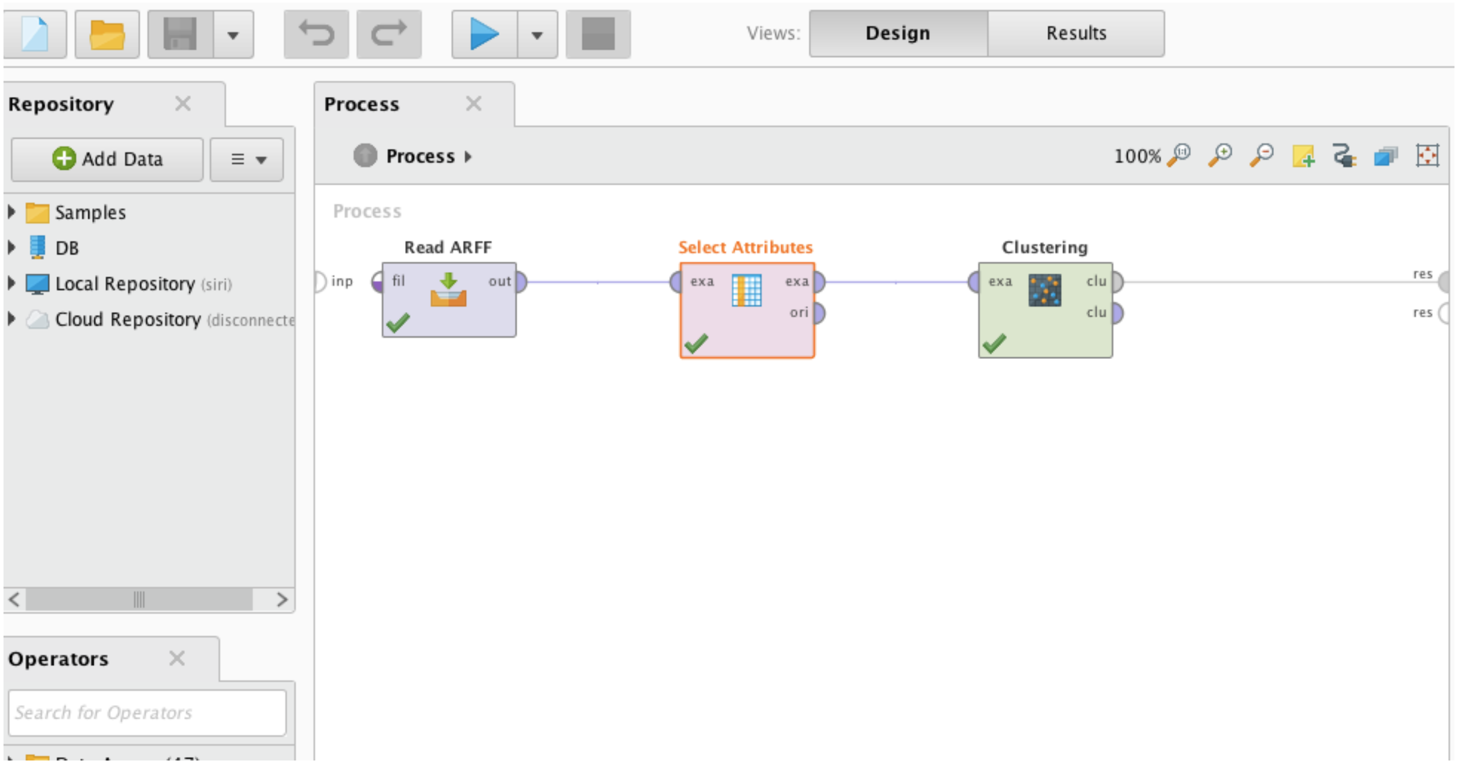
M (1,155/2,447)

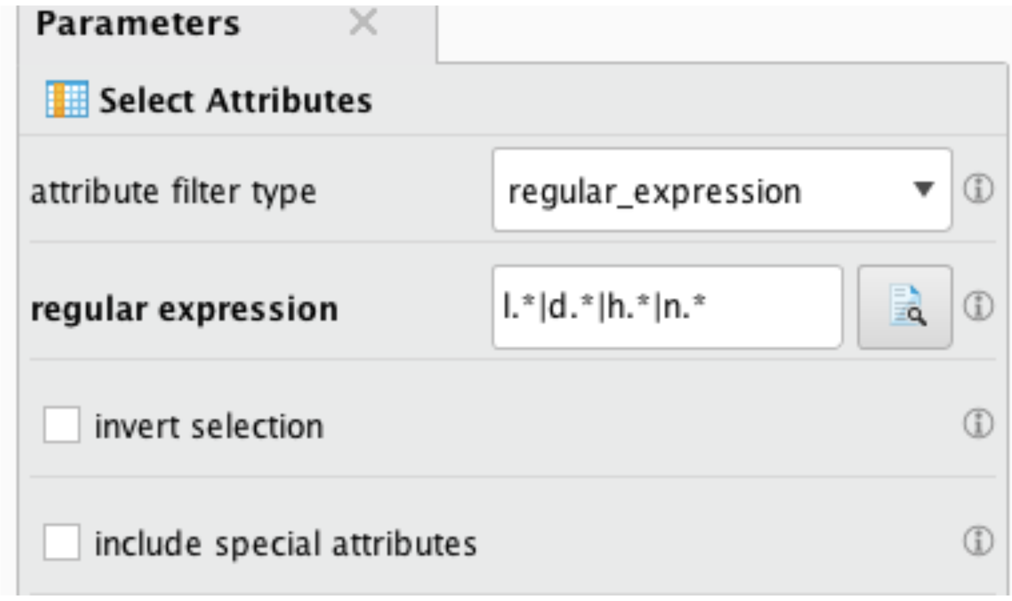
Table:

Category	%	n
M	47.2	1,155
F	42.6	1,042
I	10.2	250
Total	58.6	2,447

BONUS QUESTION:

RAPID MINER :





Parameters

Select Attributes

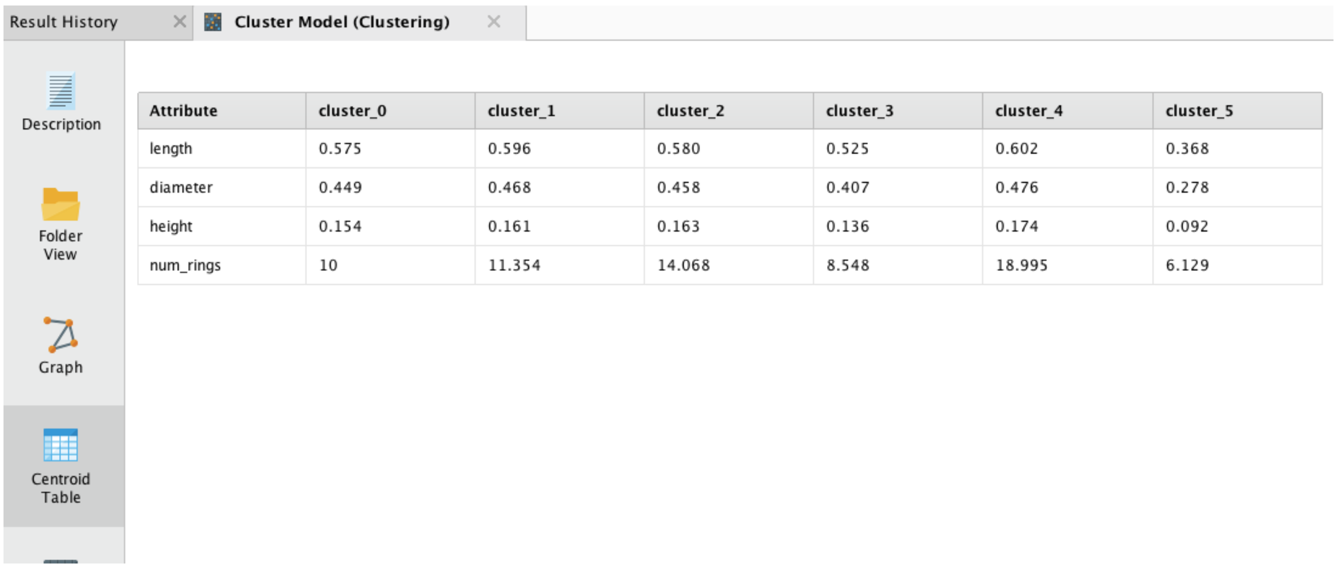
attribute filter type: regular_expression

regular expression: l.*|d.*|h.*|n.*

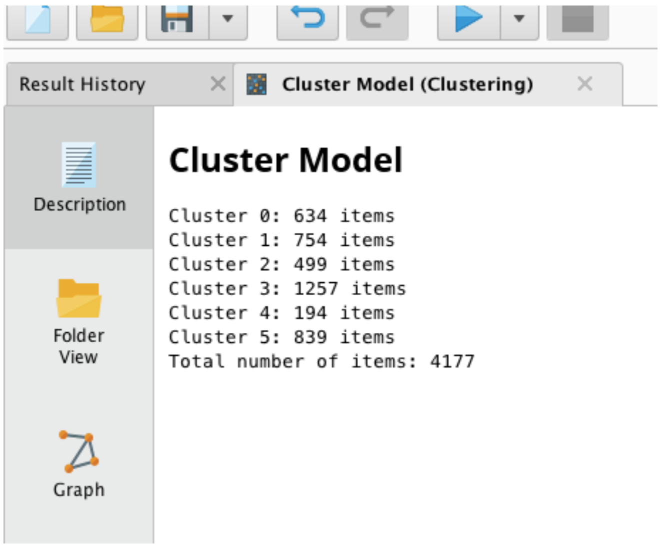
□ invert selection

□ include special attributes

The regular expression for select attributes is given as l.*ld.*l h.*l n.* which takes the attributes which start with l or d or h or n which are length, diameter, height, num_rings. The output is shown after running the nodes. (In the screenshot below only length, diameter, height, num_rings are taken in the centroid table)

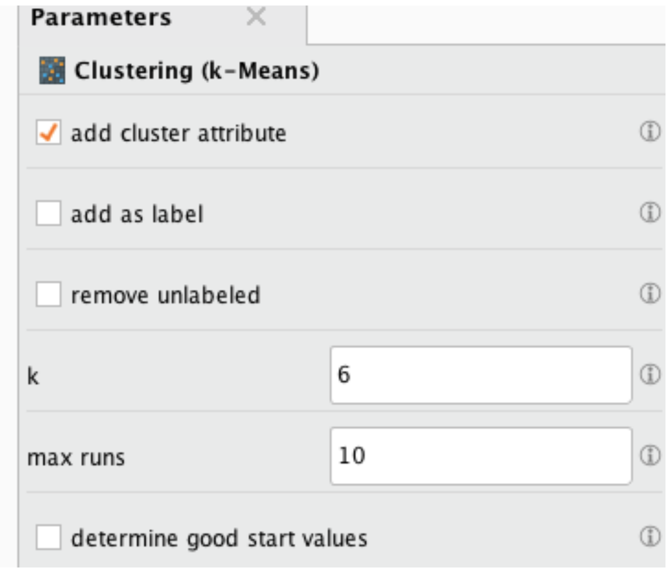


Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
length	0.575	0.596	0.580	0.525	0.602	0.368
diameter	0.449	0.468	0.458	0.407	0.476	0.278
height	0.154	0.161	0.163	0.136	0.174	0.092
num_rings	10	11.354	14.068	8.548	18.995	6.129



Cluster Model

```
Cluster 0: 634 items
Cluster 1: 754 items
Cluster 2: 499 items
Cluster 3: 1257 items
Cluster 4: 194 items
Cluster 5: 839 items
Total number of items: 4177
```



Parameters

Clustering (k-Means)

☑ add cluster attribute

□ add as label

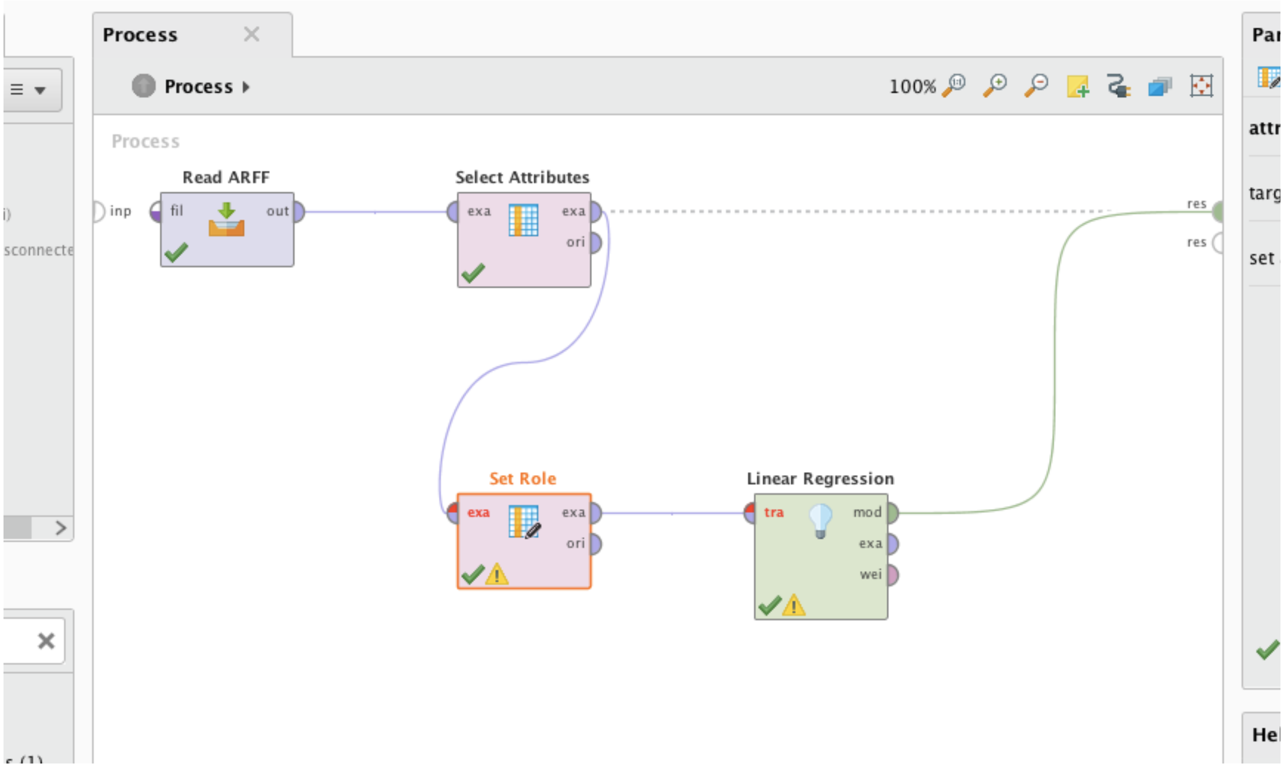
□ remove unlabeled

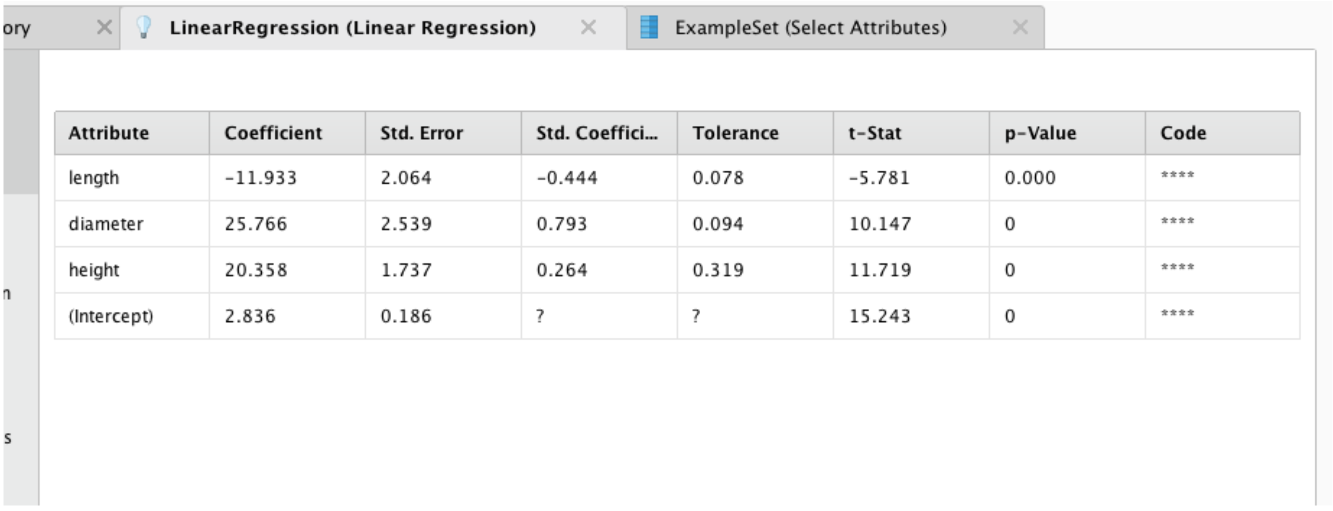
k: 6

max runs: 10

□ determine good start values

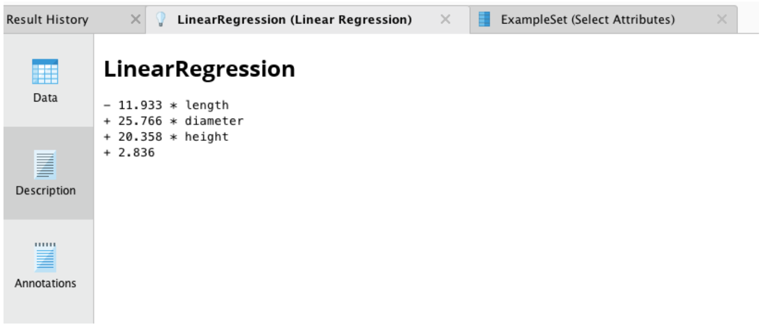
Q1.
In total, 6 clusters are formed. The number of data points in each cluster are 634 in cluster 0, 754 in cluster 1, 499 in cluster 2, 1257 in cluster 3, 194 in cluster 4, 839 in cluster 5.




ory
LinearRegression (Linear Regression)
ExampleSet (Select Attributes)

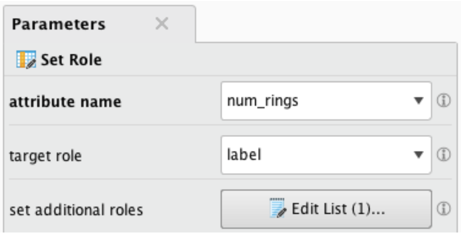
Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
length	-11.933	2.064	-0.444	0.078	-5.781	0.000	****
diameter	25.766	2.539	0.793	0.094	10.147	0	****
height	20.358	1.737	0.264	0.319	11.719	0	****
(Intercept)	2.836	0.186	?	?	15.243	0	****

Q2.
The equation for obtained in Linear regression is (-11.933 * length) + (25.766 * diameter) + (20.358*height) + (2.836)


Result History
LinearRegression (Linear Regression)
ExampleSet (Select Attributes)
Data
Description
Annotations

LinearRegression

```
- 11.933 * length
+ 25.766 * diameter
+ 20.358 * height
+ 2.836
```



attribute name	num_rings
target role	label
set additional roles	Edit List (1)...