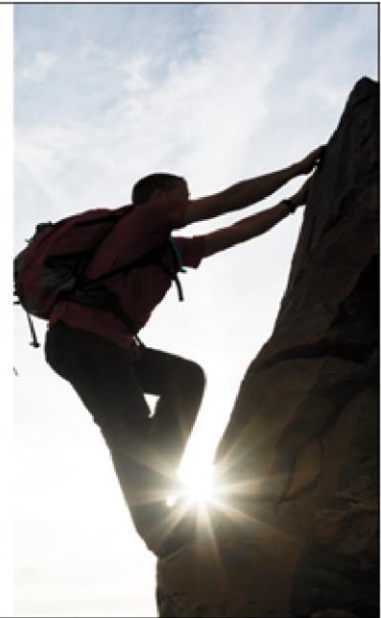


Decision Forests



Revised: 12/11/2019



Summary 1

Data Input..... 3

Analysis Options 5

Tables and Graphs..... 6

Analysis Summary 7

Tree Structure..... 9

Tree Diagram 11

Predictions and Residuals 12

Pareto Chart 14

Observed versus Predicted..... 15

Save Results 16

Using a Validation Set 17

Regression Trees 19

References 22

Summary

The *Decision Forests* procedure implements a machine-learning process to predict observations from data. It creates models of 2 forms:

1. *Classification models* that divide observations into groups based on their observed characteristics.
2. *Regression models* that predict the value of a dependent variable.

The models are constructed by creating a large number of decision trees and averaging the predictions made from those trees. Many trees are constructed using a procedure similar to that

of Classification and Regression Trees, with randomized node optimization and “bagging”. The procedure was developed by Brieman (2001).

Observations are typically divided into three sets:

1. A *training* set which is used to construct the tree.
2. A *validation* set for which the actual classification or value is known, which can be used to validate the model.
3. A *prediction* set for which the actual classification or value is not known but for which predictions are desired.

The dependent variable may be either categorical or quantitative, as may the predictor variables.

The calculations are performed by the “randomForest” package in R. To run the procedure, R must be installed on your computer together with the *randomForest* package. For information on downloading and installing R, refer to the document titled “R – Installation and Configuration”.

Sample StatFolios: *decisionForest.sgp*

Sample Data

As an example, consider the dataset described by Forina et al. (1991). These data are the results of a chemical analysis of 178 wines grown in the same region in Italy but derived from three different cultivars. Input variables included are:

- 1) Alcohol
- 2) Malic acid
- 3) Ash
- 4) Alcalinity of ash
- 5) Magnesium
- 6) Total phenols
- 7) Flavanoids
- 8) Nonflavanoid phenols
- 9) Proanthocyanins
- 10) Color intensity
- 11) Hue
- 12) OD280/OD315 of diluted wines
- 13) Proline

The output variable is class or cultivar (1-3). The data have been stored in the file *wine.sgd*.

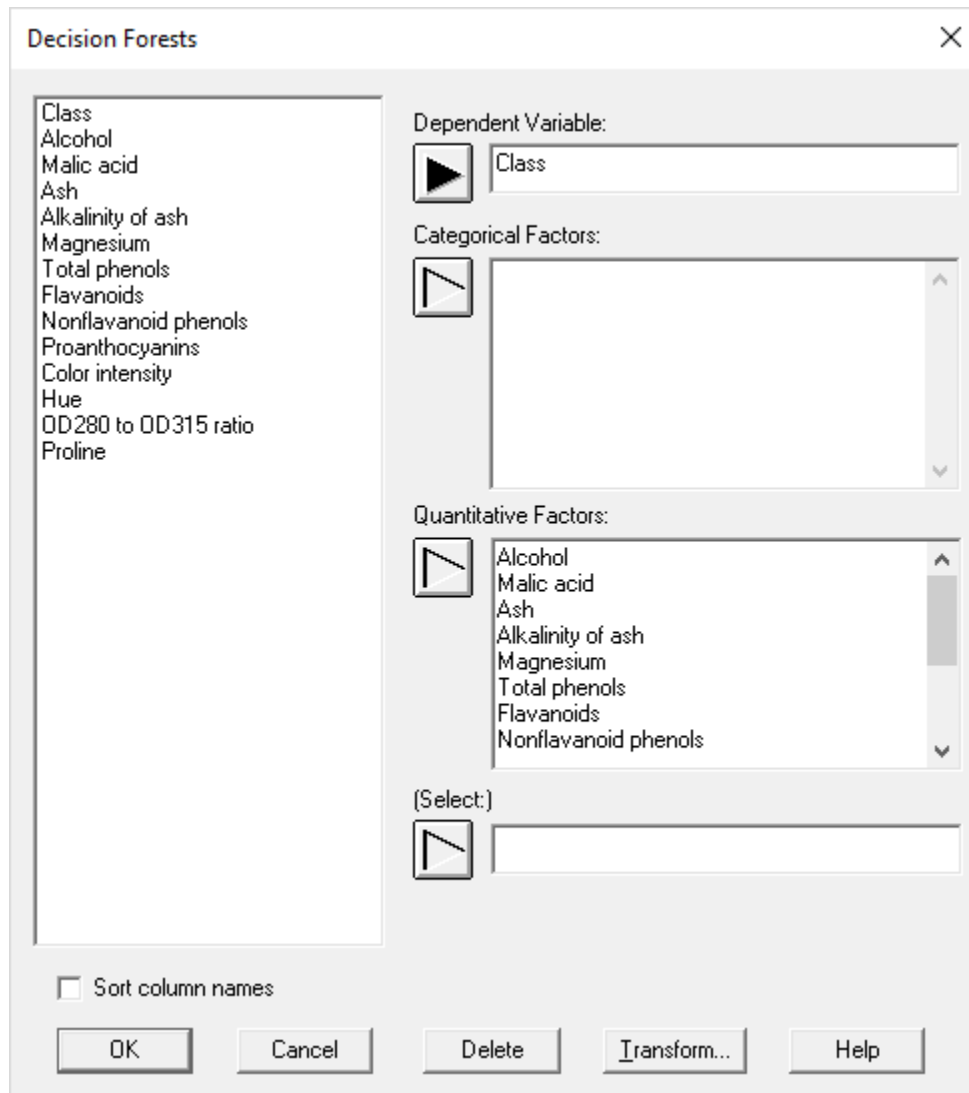
A small portion of the data is shown below:

C:\Data\webinar\winesgd														
Class	Alcohol	Malic acid	Ash	Alkalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280 to OD315 ratio	Proline	
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	
1	1	14.23	1.71	2.43	15.6	127	2.8	3.06	.28	2.29	5.64	1.04	3.92	1065
2	1	13.2	1.78	2.14	11.2	100	2.65	2.76	.26	1.28	4.38	1.05	3.4	1050
3	1	13.16	2.36	2.67	18.6	101	2.8	3.24	.3	2.81	5.68	1.03	3.17	1185
4	1	14.37	1.95	2.5	16.8	113	3.85	3.49	.24	2.18	7.8	.86	3.45	1480
5	1	13.24	2.59	2.87	21	118	2.8	2.69	.39	1.82	4.32	1.04	2.93	735
6	1	14.2	1.76	2.45	15.2	112	3.27	3.39	.34	1.97	6.75	1.05	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.5	2.52	.3	1.98	5.25	1.02	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.6	2.51	.31	1.25	5.05	1.06	3.58	1295
9	1	14.83	1.64	2.17	14	97	2.8	2.98	.29	1.98	5.2	1.08	2.85	1045
10	1	13.86	1.35	2.27	16	98	2.98	3.15	.22	1.85	7.22	1.01	3.55	1045
11	1	14.1	2.16	2.3	18	105	2.95	3.32	.22	2.38	5.75	1.25	3.17	1510
12	1	14.12	1.48	2.32	16.8	95	2.2	2.43	.26	1.57	5	1.17	2.82	1280
13	1	13.75	1.73	2.41	16	89	2.6	2.76	.29	1.81	5.6	1.15	2.9	1320
14	1	14.75	1.73	2.39	11.4	91	3.1	3.69	.43	2.81	5.4	1.25	2.73	1150
15	1	14.38	1.87	2.38	12	102	3.3	3.64	.29	2.96	7.5	1.2	3	1547
16	1	13.63	1.81	2.7	17.2	112	2.85	2.91	.3	1.46	7.3	1.28	2.88	1310
17	1	14.3	1.92	2.72	20	120	2.8	3.14	.33	1.97	6.2	1.07	2.65	1280
18	1	13.83	1.57	2.62	20	115	2.95	3.4	.4	1.72	6.6	1.13	2.57	1130
19	1	14.19	1.59	2.48	16.5	108	3.3	3.93	.32	1.86	8.7	1.23	2.82	1680
20	1	13.64	3.1	2.56	15.2	116	2.7	3.03	.17	1.66	5.1	.96	3.36	845
21	1	14.06	1.63	2.28	16	126	3	3.17	.24	2.1	5.65	1.09	3.71	780
22	1	12.93	3.8	2.65	18.6	102	2.41	2.41	.25	1.98	4.5	1.03	3.52	770
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	.27	1.69	3.8	1.11	4	1035
24	1	12.85	1.6	2.52	17.8	95	2.48	2.37	.26	1.46	3.93	1.09	3.63	1015
25	1	13.5	1.81	2.61	20	96	2.53	2.61	.28	1.66	3.52	1.12	3.82	845

A classification model is desired that uses the 13 quantitative variables to determine the probable cultivar of each wine.

Data Input

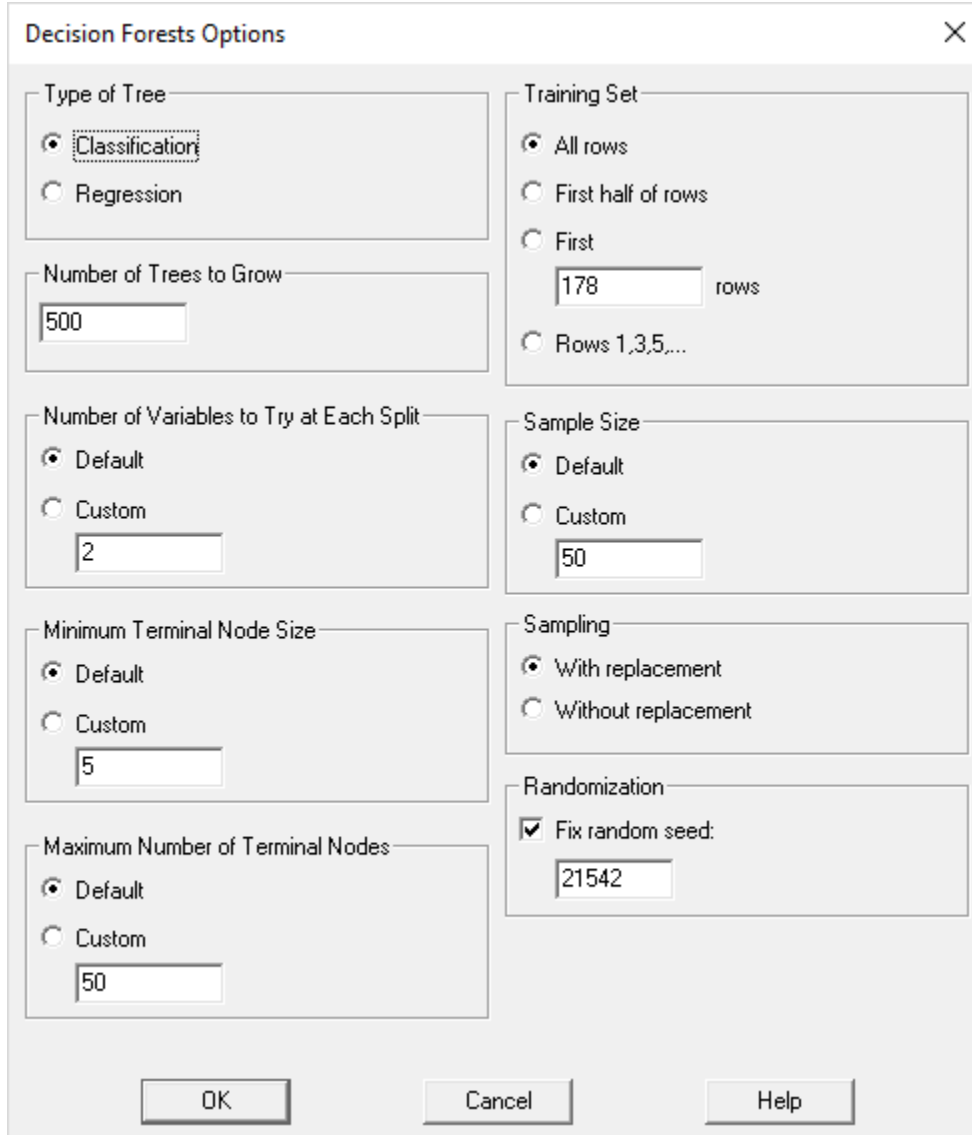
When the *Decision Forests* procedure is selected from the Statgraphics menu, a data input dialog box is displayed. 13 quantitative factors will be used to construct a classification model for *class*:



- **Dependent variable:** name of the column containing the class or value of the variable to be predicted. If fitting a classification model, this variable may be either categorical or quantitative. If fitting a regression model, this variable must be quantitative.
- **Categorical factors:** names of the columns containing the categorical variables (if any) that will be used to predict the dependent variable.
- **Quantitative factors:** names of the columns containing the continuous quantitative variables (if any) that will be used to predict the dependent variable.
- **Select:** optional Boolean column or expression identifying the cases (rows of the Databook) to be included in the analysis.

Analysis Options

The *Analysis Options* dialog box sets various options for fitting the model:



Decision Forests Options

Type of Tree

- Classification
- Regression

Number of Trees to Grow

500

Number of Variables to Try at Each Split

- Default
- Custom

2

Minimum Terminal Node Size

- Default
- Custom

5

Maximum Number of Terminal Nodes

- Default
- Custom

50

Training Set

- All rows
- First half of rows
- First

178 rows

- Rows 1,3,5,...

Sample Size

- Default
- Custom

50

Sampling

- With replacement
- Without replacement

Randomization

- Fix random seed:

21542

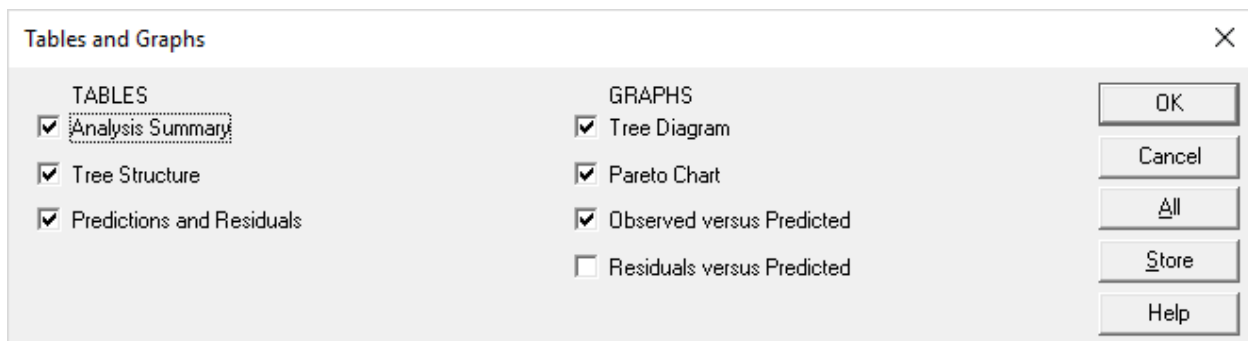
OK Cancel Help

- **Type of Tree:** Classification trees are used to predict the value of categorical variables. Regression trees are used to predict the value of continuous quantitative variables.
- **Number of Trees to Grow:** number of trees in the forest. Predictions are based on combining the results of these trees. Enter a large enough number to be sure that every input row is sampled at least a few times.
- **Number of Variables to Try at Each Split:** number of predictors sampled for splitting at each node. The default value is \sqrt{p} for classification trees and $p/3$ for regression trees, where p is the number of predictor variables.

- **Minimum Terminal Node Size:** smallest number of cases allowed at each terminal node. A larger number results in smaller trees. The default value is 1 for classification trees and 5 for regression trees.
- **Maximum Number of Terminal Nodes:** largest number of terminal nodes allowed in any tree. Default allows unlimited number of terminal nodes.
- **Training Set:** observations to be included in the training set used to fit the tree. All other rows are used for validation.
- **Sample Size:** size of samples to draw when creating each tree.
- **Sampling:** whether random sampling of observation for each tree should be drawn with or without replacement.
- **Randomization:** whether to fix the seed of the random number generator using the value indicated. If the seed is fixed, the same results will be obtained each time the procedure is run (assuming no other options are changed).

Tables and Graphs

The following tables and graphs may be created:



Tables and Graphs		×
TABLES		
<input checked="" type="checkbox"/>	Analysis Summary	<input type="button" value="OK"/> <input type="button" value="Cancel"/> <input type="button" value="All"/> <input type="button" value="Store"/> <input type="button" value="Help"/>
<input checked="" type="checkbox"/>	Tree Structure	
<input checked="" type="checkbox"/>	Predictions and Residuals	
GRAPHS		
<input checked="" type="checkbox"/>	Tree Diagram	
<input checked="" type="checkbox"/>	Pareto Chart	
<input checked="" type="checkbox"/>	Observed versus Predicted	
<input type="checkbox"/>	Residuals versus Predicted	

Analysis Summary

The *Analysis Summary* begins with a list of the R commands that were executed.

```

Decision Forests

d<-
read.csv("C:\\\\Users\\Neil\\AppData\\Local\\Temp\\data.csv",dec=".",sep=",",
",stringsAsFactors=TRUE)
setwd("C:\\Users\\Neil\\AppData\\Local\\Temp\\")
library("randomForest")

## Warning: package 'randomForest' was built under R version 3.2.5
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
library("ggplot2")
## Warning: package 'ggplot2' was built under R version 3.2.5
##
## Attaching package: 'ggplot2'
## The following object is masked from 'package:randomForest':
##
##   margin
library("RColorBrewer")
library(igraph)
## Warning: package 'igraph' was built under R version 3.2.5
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:stats':
##
##   decompose, spectrum
## The following object is masked from 'package:base':
##
##   union

library(tools)
set.seed(21542)
forest=randomForest(factor(Class)~Alcohol+Malic.acid+Ash+Alkalinity.of.ash+Magnesium+
Total.phernols+Flavanoids+Nonflavanoid.phenols+Proanthocyanins+Color.intensity+Hue+OD
280.to.OD315.ratio+Proline,data=d,keep.forest=TRUE,importance=TRUE,ntree=500)

```

The call to *randomForest* at the bottom of the output shows the model to be fit and any selected options. This section is followed by output generated by R:

```

##
## Call:
## randomForest(formula = factor(Class) ~ Alcohol + Malic.acid + Ash +
Alkalinity.of.ash + Magnesium + Total.phernols + Flavanoids +
Nonflavanoid.phenols + Proanthocyanins + Color.intensity + Hue +
OD280.to.OD315.ratio + Proline, data = d, keep.forest = TRUE, importance = TRUE,
ntree = 500)

```

```
##                               Type of random forest: classification
##                               Number of trees: 500
## No. of variables tried at each split: 3
##
##                               OOB estimate of error rate: 1.69%
## Confusion matrix:
##      1  2  3 class.error
## 1 59  0  0 0.00000000
## 2  1 68  2 0.04225352
## 3  0  0 48 0.00000000

importance(forest,type=1)

##                               MeanDecreaseAccuracy
## Alcohol                               20.322753
## Malic.acid                             10.500791
## Ash                                     6.749147
## Alkalinity.of.ash                       10.302894
## Magnesium                               11.189854
## Total.phernols                          12.958605
## Flavanoids                              23.570080
## Nonflavanoid.phenols                     4.658361
## Proanthocyanins                          8.383680
## Color.intensity                          26.638017
## Hue                                      18.212198
## OD280.to.OD315.ratio                     19.167853
## Proline                                  25.118020

importance(forest,2)

##                               MeanDecreaseGini
## Alcohol                               15.135820
## Malic.acid                             3.426725
## Ash                                     1.477518
## Alkalinity.of.ash                       3.108055
## Magnesium                               3.391125
## Total.phernols                          6.119080
## Flavanoids                              18.243380
## Nonflavanoid.phenols                     1.360557
## Proanthocyanins                          2.819832
## Color.intensity                          18.973468
## Hue                                      9.910013
## OD280.to.OD315.ratio                     13.472486
## Proline                                  18.955402

tree=getTree(forest,k=1)
newd<-
read.csv("C:\\\\Users\\Neil\\AppData\\Local\\Temp\\newdata.csv",dec=".",sep="," ,stringsAsFactors=TRUE)
pd<-predict(forest,newd)
write.table(forest$importance,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\importance.csv",sep=",")
write.table(forest$confusion,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\confusion.csv",sep="," ,row.names=FALSE)
write.table(tree,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\tree.csv",sep="," ,row.names=FALSE)
write.table(pd,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\predict.csv",sep=",")
```

Of particular interest are:

1. **OOB estimate of error rate:** shows the “Out Of Box” error rate. This shows the percentage of observations in the training set that were misclassified by the random forest.
2. **Confusion matrix:** a p by p matrix indicating how often observations in the training set were misclassified by class. Rows of the matrix indicate the actual class. Columns of the matrix indicate the class predicted by the random forest.
3. **Importance:** a measure of importance associated with each predictor variable. In the output above, it shows both the mean decrease in accuracy when the values of each variable are randomly permuted, and the mean decrease in the Gini index over all classes due to inclusion of the indicated variable. The larger the number, the more important the variable.

Tree Structure

Each tree in the forest has a different structure. The structure of a specific tree may be displayed in a tabular form:

Tree Structure					
<i>Node</i>	<i>Label</i>	<i>Class</i>	<i>Split point</i>	<i>Left daughter</i>	<i>Right daughter</i>
1	Alcohol		12.78	2	3
2	Color intensity		4.8	4	5
3	OD280 to OD315 ratio		2.49	6	7
4	Proanthocyanins		0.815	8	9
5	Proline		492.5	10	11
6	Ash		2.335	12	13
7	Proline		591.0	14	15
8	Hue		0.9	16	17
9	<terminal>	2			
10	<terminal>	2			
11	<terminal>	3			
12	Flavonoids		1.42	18	19
13	<terminal>	3			
14	<terminal>	2			
15	Magnesium		135.5	20	21
16	<terminal>	3			
17	<terminal>	2			
18	<terminal>	3			
19	<terminal>	2			
20	<terminal>	1			
21	<terminal>	2			

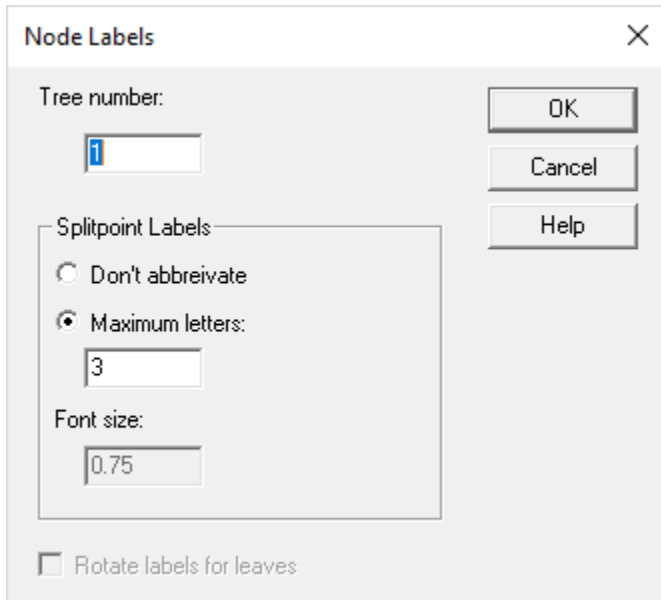
The table includes:

- **Node:** a number assigned to each node by R.

- **Label:** if not a terminating node, the variable involved in the decision to split the node.
- **Class:** if a terminating node, the predicted class at that node.
- **Split point:** if not a terminating node, the value of the indicated variable below which observations will move left rather than right.
- **Left daughter:** if not a terminating node, the subsequent node number if moving along the left branch.
- **Right daughter:** if not a terminating node, the subsequent node number if moving along the right branch.

For example, at node #1, observations travel along the branch to the left if $Alcohol \leq 12.78$ and along the branch to the right otherwise. When moving left, the next node is node #2. When moving right, the next node is #3. Observations ending up at node #9 are predicted to belong to class 2.

Pane Options



The image shows a dialog box titled "Node Labels" with a close button (X) in the top right corner. It contains the following controls:

- Tree number:** A text input field containing the number "1".
- Splitpoint Labels:** A section containing two radio buttons: "Don't abbreviate" (unselected) and "Maximum letters:" (selected). Below "Maximum letters:" is a text input field containing the number "3".
- Font size:** A text input field containing the value "0.75".
- Rotate labels for leaves:** A checkbox that is currently unchecked.

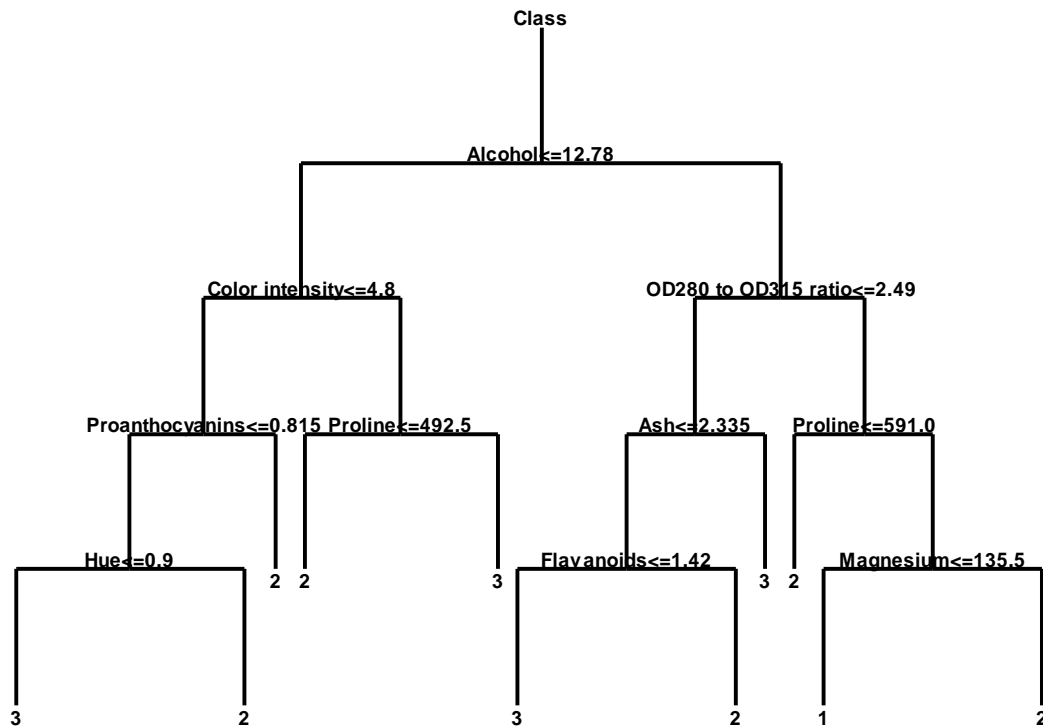
On the right side of the dialog box, there are three buttons: "OK", "Cancel", and "Help".

Tree number: number of the tree in the forest displayed by the table.

Labels: whether to abbreviate class labels for categorical factors.

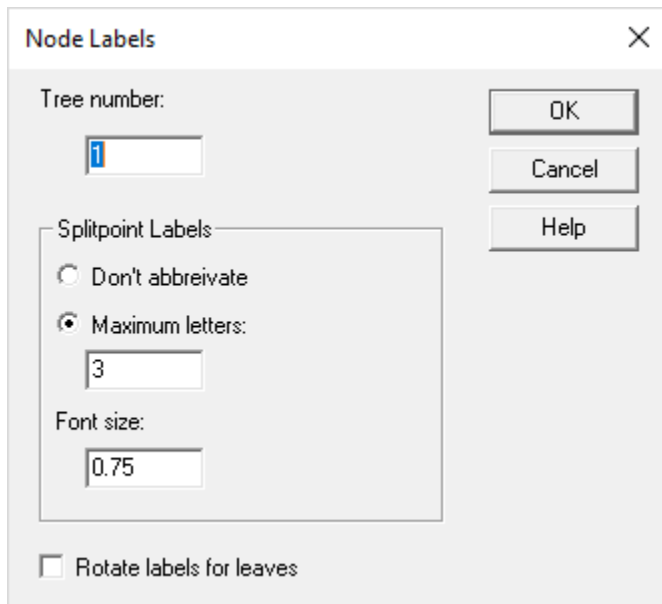
Tree Diagram

Shows a diagram of a selected tree in the forest:



The decision criterion for moving left rather than right is shown at each non-terminating node. The predicted class is shown at each terminating node.

Pane Options



Tree number: number of the tree in the forest displayed by the graph.

Labels: whether to abbreviate class labels for categorical factors.

Font size: scaling factor for all labels on the graph. Smaller values result in smaller text.

Rotate labels for leaves: if checked, labels for the leaves will be oriented vertically.

Predictions and Residuals

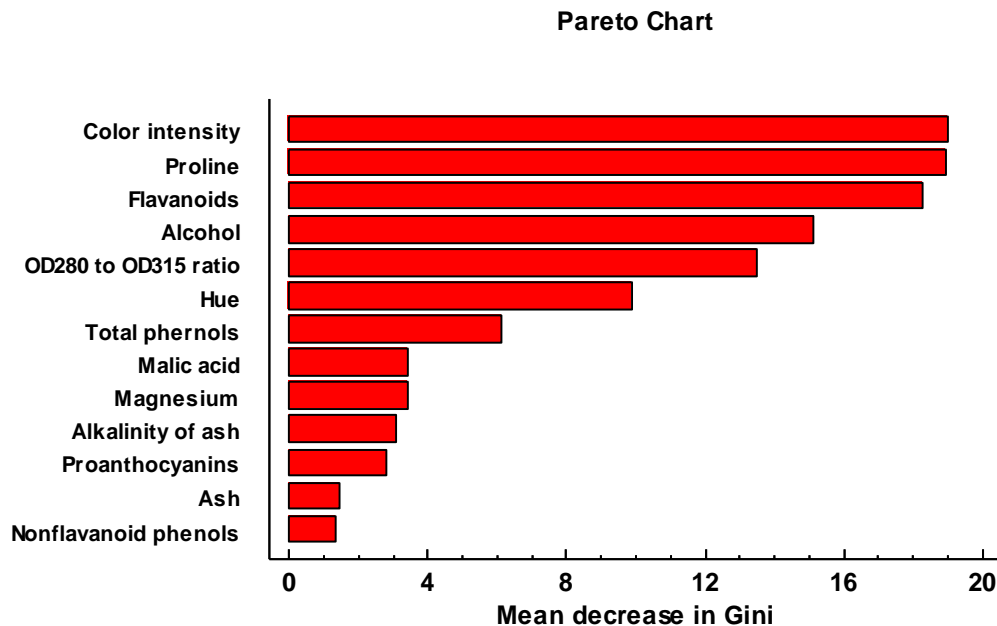
This table shows the predicted values for each observation:

Predictions and Residuals		
Training set n=178		
<i>Row</i>	<i>Predicted</i>	<i>Observed</i>
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
...

- **Predicted:** predicted value for the observation in the indicated row.
- **Observed:** observed value of the observation in the indicated row.

Pareto Chart

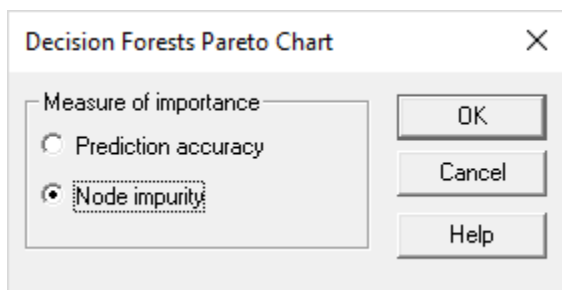
This graph plots the importance of each variable, sorted from largest to smallest:



For the sample data, *color intensity* and *proline* are the 2 most important variables.

Pane Options

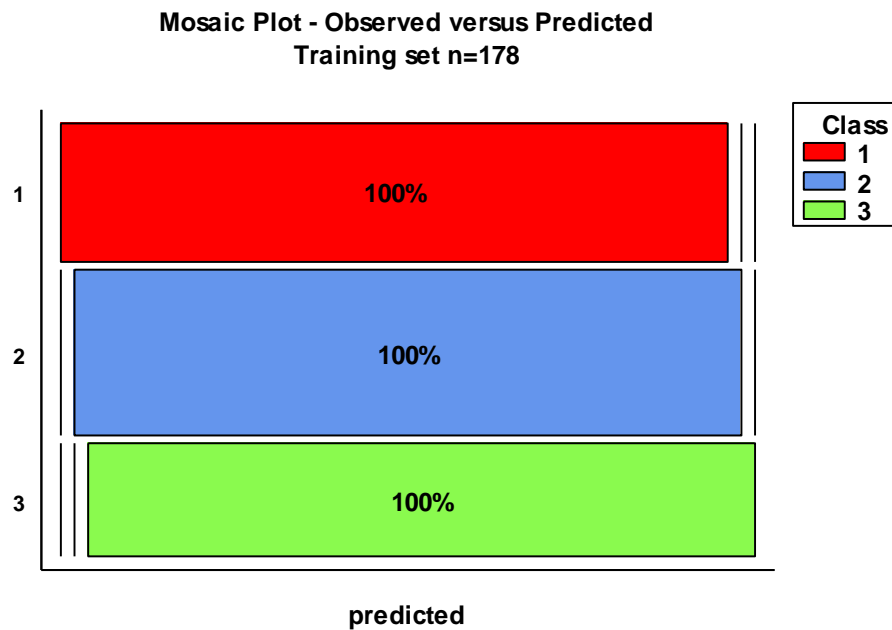
This dialog box specifies the importance measure to be plotted.



- **Prediction accuracy:** prediction error for out-of-bag portion of the data (error rate for classification trees, MSE for regression trees).
- **Node impurity:** decrease in node impurities by splitting on the variables. For classification trees, impurity is measured using the Gini index. For regression trees, impurity is measured using residual sum of squares.

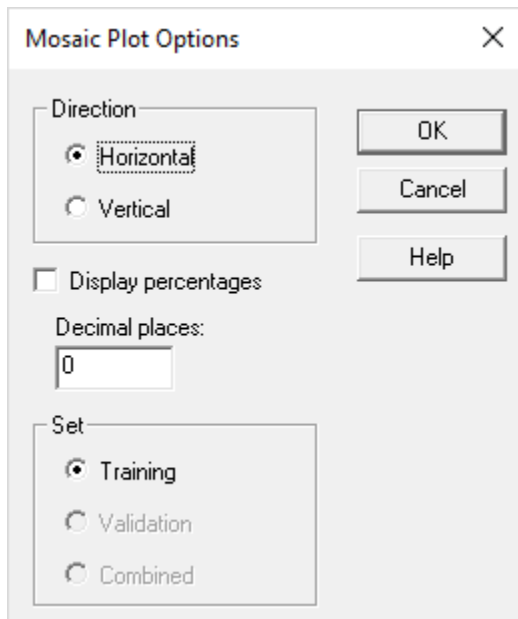
Observed versus Predicted

When fitting a classification tree, this graph creates a mosaic plot.



By default, the mosaic plot contains a row for each level of the dependent variable. Bars are drawn in each row with length proportional to the number of times observations at that level were predicted to be of each class. The plot above shows that all 3 classes were predicted correctly all of the time.

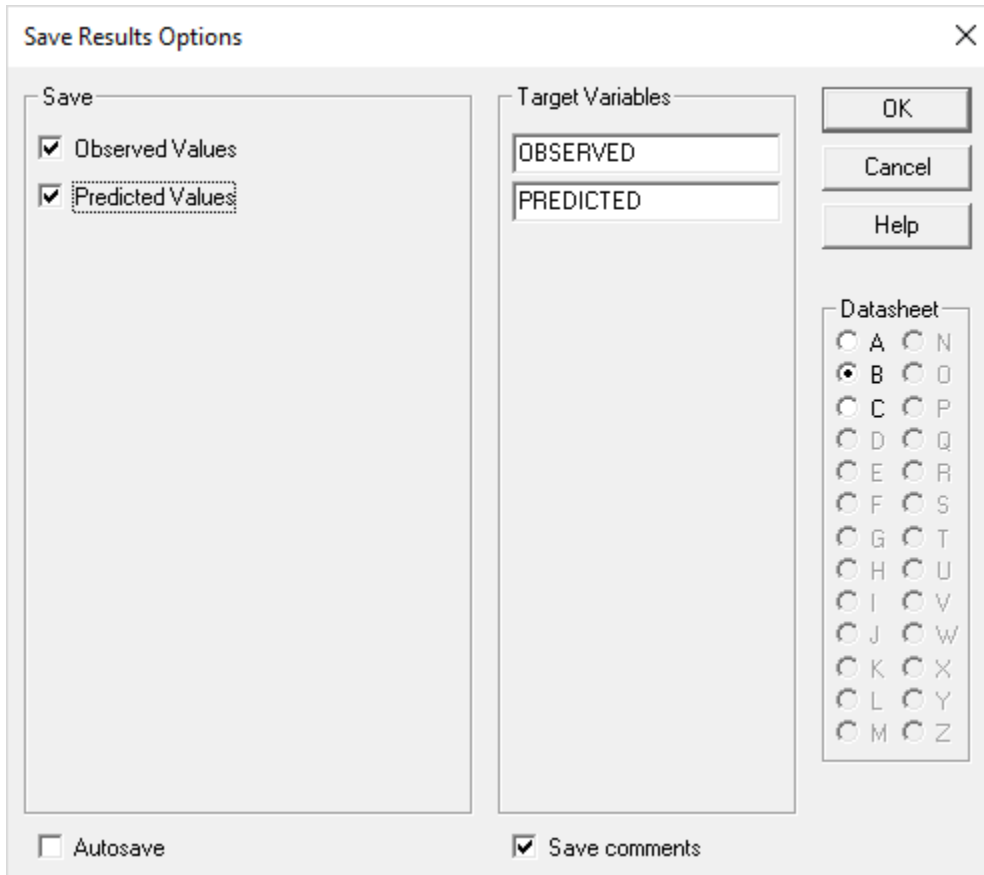
Pane Options



- **Direction:** orientation of the bars.
- **Display percentage:** whether the plot should display the percentage corresponding to each bar.
- **Set:** set of points to be displayed on the plot.

Save Results

Selected output may be saved in a Statgraphics datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



The dialog box is titled "Save Results Options" and contains the following elements:

- Save:** A section with two checked checkboxes: "Observed Values" and "Predicted Values".
- Target Variables:** Two text input fields containing "OBSERVED" and "PREDICTED".
- Datasheet:** A grid of radio buttons labeled with letters A through Z. The radio button for "B" is selected.
- Buttons:** "OK", "Cancel", and "Help" buttons are located on the right side.
- Checkboxes:** "Autosave" (unchecked) and "Save comments" (checked) are located at the bottom.

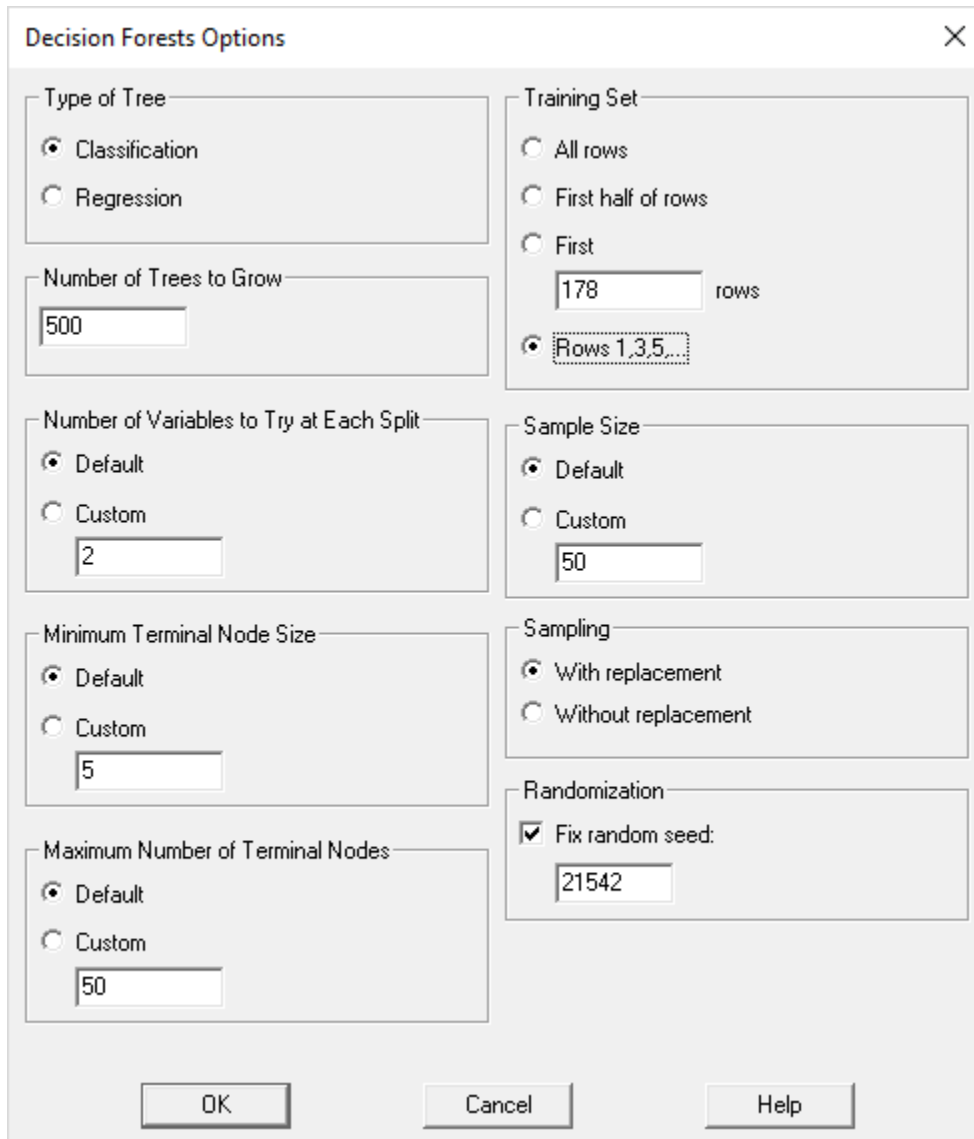
To save results, select:

- **Save:** select the items to be saved.
- **Target Variables:** enter names for the columns to be created.
- **Datasheet:** the datasheet into which the results will be saved.
- **Autosave:** if checked, the results will be saved automatically each time a saved StatFolio is loaded.
- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

Using a Validation Set

It is common practice when building a machine learning model to divide the data into 2 sets: a training set used to build the model and a validation set to evaluate how well the method

performs on data not used to develop the model. The *Analysis Options* dialog box lets you withhold selected observations. For example, the settings below place every second row beginning with row 1 into the training set and put the other rows into the validation set:



The image shows a dialog box titled "Decision Forests Options" with a close button (X) in the top right corner. The dialog is divided into several sections:

- Type of Tree:** Radio buttons for "Classification" (selected) and "Regression".
- Number of Trees to Grow:** A text box containing the value "500".
- Number of Variables to Try at Each Split:** Radio buttons for "Default" (selected) and "Custom", with a text box containing "2" below.
- Minimum Terminal Node Size:** Radio buttons for "Default" (selected) and "Custom", with a text box containing "5" below.
- Maximum Number of Terminal Nodes:** Radio buttons for "Default" (selected) and "Custom", with a text box containing "50" below.
- Training Set:** Radio buttons for "All rows", "First half of rows", "First", and "Rows 1,3,5..." (selected). A text box containing "178" and the label "rows" is positioned between "First" and "Rows 1,3,5...".
- Sample Size:** Radio buttons for "Default" (selected) and "Custom", with a text box containing "50" below.
- Sampling:** Radio buttons for "With replacement" (selected) and "Without replacement".
- Randomization:** A checked checkbox for "Fix random seed:" and a text box containing "21542" below.

At the bottom of the dialog are three buttons: "OK", "Cancel", and "Help".

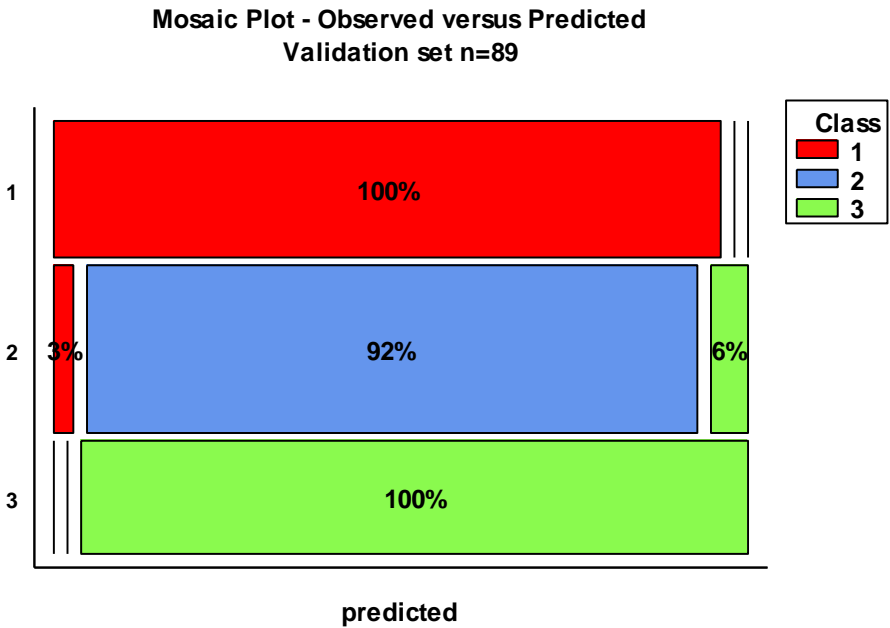
When building a classification model, it is important that all categorical variables have the same unique values in both sets.

When a validation set is present, the *Analysis Summary* table will display statistics for both the training and validation sets:

```
##          OOB estimate of error rate: 0%
## Confusion matrix:
##    1  2  3 class.error
## 1 30  0  0          0
```

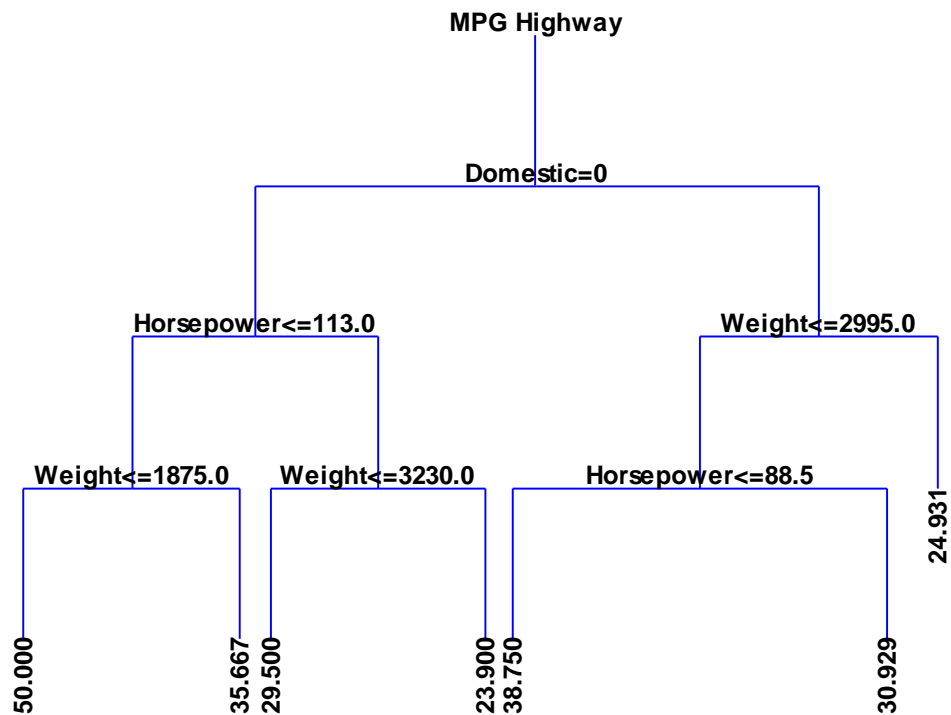
```
## 2  0 35  0          0
## 3  0  0 24          0
##
##          Test set error rate: 3.37%
## Confusion matrix:
##      1  2  3 class.error
## 1 29  0  0 0.00000000
## 2  1 33  2 0.08333333
## 3  0  0 24 0.00000000
```

The confusion matrix at the top applies to the training set, which was predicted with no errors. The confusion matrix at the bottom applies to the test or validation set, for which the error rate was 3.37%. When plotting the mosaic plot, you may use Pane Options to select either set:



Regression Trees

When the dependent variable is continuous, a regression tree may be fit instead of a classification tree. In such cases, the terminating nodes will give a quantitative prediction based on the median result of all trees in the forest:



There are 7 terminating nodes.

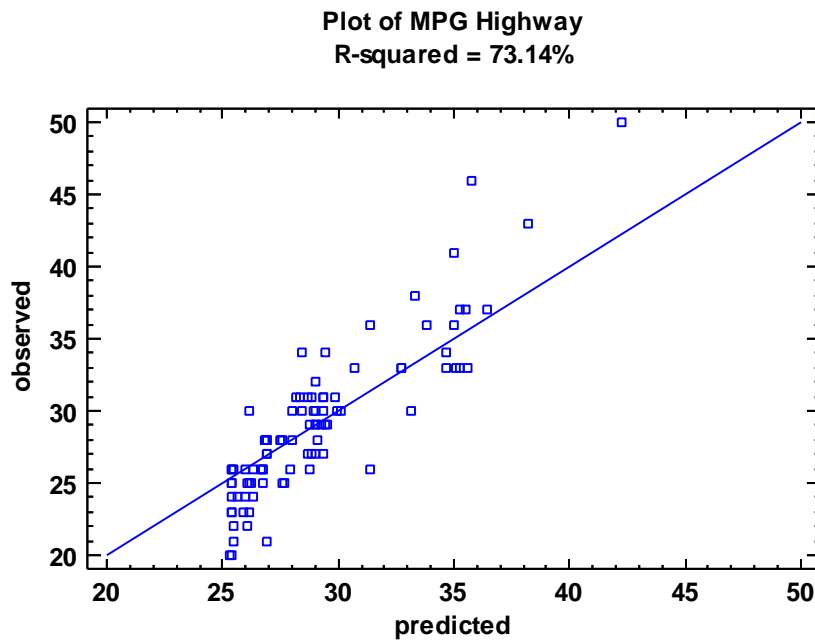
There are a few differences between a regression tree and a classification tree:

1. When displaying the report of observed and predicted values, residuals are also calculated and displayed.

Predictions and Residuals			
Training set n=93			
Row	Predicted	Observed	Residual
1	28.8036	31.0	2.19639
2	26.1999	25.0	-1.19993
3	26.765	26.0	-0.765047
4	26.7022	26.0	-0.702172
5	26.1822	30.0	3.81781
...

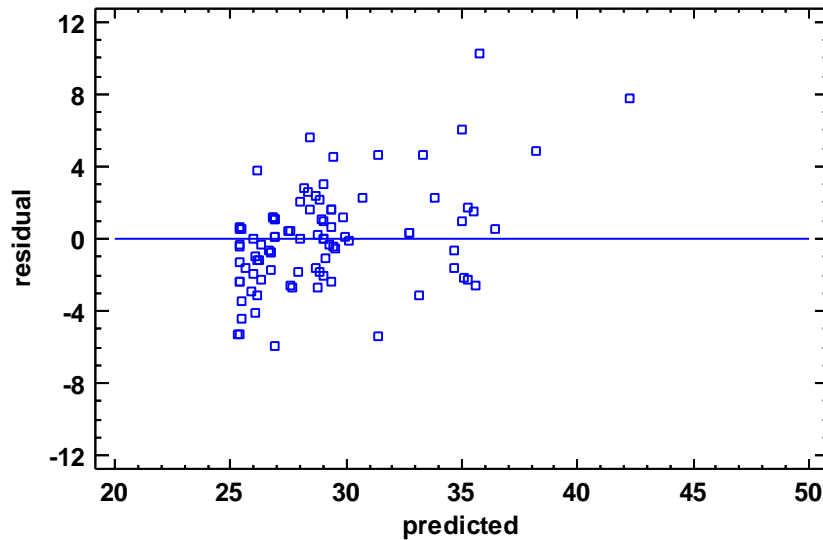
The predicted values equal the average of all observations in the training set that arrive at a given leaf.

2. No table of node probabilities or classification table is created.
3. The graph of *Observed versus Predicted* values produces a scatterplot rather than a mosaic chart:



4. An additional graph is available plotting the *Residuals versus Predicted* values:

Residual Plot



References

Brieman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1998) Classification and Regression Trees. Wadsworth.

Breiman, L. (2001). "Random Forests". Machine Learning. **45** (1): 5–32.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) "Modeling wine preferences by data mining from physicochemical properties". In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

R Package "randomForests" (2015) <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>