IBM SPSS Decision Trees 28

Note

Before using this information and the product it supports, read the information in "Notices" on page 19.

Product Information

This edition applies to version 28, release 0, modification 0 of IBM® SPSS® Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Decision Trees

The following Decision Trees features are included in SPSS Statistics Professional Edition or the Decision Trees option.

Creating Decision Trees

The Decision Tree procedure creates a tree-based classification model. It classifies cases into groups or predicts values of a dependent (target) variable based on values of independent (predictor) variables. The procedure provides validation tools for exploratory and confirmatory classification analysis.

The procedure can be used for:

Segmentation. Identify persons who are likely to be members of a particular group.

Stratification. Assign cases into one of several categories, such as high-, medium-, and low-risk groups.

Prediction. Create rules and use them to predict future events, such as the likelihood that someone will default on a loan or the potential resale value of a vehicle or home.

Data reduction and variable screening. Select a useful subset of predictors from a large set of variables for use in building a formal parametric model.

Interaction identification. Identify relationships that pertain only to specific subgroups and specify these in a formal parametric model.

Category merging and discretizing continuous variables. Recode group predictor categories and continuous variables with minimal loss of information.

Example. A bank wants to categorize credit applicants according to whether or not they represent a reasonable credit risk. Based on various factors, including the known credit ratings of past customers, you can build a model to predict if future customers are likely to default on their loans.

A tree-based analysis provides some attractive features:

- It allows you to identify homogeneous groups with high or low risk.
- It makes it easy to construct rules for making predictions about individual cases.

Data Considerations

Data. The dependent and independent variables can be:

- *Nominal*. A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, postal code, and religious affiliation.
- *Ordinal*. A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- *Scale*. A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

Frequency weights If weighting is in effect, fractional weights are rounded to the closest integer; so, cases with a weight value of less than 0.5 are assigned a weight of 0 and are therefore excluded from the analysis.

Assumptions. This procedure assumes that the appropriate measurement level has been assigned to all analysis variables, and some features assume that all values of the dependent variable included in the analysis have defined value labels.

- **Measurement level.** Measurement level affects the tree computations; so, all variables should be assigned the appropriate measurement level. By default, numeric variables are assumed to be scale and string variables are assumed to be nominal, which may not accurately reflect the true measurement level. An icon next to each variable in the variable list identifies the variable type.

Table 1. Measurement level icons

Icon	Measurement level
	Scale
	Nominal
	Ordinal

You can temporarily change the measurement level for a variable by right-clicking the variable in the source variable list and selecting a measurement level from the pop-up menu.

- **Value labels.** The dialog box interface for this procedure assumes that either all nonmissing values of a categorical (nominal, ordinal) dependent variable have defined value labels or none of them do. Some features are not available unless at least two nonmissing values of the categorical dependent variable have value labels. If at least two nonmissing values have defined value labels, any cases with other values that do not have value labels will be excluded from the analysis.

To Obtain Decision Trees

1. From the menus choose:

   Analyze > **Classify** > **Tree...**
2. Select a dependent variable.
3. Select one or more independent variables.
4. Select a growing method.

Optionally, you can:

- Change the measurement level for any variable in the source list.
- Force the first variable in the independent variables list into the model as the first split variable.
- Select an influence variable that defines how much influence a case has on the tree-growing process. Cases with lower influence values have less influence; cases with higher values have more. Influence variable values must be positive.
- Validate the tree.
- Customize the tree-growing criteria.
- Save terminal node numbers, predicted values, and predicted probabilities as variables.
- Save the model in XML (PMML) format.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data. Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually. Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Changing Measurement Level

1. Right-click the variable in the source list.
2. Select a measurement level from the pop-up menu.

This changes the measurement level temporarily for use in the Decision Tree procedure.

Growing Methods

The available growing methods are:

CHAID. Chi-squared Automatic Interaction Detection. At each step, CHAID chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable.

Exhaustive CHAID. A modification of CHAID that examines all possible splits for each predictor.

CRT. Classification and Regression Trees. CRT splits the data into segments that are as homogeneous as possible with respect to the dependent variable. A terminal node in which all cases have the same value for the dependent variable is a homogeneous, "pure" node.

QUEST. Quick, Unbiased, Efficient Statistical Tree. A method that is fast and avoids other methods' bias in favor of predictors with many categories. QUEST can be specified only if the dependent variable is nominal.

There are benefits and limitations with each method, including:

Table 2. Features of growing method			
Feature	**CHAID***	**CRT**	**QUEST**
Chi-square-based**	X		
Surrogate independent (predictor) variables		X	X
Tree pruning		X	X
Multiway node splitting	X		
Binary node splitting		X	X
Influence variables	X	X	
Prior probabilities		X	X
Misclassification costs	X	X	X
Fast calculation	X		X

*Includes Exhaustive CHAID.

**QUEST also uses a chi-square measure for nominal independent variables.

Selecting Categories

For categorical (nominal, ordinal) dependent variables, you can:

- Control which categories are included in the analysis.
- Identify the target categories of interest.

Including/Excluding Categories

You can limit the analysis to specific categories of the dependent variable.

- Cases with values of the dependent variable in the Exclude list are not included in the analysis.

- For nominal dependent variables, you can also include user-missing categories in the analysis. (By default, user-missing categories are displayed in the Exclude list.)

Target Categories

Selected (checked) categories are treated as the categories of primary interest in the analysis. For example, if you are primarily interested in identifying those individuals most likely to default on a loan, you might select the "bad" credit-rating category as the target category.

- There is no default target category. If no category is selected, some classification rule options and gains-related output are not available.
- If multiple categories are selected, separate gains tables and charts are produced for each target category.
- Designating one or more categories as target categories has no effect on the tree model, risk estimate, or misclassification results.

Categories and Value Labels

This dialog box requires defined value labels for the dependent variable. It is not available unless at least two values of the categorical dependent variable have defined value labels.

To Include/Exclude Categories and Select Target Categories

1. In the main Decision Tree dialog box, select a categorical (nominal, ordinal) dependent variable with two or more defined value labels.
2. Click **Categories**.

Validation

Validation allows you to assess how well your tree structure generalizes to a larger population. Two validation methods are available: crossvalidation and split-sample validation.

Crossvalidation

Crossvalidation divides the sample into a number of subsamples, or **folds**. Tree models are then generated, excluding the data from each subsample in turn. The first tree is based on all of the cases except those in the first sample fold, the second tree is based on all of the cases except those in the second sample fold, and so on. For each tree, misclassification risk is estimated by applying the tree to the subsample excluded in generating it.

- You can specify a maximum of 25 sample folds. The higher the value, the fewer the number of cases excluded for each tree model.
- Crossvalidation produces a single, final tree model. The crossvalidated risk estimate for the final tree is calculated as the average of the risks for all of the trees.

Split-Sample Validation

With split-sample validation, the model is generated using a training sample and tested on a hold-out sample.

- You can specify a training sample size, expressed as a percentage of the total sample size, or a variable that splits the sample into training and testing samples.
- If you use a variable to define training and testing samples, cases with a value of 1 for the variable are assigned to the training sample, and all other cases are assigned to the testing sample. The variable cannot be the dependent variable, weight variable, influence variable, or a forced independent variable.
- You can display results for both the training and testing samples or just the testing sample.
- Split-sample validation should be used with caution on small data files (data files with a small number of cases). Small training sample sizes may yield poor models, since there may not be enough cases in some categories to adequately grow the tree.

To Validate a Decision Tree

1. In the main Decision Trees dialog, click **Validation**.

2. Select **Crossvalidation** or **Split-sample validation**.

Note: Both validation methods randomly assign cases to sample groups. If you want to be able to reproduce the exact same results in a subsequent analysis, you should set the random number seed (Transform menu, Random Number Generators) before running the analysis for the first time and then reset the seed to that value for the subsequent analysis.

Tree-Growing Criteria

The available growing criteria may depend on the growing method, level of measurement of the dependent variable, or a combination of the two.

Growth Limits

The Growth Limits tab allows you to limit the number of levels in the tree and control the minimum number of cases for parent and child nodes.

Maximum Tree Depth. Controls the maximum number of levels of growth beneath the root node. The **Automatic** setting limits the tree to three levels beneath the root node for the CHAID and Exhaustive CHAID methods and five levels for the CRT and QUEST methods.

Minimum Number of Cases. Controls the minimum numbers of cases for nodes. Nodes that do not satisfy these criteria will not be split.

- Increasing the minimum values tends to produce trees with fewer nodes.
- Decreasing the minimum values produces trees with more nodes.

For data files with a small number of cases, the default values of 100 cases for parent nodes and 50 cases for child nodes may sometimes result in trees with no nodes below the root node; in this case, lowering the minimum values may produce more useful results.

To Specify Growth Limits

1. In the main Decision Tree dialog, click **Criteria**.
2. Click the **Growth Limits** tab.

CHAID Criteria

For the CHAID and Exhaustive CHAID methods, you can control:

Significance Level. You can control the significance value for splitting nodes and merging categories. For both criteria, the default significance level is 0.05.

- For splitting nodes, the value must be greater than 0 and less than 1. Lower values tend to produce trees with fewer nodes.
- For merging categories, the value must be greater than 0 and less than or equal to 1. To prevent merging of categories, specify a value of 1. For a scale independent variable, this means that the number of categories for the variable in the final tree is the specified number of intervals (the default is 10). See the topic "Scale Intervals for CHAID Analysis" on page 6 for more information.

Chi-Square Statistic. For ordinal dependent variables, chi-square for determining node splitting and category merging is calculated using the likelihood-ratio method. For nominal dependent variables, you can select the method:

- **Pearson.** This method provides faster calculations but should be used with caution on small samples. This is the default method.
- **Likelihood ratio.** This method is more robust that Pearson but takes longer to calculate. For small samples, this is the preferred method.

Model Estimation. For nominal and ordinal dependent variables, you can specify:

- **Maximum number of iterations.** The default is 100. If the tree stops growing because the maximum number of iterations has been reached, you may want to increase the maximum or change one or more of the other criteria that control tree growth.
- **Minimum change in expected cell frequencies.** The value must be greater than 0 and less than 1. The default is 0.05. Lower values tend to produce trees with fewer nodes.

Adjust significance values using Bonferroni method. For multiple comparisons, significance values for merging and splitting criteria are adjusted using the Bonferroni method. This is the default.

Allow resplitting of merged categories within a node. Unless you explicitly prevent category merging, the procedure will attempt to merge independent (predictor) variable categories together to produce the simplest tree that describes the model. This option allows the procedure to resplit merged categories if that provides a better solution.

To Specify CHAID Criteria

1. In the main Decision Tree dialog, select **CHAID** or **Exhaustive CHAID** as the growing method.
2. Click **Criteria**.
3. Click the **CHAID** tab.

Scale Intervals for CHAID Analysis

In CHAID analysis, scale independent (predictor) variables are always banded into discrete groups (for example, 0–10, 11–20, 21–30, etc.) prior to analysis. You can control the initial/maximum number of groups (although the procedure may merge contiguous groups after the initial split):

- **Fixed number.** All scale independent variables are initially banded into the same number of groups. The default is 10.
- **Custom.** Each scale independent variable is initially banded into the number of groups specified for that variable.

To Specify Intervals for Scale Independent Variables

1. In the main Decision Tree dialog box, select one or more scale independent variables.
2. For the growing method, select **CHAID** or **Exhaustive CHAID**.
3. Click **Criteria**.
4. Click the **Intervals** tab.

In CRT and QUEST analysis, all splits are binary and scale and ordinal independent variables are handled the same way; so, you cannot specify a number of intervals for scale independent variables.

CRT Criteria

The CRT growing method attempts to maximize within-node homogeneity. The extent to which a node does not represent a homogenous subset of cases is an indication of **impurity**. For example, a terminal node in which all cases have the same value for the dependent variable is a homogenous node that requires no further splitting because it is "pure."

You can select the method used to measure impurity and the minimum decrease in impurity required to split nodes.

Impurity Measure. For scale dependent variables, the least-squared deviation (LSD) measure of impurity is used. It is computed as the within-node variance, adjusted for any frequency weights or influence values.

For categorical (nominal, ordinal) dependent variables, you can select the impurity measure:

- **Gini.** Splits are found that maximize the homogeneity of child nodes with respect to the value of the dependent variable. Gini is based on squared probabilities of membership for each category of the dependent variable. It reaches its minimum (zero) when all cases in a node fall into a single category. This is the default measure.

- **Twoing.** Categories of the dependent variable are grouped into two subclasses. Splits are found that best separate the two groups.
- **Ordered twoing.** Similar to twoing except that only adjacent categories can be grouped. This measure is available only for ordinal dependent variables.

Minimum change in improvement. This is the minimum decrease in impurity required to split a node. The default is 0.0001. Higher values tend to produce trees with fewer nodes.

To Specify CRT Criteria

1. For the growing method, select **CRT**.
2. Click **Criteria**.
3. Click the **CRT** tab.

QUEST Criteria

For the QUEST method, you can specify the significance level for splitting nodes. An independent variable cannot be used to split nodes unless the significance level is less than or equal to the specified value. The value must be greater than 0 and less than 1. The default is 0.05. Smaller values will tend to exclude more independent variables from the final model.

To Specify QUEST Criteria

1. In the main Decision Tree dialog box, select a nominal dependent variable.
2. For the growing method, select **QUEST**.
3. Click **Criteria**.
4. Click the **QUEST** tab.

Pruning Trees

With the CRT and QUEST methods, you can avoid overfitting the model by **pruning** the tree: the tree is grown until stopping criteria are met, and then it is trimmed automatically to the smallest subtree based on the specified maximum difference in risk. The risk value is expressed in standard errors. The default is 1. The value must be non-negative. To obtain the subtree with the minimum risk, specify 0.

To Prune a Tree

1. In the main Decision Tree dialog box, for the growing method, select **CRT** or **QUEST**.
2. Click **Criteria**.
3. Click the **Pruning** tab.

Pruning versus Hiding Nodes

When you create a pruned tree, any nodes pruned from the tree are not available in the final tree. You can interactively hide and show selected child nodes in the final tree, but you cannot show nodes that were pruned in the tree creation process. See the topic "Tree Editor" on page 15 for more information.

Surrogates

CRT and QUEST can use **surrogates** for independent (predictor) variables. For cases in which the value for that variable is missing, other independent variables having high associations with the original variable are used for classification. These alternative predictors are called surrogates. You can specify the maximum number of surrogates to use in the model.

- By default, the maximum number of surrogates is one less than the number of independent variables. In other words, for each independent variable, all other independent variables may be used as surrogates.
- If you don't want the model to use surrogates, specify 0 for the number of surrogates.

To Specify Surrogates

1. In the main Decision Tree dialog box, for the growing method, select **CRT** or **QUEST**.

2. Click **Criteria**.
3. Click the **Surrogates** tab.

Options

Available options may depend on the growing method, the level of measurement of the dependent variable, and/or the existence of defined value labels for values of the dependent variable.

Misclassification Costs

For categorical (nominal, ordinal) dependent variables, misclassification costs allow you to include information about the relative penalty associated with incorrect classification. For example:

- The cost of denying credit to a creditworthy customer is likely to be different from the cost of extending credit to a customer who then defaults on the loan.
- The cost of misclassifying an individual with a high risk of heart disease as low risk is probably much higher than the cost of misclassifying a low-risk individual as high-risk.
- The cost of sending a mass mailing to someone who isn't likely to respond is probably fairly low, while the cost of not sending the mailing to someone who is likely to respond is relatively higher (in terms of lost revenue).

Misclassification Costs and Value Labels

This dialog box is not available unless at least two values of the categorical dependent variable have defined value labels.

To Specify Misclassification Costs

1. In the main Decision Tree dialog box, select a categorical (nominal, ordinal) dependent variable with two or more defined value labels.
2. Click **Options**.
3. Click the **Misclassification Costs** tab.
4. Click **Custom**.
5. Enter one or more misclassification costs in the grid. Values must be non-negative. (Correct classifications, represented on the diagonal, are always 0.)

Fill Matrix. In many instances, you may want costs to be symmetric—that is, the cost of misclassifying A as B is the same as the cost of misclassifying B as A. The following controls can make it easier to specify a symmetric cost matrix:

- **Duplicate Lower Triangle.** Copies values in the lower triangle of the matrix (below the diagonal) into the corresponding upper-triangular cells.
- **Duplicate Upper Triangle.** Copies values in the upper triangle of the matrix (above the diagonal) into the corresponding lower-triangular cells.
- **Use Average Cell Values.** For each cell in each half of the matrix, the two values (upper- and lower-triangular) are averaged and the average replaces both values. For example, if the cost of misclassifying A as B is 1 and the cost of misclassifying B as A is 3, then this control replaces both of those values with the average (1+3)/2 = 2.

Profits

For categorical dependent variables, you can assign revenue and expense values to levels of the dependent variable.

- Profit is computed as revenue minus expense.
- Profit values affect average profit and ROI (return on investment) values in gains tables. They do not affect the basic tree model structure.

- Revenue and expense values must be numeric and must be specified for all categories of the dependent variable displayed in the grid.

Profits and Value Labels

This dialog box requires defined value labels for the dependent variable. It is not available unless at least two values of the categorical dependent variable have defined value labels.

To Specify Profits

1. In the main Decision Tree dialog box, select a categorical (nominal, ordinal) dependent variable with two or more defined value labels.
2. Click **Options**.
3. Click the **Profits** tab.
4. Click **Custom**.
5. Enter revenue and expense values for all dependent variable categories listed in the grid.

Prior Probabilities

For CRT and QUEST trees with categorical dependent variables, you can specify prior probabilities of group membership. **Prior probabilities** are estimates of the overall relative frequency for each category of the dependent variable prior to knowing anything about the values of the independent (predictor) variables. Using prior probabilities helps to correct any tree growth caused by data in the sample that is not representative of the entire population.

Obtain from training sample (empirical priors). Use this setting if the distribution of dependent variable values in the data file is representative of the population distribution. If you are using split-sample validation, the distribution of cases in the training sample is used.

Note: Since cases are randomly assigned to the training sample in split-sample validation, you won't know the actual distribution of cases in the training sample in advance. See the topic "Validation" on page 4 for more information.

Equal across categories. Use this setting if categories of the dependent variable are represented equally in the population. For example, if there are four categories, approximately 25% of the cases are in each category.

Custom. Enter a non-negative value for each category of the dependent variable listed in the grid. The values can be proportions, percentages, frequency counts, or any other values that represent the distribution of values across categories.

Adjust priors using misclassification costs. If you define custom misclassification costs, you can adjust prior probabilities based on those costs. See the topic "Misclassification Costs" on page 8 for more information.

Profits and Value Labels

This dialog box requires defined value labels for the dependent variable. It is not available unless at least two values of the categorical dependent variable have defined value labels.

To Specify Prior Probabilities

1. In the main Decision Tree dialog box, select a categorical (nominal, ordinal) dependent variable with two or more defined value labels.
2. For the growing method, select **CRT** or **QUEST**.
3. Click **Options**.
4. Click the **Prior Probabilities** tab.

Scores

For CHAID and Exhaustive CHAID with an ordinal dependent variable, you can assign custom scores to each category of the dependent variable. Scores define the order of and distance between categories of

the dependent variable. You can use scores to increase or decrease the relative distance between ordinal values or to change the order of the values.

- **Use ordinal rank for each category.** The lowest category of the dependent variable is assigned a score of 1, the next highest category is assigned a score of 2, and so on. This is the default.
- **Custom.** Enter a numeric score value for each category of the dependent variable listed in the grid.

Example

Table 3. Custom score values		
Value Label	**Original Value**	**Score**
Unskilled	1	1
Skilled manual	2	4
Clerical	3	4.5
Professional	4	7
Management	5	6

- The scores increase the relative distance between *Unskilled* and *Skilled manual* and decrease the relative distance between *Skilled manual* and *Clerical*.
- The scores reverse the order of *Management* and *Professional*.

Scores and Value Labels

This dialog box requires defined value labels for the dependent variable. It is not available unless at least two values of the categorical dependent variable have defined value labels.

To Specify Scores

1. In the main Decision Tree dialog box, select an ordinal dependent variable with two or more defined value labels.
2. For the growing method, select **CHAID** or **Exhaustive CHAID**.
3. Click **Options**.
4. Click the **Scores** tab.

Missing Values

The Missing Values tab controls the handling of nominal, user-missing, independent (predictor) variable values.

- Handling of ordinal and scale user-missing independent variable values varies between growing methods.
- Handling of nominal dependent variables is specified in the Categories dialog box. See the topic "Selecting Categories" on page 3 for more information.
- For ordinal and scale dependent variables, cases with system-missing or user-missing dependent variable values are always excluded.

Treat as missing values. User-missing values are treated like system-missing values. The handling of system-missing values varies between growing methods.

Treat as valid values. User-missing values of nominal independent variables are treated as ordinary values in tree growing and classification.

Method-Dependent Rules

If some, but not all, independent variable values are system- or user-missing:

- For CHAID and Exhaustive CHAID, system- and user-missing independent variable values are included in the analysis as a single, combined category. For scale and ordinal independent variables, the

algorithms first generate categories using valid values and then decide whether to merge the missing category with its most similar (valid) category or keep it as a separate category.

- For CRT and QUEST, cases with missing independent variable values are excluded from the tree-growing process but are classified using surrogates if surrogates are included in the method. If nominal user-missing values are treated as missing, they are also handled in this manner. See the topic "Surrogates" on page 7 for more information.

To Specify Nominal, Independent User-Missing Treatment

1. In the main Decision Tree dialog box, select at least one nominal independent variable.
2. Click **Options**.
3. Click the **Missing Values** tab.

Saving Model Information

You can save information from the model as variables in the working data file, and you can also save the entire model in XML (PMML) format to an external file.

Saved Variables

Terminal node number. The terminal node to which each case is assigned. The value is the tree node number.

Predicted value. The class (group) or value for the dependent variable predicted by the model.

Predicted probabilities. The probability associated with the model's prediction. One variable is saved for each category of the dependent variable. Not available for scale dependent variables.

Sample assignment (training/testing). For split-sample validation, this variable indicates whether a case was used in the training or testing sample. The value is 1 for the training sample and 0 for the testing sample. Not available unless you have selected split-sample validation. See the topic "Validation" on page 4 for more information.

Export Tree Model as XML

You can save the entire tree model in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Training sample. Writes the model to the specified file. For split-sample validated trees, this is the model for the training sample.

Test sample. Writes the model for the test sample to the specified file. Not available unless you have selected split-sample validation.

Output

Available output options depend on the growing method, the measurement level of the dependent variable, and other settings.

Tree Display

You can control the initial appearance of the tree or completely suppress the tree display.

Tree. By default, the tree diagram is included in the output displayed in the Viewer. Deselect this option to exclude the tree diagram from the output.

Display. These options control the initial appearance of the tree diagram in the Viewer. All of these attributes can also be modified by editing the generated tree.

- **Orientation.** The tree can be displayed top down with the root node at the top, left to right, or right to left.
- **Node contents.** Nodes can display tables, charts, or both. For categorical dependent variables, tables display frequency counts and percentages, and the charts are bar charts. For scale dependent variables,

tables display means, standard deviations, number of cases, and predicted values, and the charts are histograms.

- **Scale.** By default, large trees are automatically scaled down in an attempt to fit the tree on the page. You can specify a custom scale percentage of up to 200%.
- **Independent variable statistics.** For CHAID and Exhaustive CHAID, statistics include F value (for scale dependent variables) or chi-square value (for categorical dependent variables) as well as significance value and degrees of freedom. For CRT, the improvement value is shown. For QUEST, F, significance value, and degrees of freedom are shown for scale and ordinal independent variables; for nominal independent variables, chi-square, significance value, and degrees of freedom are shown.
- **Node definitions.** Node definitions display the value(s) of the independent variable used at each node split.

Tree in table format. Summary information for each node in the tree, including parent node number, independent variable statistics, independent variable value(s) for the node, mean and standard deviation for scale dependent variables, or counts and percentages for categorical dependent variables.

To Control the Initial Tree Display

1. In the main Decision Tree dialog, click **Output**.
2. Click the **Tree** tab.

Statistics

Available statistics tables depend on the measurement level of the dependent variable, the growing method, and other settings.

Model

Summary. The summary includes the method used, the variables included in the model, and the variables specified but not included in the model.

Risk. Risk estimate and its standard error. A measure of the tree's predictive accuracy.

- For categorical dependent variables, the risk estimate is the proportion of cases incorrectly classified after adjustment for prior probabilities and misclassification costs.
- For scale dependent variables, the risk estimate is within-node variance.

Classification table. For categorical (nominal, ordinal) dependent variables, this table shows the number of cases classified correctly and incorrectly for each category of the dependent variable. Not available for scale dependent variables.

Cost, prior probability, score, and profit values. For categorical dependent variables, this table shows the cost, prior probability, score, and profit values used in the analysis. Not available for scale dependent variables.

Independent Variables

Importance to model. For the CRT growing method, ranks each independent (predictor) variable according to its importance to the model. Not available for QUEST or CHAID methods.

Surrogates by split. For the CRT and QUEST growing methods, if the model includes surrogates, lists surrogates for each split in the tree. Not available for CHAID methods. See the topic "Surrogates" on page 7 for more information.

Node Performance

Summary. For scale dependent variables, the table includes the node number, the number of cases, and the mean value of the dependent variable. For categorical dependent variables with defined profits, the table includes the node number, the number of cases, the average profit, and the ROI (return on investment) values. Not available for categorical dependent variables without defined profits. See the topic "Profits" on page 8 for more information.

By target category. For categorical dependent variables with defined target categories, the table includes the percentage gain, the response percentage, and the index percentage (lift) by node or percentile group.

A separate table is produced for each target category. Not available for scale dependent variables or categorical dependent variables without defined target categories. See the topic "Selecting Categories" on page 3 for more information.

Rows. The node performance tables can display results by terminal nodes, percentiles, or both. If you select both, two tables are produced for each target category. Percentile tables display cumulative values for each percentile, based on sort order.

Percentile increment. For percentile tables, you can select the percentile increment: 1, 2, 5, 10, 20, or 25.

Display cumulative statistics. For terminal node tables, displays additional columns in each table with cumulative results.

To Select Statistics Output

1. In the main Decision Tree dialog, click **Output**.
2. Click the **Statistics** tab.

Charts

Available charts depend on the measurement level of the dependent variable, the growing method, and other settings.

Independent variable importance to model. Bar chart of model importance by independent variable (predictor). Available only with the CRT growing method.

Node Performance

Gain. Gain is the percentage of total cases in the target category in each node, computed as: (node target n / total target n) x 100. The gains chart is a line chart of cumulative percentile gains, computed as: (cumulative percentile target n / total target n) x 100. A separate line chart is produced for each target category. Available only for categorical dependent variables with defined target categories. See the topic "Selecting Categories" on page 3 for more information.

The gains chart plots the same values that you would see in the *Gain Percent* column in the gains for percentiles table, which also reports cumulative values.

Index. Index is the ratio of the node response percentage for the target category compared to the overall target category response percentage for the entire sample. The index chart is a line chart of cumulative percentile index values. Available only for categorical dependent variables. Cumulative percentile index is computed as: (cumulative percentile response percent / total response percent) x 100. A separate chart is produced for each target category, and target categories must be defined.

The index chart plots the same values that you would see in the *Index* column in the gains for percentiles table.

Response. The percentage of cases in the node in the specified target category. The response chart is a line chart of cumulative percentile response, computed as: (cumulative percentile target n / cumulative percentile total n) x 100. Available only for categorical dependent variables with defined target categories.

The response chart plots the same values that you would see in the *Response* column in the gains for percentiles table.

Mean. Line chart of cumulative percentile mean values for the dependent variable. Available only for scale dependent variables.

Average profit. Line chart of cumulative average profit. Available only for categorical dependent variables with defined profits. See the topic "Profits" on page 8 for more information.

The average profit chart plots the same values that you would see in the *Profit* column in the gain summary for percentiles table.

Return on investment (ROI). Line chart of cumulative ROI (return on investment). ROI is computed as the ratio of profits to expenses. Available only for categorical dependent variables with defined profits.

The ROI chart plots the same values that you would see in the *ROI* column in the gain summary for percentiles table.

Percentile increment. For all percentile charts, this setting controls the percentile increments displayed on the chart: 1, 2, 5, 10, 20, or 25.

To Select Chart Output

1. In the main Decision Tree dialog, click **Output**.
2. Click the **Plots** tab.

Selection and Scoring Rules

The Rules tab provides the ability to generate selection or classification/prediction rules in the form of command syntax, SQL, or simple (plain English) text. You can display these rules in the Viewer and/or save the rules to an external file.

Syntax. Controls the form of the selection rules in both output displayed in the Viewer and selection rules saved to an external file.

- **IBM SPSS Statistics.** Command syntax language. Rules are expressed as a set of commands that define a filter condition that can be used to select subsets of cases or as COMPUTE statements that can be used to score cases.
- **SQL.** Standard SQL rules are generated to select or extract records from a database or assign values to those records. The generated SQL rules do not include any table names or other data source information.
- **Simple text.** Plain English pseudo-code. Rules are expressed as a set of logical "if...then" statements that describe the model's classifications or predictions for each node. Rules in this form can use defined variable and value labels or variable names and data values.

Type. For IBM SPSS Statistics and SQL rules, controls the type of rules generated: selection or scoring rules.

- **Assign values to cases.** The rules can be used to assign the model's predictions to cases that meet node membership criteria. A separate rule is generated for each node that meets the node membership criteria.
- **Select cases.** The rules can be used to select cases that meet node membership criteria. For IBM SPSS Statistics and SQL rules, a single rule is generated to select all cases that meet the selection criteria.

Include surrogates in IBM SPSS Statistics and SQL rules. For CRT and QUEST, you can include surrogate predictors from the model in the rules. Rules that include surrogates can be quite complex. In general, if you just want to derive conceptual information about your tree, exclude surrogates. If some cases have incomplete independent variable (predictor) data and you want rules that mimic your tree, include surrogates. See the topic "Surrogates" on page 7 for more information.

Nodes. Controls the scope of the generated rules. A separate rule is generated for each node included in the scope.

- **All terminal nodes.** Generates rules for each terminal node.
- **Best terminal nodes.** Generates rules for the top *n* terminal nodes based on index values. If the number exceeds the number of terminal nodes in the tree, rules are generated for all terminal nodes. (See note below.)
- **Best terminal nodes up to a specified percentage of cases.** Generates rules for terminal nodes for the top *n* percentage of cases based on index values. (See note below.)
- **Terminal nodes whose index value meets or exceeds a cutoff value.** Generates rules for all terminal nodes with an index value greater than or equal to the specified value. An index value greater than 100 means that the percentage of cases in the target category in that node exceeds the percentage in the root node. (See note below.)
- **All nodes.** Generates rules for all nodes.

Note 1: Node selection based on index values is available only for categorical dependent variables with defined target categories. If you have specified multiple target categories, a separate set of rules is generated for each target category.

Note 2: For IBM SPSS Statistics and SQL rules for selecting cases (not rules for assigning values), **All nodes** and **All terminal nodes** will effectively generate a rule that selects all cases used in the analysis.

Export rules to a file. Saves the rules in an external text file.

You can also generate and save selection or scoring rules interactively, based on selected nodes in the final tree model. See the topic "Case Selection and Scoring Rules" on page 17 for more information.

Note: If you apply rules in the form of command syntax to another data file, that data file must contain variables with the same names as the independent variables included in the final model, measured in the same metric, with the same user-defined missing values (if any).

To Specify Selection or Scoring Rules

1. In the main Decision Tree dialog, click **Output**.
2. Click the **Rules** tab.

Tree Editor

With the Tree Editor, you can:

- Hide and show selected tree branches.
- Control display of node content, statistics displayed at node splits, and other information.
- Change node, background, border, chart, and font colors.
- Change font style and size.
- Change tree alignment.
- Select subsets of cases for further analysis based on selected nodes.
- Create and save rules for selecting or scoring cases based on selected nodes.

To edit a tree model:

1. Double-click the tree model in the Viewer window.

   or

2. From the Edit menu or the right-click pop-up menu choose:

   Edit

Hiding and Showing Nodes

To hide (collapse) all the child nodes in a branch beneath a parent node:

1. Click the minus sign (–) in the small box below the lower right corner of the parent node.

   All nodes beneath the parent node on that branch will be hidden.

   To show (expand) the child nodes in a branch beneath a parent node:

2. Click the plus sign (+) in the small box below the lower right corner of the parent node.

Note: Hiding the child nodes on a branch is not the same as pruning a tree. If you want a pruned tree, you must request pruning before you create the tree, and pruned branches are not included in the final tree. See the topic "Pruning Trees" on page 7 for more information.

Selecting Multiple Nodes

You can select cases, generate scoring and selections rules, and perform other actions based on the currently selected node(s). To select multiple nodes:

1. Click a node you want to select.
2. Ctrl-click the other nodes you want to select.

You can multiple-select sibling nodes and/or parent nodes in one branch and child nodes in another branch. You cannot, however, use multiple selection on a parent node and a child/descendant of the same node branch.

Working with Large Trees

Tree models may sometimes contain so many nodes and branches that it is difficult or impossible to view the entire tree at full size. There are a number of features that you may find useful when working with large trees:

- **Tree map.** You can use the tree map, a much smaller, simplified version of the tree, to navigate the tree and select nodes. See the topic "Tree Map" on page 16 for more information.
- **Scaling.** You can zoom out and zoom in by changing the scale percentage for the tree display. See the topic "Scaling the Tree Display" on page 16 for more information.
- **Node and branch display.** You can make a tree more compact by displaying only tables or only charts in the nodes and/or suppressing the display of node labels or independent variable information. See the topic "Controlling Information Displayed in the Tree" on page 17 for more information.

Tree Map

The tree map provides a compact, simplified view of the tree that you can use to navigate the tree and select nodes.

To use the tree map window:

1. From the Tree Editor menus choose:

    View > Tree Map

- The currently selected node is highlighted in both the Tree Model Editor and the tree map window.
- The portion of the tree that is currently in the Tree Model Editor view area is indicated with a red rectangle in the tree map. Right-click and drag the rectangle to change the section of the tree displayed in the view area.
- If you select a node in the tree map that isn't currently in the Tree Editor view area, the view shifts to include the selected node.
- Multiple node selection works the same in the tree map as in the Tree Editor: Ctrl-click to select multiple nodes. You cannot use multiple selection on a parent node and a child/descendant of the same node branch.

Scaling the Tree Display

By default, trees are automatically scaled to fit in the Viewer window, which can result in some trees that are initially very difficult to read. You can select a preset scale setting or enter your own custom scale value of between 5% and 200%.

To change the scale of the tree:

1. Select a scale percentage from the drop-down list on the toolbar, or enter a custom percentage value.

    or

2. From the Tree Editor menus choose:

    View > Scale...

You can also specify a scale value before you create the tree model. See the topic "Output" on page 11 for more information.

Node Summary Window

The node summary window provides a larger view of the selected nodes. You can also use the summary window to view, apply, or save selection or scoring rules based on the selected nodes.

- Use the View menu in the node summary window to switch between views of a summary table, chart, or rules.
- Use the Rules menu in the node summary window to select the type of rules you want to see. See the topic "Case Selection and Scoring Rules" on page 17 for more information.
- All views in the node summary window reflect a combined summary for all selected nodes.

To use the node summary window:

1. Select the nodes in the Tree Editor. To select multiple nodes, use Ctrl-click.
2. From the menus choose:

   View > **Summary**

Controlling Information Displayed in the Tree

The Options menu in the Tree Editor allows you to control the display of node contents, independent variable (predictor) names and statistics, node definitions, and other settings. Many of these settings can be also be controlled from the toolbar.

Changing Tree Colors and Text Fonts

You can change the following colors in the tree:

- Node border, background, and text color
- Branch color and branch text color
- Tree background color
- Predicted category highlight color (categorical dependent variables)
- Node chart colors

You can also change the type font, style, and size for all text in the tree.

Note: You cannot change color or font attributes for individual nodes or branches. Color changes apply to all elements of the same type, and font changes (other than color) apply to all chart elements.

To change colors and text font attributes:

1. Use the toolbar to change font attributes for the entire tree or colors for different tree elements. (ToolTips describe each control on the toolbar when you put the mouse cursor on the control.)

   or
2. Double-click anywhere in the Tree Editor to open the Properties window, or from the menus choose:

   View > **Properties**
3. For border, branch, node background, predicted category, and tree background, click the **Color** tab.
4. For font colors and attributes, click the **Text** tab.
5. For node chart colors, click the **Node Charts** tab.

Case Selection and Scoring Rules

You can use the Tree Editor to:

- Select subsets of cases based on the selected node(s). See the topic "Filtering Cases" on page 18 for more information.
- Generate case selection rules or scoring rules in IBM SPSS Statistics command syntax or SQL format. See the topic "Saving Selection and Scoring Rules" on page 18 for more information.

You can also automatically save rules based on various criteria when you run the Decision Tree procedure to create the tree model. See the topic "Selection and Scoring Rules" on page 14 for more information.

Filtering Cases

If you want to know more about the cases in a particular node or group of nodes, you can select a subset of cases for further analysis based on the selected nodes.

1. Select the nodes in the Tree Editor. To select multiple nodes, use Ctrl-click.
2. From the menus choose:

   Rules > **Filter Cases...**

3. Enter a filter variable name. Cases from the selected nodes will receive a value of 1 for this variable. All other cases will receive a value of 0 and will be excluded from subsequent analysis until you change the filter status.
4. Click **OK**.

Saving Selection and Scoring Rules

You can save case selection or scoring rules in an external file and then apply those rules to a different data source. The rules are based on the selected nodes in the Tree Editor.

Syntax. Controls the form of the selection rules in both output displayed in the Viewer and selection rules saved to an external file.

- **IBM SPSS Statistics.** Command syntax language. Rules are expressed as a set of commands that define a filter condition that can be used to select subsets of cases or as COMPUTE statements that can be used to score cases.
- **SQL.** Standard SQL rules are generated to select/extract records from a database or assign values to those records. The generated SQL rules do not include any table names or other data source information.

Type. You can create selection or scoring rules.

- **Select cases.** The rules can be used to select cases that meet node membership criteria. For IBM SPSS Statistics and SQL rules, a single rule is generated to select all cases that meet the selection criteria.
- **Assign values to cases.** The rules can be used to assign the model's predictions to cases that meet node membership criteria. A separate rule is generated for each node that meets the node membership criteria.

Include surrogates. For CRT and QUEST, you can include surrogate predictors from the model in the rules. Rules that include surrogates can be quite complex. In general, if you just want to derive conceptual information about your tree, exclude surrogates. If some cases have incomplete independent variable (predictor) data and you want rules that mimic your tree, include surrogates. See the topic "Surrogates" on page 7 for more information.

To save case selection or scoring rules:

1. Select the nodes in the Tree Editor. To select multiple nodes, use Ctrl-click.
2. From the menus choose:

   Rules > **Export...**

3. Select the type of rules you want and enter a filename.

Note: If you apply rules in the form of command syntax to another data file, that data file must contain variables with the same names as the independent variables included in the final model, measured in the same metric, with the same user-defined missing values (if any).

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© Copyright IBM Corp. 2021. Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. 1989 - 2021. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

trees *(continued)*
  charts 13
  colors 17
  controlling node size 5
  controlling tree display 11, 17
  crossvalidation 4
  CRT method 6
  editing 15
  fonts 17
  generating rules 14, 17
  hiding branches and nodes 15
  index values 12
  intervals for scale independent variables 6
  limiting number of levels 5
  misclassification costs 8
  misclassification table 12
  missing values 10
  node chart colors 17
  predictor importance 12
  prior probability 9
  profits 8
  pruning 7
  risk estimates 12
  saving model variables 11
  scaling tree display 16
  scores 9
  selecting multiple nodes 15
  showing and hiding branch statistics 11
  split-sample validation 4
  terminal node statistics 12
  text attributes 17
  tree contents in a table 11
  tree map 16
  tree orientation 11
  working with large trees 16
twoing 6

V

validation
  trees 4

W

weighting cases
  fractional weights in decision trees 1

IBM ®