

TIBCO Statistica™

Quick Reference

Software Release 13.5

November 2018

Two-Second Advantage®



Important Information

SOME TIBCO SOFTWARE EMBEDS OR BUNDLES OTHER TIBCO SOFTWARE. USE OF SUCH EMBEDDED OR BUNDLED TIBCO SOFTWARE IS SOLELY TO ENABLE THE FUNCTIONALITY (OR PROVIDE LIMITED ADD-ON FUNCTIONALITY) OF THE LICENSED TIBCO SOFTWARE. THE EMBEDDED OR BUNDLED SOFTWARE IS NOT LICENSED TO BE USED OR ACCESSED BY ANY OTHER TIBCO SOFTWARE OR FOR ANY OTHER PURPOSE.

USE OF TIBCO SOFTWARE AND THIS DOCUMENT IS SUBJECT TO THE TERMS AND CONDITIONS OF A LICENSE AGREEMENT FOUND IN EITHER A SEPARATELY EXECUTED SOFTWARE LICENSE AGREEMENT, OR, IF THERE IS NO SUCH SEPARATE AGREEMENT, THE CLICKWRAP END USER LICENSE AGREEMENT WHICH IS DISPLAYED DURING DOWNLOAD OR INSTALLATION OF THE SOFTWARE (AND WHICH IS DUPLICATED IN THE LICENSE FILE) OR IF THERE IS NO SUCH SOFTWARE LICENSE AGREEMENT OR CLICKWRAP END USER LICENSE AGREEMENT, THE LICENSE(S) LOCATED IN THE "LICENSE" FILE(S) OF THE SOFTWARE. USE OF THIS DOCUMENT IS SUBJECT TO THOSE TERMS AND CONDITIONS, AND YOUR USE HEREOF SHALL CONSTITUTE ACCEPTANCE OF AND AN AGREEMENT TO BE BOUND BY THE SAME.

ANY SOFTWARE ITEM IDENTIFIED AS THIRD PARTY LIBRARY IS AVAILABLE UNDER SEPARATE SOFTWARE LICENSE TERMS AND IS NOT PART OF A TIBCO PRODUCT. AS SUCH, THESE SOFTWARE ITEMS ARE NOT COVERED BY THE TERMS OF YOUR AGREEMENT WITH TIBCO, INCLUDING ANY TERMS CONCERNING SUPPORT, MAINTENANCE, WARRANTIES, AND INDEMNITIES. DOWNLOAD AND USE OF THESE ITEMS IS SOLELY AT YOUR OWN DISCRETION AND SUBJECT TO THE LICENSE TERMS APPLICABLE TO THEM. BY PROCEEDING TO DOWNLOAD, INSTALL OR USE ANY OF THESE ITEMS, YOU ACKNOWLEDGE THE FOREGOING DISTINCTIONS BETWEEN THESE ITEMS AND TIBCO PRODUCTS.

This document contains confidential information that is subject to U.S. and international copyright laws and treaties. No part of this document may be reproduced in any form without the written authorization of TIBCO Software Inc.

TIBCO, Better Decisioning, Data Health Check, Data Science, Decisioning Platform, Electronic Statistics Textbook, Information Bus, Live Score, Making the World Productive, Messaging Appliance, Predictive Claims Flow, Process Data Explorer, Process Tree Viewer, Rendezvous, Statistica, Statsoft, Statsoft Iberica, The Power of Now, TIB, TIBCO Rendezvous, and Two-Second Advantage are either registered trademarks or trademarks of TIBCO Software Inc. in the United States and/or other countries.

Enterprise Java Beans (EJB), Java Platform Enterprise Edition (Java EE), Java 2 Platform Enterprise Edition (J2EE), and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle Corporation in the U.S. and other countries.

All other product and company names and marks mentioned in this document are the property of their respective owners and are mentioned for identification purposes only.

THIS SOFTWARE MAY BE AVAILABLE ON MULTIPLE OPERATING SYSTEMS. HOWEVER, NOT ALL OPERATING SYSTEM PLATFORMS FOR A SPECIFIC SOFTWARE VERSION ARE RELEASED AT THE SAME TIME. SEE THE README FILE FOR THE AVAILABILITY OF THIS SOFTWARE VERSION ON A SPECIFIC OPERATING SYSTEM PLATFORM.

THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT.

THIS DOCUMENT COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THIS DOCUMENT. TIBCO SOFTWARE INC. MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED IN THIS DOCUMENT AT ANY TIME.

THE CONTENTS OF THIS DOCUMENT MAY BE MODIFIED AND/OR QUALIFIED, DIRECTLY OR INDIRECTLY, BY OTHER DOCUMENTATION WHICH ACCOMPANIES THIS SOFTWARE, INCLUDING BUT NOT LIMITED TO ANY RELEASE NOTES AND "READ ME" FILES.

Contents

TIBCO Documentation and Support Services	4
Chapter One Statistica: A general overview of features.....	5
Analytic Facilities	6
Unique Features	6
The General Philosophy of the Statistica Approach	7
Step-by-Step Examples	10
Using Statistica Data Miner Recipes (SDMR)	52
Data Management	59
Enterprise Installations	79
Chapter Three: User Interface	101
General Features	102
Chapter Four Six Channels for Output from Analyses	116
Overview	117
Chapter Five Statistica Documents	133
Workbooks	134
Chapter Six Graphs	148
Chapter Seven Customizing Statistica	166
Chapter Eight Statistica Visual Basic	171
Chapter Nine Statistica Query	186
Chapter Ten Programming Statistica from .NET	190
Appendix A Statistica Enterprise Server	194

TIBCO Documentation and Support Services

How to Access TIBCO Documentation

Documentation for TIBCO products is available on the TIBCO Product Documentation website, mainly in HTML and PDF formats.

The TIBCO Product Documentation website is updated frequently and is more current than any other documentation included with the product. To access the latest documentation, visit <https://docs.tibco.com>.

Product-Specific Documentation

Documentation for TIBCO Statistica™ is available on the [TIBCO Statistica™ Product Documentation page](#).

The following documents for this product can be found on the TIBCO Documentation site:

- *TIBCO Statistica™ Installation*
- *TIBCO Statistica™ Quick Reference*
- *TIBCO Statistica™ Release Notes*
- *TIBCO Statistica™ Admin Guide*
- *TIBCO Statistica™ Product Traceability*
- *TIBCO Statistica™ Configuration for Windows Server 2012*
- *TIBCO Statistica™ Data Entry Admin Guide*
- *TIBCO Statistica™ Server Administrator's Guide*
- *TIBCO Statistica™ Options Configuration*

How to Contact TIBCO Support

You can contact TIBCO Support in the following ways:

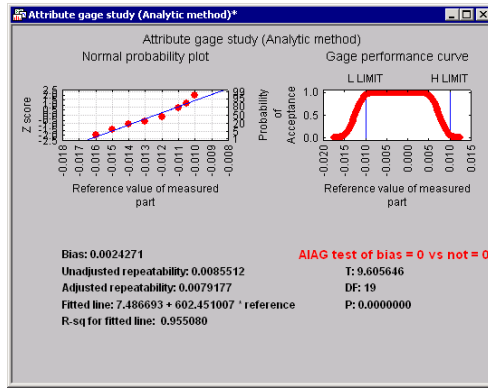
- For an overview of TIBCO Support, visit <http://www.tibco.com/services/support>.
- For accessing the Support Knowledge Base and getting personalized content about products you are interested in, visit the TIBCO Support portal at <https://support.tibco.com>.
- For creating a Support case, you must have a valid maintenance or support contract with TIBCO. You also need a user name and password to log in to <https://support.tibco.com>. If you do not have a user name, you can request one by clicking Register on the website.

How to Join TIBCO Community

TIBCO Community is the official channel for TIBCO customers, partners, and employee subject matter experts to share and access their collective experience. TIBCO Community offers access to Q&A forums, product wikis, and best practices. It also offers access to extensions, adapters, solution accelerators, and tools that extend and enable customers to gain full value from TIBCO products. In addition, users can submit and vote on feature requests from within the [TIBCO Ideas Portal](#). For a free registration, go to <https://community.tibco.com>.

Chapter One

Statistica: A general overview of features



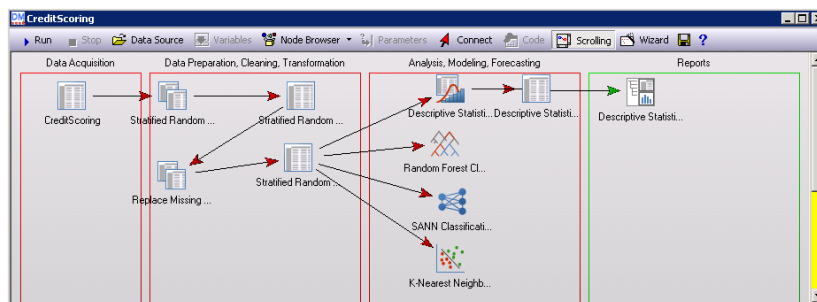
All important results to determine the bias and repeatability (of measurements) of the attribute gage are summarized on a single page.

Example 6: Statistica Data Miner

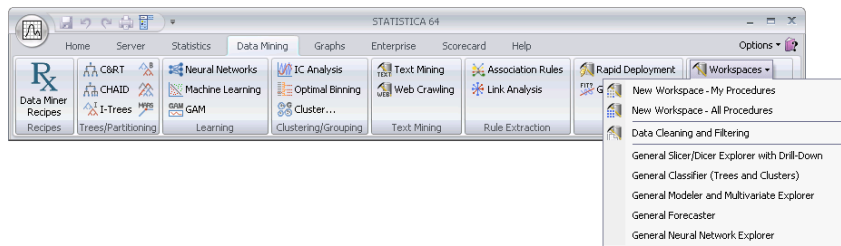
Statistica Data Miner (SDM) is a comprehensive system for predictive modeling that offers a wide variety of analytic techniques and model building, validation, and model deployment options.

The default, and perhaps the industry standard, type of user interface provided in SDM follows the general interactive data mining workspace approach that enables users to build models by dragging icons representing steps of data acquisition, data preparation, modeling, and deployment and connect them with arrows.

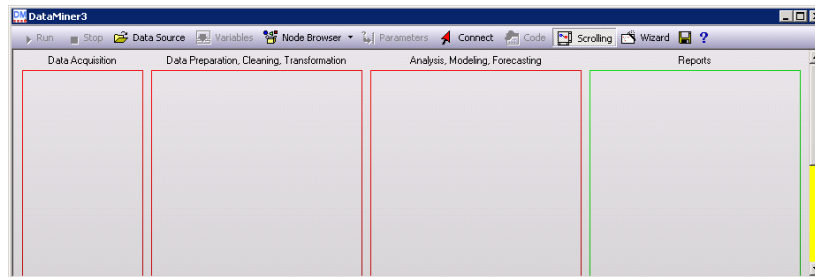
The workspace user interface option in SDM represents a powerful alternative to the traditional interactive data analysis user interface, and it can be used not only as a tool for developing and testing predictive data mining modes, but also as a powerful general tool to be used for visual programming of analytic workflows for many types of analyses.







1. To open a new (blank) data mining workspace, select the **Data Mining** tab.
2. In the Tools group, click Workspaces and from the menu, select either **New Workspace-My Procedures** or **New Workspace - All Procedures**.



A blank data mining workspace will be displayed.



3. Now, click  **Data Source** on the toolbar to display the Select Data Source dialog, used to select a data file for analysis.
4. Next, the **Select dependent variables and predictors** dialog is displayed. Click the  **Variables** button to display the variable selection dialog, used to specify the dependent variables and predictors.
5. Then, click  **Node Browser** to create analytic nodes, and connect them with  **Connect** arrows to specify the desired project workflow.

The following section includes a step-by-step example of Data Miner Recipes – an innovative user interface for data mining introduced by Statistica – which offers a powerful alternative to the workspace-based approach to model building, and can be used by both novices and advanced analysts.

Overview

This example pertains to Statistica Data Miner Recipes, a Statistica product that offers a wide selection of methods for predictive data mining.

A general trend in data mining is the increasing emphasis on solutions based on simple analytic processes rather than the creation of ever-more sophisticated general analytic tools. Statistica Data Miner Recipes (SDMR) offers an easy-to-use alternative to the traditional data miner workspace user interface for building predictive data mining models.

This approach provides an intuitive graphical interface to enable those with limited data mining experience to execute a recipe-like step-by-step analytic process. With these intuitive dialogs, you can perform various data mining tasks such as regression, classification, and clustering. Other recipes can be built quickly as custom solutions.

Completed recipes can be saved and deployed as project files to score new data. The project files can be generated as C/C++ language or PMML script, or sent to Statistica Enterprise.

The SDMR user interface can also be used by advanced analysts to automate and store specific data mining algorithms.

SDMR spans the entire data mining process – from querying external databases to the final deployment of solutions – and, in general, consists of the following steps.

1. Identifies the data from which to learn
 - Connects to ODBC or OLEDB compliant databases
 - Connects to Statistica data files
2. Cleans data and removes the redundant predictors
 - Flexible and efficient methods for sampling the data (simple, stratified, systematic, etc.)
 - More flexible ways to identify and recode the missing data
 - Identification of outliers
 - Transform the data prior to performing the subsequent steps
 - Identify and eliminate redundant predictors
3. Identifies important predictors from a large pool of predictors that are strongly related to the dependent (outcome or target) variable of interest
 - Feature selection for very large data sets (for example, thousands of variables)
 - Detection of important interactions among the predictors by using tree-based methods
4. Generates a pool of eligible models
 - Leverage the comprehensive selection of cutting edge techniques for predictive data mining available in SDMR
 - Offload computationally expensive tasks to Statistica Enterprise Server, freeing your local computer for other tasks
5. Performs automatic competitive evaluation of models to identify the optimum model with respect to performance and complexity
6. Deploys the model to score new data using the inbuilt efficient deployment engine

Statistica Data Miner Recipes provides the solution that maps the steps of the data mining workflow into a results-oriented user interface. From data cleaning to model validation, SDMR guides your analysis from start to finish so that you can get actionable results and answers quickly.

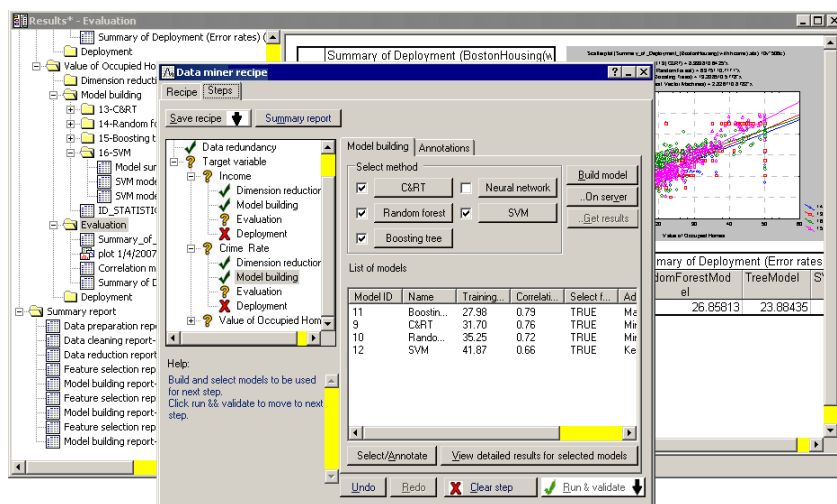
At the same time, SDMR still applies the most comprehensive collection of data mining algorithms in a single package without requiring the user to know the details of those algorithms.

Statistica Data Miner Recipes contains the largest collection of data mining methods and algorithms in a single package or library. In most general terms, these algorithms borrow insights and methodologies from various domains such as statistics, engineering, artificial intelligence, cognitive science, etc., to learn patterns from data that can be used to make predictions (about insurance or credit risk, process or product quality, equipment failure, medical diagnoses, and so on).

In practice, specific domains and types of data are best analyzed using particular types of methods and algorithms. For example, the data mining techniques that work best for modeling insurance loss data are different from those that work best for predicting emissions from a furnace.

However, there is a typical workflow – from the definition of the data and analysis problem through sampling, model building, and evaluation – that is applicable to all predictive data mining.

Data Miner Recipes enable those without extensive experience with data mining tools to move very quickly from the definition of a problem to tangible and actionable results.



In this approach, you simply follow a recipe-like user interface to complete the necessary steps to move to a solution. In fact, most of these steps are entirely automated so that the only required input is to define the data and variables for the analyses, while the program automatically does the rest – determines learning and testing samples, performs feature selection, tries various data mining algorithms and methods, and evaluates results to select the best data mining model. These computations and analyses can be performed with either the desktop Statistica Data Miner software or, if available, on the Statistica Data Miner Server.

Data Miner Recipes Project Files

When you save a Data Miner Recipes project at any stage of completion, two separate files are created:

- A Data Miner Recipes file with the file name extension `.dmrproj`
- A Statistica Workbook file by the same name, but with the file name extension `.stw`, containing results and detailed information for each step of the recipe

It is important that both files reside in the same file directory. So, if you want to copy a Data Miner Recipe project called `MyDataMinerProject` to a new file directory, email it to a colleague, or check it into the Statistica Document Management System, then both files – `MyDataMinerProject.dmrproj` and `MyDataMinerProject.stw` – must be copied to the new destination.

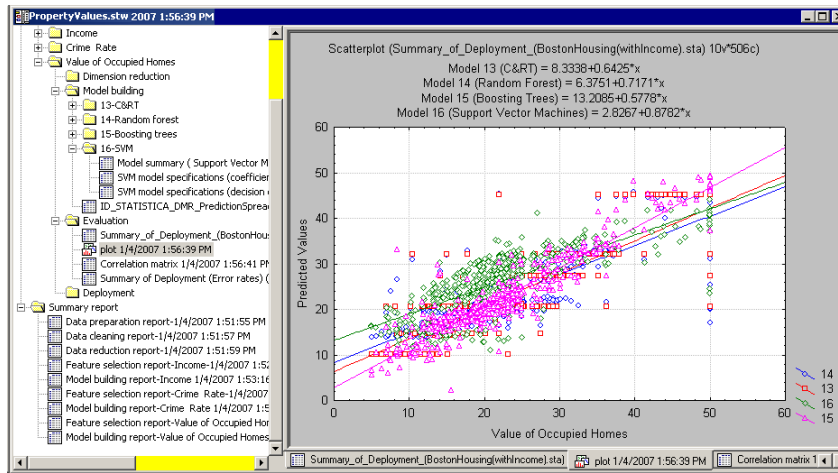
Following are additional details about these two files.

Data Miner Recipes file (.dmrproj). The Data Miner Recipes are XML (extensible markup language) format files that contain all information regarding users' choices (or choices automatically made by the program), including:

- Data file information (or data connection information)
- Variable selections and variable metadata (for example, defining continuous and categorical predictors and outcomes)
- Choices about data preprocessing steps (for example, missing data handling, filtering of duplicate records, transformations, etc.)
- Final variable selections based on the application of feature selection algorithms
- Results from model building and final evaluation and choices of models
- All information necessary to deploy predictive models and to predict new cases (for example, to score databases, compute component scores, inferred sensor values, predicted risk or failure probabilities, etc.)

Therefore, when deploying Data Miner Recipes to the Statistica Enterprise software to automatically compute predicted values in an enterprise application (automated credit scoring, multivariate control charting and failure analysis, etc.), all information necessary to compute predicted values, classifications, or classification probabilities (for example, probability of default, loss) is contained inside these XML format files.

Data Miner Recipes Workbook file (.stw). These files contain detailed information describing the results for each step.



The results stored in this workbook provide complete documentation for the computations and analyses performed as the Data Miner Recipe was (or is in the process of being) completed. Therefore, if the data mining analyses are performed in a regulated (for example, FDA, ISO, etc.) environment, or if data mining is part of an organization's mission critical activities performed under the guidance and in compliance with specific standard operating procedures (SOPs), then it is usually recommended that this file be stored in the Statistica Document Management System along with the Data Miner Recipe project file (.dmrproj).

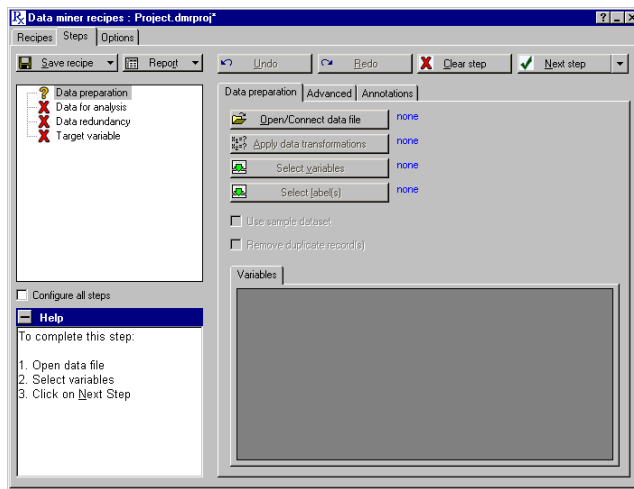
Using Statistica Data Miner Recipes (SDMR)

This example illustrates how quickly and efficiently data mining projects can be completed using Statistica Data Miner Recipes, even if the best solution to the (prediction) problem emerges only after (automatically) comparing the efficacy of various advanced data mining algorithms.

In this example, we will explore the use of SDMR for credit scoring applications. The example is based on the data file CreditScoring.sta, which contains observations on 18 variables for 1,000 past applicants for credit.

Each applicant was rated as good credit (700 cases) or bad credit (300 cases). We want to develop a credit scoring model that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on the values of one or more of the predictor variables. An additional **Train/Test** indicator variable is also included in the data file for validation purposes.

1. In Statistica, select the **Data Mining** tab.
2. In the **Recipes** group, click **Data Miner Recipes** to display the **Data miner recipes** dialog.
3. On the **Recipes** tab, click the **New** button to create a new project. The **Steps** tab will be selected automatically.



The step-node panel is located in the upper-left area of the **Steps** tab. It contains four major nodes: **Data preparation**, **Data for analysis**, **Data redundancy**, and **Target variable**.

Nodes (steps). Each node (or step) can exist in one of four states, depending on whether all required options have been specified. Each state is represented by an icon: a red **X** indicates a wait state, meaning a step cannot be started because it is dependent on a previous step that has not been completed; a yellow **?** indicates a ready state, meaning you are ready to start the step because previous steps have been completed; a green **✓** indicates a completed step. Note that you must click the **Next step** button to change the yellow **?** (ready state) to the green **✓** (completed state).

The change will be made only if the step has been successfully completed (such as all required information has been specified). Lastly, if you have opened a data set and selected variables, and you do not want to proceed step by step through all the options, you can select the **Configure all steps** check box on the **Steps** tab. The steps will now be represented by a navy **•** icon.

You can select any of the steps and modify the options, or you can leave all options at their defaults. Then, click the **Next step** arrow, and from the drop-down list, select **Run to completion**. Statistica Data Miner Recipes will run the analysis and create the model results.

Options tab. The **Options** tab of Statistica Data Miner Recipes is used to set global options for recipes using very large data files. Options include specifications for sampling and for maximum file size to save in the Project Workbook. Since most of these options are applied to the Data preparation step, they should be set prior to starting work on a new recipe. Modifications to the values on this tab apply only to the current recipe unless you click the **Save defaults** button.

Data Preparation

Connecting data. On the **Data preparation** tab, click the **Open/Connect data file** button.

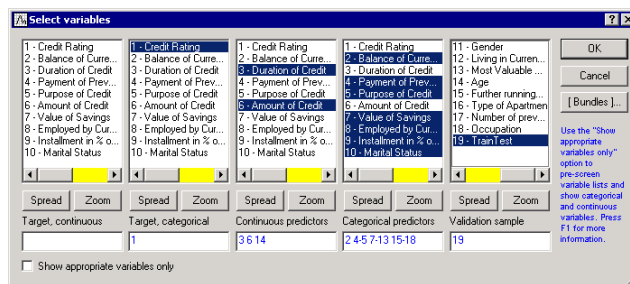
1. In the **Select Data Source** dialog, click the **Files** button to browse to and open the CreditScoring.sta data file (located in the Datasets folder installed with Statistica).
2. If the data file is already open, it will be listed in the Open Spreadsheet Documents folder; double-click it to open it, or select it and click the **OK** button.
3. On the **Data preparation** tab, click the **Select variables** button, and in the **Select variables** dialog, select:

Variable 1 - (Credit Rating) as the **Target, categorical** variable,

Variables 3, 6, and 14 as **Input, continuous** predictors

Variables 2, 4-5, 7-13, and 15-18 as **Input, categorical** predictors, and

Variable 19 - TrainTest as the **Testing sample** variable.

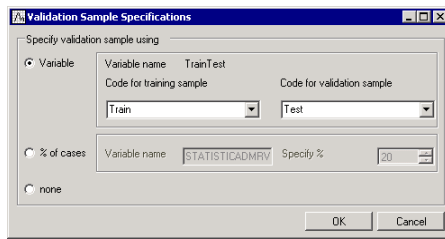


4. Then, click the **OK** button.
5. Select the **Advanced** tab in the **Data miner recipes** dialog, and select the **Use sample data** check box.
6. Select the **Stratified random sampling** option button as the sampling strategy to ensure that each class of the dependent variable Credit Rating is represented with approximately equal numbers of cases in train and validation sets.
7. Then click the **More options** button to display the **Stratified sampling** dialog.
8. Click the **Strata variables** button, select Credit Rating as the strata variable, and click **OK** in this dialog and in the **Stratified sampling** dialog.
9. Click the **Next step** button for the **Data preparation** step to ensure that this step has been successfully completed (in the step-node panel next to **Data preparation**, the yellow **?** changes to a green **✓**).

Data for Analysis

After the **Data preparation** step is completed, the **Data for analysis step** will be selected automatically.

1. On the **Data for analysis** tab, click the **Select testing sample** button, and in the **Testing Sample Specifications** dialog, select the **Variable** option button.
2. Verify that the category (value) Train is entered in the **Code for training sample** field and Test is entered in the **Code for testing sample** field.



3. Then, click the **OK** button. The models will be fitted using the training sample and evaluated using the observations in the testing sample. By using observations that did not participate in the model fitting computations, the goodness-of-fit statistics computed for (predicted values derived from) the different data mining models (algorithms) can be used to evaluate the predictive validity of each model and, hence, can be used to compare models and to choose one or more over others.

Descriptive statistics. This step will also compute descriptive statistics for all variables selected in the analysis. Descriptive stats provide useful information about ranges and distributions of the data used for the project.

1. Click the **Next step** button to ensure that this step is successfully complete.

Data Redundancy

Now, the **Data redundancy** step will be selected. The purpose of this step is to eliminate highly redundant predictors. For example, if the data set contained two measures for weight, one in kilogram the other in pounds, then those two measures would be redundant.

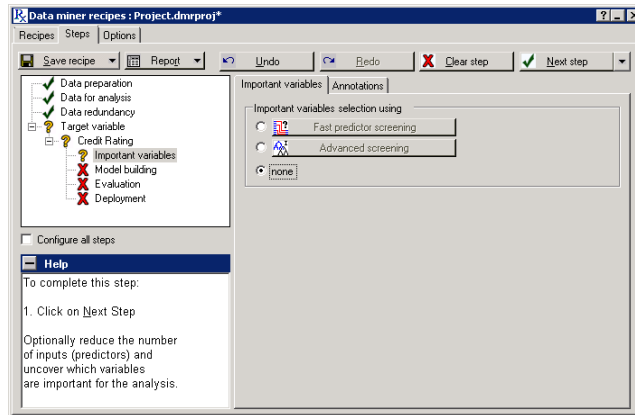
1. On the **Data redundancy** tab, select the **Correlation coefficient** option button, and specify the **Criterion value** as 0.8.
2. Click the **Next step** button to eliminate the redundant predictors that are highly correlated ($r \geq 0.8$). Since there is no redundancy in the data set we are using in this example, a message dialog will be displayed stating this.



3. Click the **OK** button. The data cleaning and preprocessing for model building is now complete.

Target Variable: Building Predictive Model

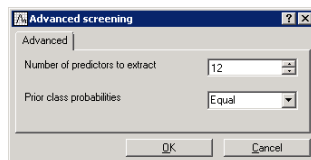
Next, we need to build predictive models for the target in this example. In the step-node panel, the **Target variable** node has a branching structure with the parent node connecting to four child nodes including **Important variables**, **Model building**, **Evaluation**, and **Deployment**.



Dimension reduction. The **Important variables** node is selected automatically. In this step, the goal is to reduce the dimensionality of the prediction problem, such as to select a subset of inputs that is most likely related to the target variable (in this example Credit Rating) and, thus, is most likely to yield accurate and useful predictive models. This type of analytic strategy is also sometimes called feature selection.

Two strategies are available. When the **Fast predictor screening** option button is selected, the program will screen through thousands of inputs and find the ones that are strongly related to the dependent variable of interest. When the **Advanced screening** option button is selected, tree methods are used to detect important interactions among the predictors.

For this example, select the **Advanced screening** option button as the feature selection strategy, and then click the **Advanced screening** button to display the **Advanced screening** dialog. Enter 12 in the **Number of predictors to extract** field, and select **Equal** in the **Prior class probabilities** field.



Click the **OK** button in this dialog, and then click the **Next step** button to complete this step. To review a summary of the analysis thus far, on the **Steps** tab, click the **Report** button, and from the drop-down list, select **Summary report** to display the Results workbook.

	1	2										
Important variables selection using	Advanced screening											
Table	ID_Advance@Setting											
Number of predictors selected	Number of predictors to extract=12											
Selected variable(s)	<table border="1"> <tr><td>Duration of Credit</td></tr> <tr><td>Balance of Current Account</td></tr> <tr><td>Purpose of Credit</td></tr> <tr><td>Employed by Current Employer for Value of Savings</td></tr> <tr><td>Payment of Previous Credits</td></tr> <tr><td>Amount of Credit</td></tr> <tr><td>Living in Current Household for Installment in % of Available Income</td></tr> <tr><td>Age</td></tr> <tr><td>Type of Apartment</td></tr> <tr><td>Most Valuable Assets</td></tr> </table>		Duration of Credit	Balance of Current Account	Purpose of Credit	Employed by Current Employer for Value of Savings	Payment of Previous Credits	Amount of Credit	Living in Current Household for Installment in % of Available Income	Age	Type of Apartment	Most Valuable Assets
Duration of Credit												
Balance of Current Account												
Purpose of Credit												
Employed by Current Employer for Value of Savings												
Payment of Previous Credits												
Amount of Credit												
Living in Current Household for Installment in % of Available Income												
Age												
Type of Apartment												
Most Valuable Assets												
Table	Step options											
Date and time	6/7/2010 1:03:43 PM											

These predictors will be further examined using various cutting-edge data mining and machine learning algorithms available in SDMR.

Building models. The **Data miner recipes** dialog was minimized so that you could see the Results workbook. Click the **Data miner recipes** button located on the Analysis Bar (in the lower-left corner of the application) to display the dialog again. Now, the **Model building** node is selected. In this step, you can build a variety of models for the selected inputs. On the **Model building** tab, **C&RT**, **Boosted tree**, and **Neural network** are selected by default as the models or algorithms that will automatically be tried against the data.

The computations for building predictive models can be performed either locally (on your computer) or on the Statistica Enterprise Server. However, the latter option is available only if you have a valid Statistica Enterprise Server account and you are connected to the server installation at your site. For this example, click the **Build model** button to perform the computations locally on your computer. This will take a few moments; when finished, click the **Next step** button to complete this step.

Evaluating and selecting models. Now, the **Evaluation** node is selected. On the **Evaluation** tab in the **Select model(s)** field, ensure that all models are selected (each check box is selected). Click the **Evaluate models** button to perform the competitive evaluation of models for identifying the best performing model in terms of performance in the validation sample.

Notice that the Boosting Trees model has the minimum error rate of 31.48%. In other words, 68.52% of the cases in the validation sample are correctly predicted by this model. Note that your results may vary slightly because these advanced data mining methods randomly split the data into subsets during training to produce reliable estimates of the error rates.

The following spreadsheet shows the classification performance of the best model on the validation data set. The columns represent the predicted class frequencies, as predicted by the Boosting Trees model, and the rows represent the actual or observed classes in the validation sample.