Data Mining

SPSS Clementine 12.0

1. Clementine Overview

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- Types of Models
- Clementine Interface
- References

Introduction

Introduction

- Three of the common data mining tools
  - SPSS Clementine
  - SAS-Enterprise Miner
  - STATISTICA Data Miner
- The most popular and the oldest data mining tool package on the market today is **SPSS Clementine**

Introduction

- **SPSS Clementine**
  - the most mature among the major data mining packages on the market today.
- Since 1993, many thousands of data miners have used Clementine to create very powerful models for business.
- It was the first data mining package to use the **graphical programming** user interface.
- It enables you to quickly develop data mining models and deploy them in business processes to improve decision making.

Introduction

- The Clementine package was integrated with CRISP-DM to help guide the modeling process flow.
- Clementine supports the entire data mining process.

Clementine Documentation

- Clementine User's Guide
  - General introduction to using Clementine.
- Clementine Source, Process, and Output Nodes
  - Descriptions of all the nodes used to read, process, and output data in different formats.
- Clementine Modeling Nodes
  - Descriptions a variety of modeling methods in Clementine
- Clementine Applications Guide
  - The examples in this guide provide brief, targeted introductions to specific modeling methods and techniques.
- Clementine Algorithms Guide
  - Descriptions of the mathematical foundations of the modeling methods used in Clementine.
- CRISP-DM 1.0 Guide
  - Step-by-step guide to data mining using the CRISP-DM methodology.

Application Examples

- While the data mining tools in Clementine can help solve a wide variety of business and organizational problems, the application examples provide brief, targeted introductions to specific modeling methods and techniques.
- You can access the examples by choosing Application Examples from the Help menu in Clementine Client.
- The data files and sample streams are installed in the Demos folder under the product installation directory.

Demos Folder

- The data files and sample streams used with the application examples are installed in the Demos folder under the product installation directory.
- This folder can also be accessed from the Clementine 12.0 program group under SPSS Inc on the Windows Start menu.

Changing the Temp Directory

- Some operations performed by Clementine may require temporary files to be created.
- By default, Clementine uses the system temporary directory to create temp files.

Changing the Temp Directory

- To alter the location of the temporary directory using the following steps:
  - Create a new directory called clem and subdirectory called *servertemp*.
  - Edit options.cfg, located in the /config directory of your Clementine installation directory.
  - Edit the *temp_directory* parameter in this file to read:

    temp_directory, "C:/clem/servertemp".
  - After doing this, you must restart the Clementine Server service. You can do this by clicking the Services tab on your Windows Control Panel. Just stop the service and then start it to activate the changes you made. Restarting the machine will also restart the service.
  - All temp files will now be written to this new directory.

Types of Models

Types of Models

- In Clementine 12.0, models are packaged into following modules:
  - **Classification** module
    - The Classification module helps organizations for prediction
  - **Association** module
    - Association models associate a particular conclusion (such as the decision to buy something) with a set of conditions.
  - **Segmentation** module
    - The Segmentation module is recommended in cases where the specific result is unknown

Classification Module

- Classification module is included:
  - **Decision Trees**
    - **C&R Tree**
    - **QUEST**
    - **CHAID**
    - **C5.0**
  - **Decision List**
  - **Neural Networks**
  - **Regression**
    - **Linear regression**
    - **Logistic regression**
  - **Bayesian Network**
  - **Support Vector Machine (SVM)**

Association Module

- Association module is included:
  - **Generalized Rule Induction (GRI)**
  - **Apriori model**
  - **CARMA model**
  - **Sequence model**

Segmentation Module

- Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong.
- This following models are included:
  - **K-Means**
  - **Kohonen**
  - **TwoStep**
  - **Anomaly Detection**

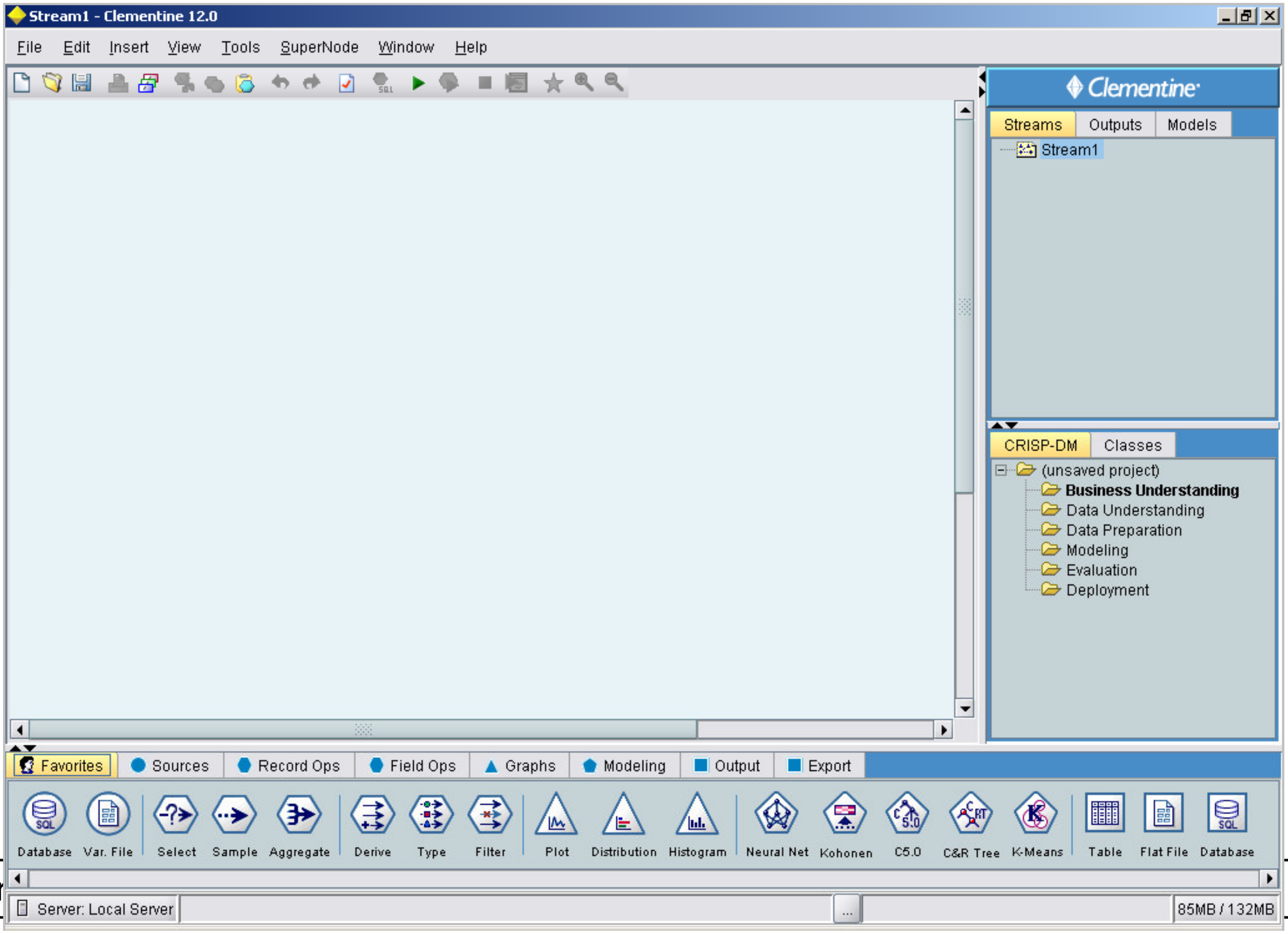
Clementine Modules

- The following are optional add-on modules for Clementine:
  - **Text Mining option** for mining unstructured data
  - **Web Mining option**
  - **Database Modeling** for integration with other data mining packages

Clementine Interface

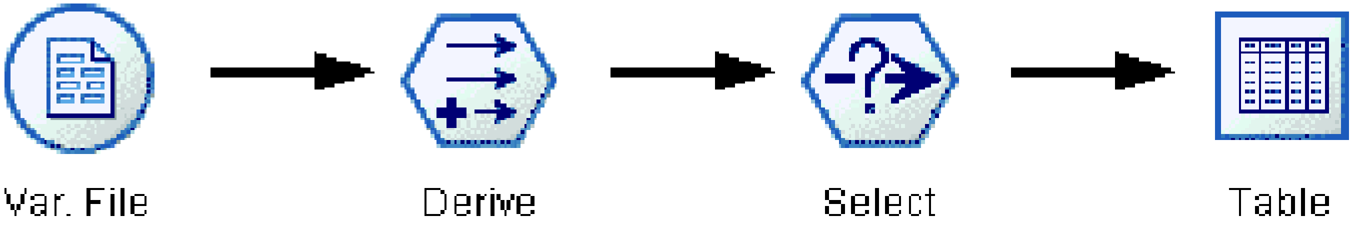
Starting Clementine

- To start the application, choose Clementine 12.0 from the SPSS Inc program group on the Windows Start menu.



Data Stream

- The Clementine interface employs an intuitive **visual programming** interface to permit you to draw logical data flows the way you think of them.
- Working with Clementine is a three-step process of working with data.
  - First, you read data into Clementine,
  - Then, run the data through a series of manipulations,
  - And finally, send the data to a destination.
- This sequence of operations is known as a **data stream**

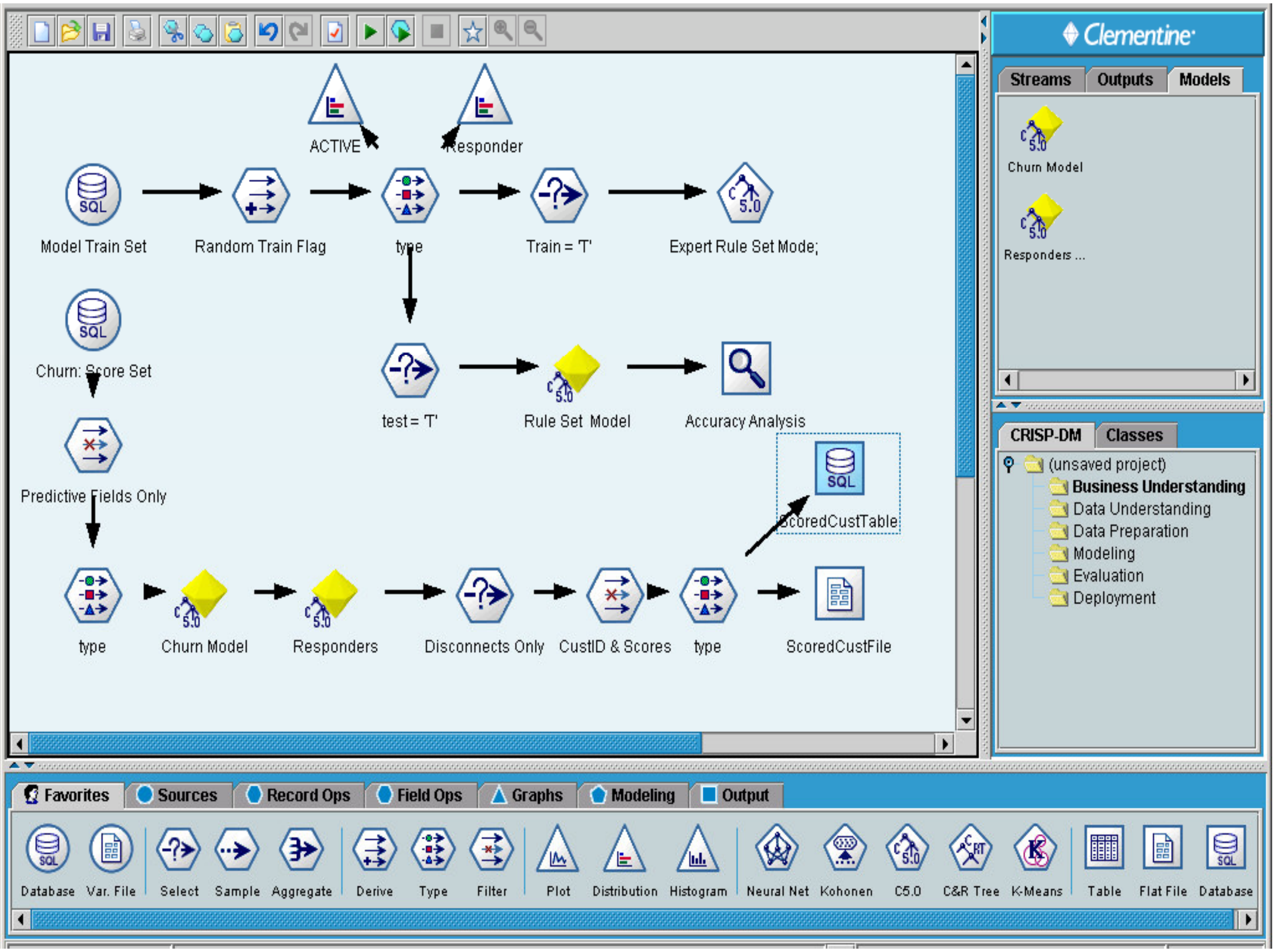


Stream Canvas

- **Stream Canvas**
  - the largest area of the Clementine window and is where you will build and manipulate data streams.
  - You can work on multiple streams in the same canvas or multiple stream files
  - Streams are created by drawing diagrams of data operations on the main canvas in the interface.

Clementine Interface

- Example of canvas:

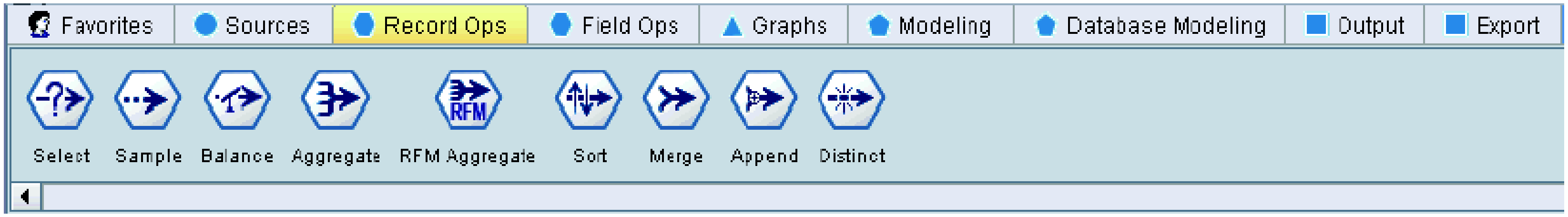


Clementine Interface

- **Node**
  - Each operation is represented by an icon or **node**
  - The nodes are linked together in a stream representing the flow of data through each operation.
- **Streams manager**
  - streams are stored in the **Streams manager**, at the upper right of the Clementine window.

Nodes Palette

- **Nodes Palettes**
  - The palettes are groups of processing nodes listed across the bottom of the screen.
  - The palettes include Favorites, Sources, Record Ops (operations), Field Ops, Graphs, Modeling, Output, and Export. When you click on one of the palette names, the list of nodes in that palette is displayed below.
- the **Record Ops** palette tab.



- To add nodes to the canvas, **double-click icons** from the Nodes Palette or **drag and drop** them onto the canvas.

Nodes Palette Tabs

- **Nodes Palette Tabs**
  - **Sources**
    - Nodes bring data into Clementine.
  - **Record Ops**
    - Nodes perform operations on data records, such as selecting, merging, and appending.
  - **Field Ops**
    - Nodes perform operations on data fields, such as filtering, deriving new fields, and determining the data type for given fields.
  - **Graphs**
    - Nodes graphically display data before and after modeling. Graphs include plots, histograms, web nodes, and evaluation charts.

Nodes Palette Tabs

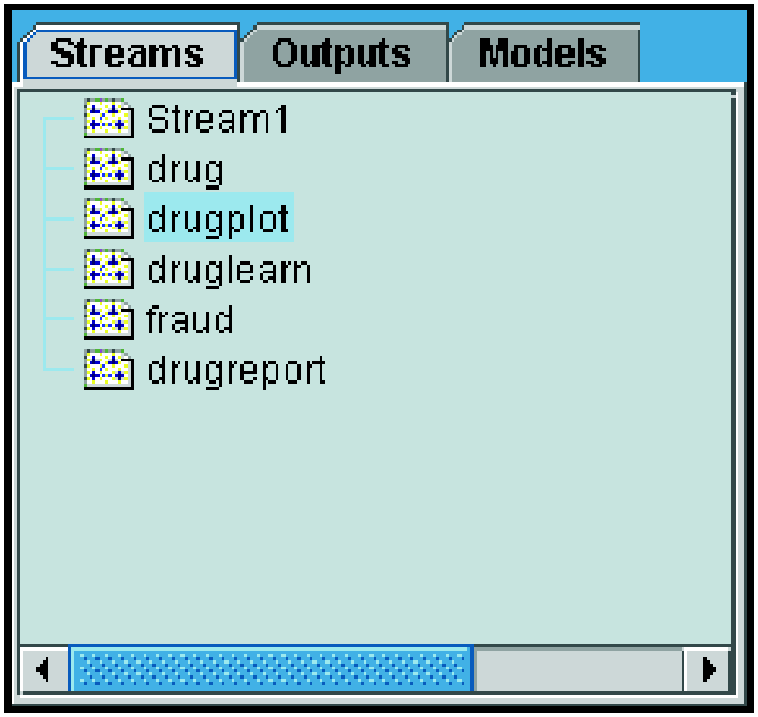
- **Nodes Palette Tabs**
  - **Modeling**
    - Nodes use the modeling algorithms available in Clementine, such as neural nets, decision trees, clustering algorithms, and data sequencing.
  - **Output**
    - Nodes produce a variety of output for data, charts, and model results, which can be viewed in Clementine or sent directly to another application, such as SPSS or Excel.

Clementine Managers

- The upper right window holds three managers:
  - Streams
  - Outputs
  - Models
- You can switch the display of the manager window to show any of these items.

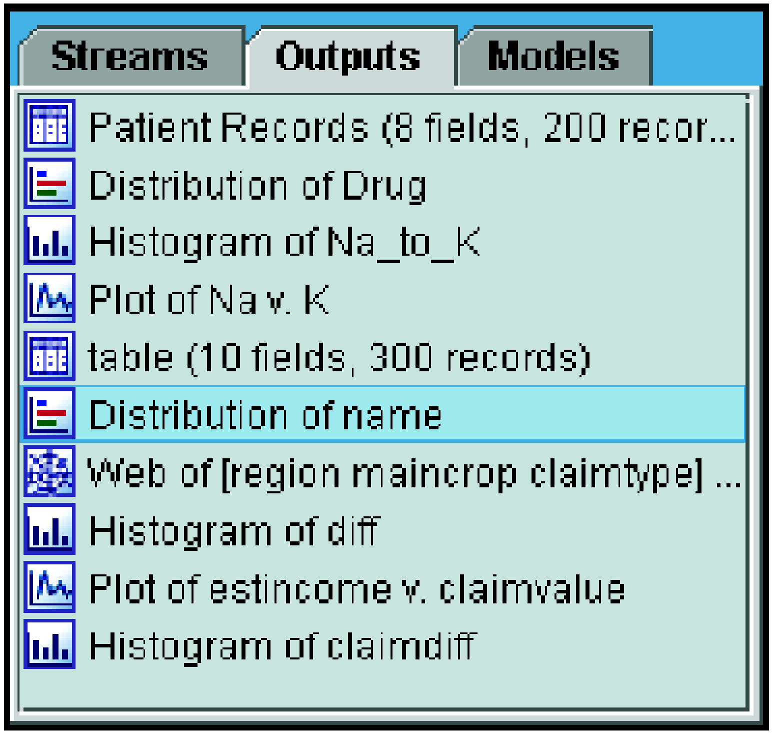
Clementine Managers

- **Streams tab**
  - The **Streams tab** will display the active streams, which you can choose to work on in a given session.
  - You can use the **Streams tab** to open, rename, save, and delete the streams created in a session.



Clementine Managers

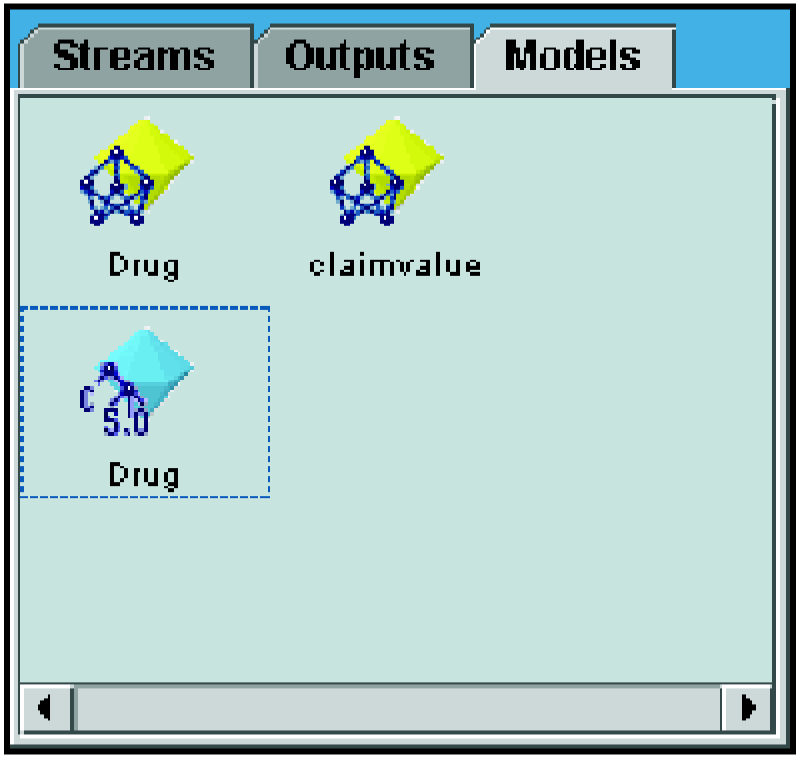
- **Outputs tab**
  - The **Output tab** will show all graphs and tables output during a session.
  - You can display, save, rename, and close the tables, graphs, and reports listed on this tab.



Patient Records (8 fields, 200 recor...
Distribution of Drug
Histogram of Na_to_K
Plot of Na v. K
table (10 fields, 300 records)
Distribution of name
Web of [region maincrop claimtype] ...
Histogram of diff
Plot of estincome v. claimvalue
Histogram of claimdiff

Clementine Managers

- **Models tab**
  - The Models tab will show all the trained models you have created in the session.
  - These models can be browsed directly from the Models tab or added to the stream in the canvas.

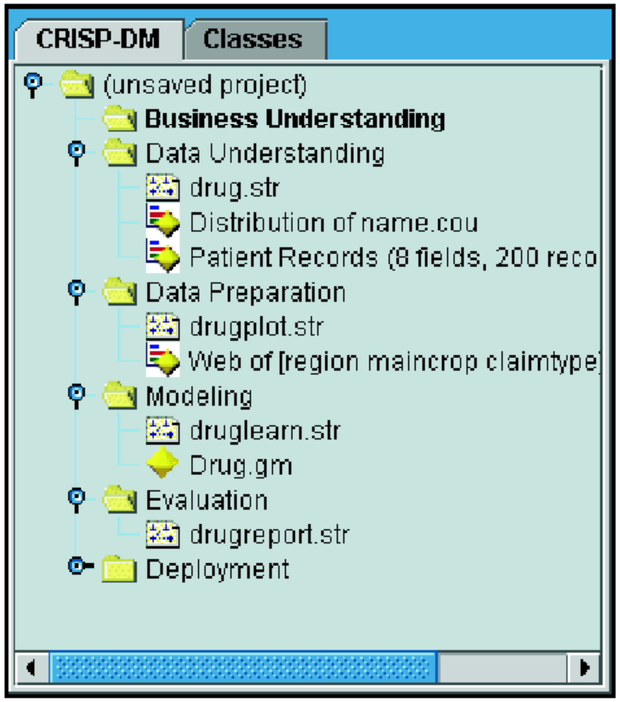


Clementine Projects

- On the lower right side of the window is the **projects tool**, used to create and manage data mining projects.
- There are two ways to view projects you create in Clementine:
  - The CRISP-DM view
  - The Classes view

Clementine Projects

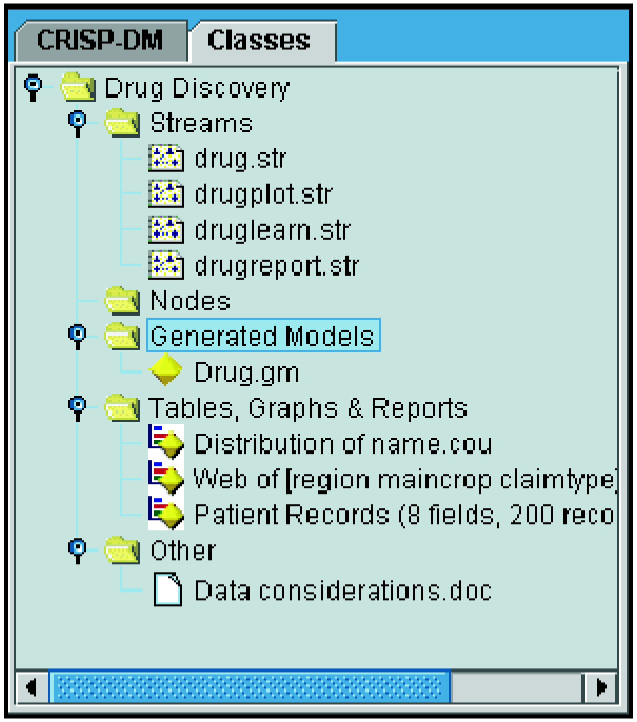
- The **CRISP-DM tab** provides a way to organize projects according to the CRISP-DM methodology.
- For both experienced and first-time data miners, using the CRISP-DM tool will help you to better organize and communicate your efforts.



- (unsaved project)
  - **Business Understanding**
  - Data Understanding
    - drug.str
    - Distribution of name.cou
    - Patient Records (8 fields, 200 reco
  - Data Preparation
    - drugplot.str
    - Web of [region maincrop claimtype]
  - Modeling
    - druglearn.str
    - Drug.gm
  - Evaluation
    - drugreport.str
  - Deployment

Clementine Projects

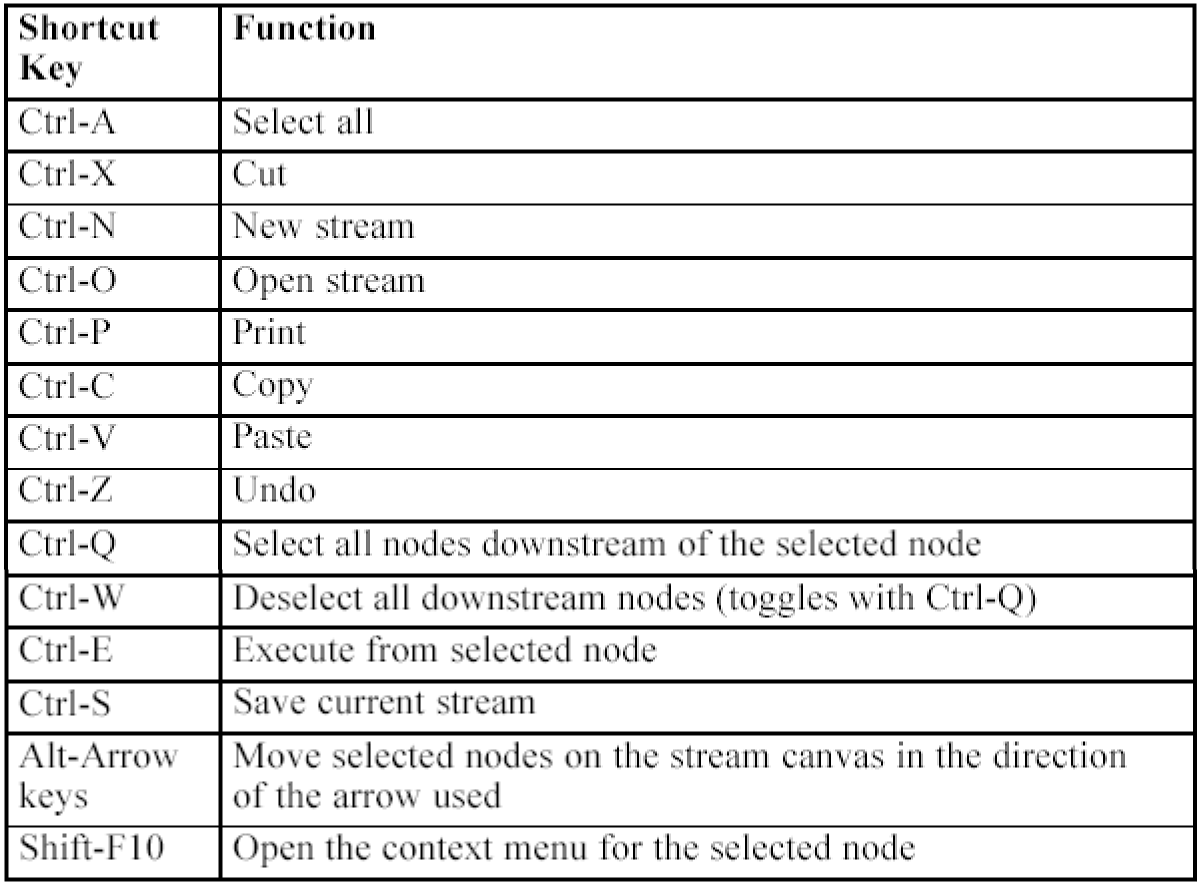
- The **Classes tab** provides a way to organize your work in Clementine categorically—by the types of objects you create.
- This view is useful when taking inventory of data, streams, and models.



Clementine Toolbars

📄	Create new stream	📂	Open stream
💾	Save stream	🖨	Print current stream
🗗	Open Clementine Application Templates (CATs)	✂	Cut & move to clipboard
⬡	Copy to clipboard	📋	Paste selection
↶	Undo last action	↷	Redo
✓	Edit stream properties	▶	Execute current stream
▶	Execute stream selection	■	Stop stream (Active only during stream execution)
☆	Add SuperNode	🔍	Zoom in (SuperNodes only)
🔍	Zoom out (SuperNodes only)		

Using Shortcut Keys



Shortcut Key	Function
Ctrl-A	Select all
Ctrl-X	Cut
Ctrl-N	New stream
Ctrl-O	Open stream
Ctrl-P	Print
Ctrl-C	Copy
Ctrl-V	Paste
Ctrl-Z	Undo
Ctrl-Q	Select all nodes downstream of the selected node
Ctrl-W	Deselect all downstream nodes (toggles with Ctrl-Q)
Ctrl-E	Execute from selected node
Ctrl-S	Save current stream
Alt-Arrow keys	Move selected nodes on the stream canvas in the direction of the arrow used
Shift-F10	Open the context menu for the selected node

Automating Clementine

- Since advanced data mining can be a complex and sometimes lengthy process, Clementine includes several types of coding and automation support. They are:
  - **Clementine Language for Expression Manipulation (CLEM)**
  - **Scripting**
  - **Batch mode**

Automating Clementine

- **Clementine Language for Expression Manipulation (CLEM)**
  - is a language for analyzing and manipulating the data that flows along Clementine streams.
  - Data miners use CLEM extensively in stream operations to perform tasks as simple as deriving profit from cost and revenue data or as complex as transforming Web-log data into a set of fields and records with usable information.
  - For more information, see What Is CLEM? in Chapter 7 in Clementine® 12.0 User's Guide.

Automating Clementine

- **Scripting**
  - is a powerful tool for automating processes in the user interface and working with objects in batch mode.
  - Scripts can perform the same kinds of actions that users perform with a mouse or a keyboard.
  - You can set options for nodes and perform derivations using a subset of CLEM.
  - You can also specify output and manipulate generated models.
  - For more information, see Scripting Overview in Chapter 2 in Clementine® 12.0 Scripting and Automation Guide.

Automating Clementine

- **Batch mode**
  - enables you to use Clementine in a non-interactive manner by running Clementine with no visible user interface.
  - Using scripts, you can specify stream and node operations as well as modeling parameters and deployment options.
  - For more information, see Introduction to Batch Mode in Chapter 7 in Clementine® 12.0 Scripting and Automation Guide.

References

References

- Integral Solutions Limited., **Clementine® 12.0 User's Guide**, 2007.

The end