Getting Started with IBM SPSS Modeler

The IBM SPSS Modeler is a data mining, modeling and reporting tool. It provides a nice GUI to carry out all the data mining tasks in form of Nodes and *Stream Flows*.**Nodes** are the icons or shapes that represent individual operations on the data. The nodes are linked together in a **stream** to represent the flow of data through each operation i.e. A set of actions (reading in, preprocessing, classification/association rule mining/clustering, reporting, etc.) on some input data is called a stream.

Modeler Interface -

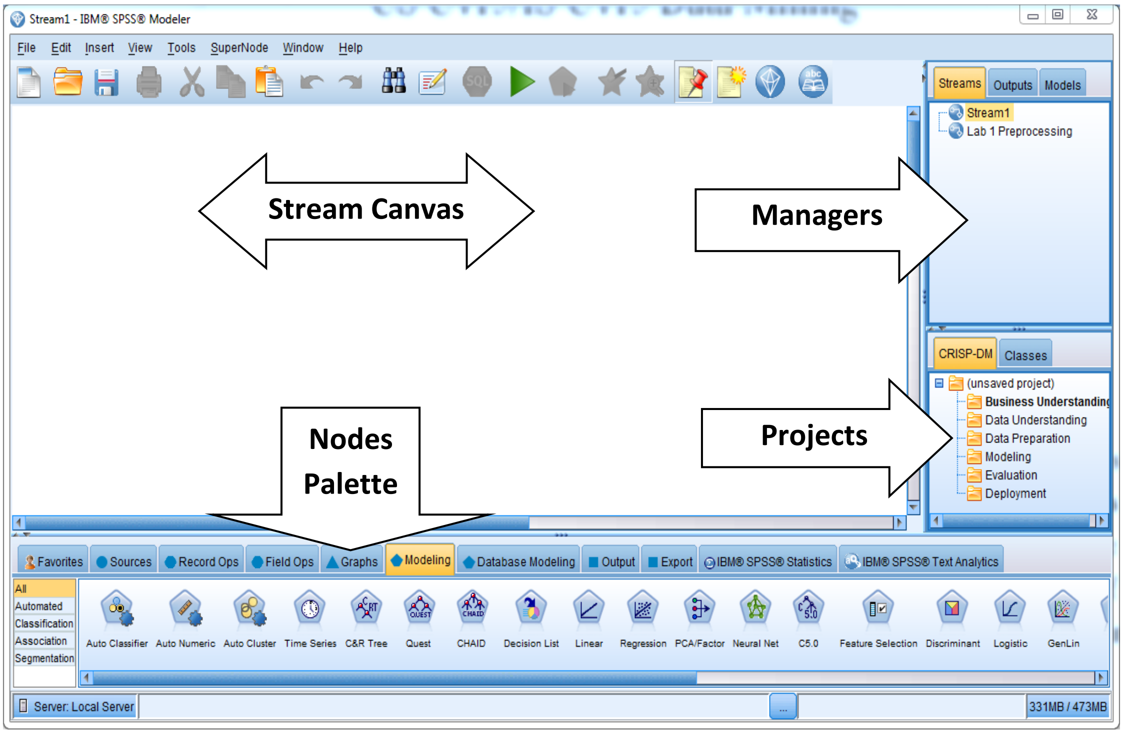

Fig 1 - SPSS Modeler Interface with main components

Modeling -

A model is a set of rules, formulas, or equations that can be used to predict an outcome based on a set of input fields or variables. For example, a financial institution might use a model to predict whether loan applicants are likely to be good or bad risks, based on information that is already known about past applicants.

To build a stream that will create a model, we need at least three elements:

- A source node that reads in data from some external source.
- A modeling node (classification, association, clustering, etc.) that generates a model nugget when the stream is run.
- [Optional] An output node if we want results in tabular or graphical form.

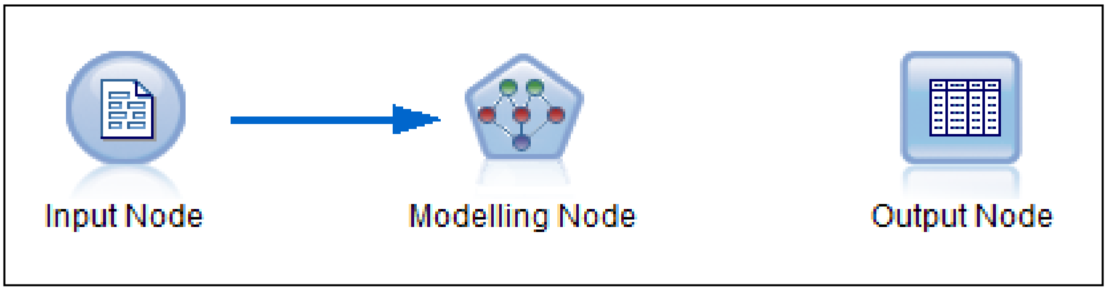

Fig 2 - An abstract stream

Source Nodes

Some important data source nodes are-

Symbol	Node type	Imports data from
SQL	Database Node	MS SQL Server, DB2, Oracle (using ODBC)
	Variable File Node	Delimited text data (*.csv files)
EXCEL	Excel Node	Microsoft Excel
<XML>	XML Node	XML files
	User Input Node	Generate synthetic data

Record Operation Nodes

Record operations nodes are used in data understanding and data preparation. Some important record operation nodes are-

Symbol	Node type	Function
-?>	Select Node	Selects or discards a subset of records from the data stream based on a specific condition
		E.g. - Display all records having an attribute value above a threshold
	Sample Node	Selects a subset of records using a sample type
		E.g. - Display every fifth record
	Aggregate Node	Replaces a sequence of input records with summarized output records
		E.g. - Find class-wise mean and standard deviation of all records

	Sort Node	Sorts record into ascending or descending order based on values of one or more fields
	Merge Node	Takes multiple input records from different sources and creates a single output record containing some or all of input fields
	Append Node	Concatenates sets of records
	Distinct Node	Removes duplicate records

Field Operation Nodes

These nodes are used to select, clean, or construct data in preparation for analysis. Some important field operation nodes are-

Symbol	Node type	Function
	Type Node	Specifies field metadata and properties. E.g. - measurement level (continuous, nominal, ordinal, or flag) for each field can be specified, options for handling missing values and system nulls can be set.
	Filter Node	Filters(discards) fields, rename fields, and maps fields from one source node to another
	Derive Node	Modifies data values or creates new fields from one or more existing fields
		E.g. - Create a new field as the multiplication of two continuous fields
	Filler Node	Replaces field values and changes storage (replace all blank values with a specific value)
	Binning Node	Creates new nominal fields based on the values of one or more existing continuous fields.
	Partition Node	Generates a partition field, which splits the data into separate subsets for the training, testing, and validation stages of model building.

Output Nodes

Output nodes provide the means to obtain information about data and models. Some important output nodes are-

Symbol	Node type	Function
	Table Node	Displays the data in tabular format, which can also be written to file

	Matrix Node	Creates a table that shows relationships between fields
	Analysis Node	Performs various comparisons between predicted values and actual values for one or more model nuggets
	Data Audit Node	Provides a comprehensive first look at the data, including summary statistics, histograms and distribution for each field, as well as information on outliers, missing values, and extremes.

Graph Nodes

These nodes are used for visualizing the data in a mathematical form. Some important graph nodes are-

Symbol	Node type	Function
	GraphboardNode	Offers many different types of graphs in one single node.
	Plot Node	Shows the relationship between numeric fields.
	Web Node	Illustrates the strength of the relationship between values of two or more symbolic (categorical) fields.

Export Nodes

These nodes provide a mechanism for exporting data in various formats to interface with other software tools. Some important export nodes are-

Symbol	Node type	Function
SQL	Database Export Node	Writes data to an ODBC-compliant relational data source.
	Flat File Export Node	Outputs data to a delimited text file.
EXCEL	Excel Export Node	Outputs data in Microsoft Excel Format (*.xls)