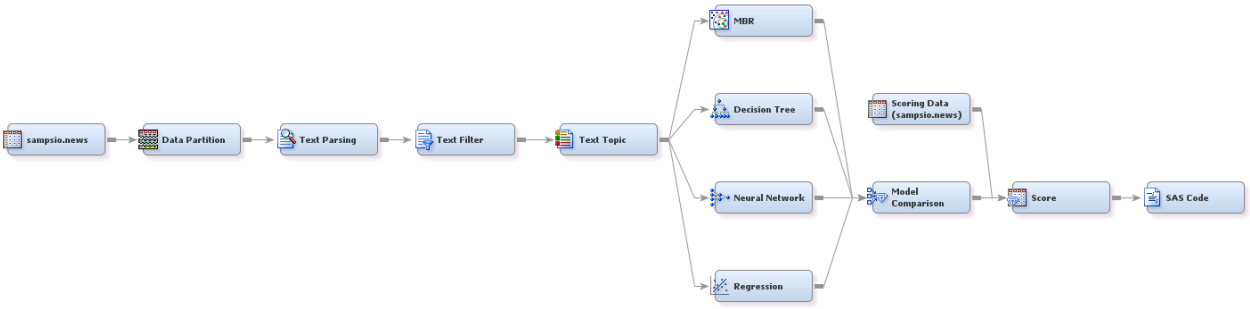
Text Mining Classification in SAS® Enterprise Miner™



Prerequisite:
This example follows the **Text Mining Exploration in SAS Enterprise Miner** example and starts at the Text Topic node, where the previous example finished.

Goal:
The goal is to classify textual articles by the article's content.

Flow:
The sampsio.news node (an Input Data node) is connected to a Data Partition node, which is configured to use 70% of the data for training and 30% for validation. Whenever predictive modeling is done, a Data Partitioning node should be included to help prevent overfitting. The proportion to use for training in the Data Partition node varies, and there is no one correct answer.

The flow splits to multiple nodes after the Text Topic node. Each of these nodes—Regression, Neural Network, Decision Tree, and Memory Based Reasoning (MBR)—is used for classification. SAS Enterprise Miner also provides other nodes for classification; for information about additional nodes that you can use, see SAS Enterprise Miner Reference Help.

In this example, default values are used for the Regression, Neural Network, Decision Tree, and MBR nodes. The defaults might not be best for every problem. In addition, multiple Decision Tree nodes, Regression nodes, and so on can be added in separate branches after the Text Topic node. Each modeling node of a particular type can use slightly different settings, enabling you to test multiple types of models at once.

Each modeling node branch is then connected to the Model Comparison node, which compares the output from each modeling node to determine the best model for the data set. You can view the output in the results of the Model Comparison node, which will indicate the champion model in addition to statistics about all the models. In this example, the defaults for the Model Comparison node were used. For categorical data, the misclassification rate is used to choose the best model, but you can change the node's property to modify this selection statistic.

The output of the Model Comparison node is directed to the Score node. The Score node applies the model that was determined to be the best to any new data that comes in, enabling a trained model to be applied to new data. In this example, the scoring data are in the input data set SAMPSIO.NEWS (although the scoring data set would be different from the input data set in reality). The Scoring Data (sampsio.news) node must have its **Role** set to **Score** and also be input into the Score node, which applies the model to the data.

The Score node is directed to the SAS Code node. The SAS Code node enables you to input custom SAS code into SAS Enterprise Miner projects. In this diagram, the SAS Code node prints out a list of articles and the newsgroups into which they were classified. You can view the SAS code that was used by clicking the ellipses next to the **Code Editor** property of this node.