**In-Class Assignment # 2**

**Stylometry: Application of Textual Analytics**

For this assignment, you will need to create and estimate a model, which predicts authorship of the 11 paragraphs, contained in the FORENSICS\_SCORE dataset. You analyze writing samples from six authors. For five authors, the writing samples in the training data have to do with technical material about statistics and SAS courses. For one of the authors (TK), the writing samples are paragraphs from his published manifesto.

Perform following steps:

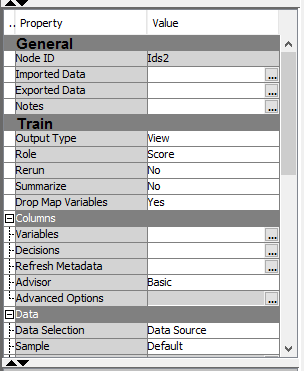
1. Create a diagram named **Forensic Linguistics**.
2. The training data that is used is contained in one SAS data set named FORENSICS, which is stored on Sakai (Resources-> Week\_08-> SAS Data) and SAS on Demand (folder Week\_08).
3. Go to **Data Sources** for the project. Right-click to open **Create Data Source**. Select **Browse** to find the SAS data set **FORENSICS**. In step 5 of the Data Source Wizard when the variables are shown, change the variable roles to correspond to the following:



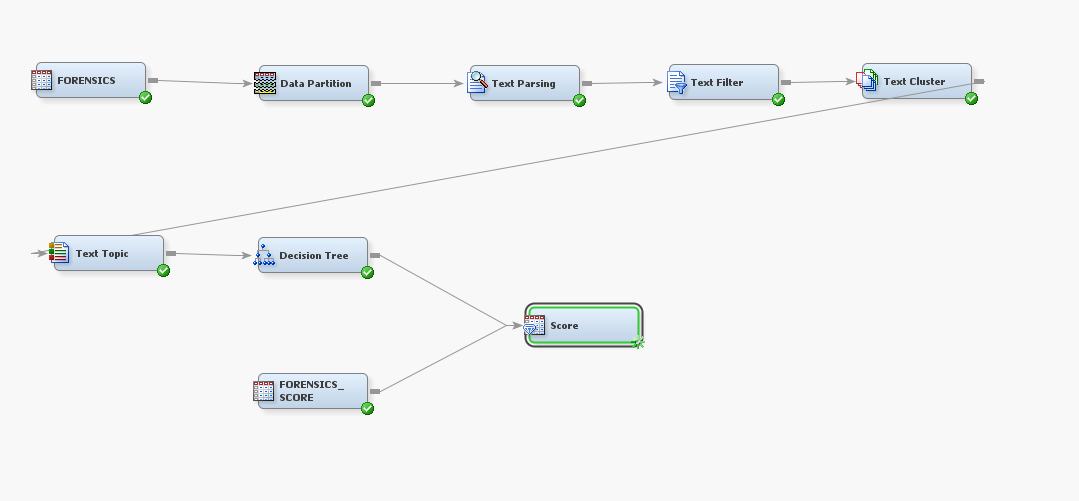
1. Connect a **Data Partition** node (from Sample tab in the EM) to the Forensics Input Data node and retain the default settings (40%/30%/30%).
2. Connect a default **Text Parsing** node to a **Data Partition** node.
3. Connect a default **Text Filter** node. Change the properties under **Weightings** to explicitly show **Log** and **Mutual Information** - these are actually the defaults that are used here because a nominal target variable is present.



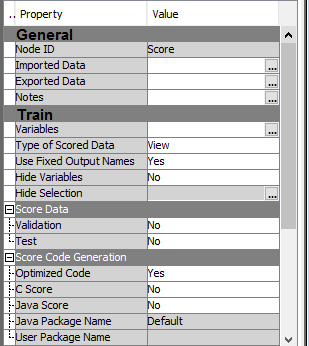
1. Attach a **Text Cluster** node to the **Text Filter** node and run it with the default settings.
2. Attach a default **Text Topic** node to the **Text Cluster** node and run it.
3. Connect a **Decision Tree** node (Model tab in the EM) to the **Text Topics** node. Change the default leaf size from **5** to **25** and change the assessment measure to **Average Square Error**.
4. Open the **FORENSICS\_SCORE** data set and designate it as a **Score** data set. This data set contains the 11 paragraphs that were drawn from TK’s interview after he was captured.



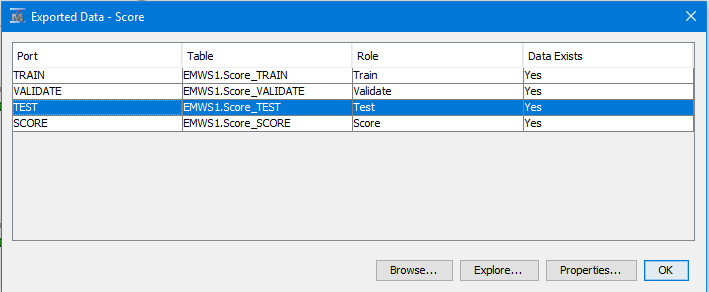
1. You want to see how accurately the tree model classifies these paragraphs. To do this, open a **Score** node (Assess tab in the EM) and attach it to both the **FORENSICS\_SCORE** data set and the **Decision Tree** node. Your complete diagram should look as shown below.



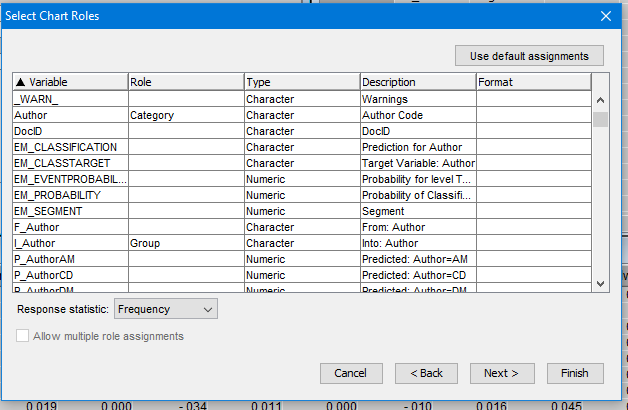
1. Run the **Score** node and click the radio button next to the **Exported data** from the properties panel.



1. Select the **Test** data and click **Explore**.



1. In the opened window, click on **Plot** button and use the **Plot Wizard** to construct a bar chart. Use the following properties:



1. Copy and paste the chart. The dominance of a single color for each author illustrates how accurate the decision tree model is for scoring the test data. You can verify by positioning the cursor over the dominant color for any of the authors.
2. Record how many documents were classified correctly as belonging to the Unabomber. Estimate recall for TK (Unabomber) as # of correctly classified to the total number of documents, belonging to that author.

***You may work in small teams on that assignment.***