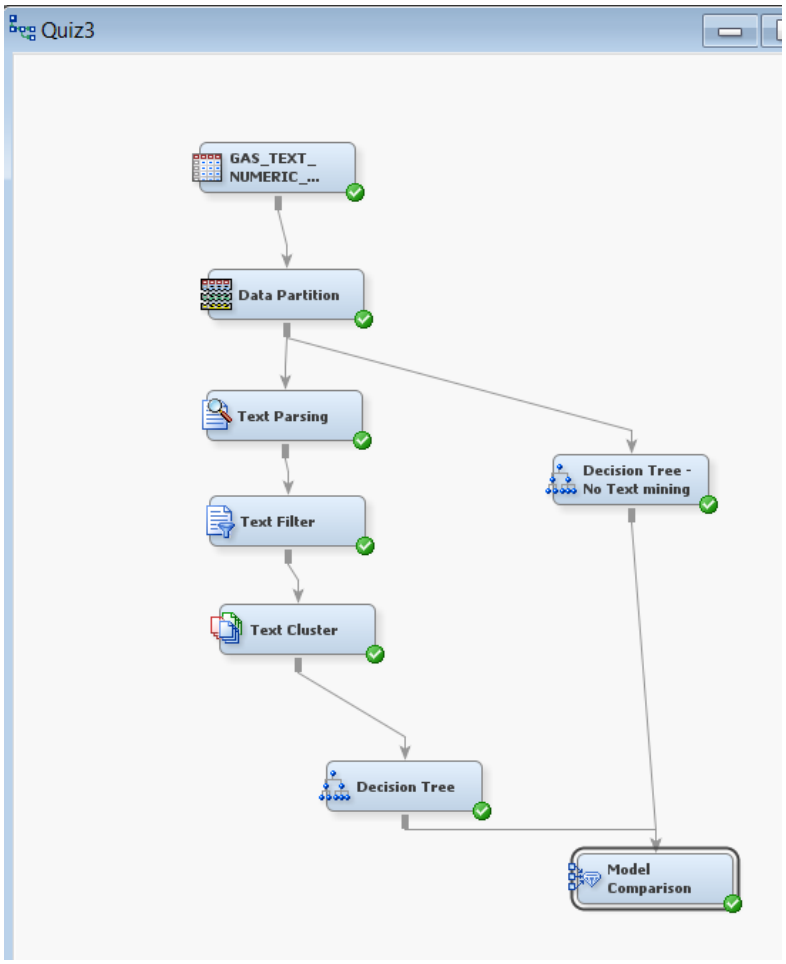
Diane Nguyen
MSBA 645
Quiz 3

Due: before 11:59pm on April 25th 2019

Worth: 100 points

While text mining customer responses can reveal valuable insights about a customer, plugging the results from text mining into a typical data mining model can often significantly improve the predictive power of the model. Organizations often want to use customer responses captured in the form of text via emails, customer survey questionnaires, and feedback on websites for building better predictive models. One way of doing this will be apply text mining to reveal groups (or clusters) of customers with similar responses or feedback. This cluster membership information about each customer may then be used as an input variable to augment the data mining model. The data used in this quiz is based on a real data set of a company that manages truck stops. The data include customers' comments. These comments were later merged with numeric variables from the company's database about these customers by matching them via the company's loyalty card number. The target is Target and the textual field is Comment_All. Ignore the other two comment fields. Use this data set to demonstrate how the use of textual data in conjunction with numeric data in a predictive model improves the performance of a predictive model. You must provide evidence to support your fining. Turn in the following:

- **Flow diagram**



- **Description of your design and the parameters you used for each node and a brief justification. Examples include partition, number of SVDs.**
  For this I partitioned the data 60/40 then followed it with a text parsing node to tokenize and stem the data getting rid of any useless parts of speech. Next I did the text filter node to reduce the number of terms through reducing a lot dimensionality that would exist. This is followed by the text cluster node which groups similar groups of words together. After this I added a decision tree that was attached to the data partition node and another one that was directly below the text cluster node to see if doing text mining improved my results or not as compared to the one with no text mining. For this I used a High SVD Resolution and had a max of 10 SVD dimensions.

- **Answer the question: if the addition of textual data to numeric data improves your classification model. Provide evidence. Determining what constitutes evidence is part of the quiz.**
  The stream that included text mining had a better misclassification rate than the one without.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Valid: Misclassification Rate	Valid: Root Average Squared Error	Valid: Cumulative Percent Response
Y	Tree	Tree	Decision Tree	Target		0.377953	0.377953	0.530121	61.36364
	Tree2	Tree2	Decision Tree - No Text mining	Target		0.448819	0.448819	0.497428	55.11811