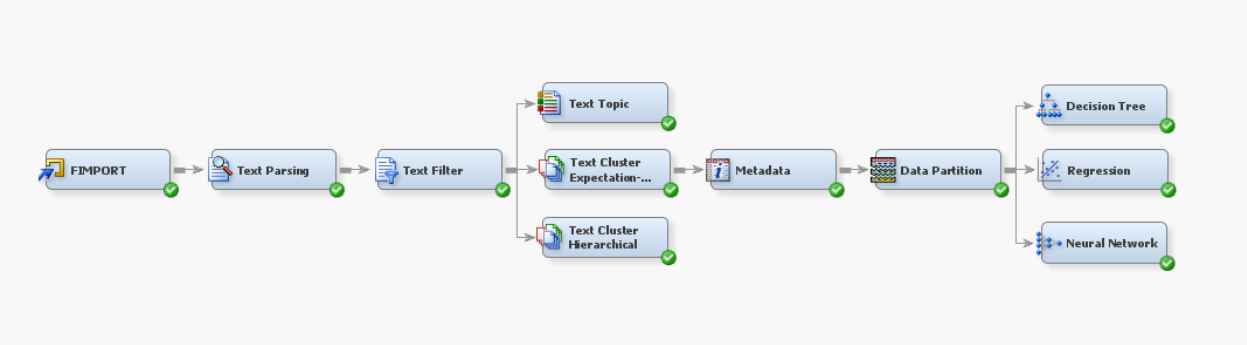
Using the dataset of the previous assignment



1- Identify five of the topics in the Amazon reviews (paste the topics in a word document)

- Orange Juice Soda (Drink)

  +juice,+orange,+soda,+drink,+tangerine
- Goat Milk Baby Formula

  +milk,+baby,+goat,+love,+formula
- Gluten Free Waffle & Pancake Mix

  +mix,+pancake,gluten,free,+waffle
- Green Iced Tea (Drink)

  +tea,+green,+ice,+green tea,+drink
- Favorite Flavor: Coffee Coconut

  +flavor,+favorite,+coffee,+coconut,+nice

Topic
+chocolate,+hot,+cocoa,+hot chocolate,+milk
+coffee,+taste,+good,+drink,+cup
+juice,+orange,+soda,+drink,+tangerine
+chip,+bag,+salt,+potato,+kettle
+dog,+food,+dog food,newman,+organic
+tea,+green,+ice,+green tea,+drink
+store,+find,+grocery,+local,+grocery store
+mix,+pancake,gluten,free,+waffle
+cup,+coffee,+work,+k-cup,keurig
+great,+price,+recommend,+order,+great product
+sauce,+sweet,+hot,+chicken,+add
+roast,+puck,+k-cup,wolfgang,+bold
+price,+good,shipping,+little,+buy
+bean,vanilla,+vanilla bean,+smell,+buy
+cookie,+eat,+snack,+chocolate,+taste
+sugar,+product,+add,+calorie,+ingredient
+cat,+eat,+food,+cat food,+grass
+milk,+baby,+goat,+love,+formula
+order,+product,+box,+receive,+package
+water,+coconut,+taste,+drink,+buy
+treat,+dog,+love,+size,+bag
+flavor,+favorite,+coffee,+coconut,+nice
+popcorn,+pop,+recommend,+taste,+white
+product,+love,+buy,+coffee,+good
+work,+product,+little,+well,+thing

2- Develop a model using decision trees, and regression analysis to predict the "Score" variable based on the text in the reviews. Compare the outcome of these models using the model fit statistics. (Paste both statistics in the word file).

- Decision Tree

Statistics Label ▲	Fit Statistics	Train	Validation
Average Squared Error	_ASE_	0.109341	0.11187
Divisor for ASE	_DIV_	34970	15030
Maximum Absolute Error	_MAX_	0.990291	1
Misclassification Rate	_MISC_	0.374607	0.38024
Root Average Squared Error	_RASE_	0.330668	0.334469
Sum of Frequencies	_NOBS_	6994	3006
Sum of Squared Errors	_SSE_	3823.656	1681.399
Total Degrees of Freedom	_DFT_	27976	.

Average Squared Error for both Training and Validation partitions is around **0.11**.

Misclassification Rate is **0.375** Training and **0.38** for Validation portion.

- Regression Analysis

Statistics Label	Fit Statistics	Train	Validation
Akaike's Information Criterion	_AIC_	13623.96	.
Average Squared Error	_ASE_	0.094811	0.09693
Average Error Function	_AVERR_	0.383642	0.393099
Degrees of Freedom for Error	_DFE_	27872	.
Model Degrees of Freedom	_DFM_	104	.
Total Degrees of Freedom	_DFT_	27976	.
Divisor for ASE	_DIV_	34970	15030
Error Function	_ERR_	13415.96	5908.271
Final Prediction Error	_FPE_	0.095518	.
Maximum Absolute Error	_MAX_	0.996645	0.995876
Mean Square Error	_MSE_	0.095164	0.09693
Sum of Frequencies	_NOBS_	6994	3006
Number of Estimate Weights	_NW_	104	.
Root Average Sum of Squares	_RASE_	0.307913	0.311335
Root Final Prediction Error	_RFPE_	0.30906	.
Root Mean Squared Error	_RMSE_	0.308487	0.311335
Schwarz's Bayesian Criterion	_SBC_	14480.82	.
Sum of Squared Errors	_SSE_	3315.53	1456.851
Sum of Case Weights Times Freq	_SUMW_	34970	15030
Misclassification Rate	_MISC_	0.348442	0.347638

Average Squared Error for Training partition is **0.095** and **0.097** for Validation partition.

Misclassification Rate is **0.348** for both Training and Validation portion.

3- What model is performing better?

The Regression Analysis Model has lower Average Squared Error and Misclassification Rate values compared to the Decision Tree Model making it perform better in predicting the “Score” variable.