**Assignment 3**

**Team Number: 04**

**Team Members: NetID:**

Akanksha Mishra axm170031

Harman Singh hxs134930

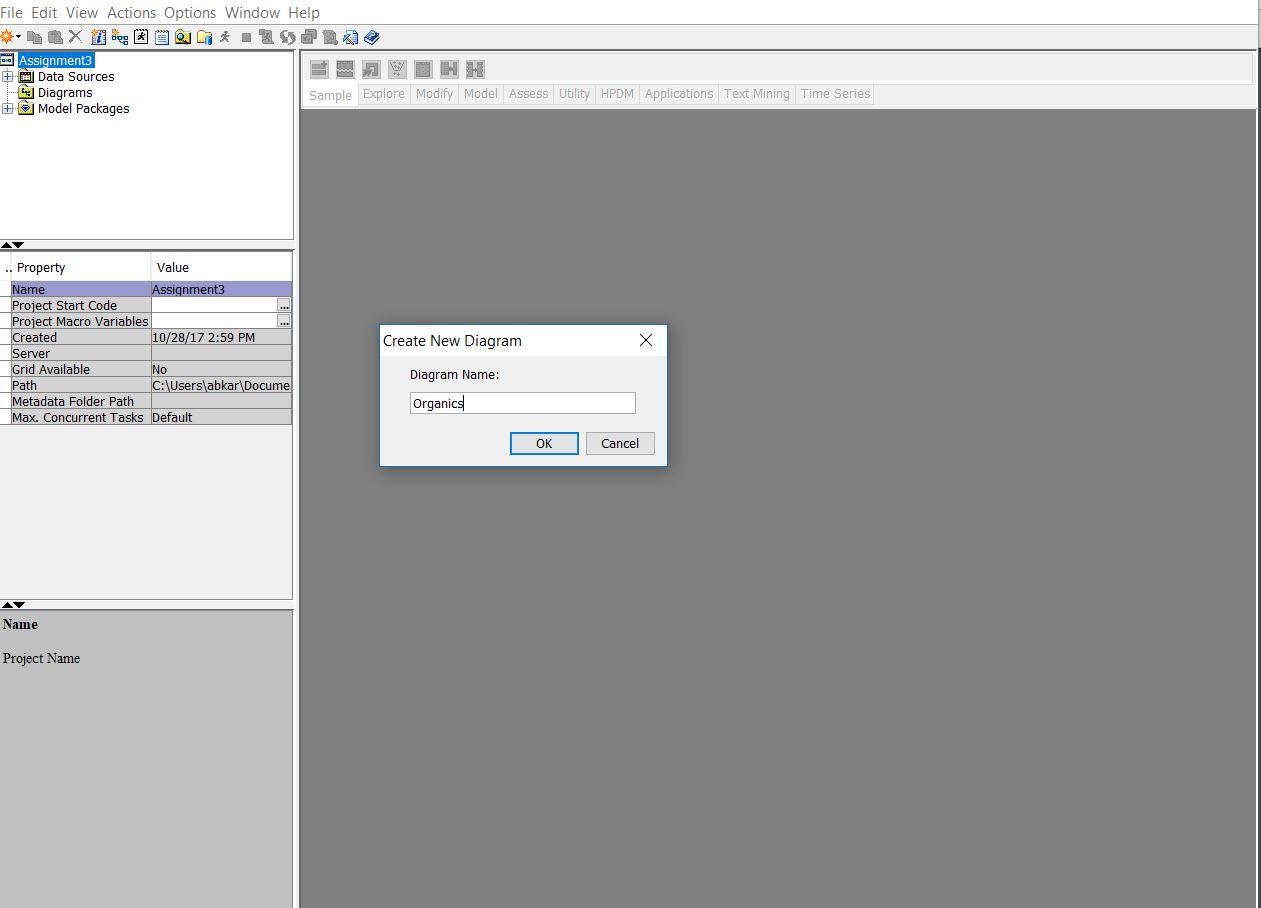
Lokesh Kumar Chaturvedi Lolla lxl163230

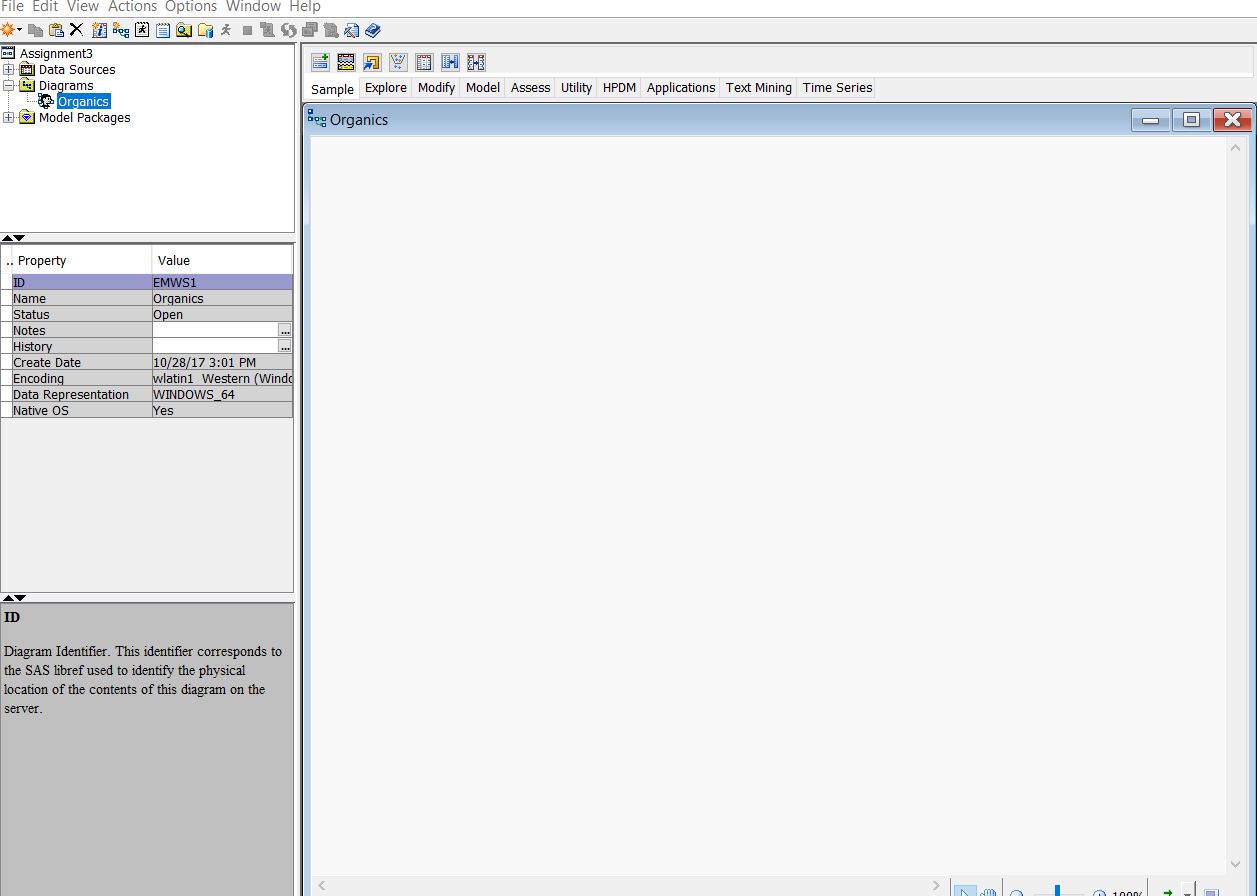
Yugandhara Rawool yer170030

**Solution 1:**

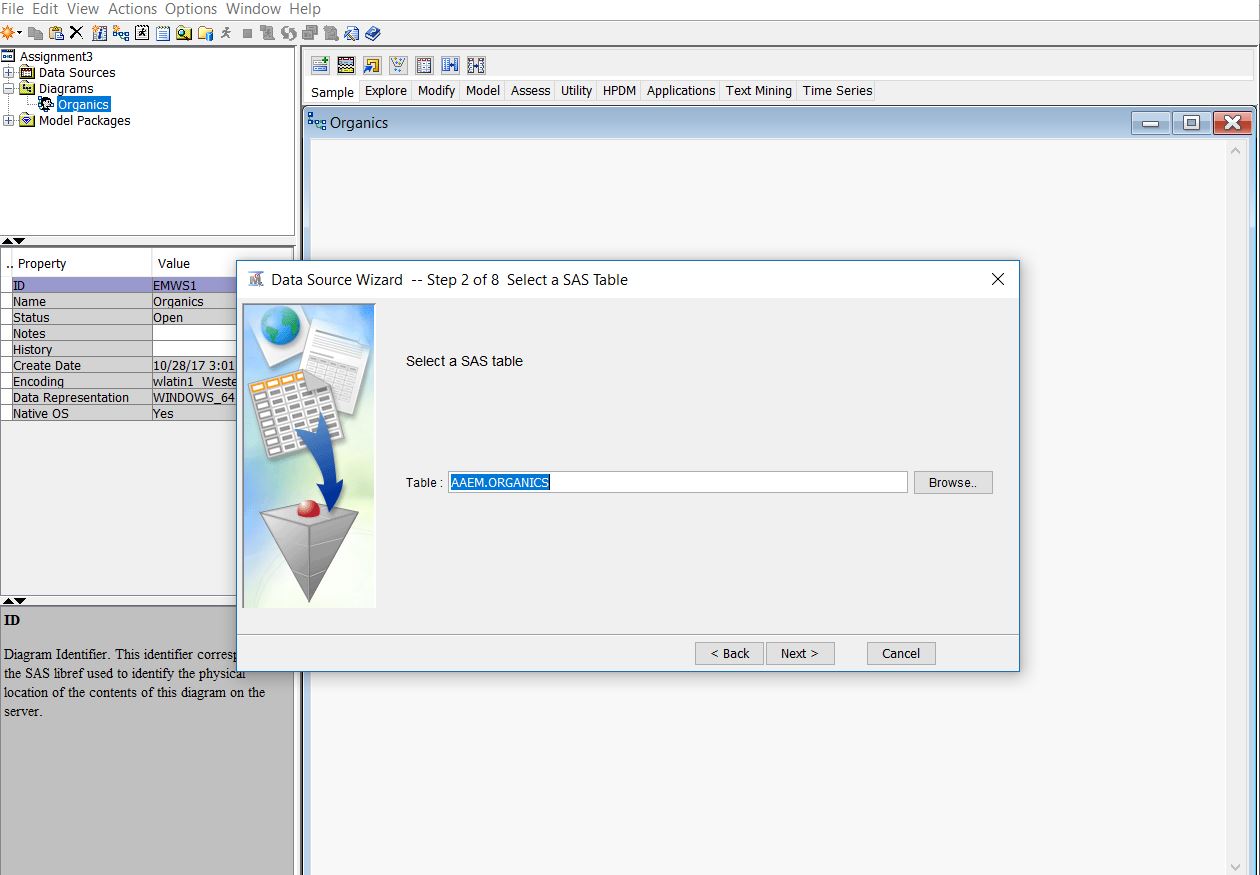
The classification using decision trees in SAS Enterprise miner is done and the screenshots are attached below.

**a).** A new diagram, Organics is created in our project titled Assignment3.

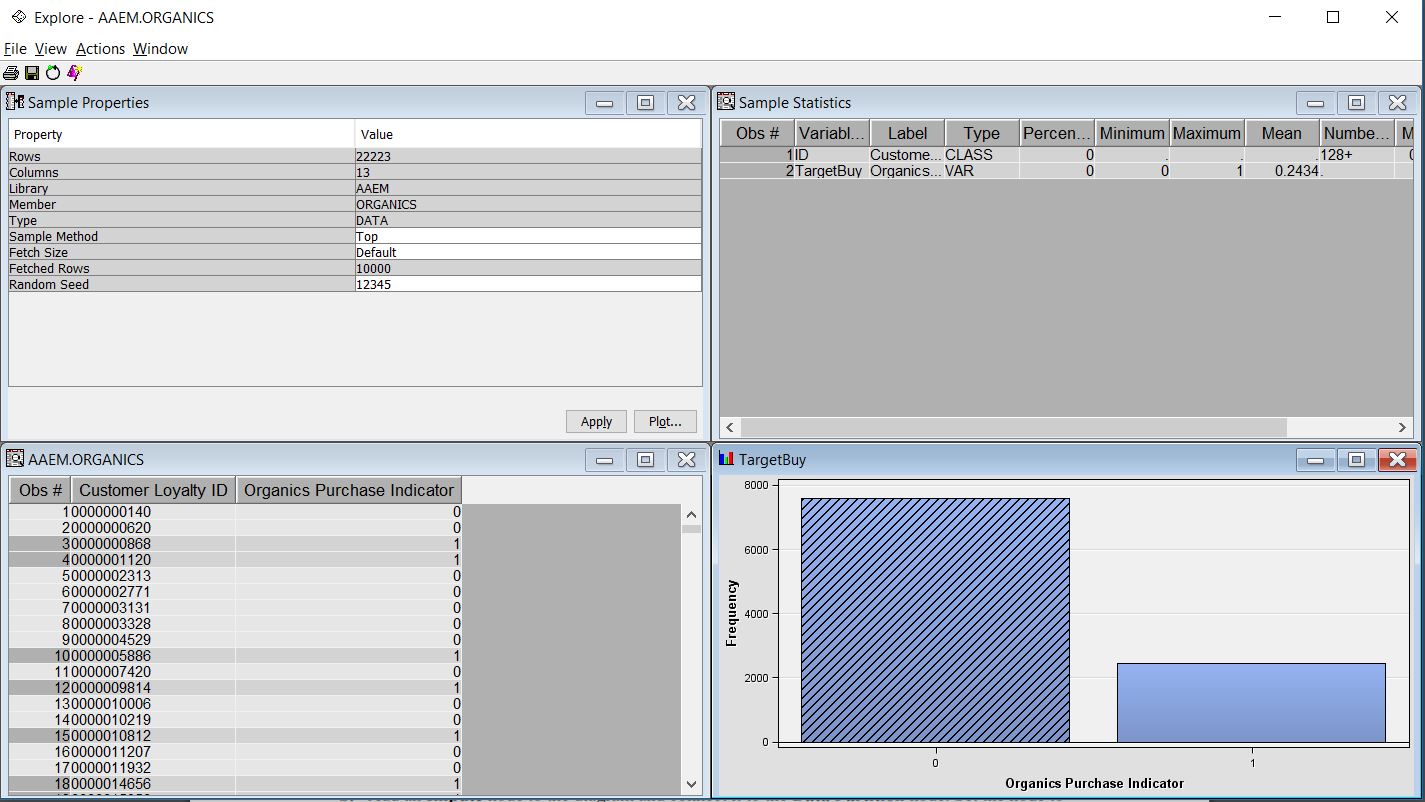




**b).** Dataset Organics is defined as the data source.

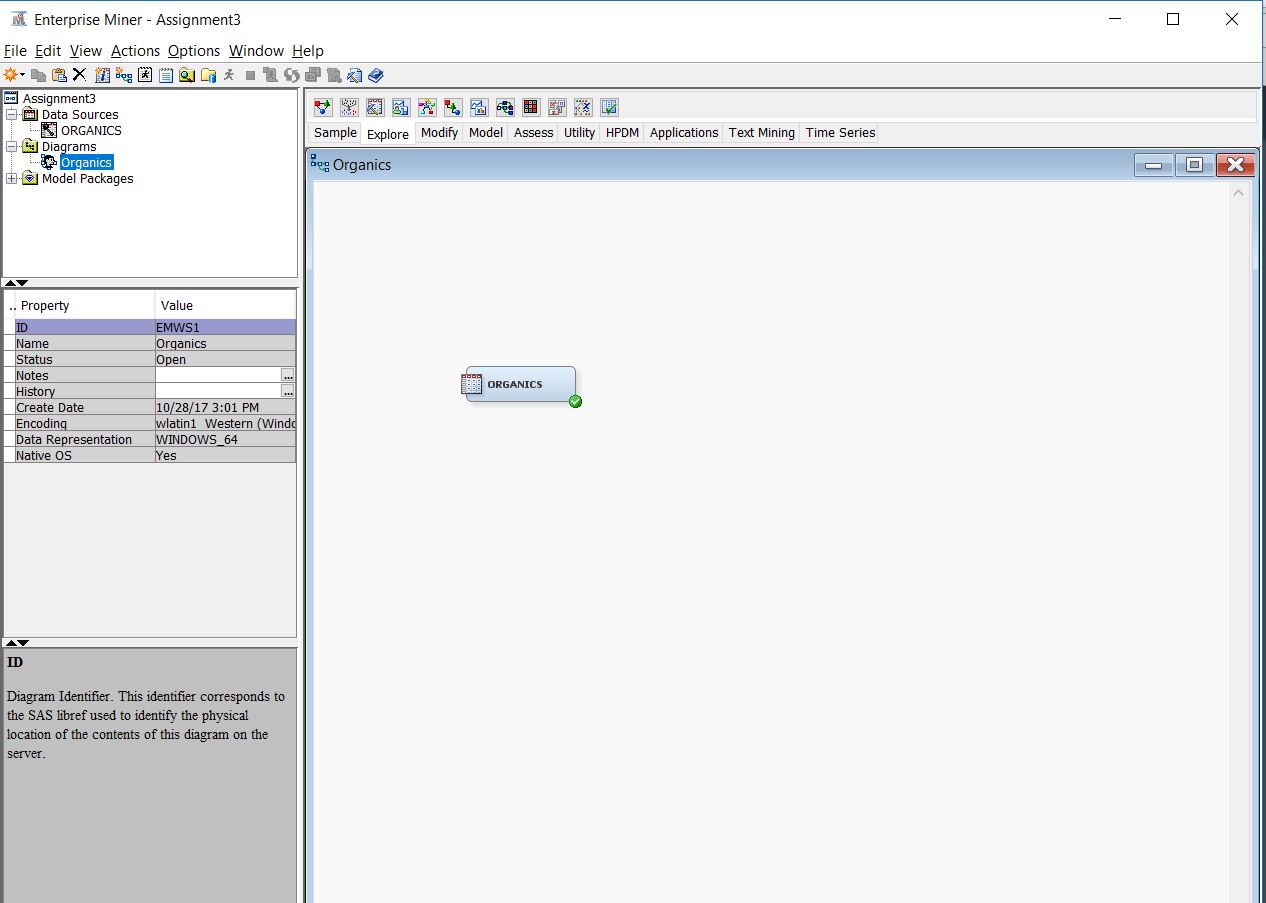


1. The model roles for the analysis variables are set as mentioned in the question.
2. The distribution of target variable is examined and the proportion of individuals who purchased organic products are 2,434 out of 10,000.

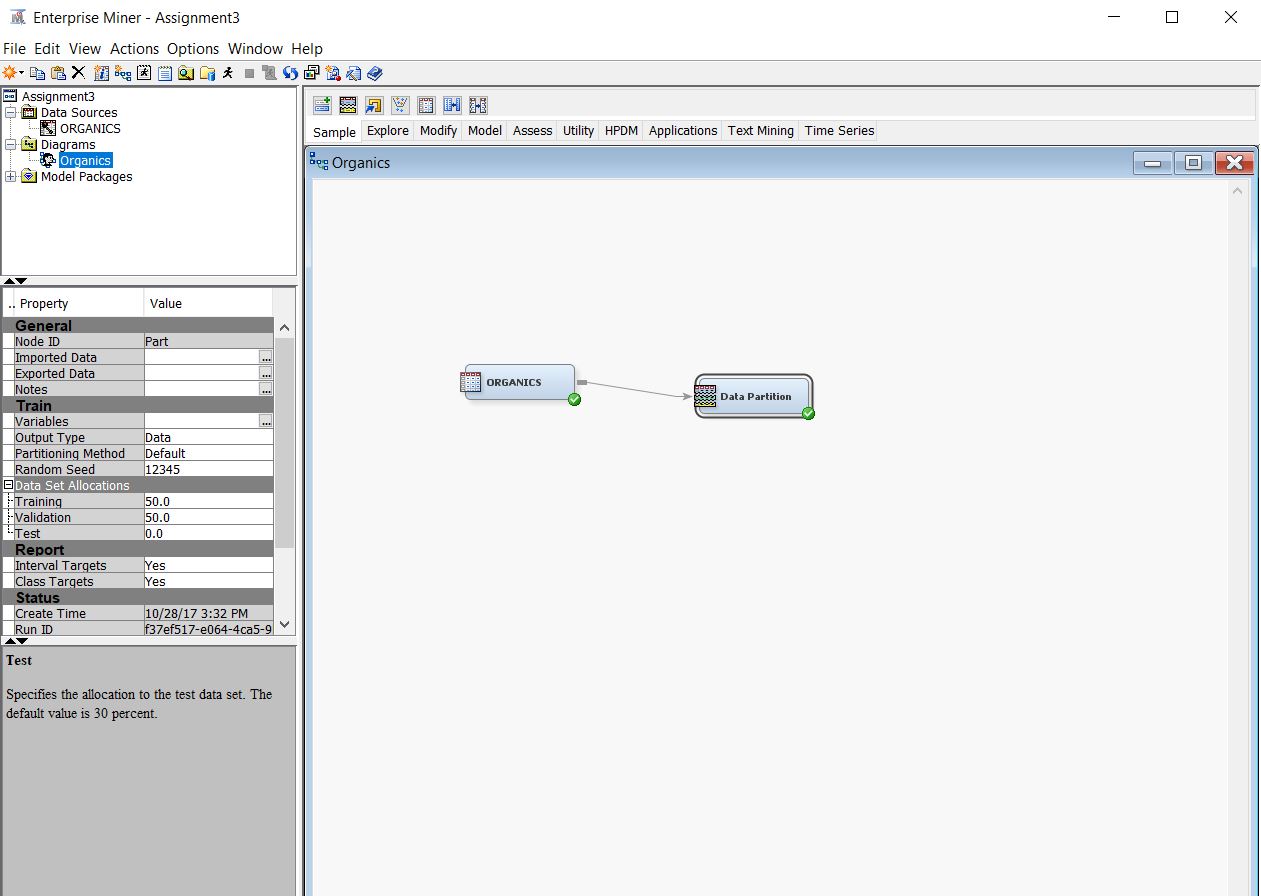


1. The model role for DemCluster is set to Rejected.
2. No, TargetAmt cannot be used as an input for a model used to predict TargetBuy because, target amount only tells us how many organic products are being sold. It doesn’t not tell us how many customers are buying organic products.
3. Organics data source definition is finished.

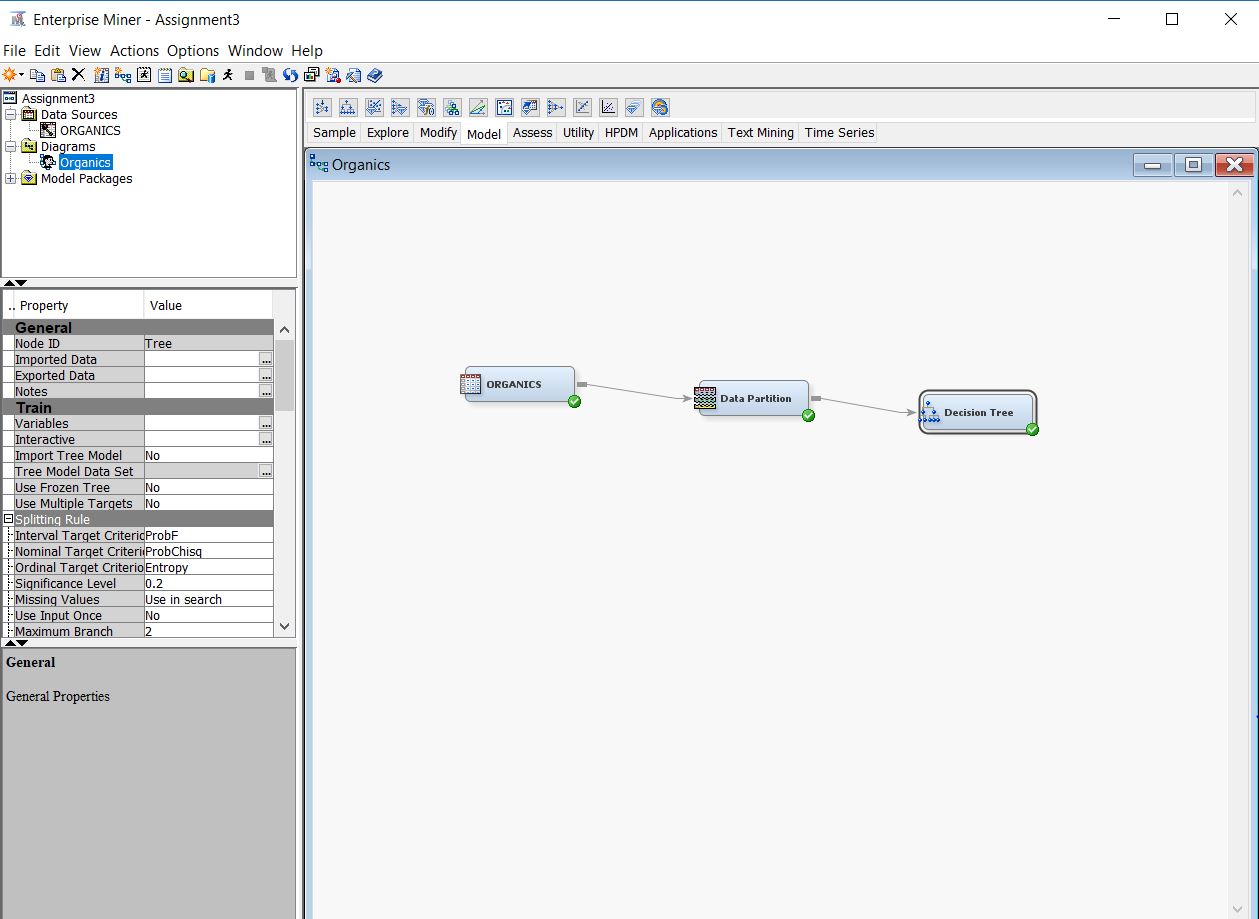
**c).** The Organics data source is added to the diagram as shown below.



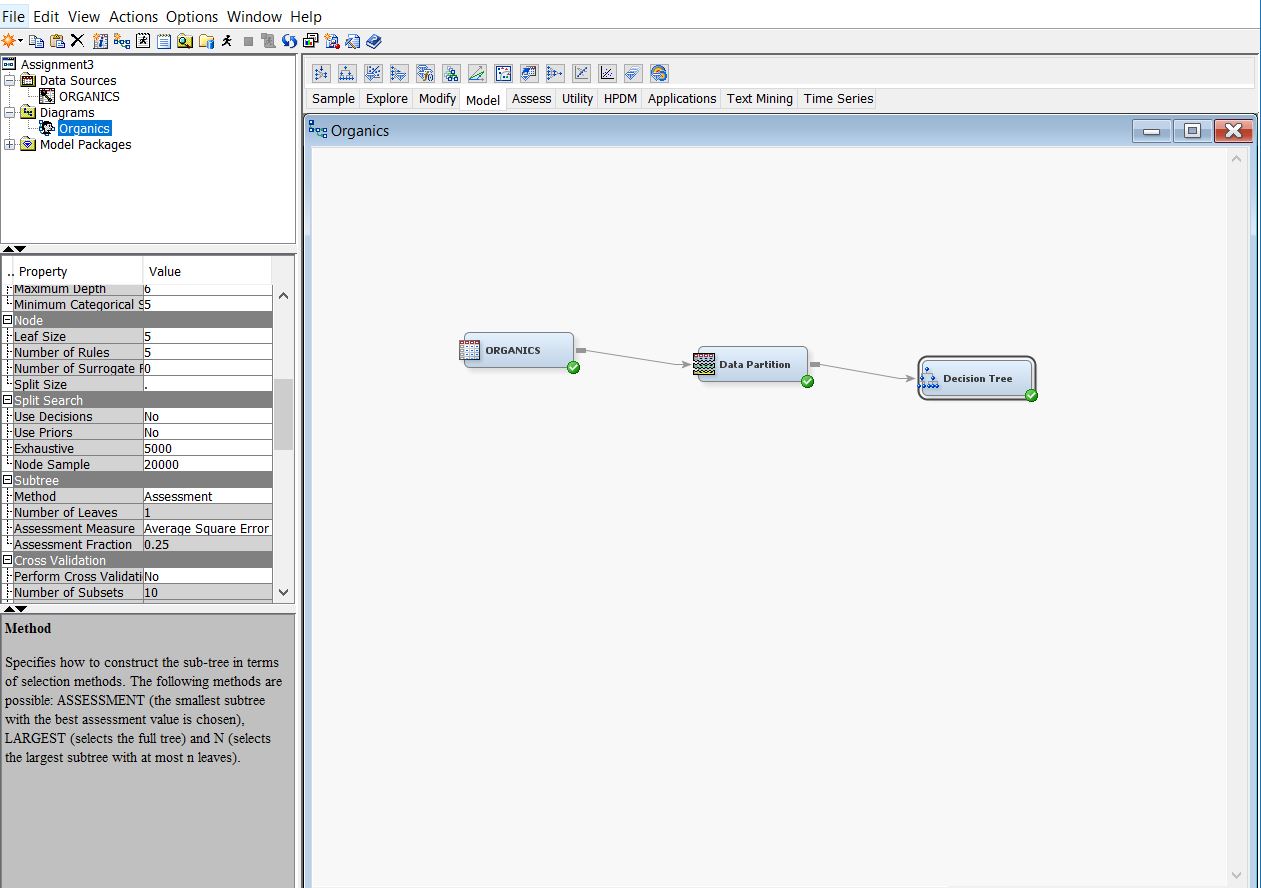
**d).** Data Partition node is added to the diagram and connected to the Data Source node. Assigned 50% of the data for training and 50% for validation as shown below.

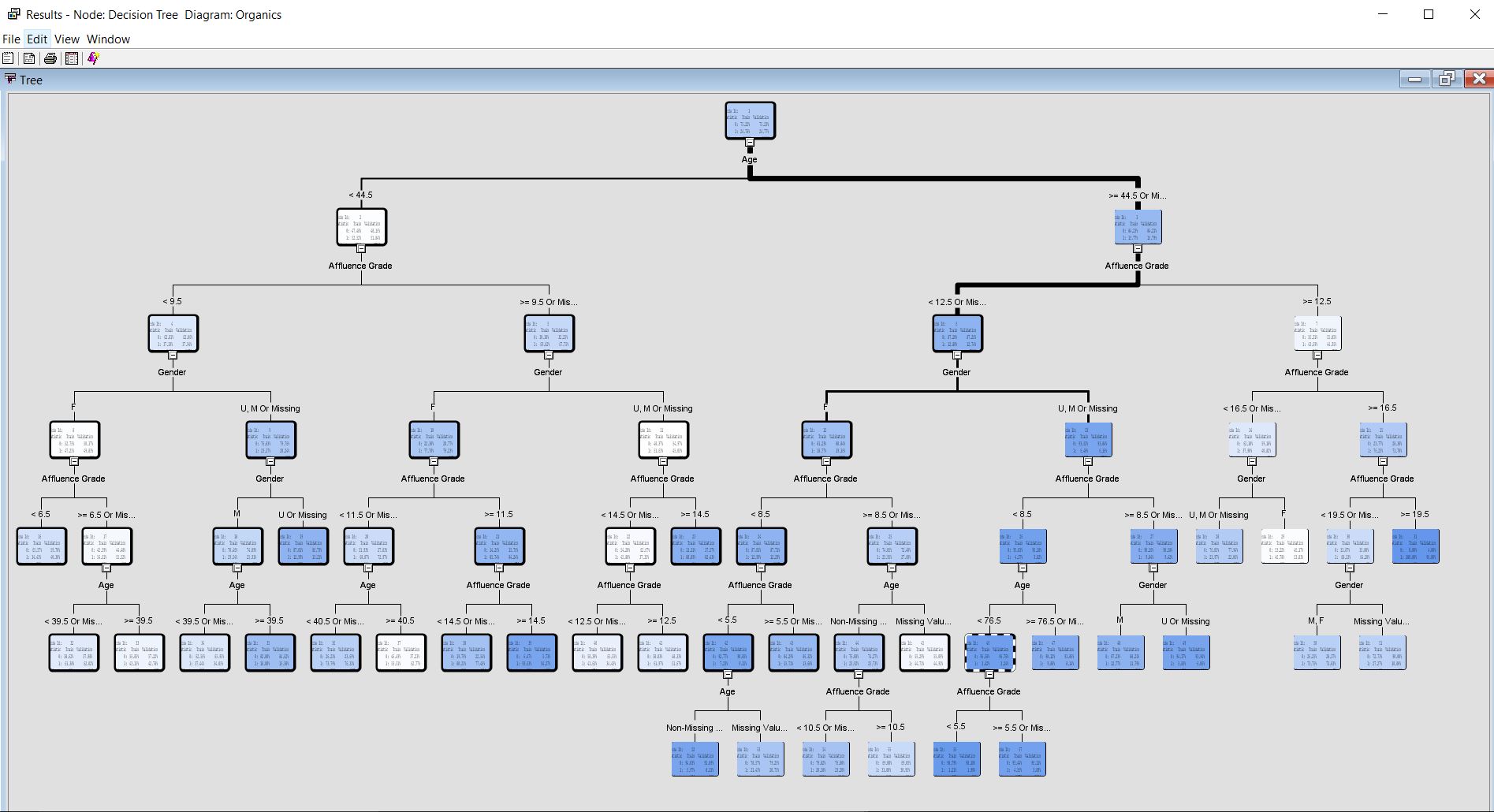


**e).** Added a Decision Tree node to the workspace and connected it to the Data Partition node as shown.



**f).** Created a decision tree model automatically. Used average square error as the model assessment statistic.

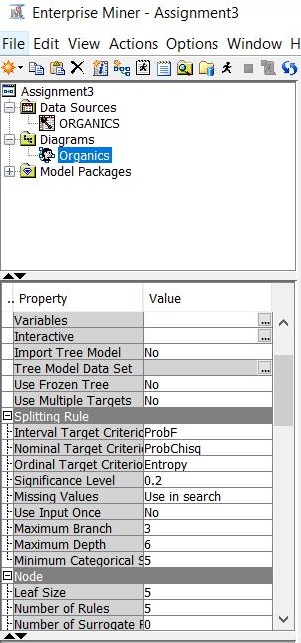
****

****

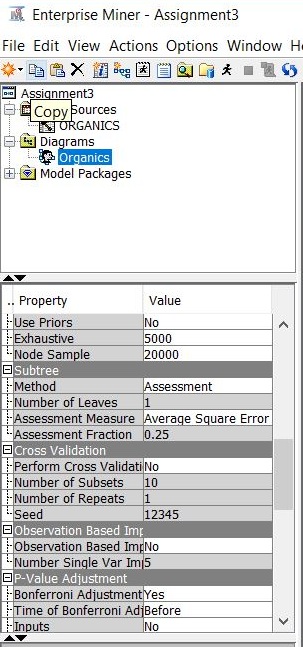
1. There are 29 leaves in the optimal tree.
2. The variable “Age” is used as the first split.

**g).** Added a second Decision Tree node to the diagram and connected it to the Data Partition node.

1. The maximum number of branches is set to 3 as shown below.



1. Used average square error as the model assessment statistic to create decision tree model.



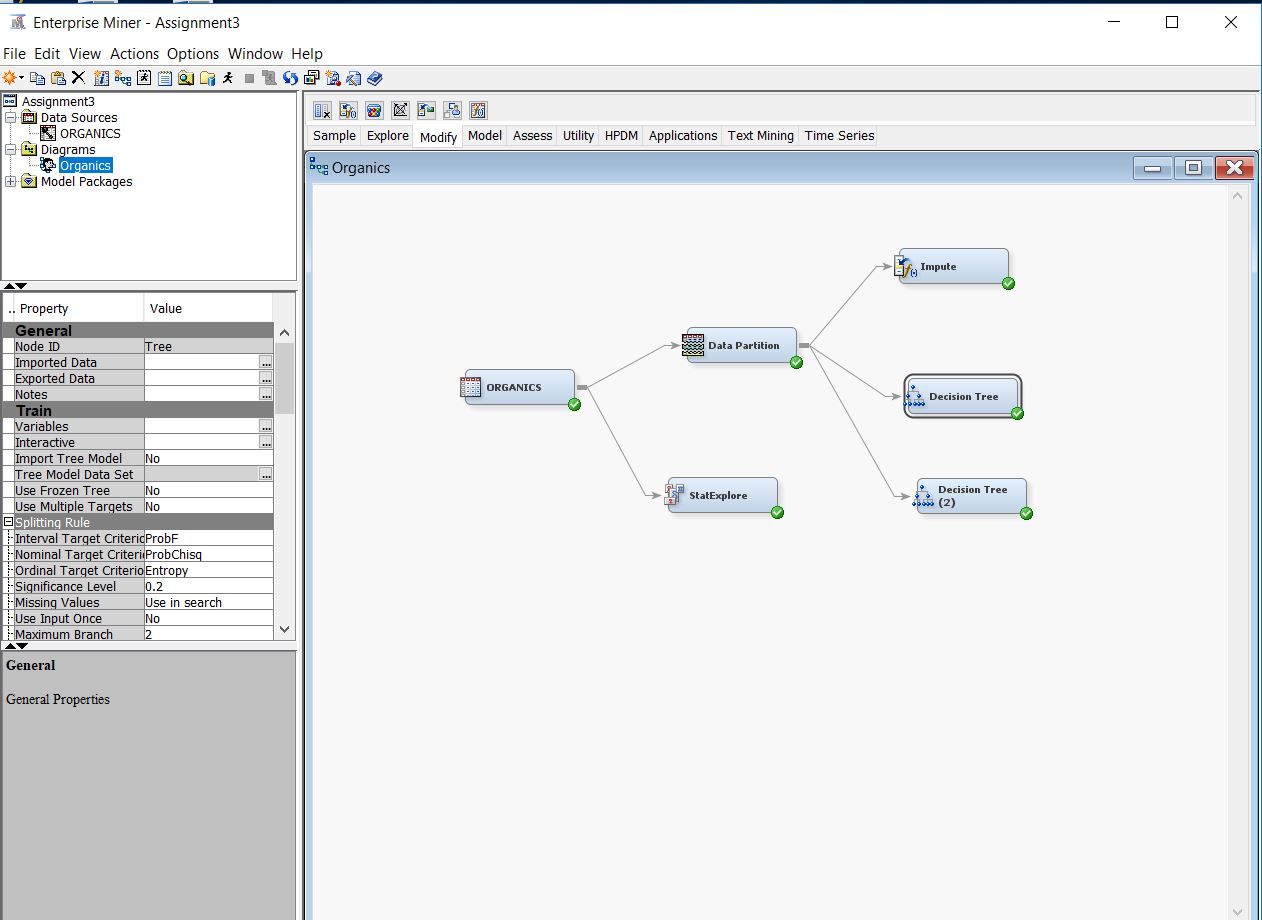
1. There are 33 leaves in the optimal tree.

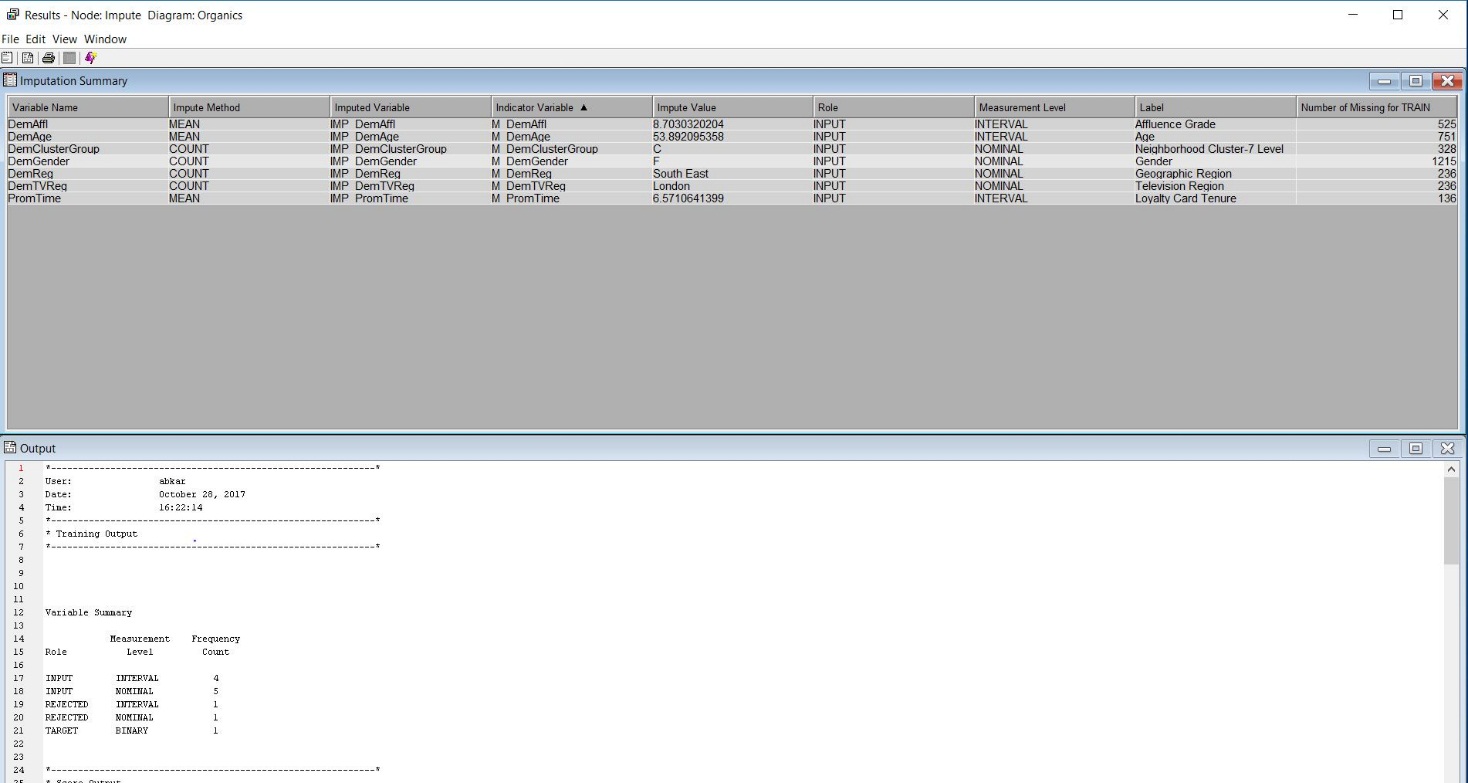
**h).** The first decision tree model is better because as the number of splits increases it will increase complexity and inherent chance of increasing error. It would be good for a training data set, but for a validation set it would eventually increase the error rate.

**Solution 2:**

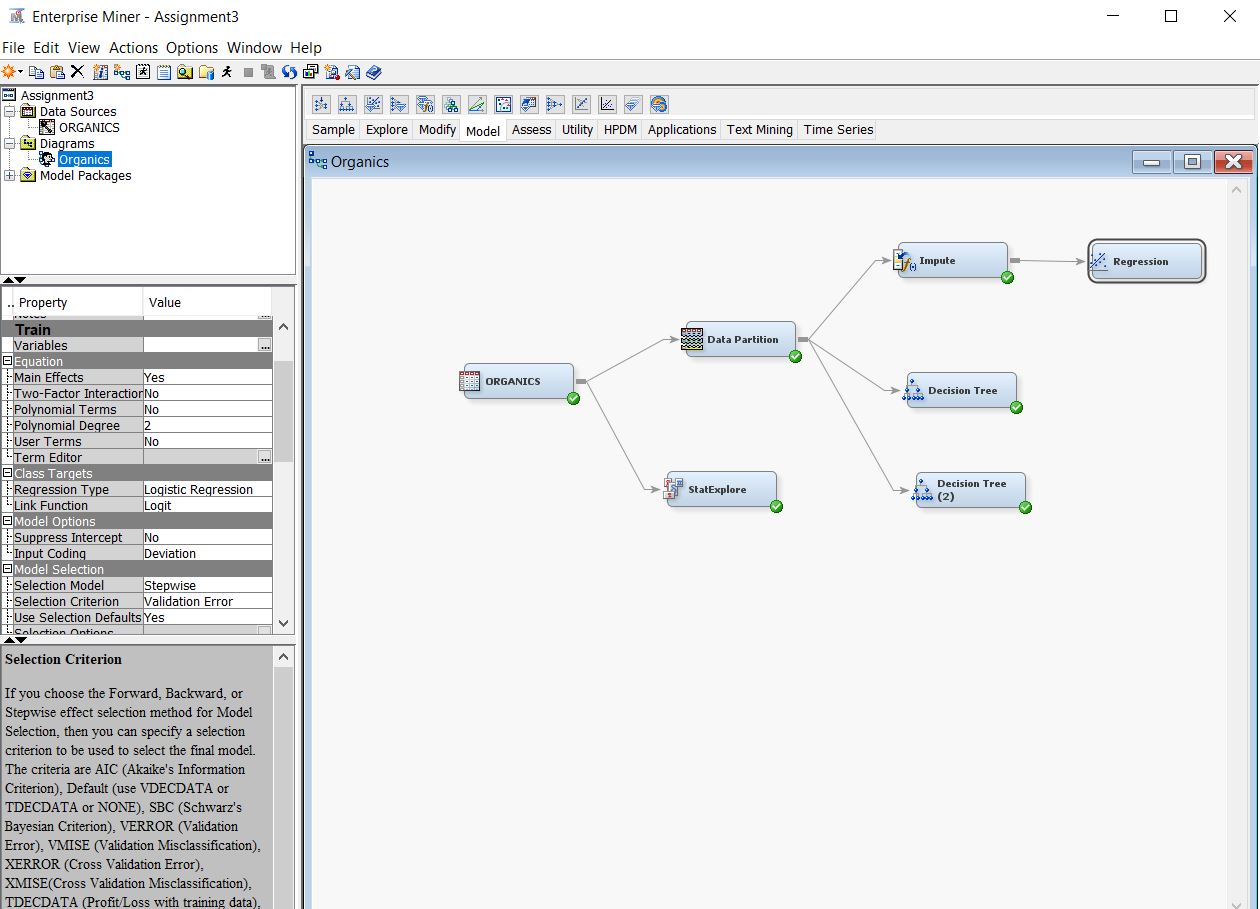
**a).** Yes, there are missing values in the data used for this problem set. However, you do not need to impute the data because for decision trees, missing values are not problematic. Surrogate splitting rules enable you to use the value of other input variables to perform a split for observations with missing values. Models such as regressions and neural networks ignore altogether observations that contain missing values, which reduces the size of the training data set. Less training data can substantially weaken the predictive power of these models. To overcome this obstacle of missing data, you can impute missing values before you fit the models.

**b).** Added an Impute node to the diagram and connected it to the Data Partition node. Set the node to impute U for unknown class variable values and the overall mean for unknown interval variable values.

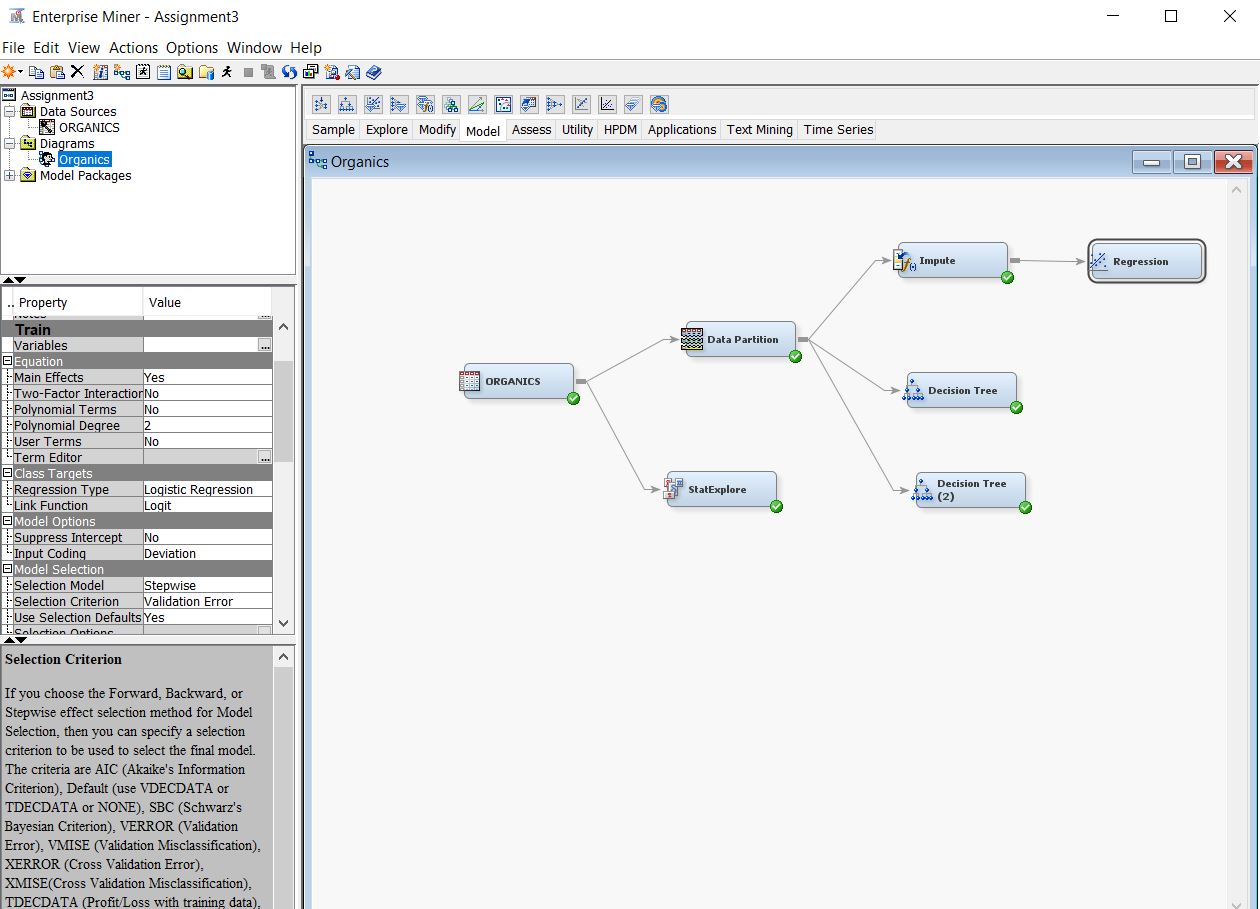




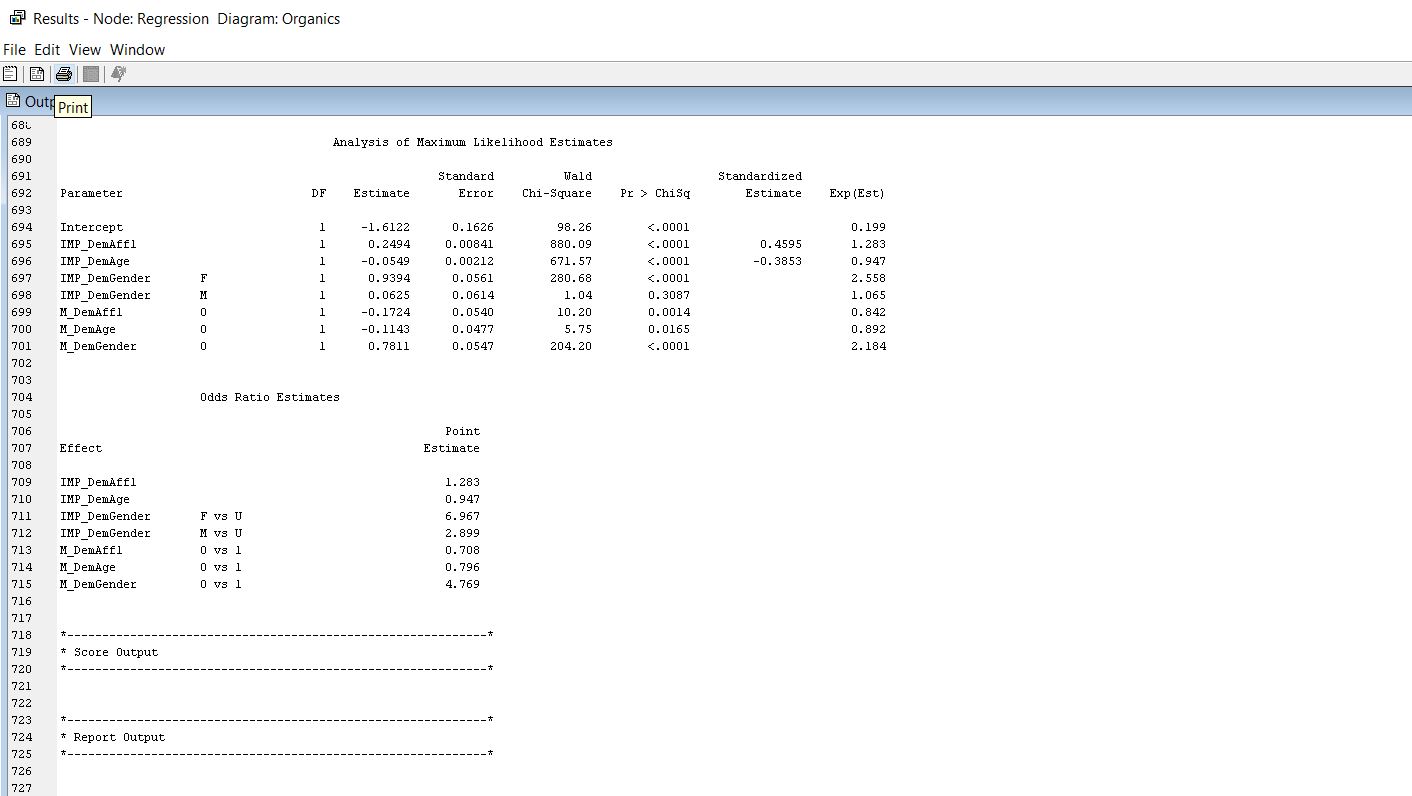
**c).** Added a Regression node to the diagram and connect edit to the Impute node.



**d).** Chose stepwise as the selection model and the validation error as the selection criterion.

****

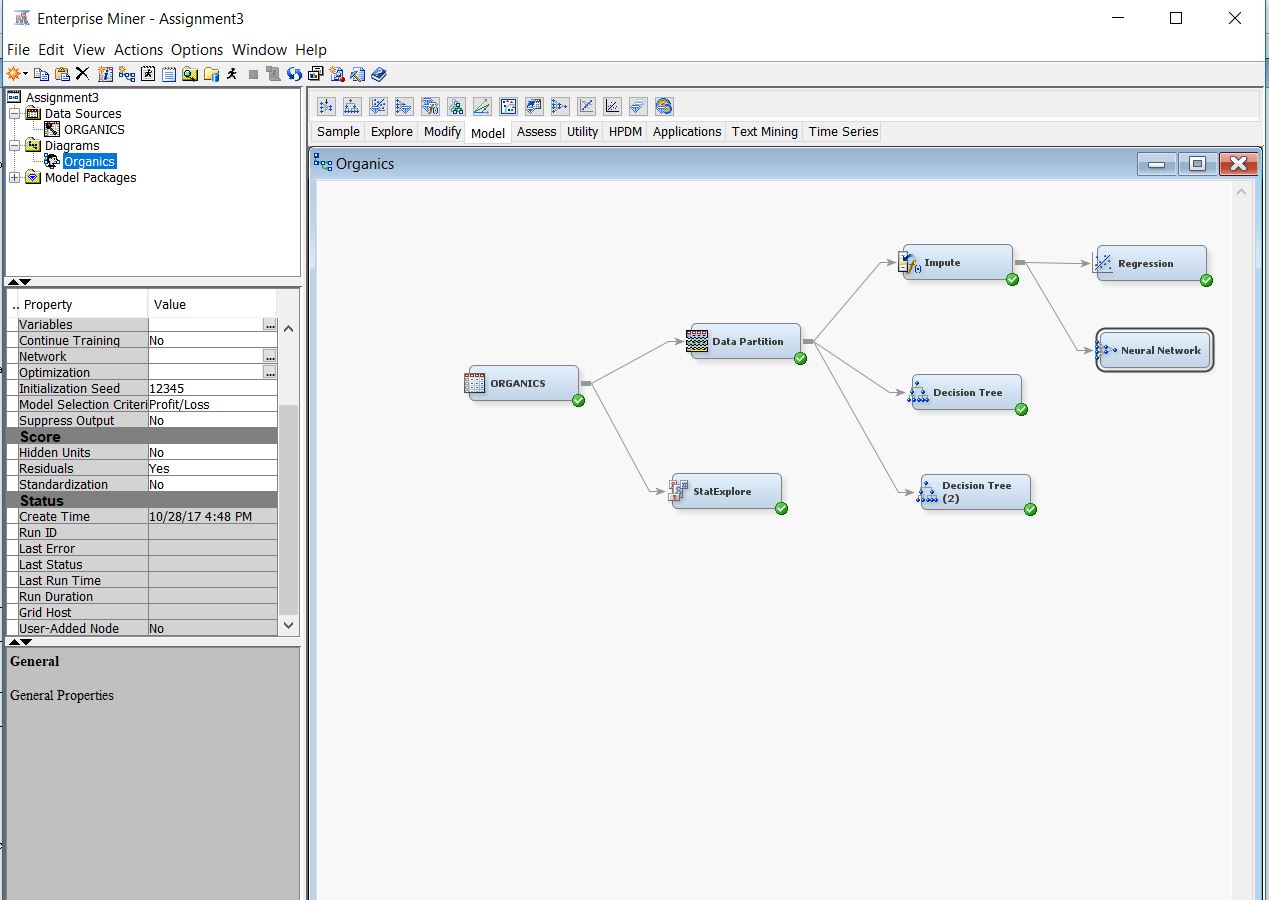
**e).** Ran the regression model and the following variables exist in the final model.



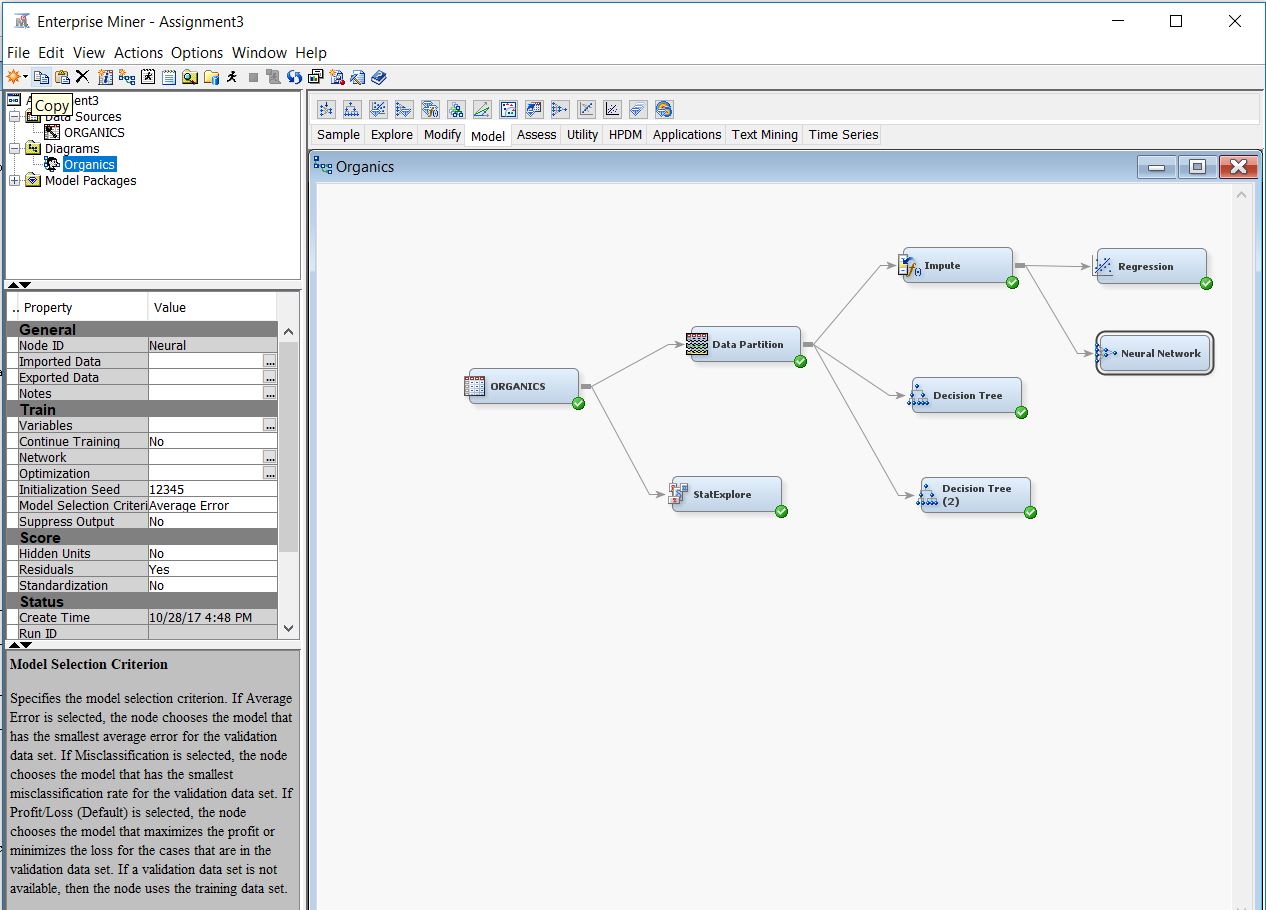
**Solution 3:**

Conducted classification using Neural Network.

**a).** Added a Neural Network tool to the Organics diagram. Connected the Impute node to the Neural Network node as shown below.



**b).** Model selection criterion is set to average error as showcased below.



**c).** The average square error for the different model we ran are as follow:

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Name | Training Data Set | Validation Data Set |
| 1 | Decision Tree 1(with branch node=2) | 0.1329 | 0.1328 |
| 2 | Decision Tree 2(with branch node=3) | 0.1330 | 0.1327 |
| 3 | Regression Model | 0.1389 | 0.1372 |
| 4 | Neural Networks | 0. 1318 | 0.1335 |

Based on this data, we can see that the average square error is lowest for neural networks in Training data set (0.1318) where as it is lowest for decision tree 2 in Validation data set (0.1327).