INDIAN DIABETES ANALYSIS

**GROUP PROJECT MIS 6324.001**

The University of Texas at Dallas

Veronica Storaci

Amruutha Chandrasekar Rao

Kashish Vermani

Adwait Joshi

Table of Contents

1. Executive Summary............................................................................................1
2. Background.........................................................................................................2
3. Introduction.........................................................................................................3
4. Motivation...........................................................................................................4
5. Description of Data Source ................................................................................5
6. Data Preparation................................................................................................6
7. BI Techniques Used. ........................................................................................7
   1. Decision Trees........................................................................................8
   2. Logistic regression .................................................................................9
   3. Neural network........................................................................................10
8. Conclusion.........................................................................................................11
9. References.........................................................................................................12

**INTRODUCTION**

In this project, we will use three models of predictive analysis to identify the traits that lead women to develop diabetes. We will use decision trees, neural networks, and logistic regression to do this. After implementing all of these models, we will share our findings and interpret our results, as well as choose the best model regarding their prediction ability.

Our objective is to find out whether we can positively identify the traits that lead to the development of diabetes or at least find a correlation between individual characteristics and the occurrence of diabetes.

**EXECUTIVE SUMMARY**

In this project we focus on trying to predict whether a person is more likely to develop diabetes based on several variables like glucose levels and age. India is a large country and supposedly has a massive amount of people suffering from type 2 diabetes. The data set we have acquired is focused only on women above the age of 21 of Indian heritage.

We will conduct tests on two types of decision trees, the two stem and three stem trees, as well as logistic regression and neural networks, and then we will find out which one of these models is the best one based on the squared error. From this we can them draw conclusions about the female population in India.

**BACKGROUND**

There are two types of diabetes that a person may have. Though doctors do not know precisely what causes diabetes, there are specific factors that might increase the risk. Type 1 diabetes is theorized to be caused by some form of autoimmune disorder, where the white blood cells in the body confuse healthy insulin cells in the body for malignant ones and start attacking them (Healthline, 2015). Type 2 diabetes is presumed to be caused by external factors, for example, diet, weight and glucose levels in the body, which means it can be prevented if precautions are taken (Healthline, 2015). Being able to predict whether someone has a higher possibility of developing diabetes can be an extreme benefit to those who have a higher chance of developing it.

The dataset we chose has sought information from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD). This dataset was exposed to some constraints that narrow down the data and focuses only on women above the age of 21 who are of Indian heritage. Meaning that any conclusion that we draw would only apply to the total population of women of Indian Heritage. Since “[India] is often known as the diabetes capital of the world, [and] has been witnessing an alarming rise in incidence of diabetes according to the International Journal of Diabetes in Developing Countries” (Malik, R. 2016), we believe that this data could help us make inferences about why diabetes happens and what are the factors that can increase the chances of a person developing the condition during their lifetime.

**MOTIVATION**

The motivation for this project is simple as a formentioned; our dataset can help us predict whether or not specific traits have any influence over whether a person will get diabetes during their lifetime. With this information, we will be able to prevent those people who could develop this disease from getting it in the first place. As mentioned before the dataset is a sample of only women of Indian heritage above the age of 21, which limits the overall assumptions that we could make for the population. Overall, the information we can gain from this dataset is massive, and even though the dataset is very niche, it can still help us accurately make assumptions about the female population in India and their chances of developing diabetes.

We found this dataset to be very interesting and even though we had started this project with a different dataset, we found it very difficult to stay motivated and could not come up with a clear objective. After a while, we started considering other datasets and found this one, since three fourth people are of Indian descent we thought it would be an interesting topic as well as relevant to the group itself. Also, there are so many implications of learning how to do predictive analysis with SAS Enterprise Miner that will help us in the future. We thought that this was the best data set that would help up learn to use the tool and master it to a degree.

**DESCRIPTION OF DATA**

Data Source: The publicly available dataset was downloaded from Kaggle.com

(<https://www.kaggle.com/uciml/pima-indians-diabetes-database>).

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Indian heritage.

**DESCRIPTION OF THE DATASET**

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, Blood Pressure, Skin Thickness, Glucose level and diabetes pedigree function. The outcome variable is whether that patient has diabetes or not (Outcome 1 or 0)

This dataset has included above mentioned nine types of different variables. These variables are described below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Attribute** | **Measurement role** |
| Pregnancies | Number of times pregnant | Numeric | Input |
| Glucose | Plasma glucose concentration a 2 hour in an oral glucose tolerance test | Numeric | Input |
| Blood Pressure | Diastolic blood pressure (mm Hg) | Numeric | Input |
| Skin Thickness | Triceps skin fold thickness (mm) | Numeric | Input |
| Insulin | 2-Hour serum insulin (mu U/ml) | Numeric | Input |
| BMI | Body mass index  (weight in kg/ (height in m) ^2) | Numeric | Input |
| Diabetes Pedigree Function | Diabetes pedigree function | Numeric | Input |
| Age | Age (years) | Numeric | Input |
| Outcome | Class variable (0 or 1) | Numeric | Target |

**DATA PREPARATION**

**SOFTWARE USED:**  SAS Enterprise Miner (Version 13.2)

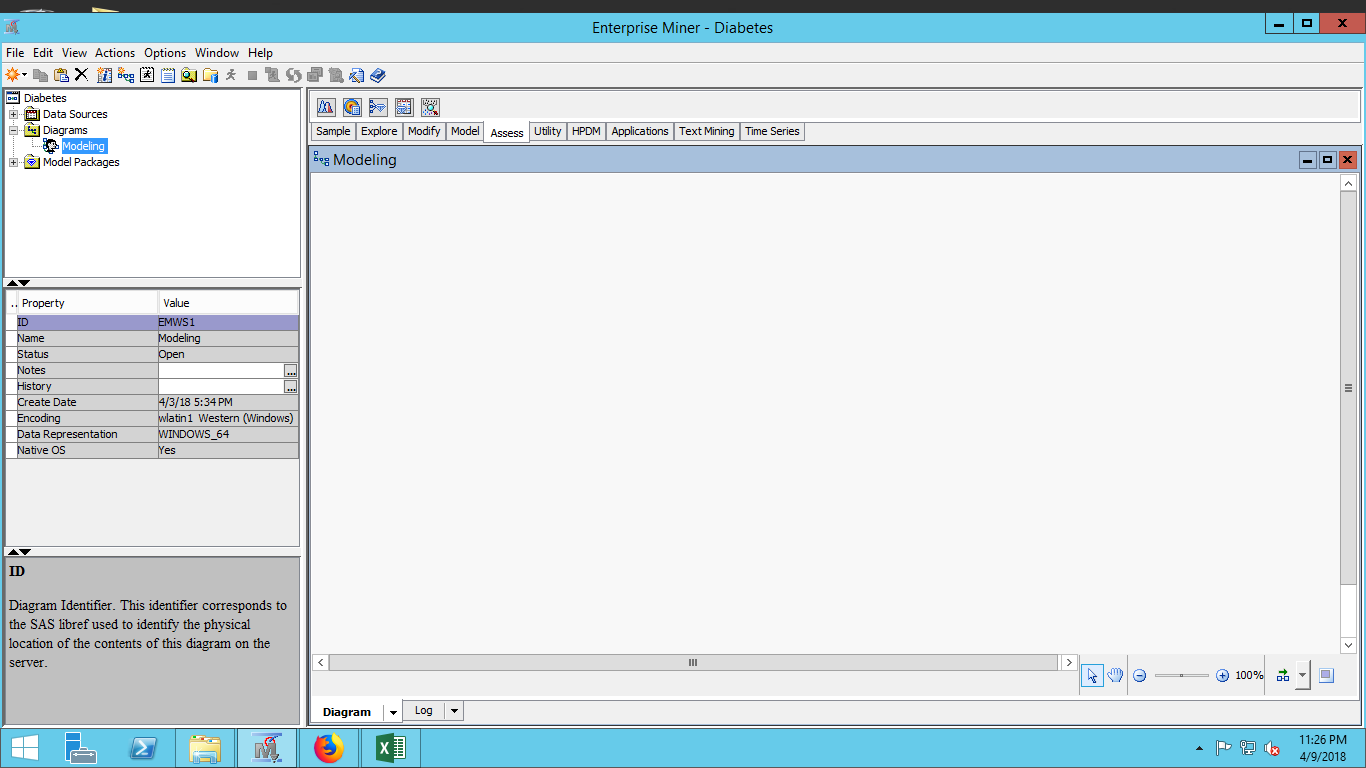
**DATA CLEANING**: There were few things that needed to be done to clean the data because it was already  clean, to begin with, there were very few instances of zero values in every input variable.

**MISSING VALUE:** Though we did not have any true missing values in our data, we did see a few instances where there were zeroes written as measurements. For example we had a couple of zero values in our Skin Thickness variable, which is not possible because no one can have skin that is 0 mm thick. We used the “Impute” node to replace those zero values with the mean value for that specific variable.

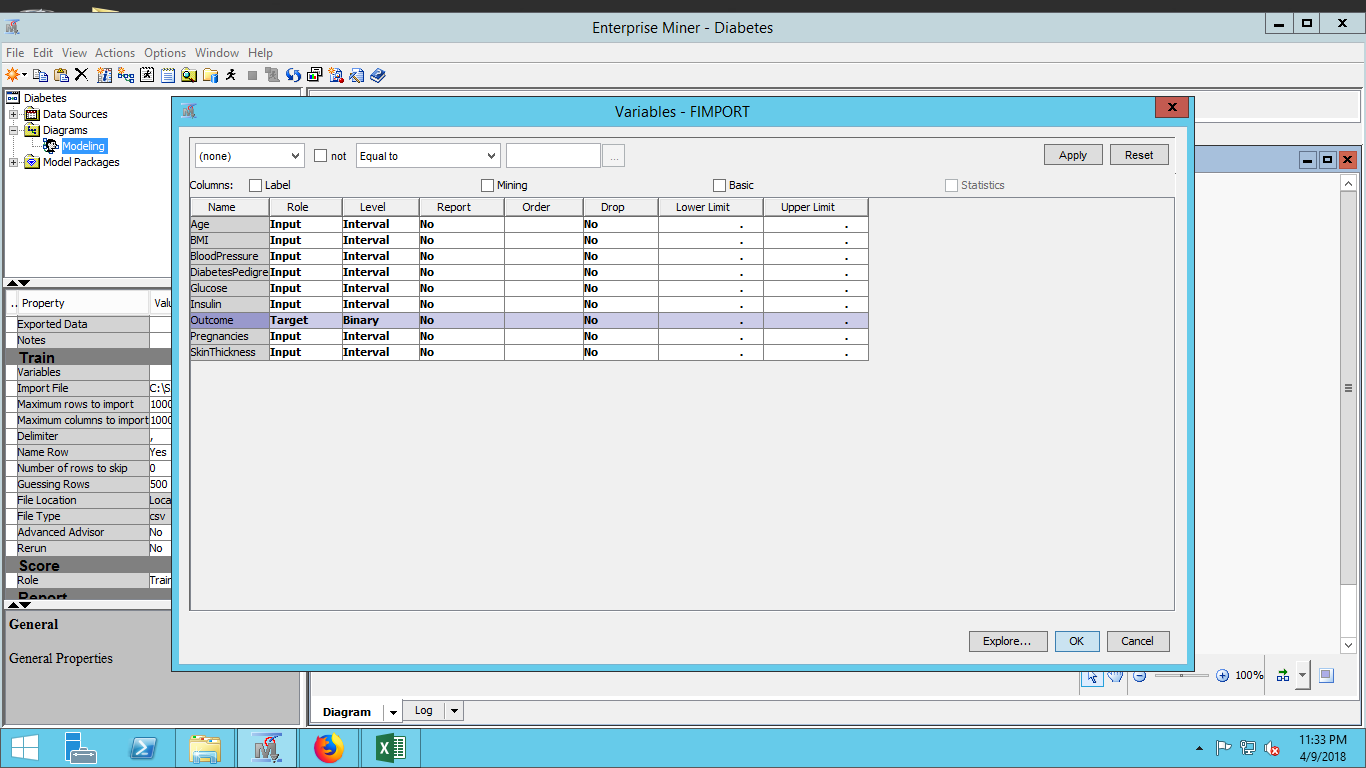
**BUSINESS INTELLIGENCE TECHNIQUES**

These are the descriptions for the BI techniques we used:

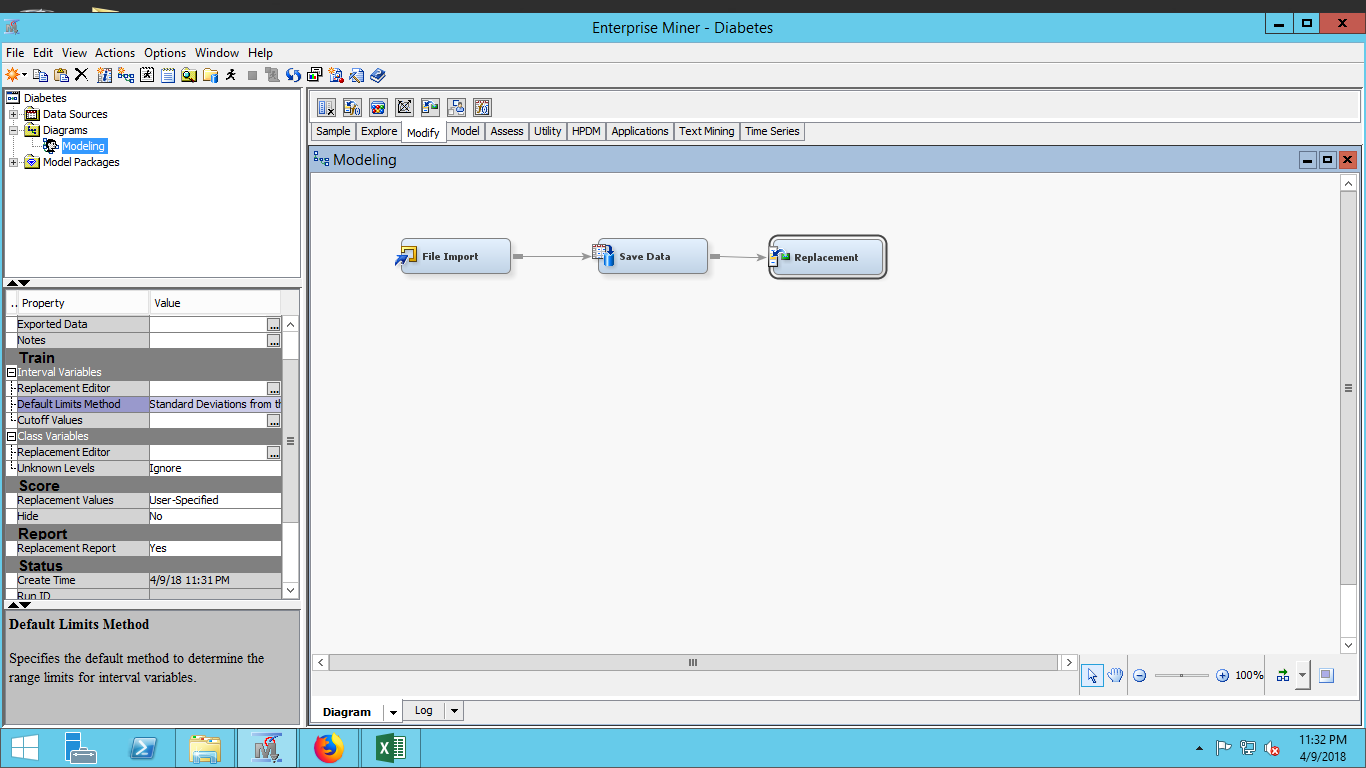
A diagram named modelling has been created in the project.

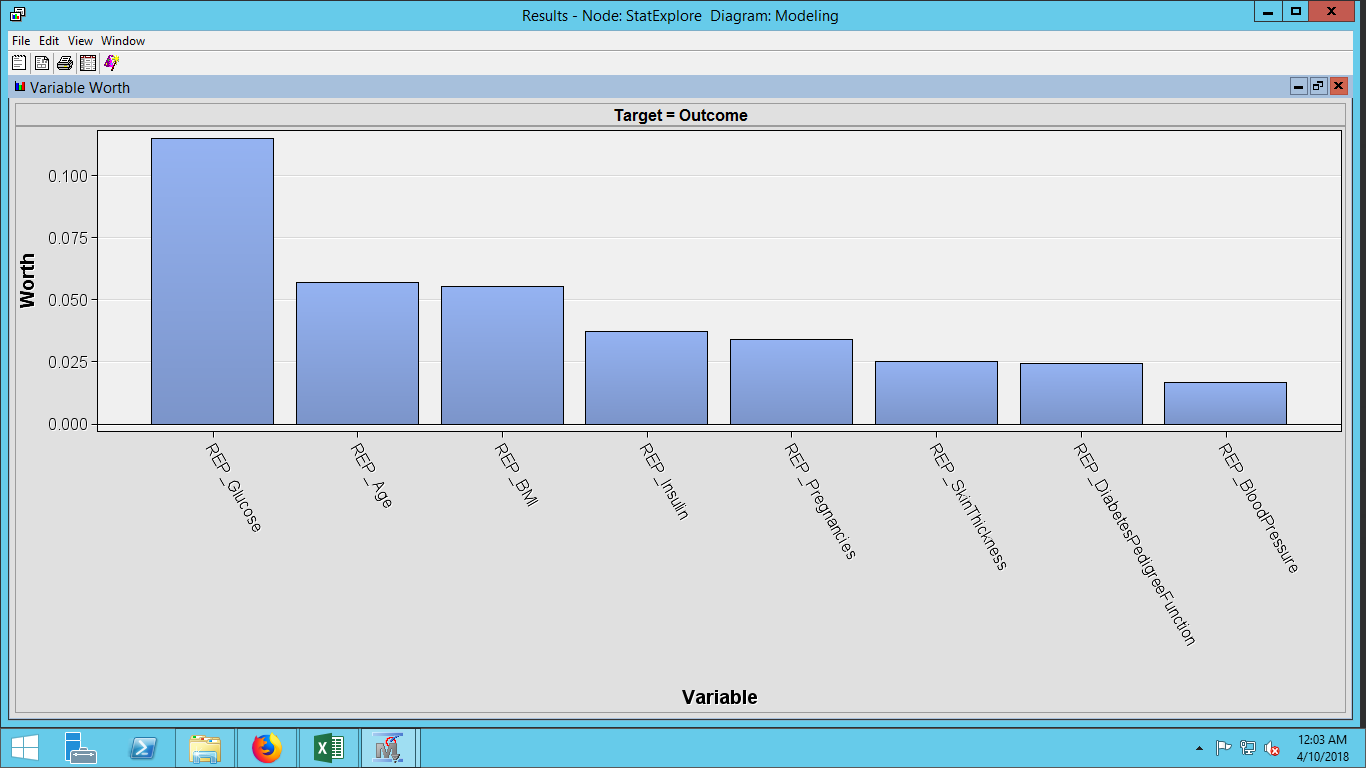
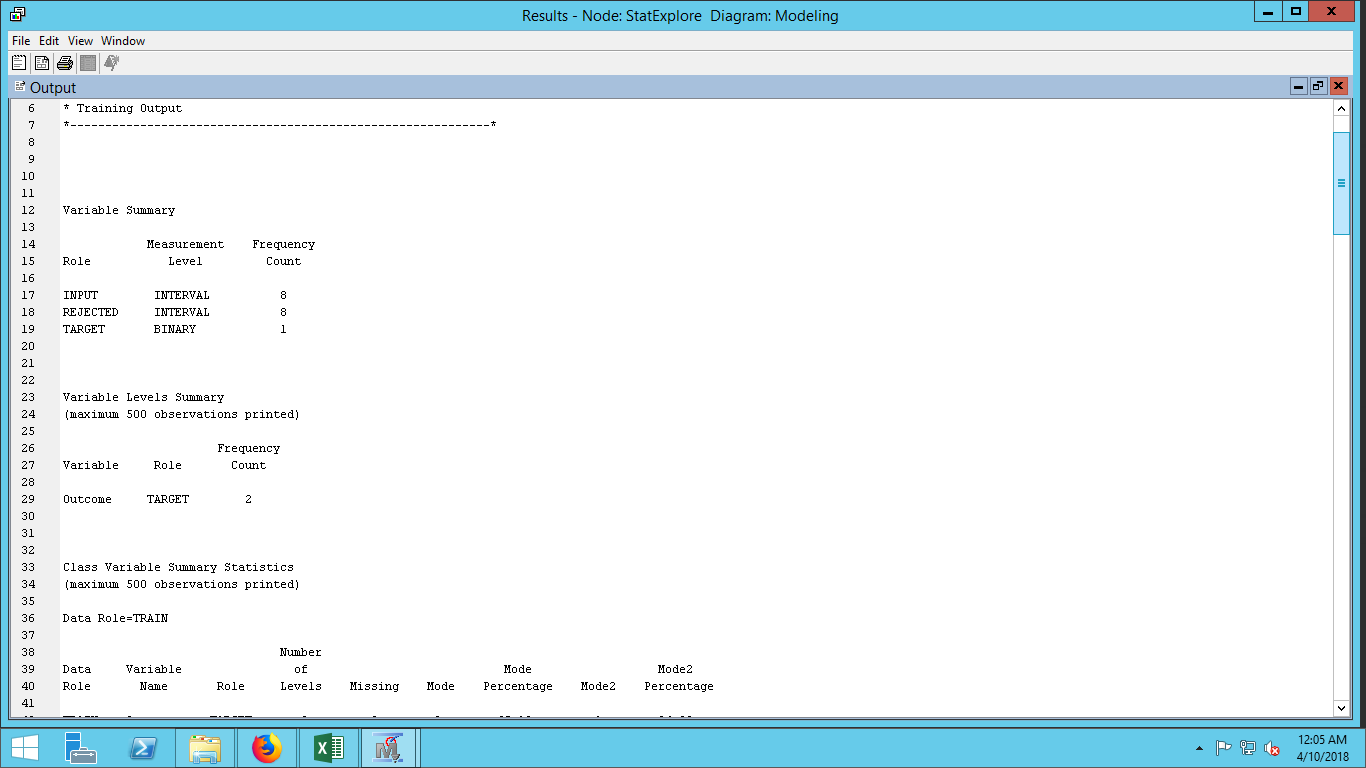


File import node has been added to the modeling diagram .This node is used to import file, we are using a comma separated file. The variables can be edited using this node and our team has set “outcome” column is set as target variable and other columns are set as interval variables.

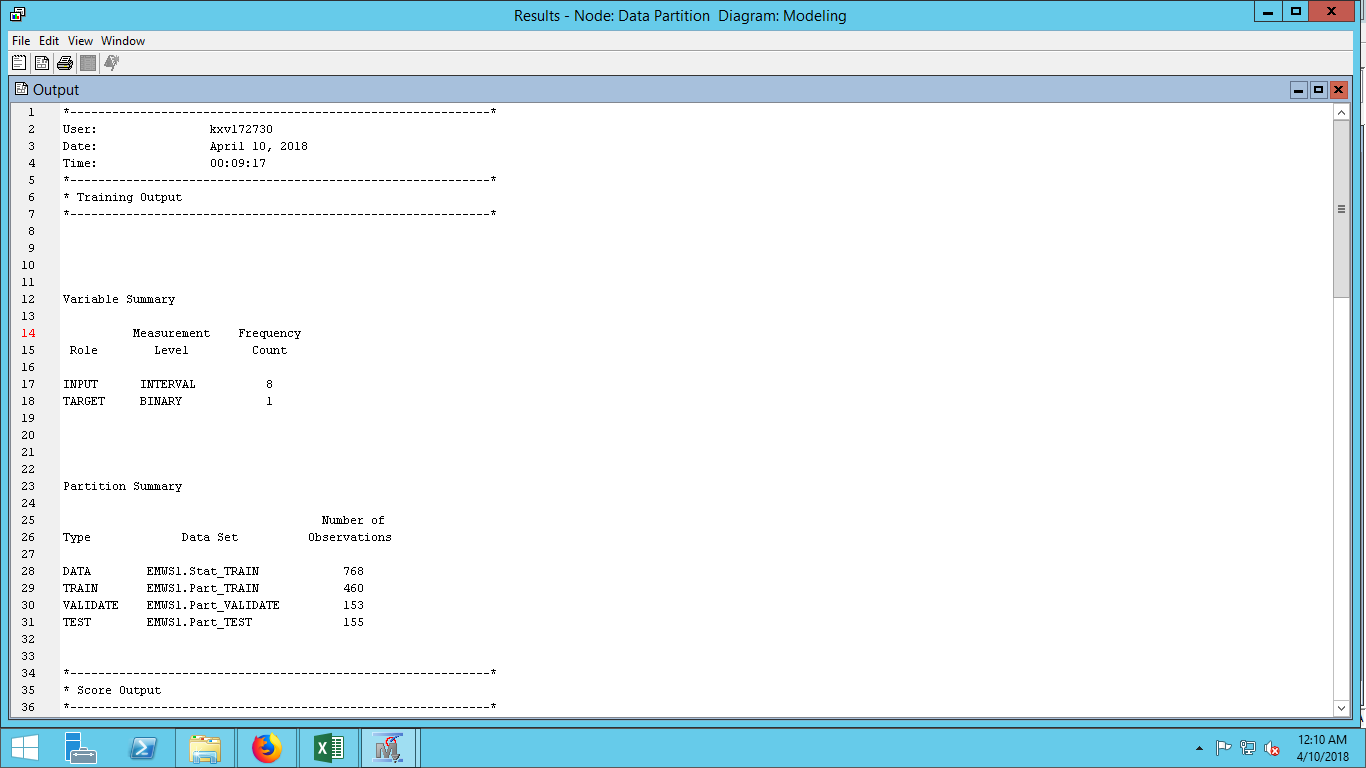


There is a save node which has been added to the diagram to save the data . The Save Data node can export JMP, Excel 2010, CSV, and tab-delimited files.

**REPLACEMENT NODE**The replacement node has been set withdefault limit methods as standard deviation from the mean and the replacement values is set as default.  
**STATSEXPLORER  
The stats explorer has been used**

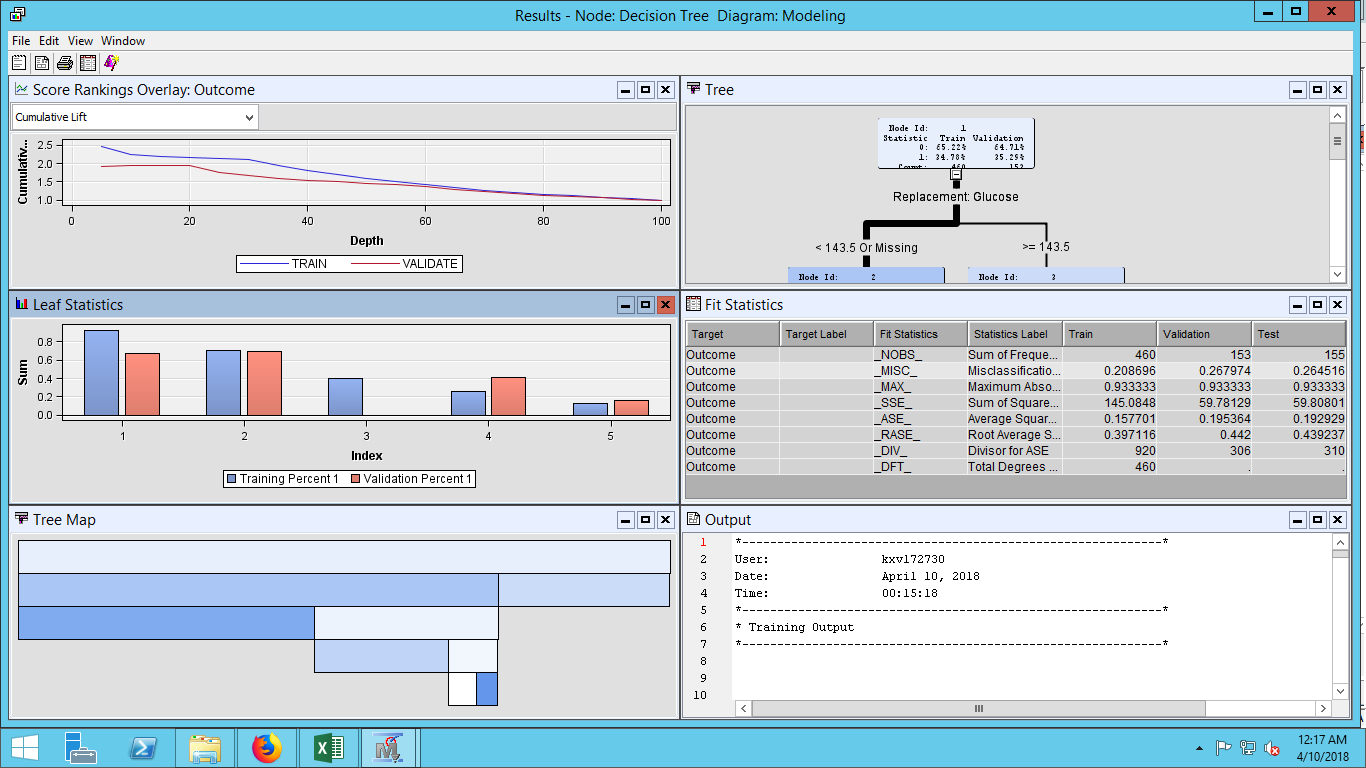
  


**DATA PARTITION**

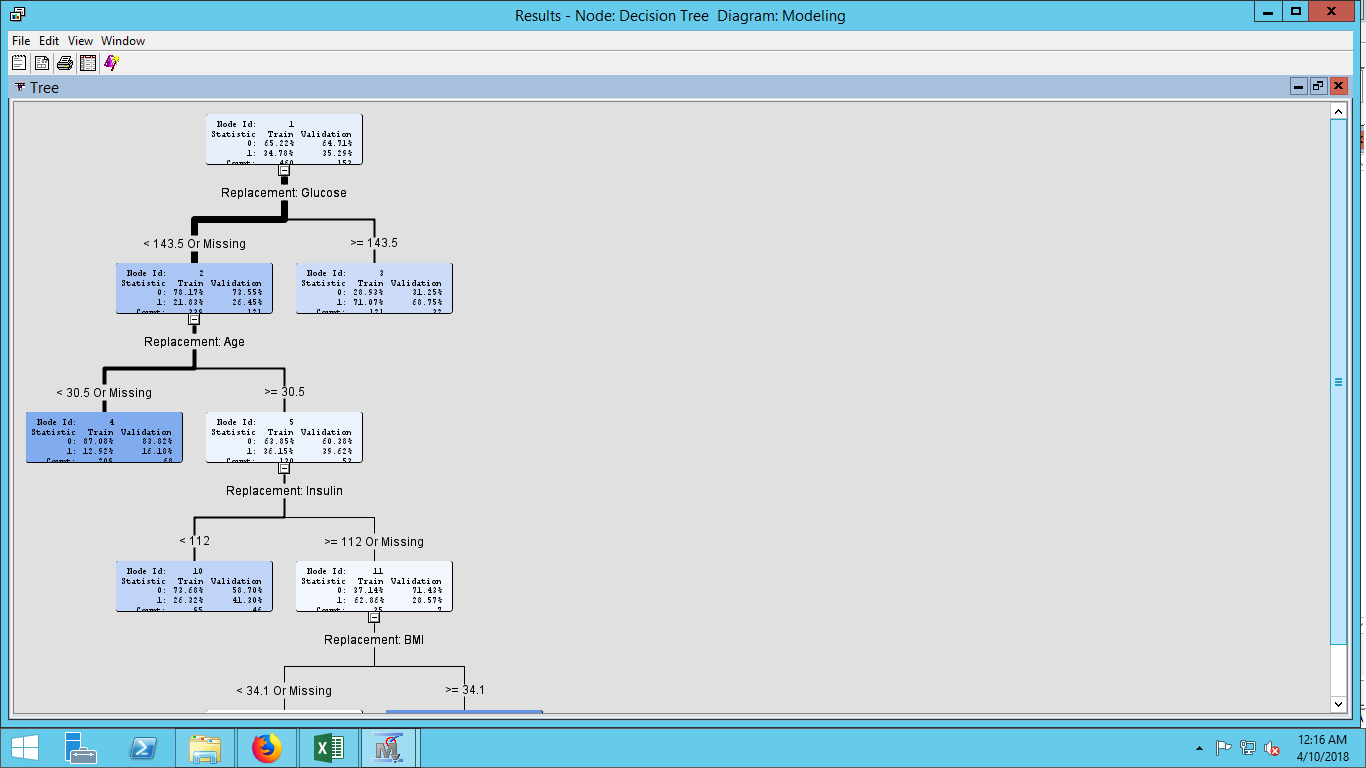
The data partition node is used to sample the data. The dataset is divided into two data tables training and validation. Our team has used 60 percent training dataset , 20 percent validation dataset and 20 testing dataset and the default partitioning of stratified is used.   
**  
MODELS FOR DATA MINING**

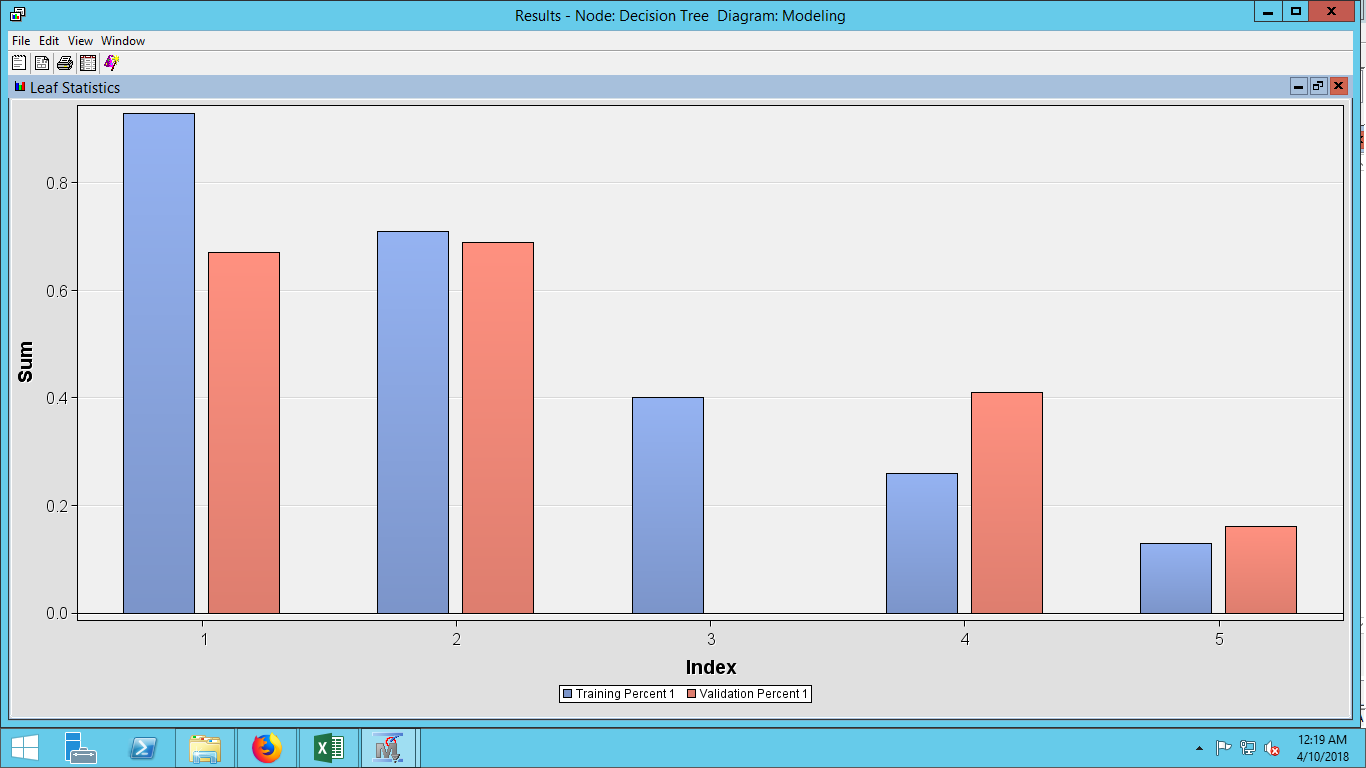
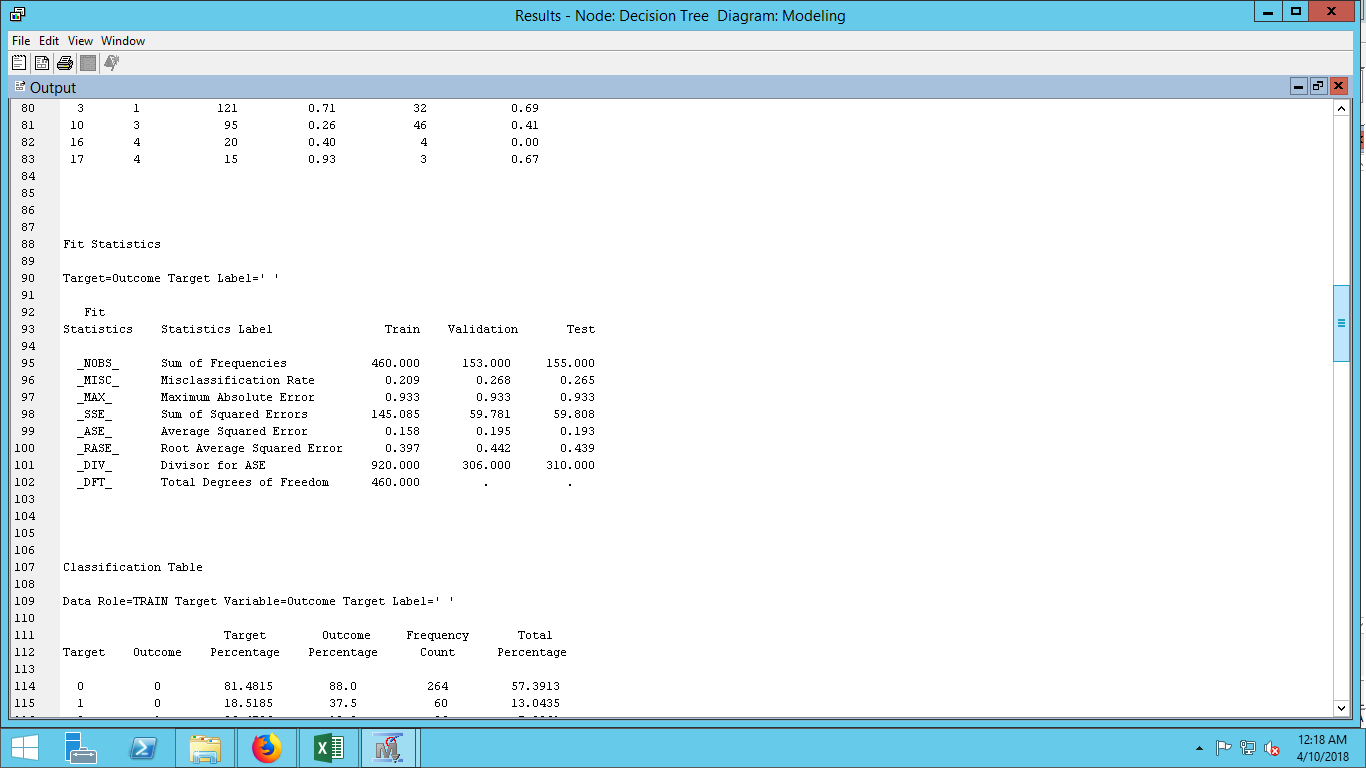
**DECISION TREES:**

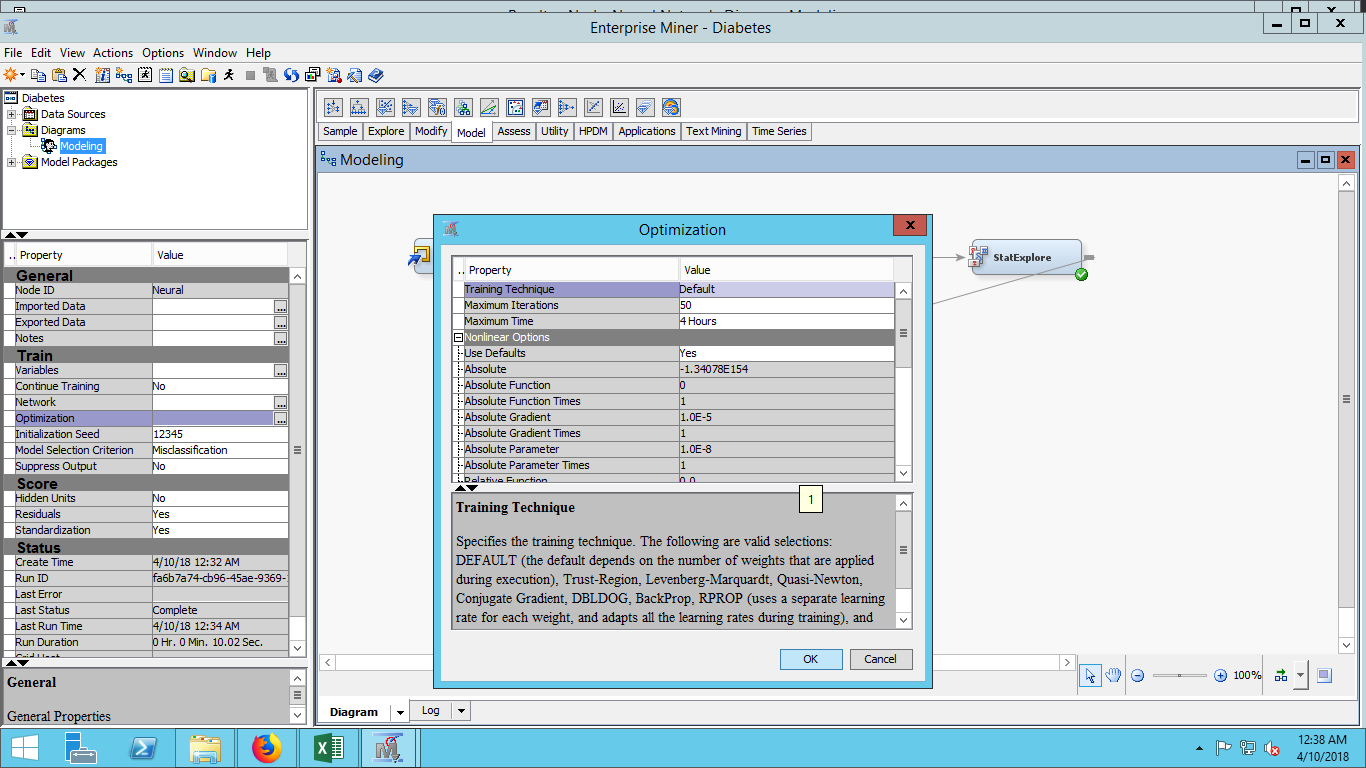
The Decision Tree node can be interactively trained and pruned. This step would involve lot of manual efforts and hence we have configured to run automatic. The minimum branch for the number of nodes from root node is set as two and the default value of 10 is set for the depth of the number of generations or the depth of the root node and the significance level of the splitting. The assessment measure which can be used to specify the optimality measure u to select the best tree is chosen as misclassification rate method.

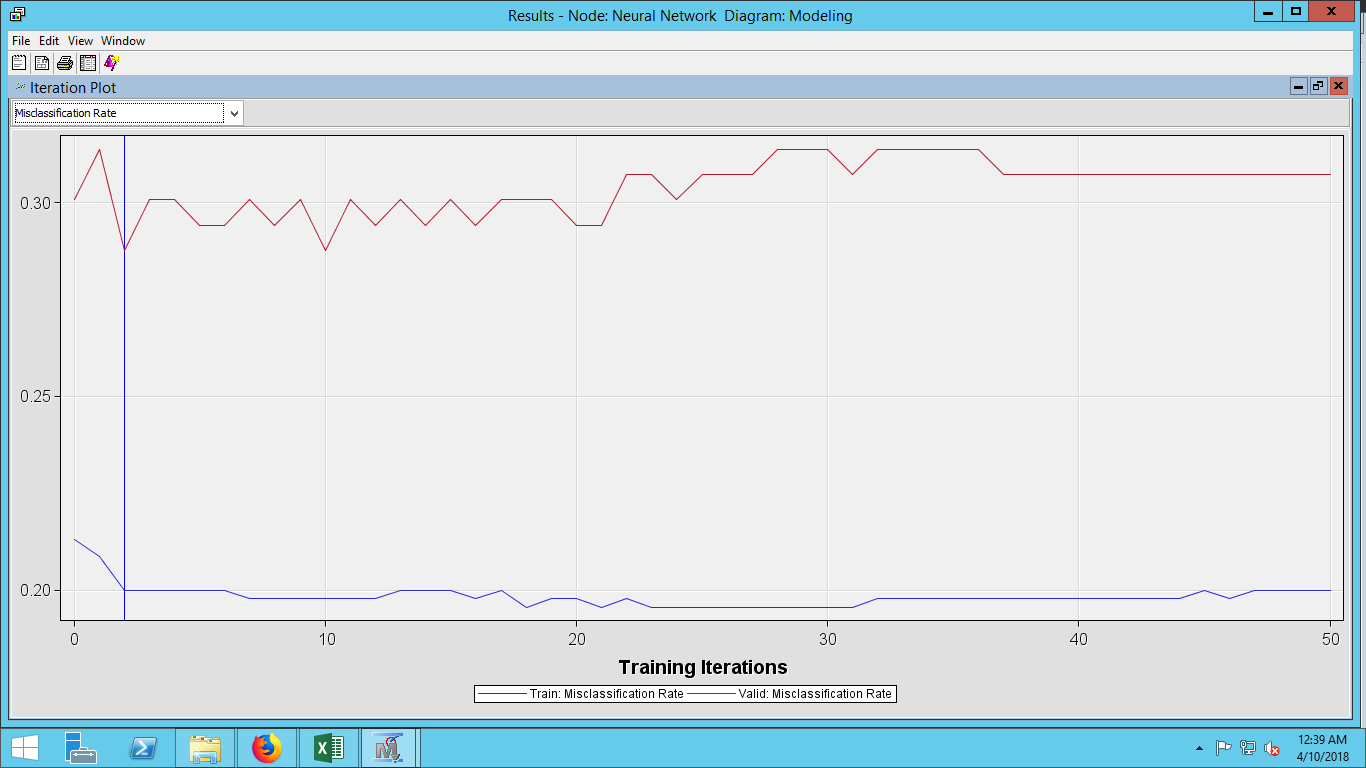
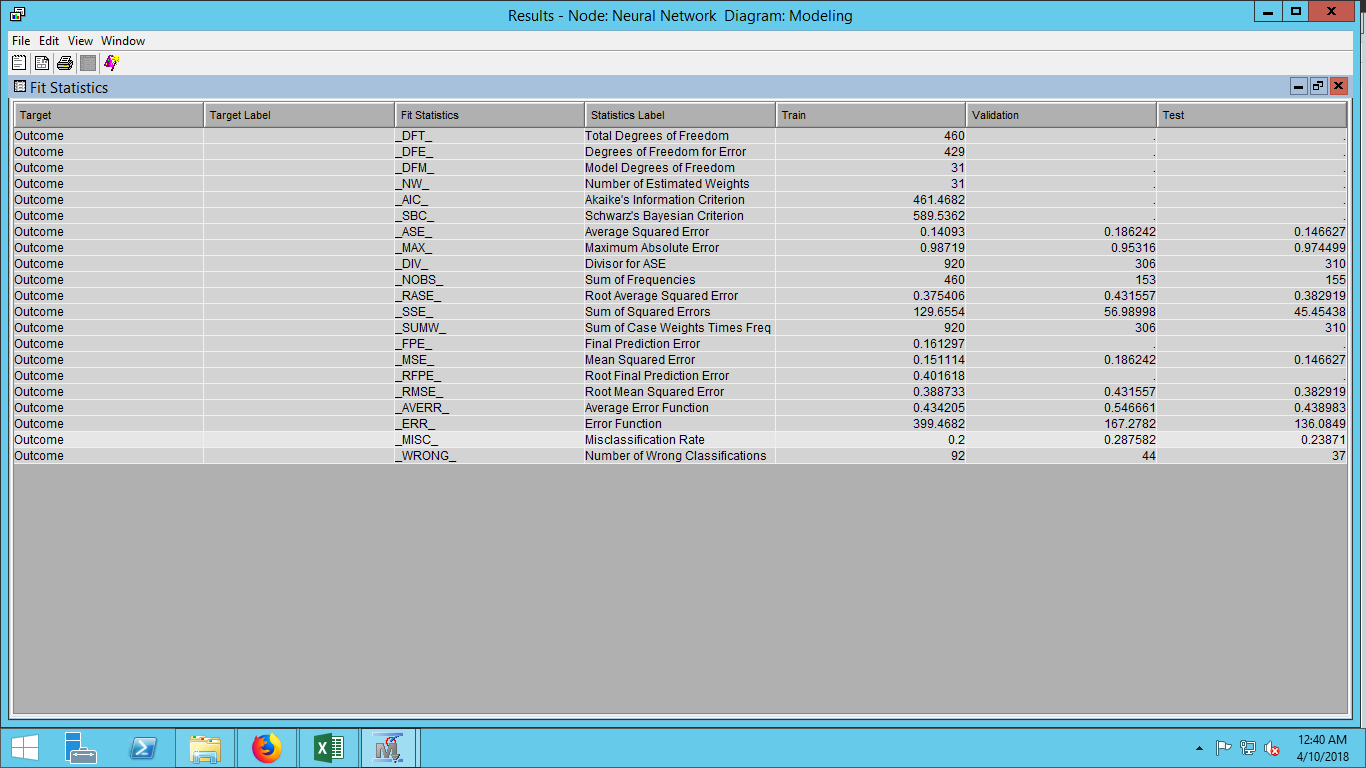


This decision tree has misclassification rate of 0.2679 and the average square error is 0.1953.

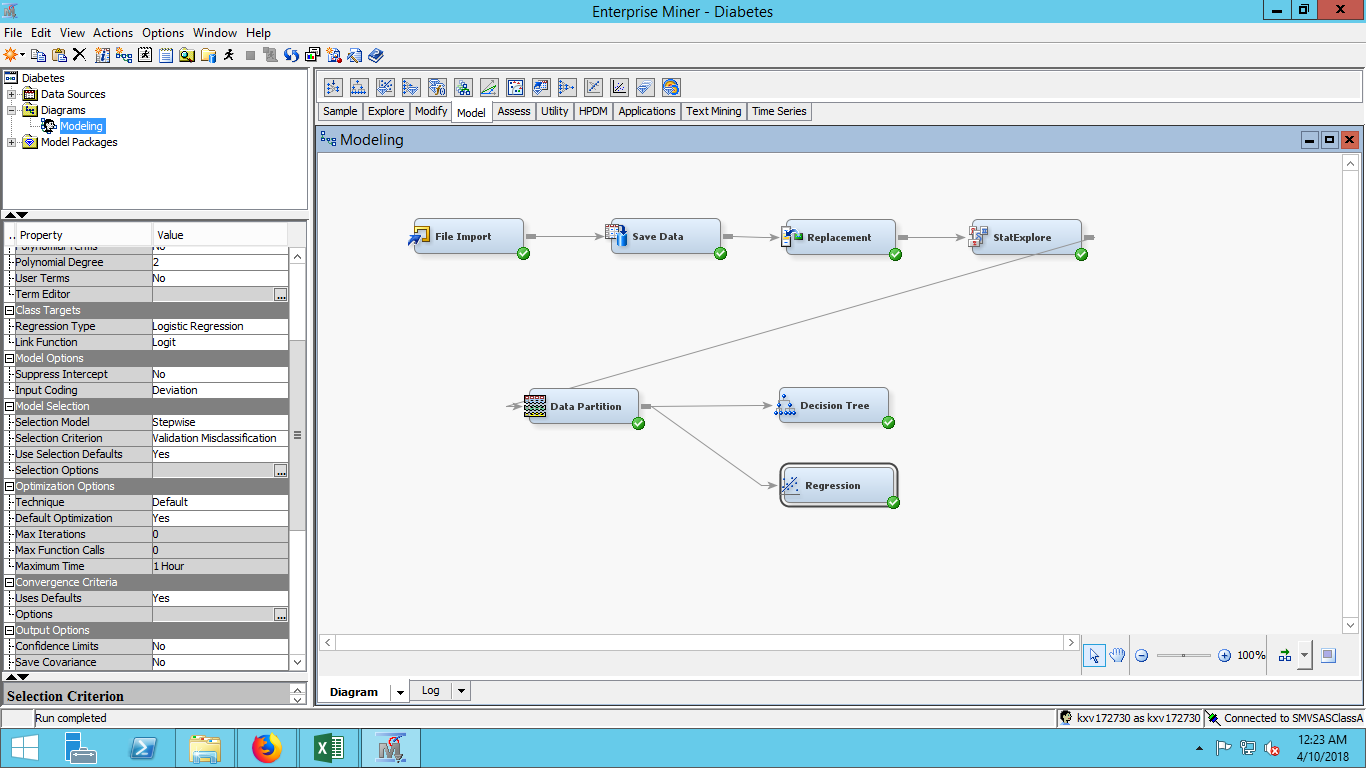
There are five leaf nodes in this decision tree.Leaf nodes are used to predict the desirable outcomes. There is a possibility of 68.75 percent that a woman above 21 years, having glucose level of 143.5 or above.

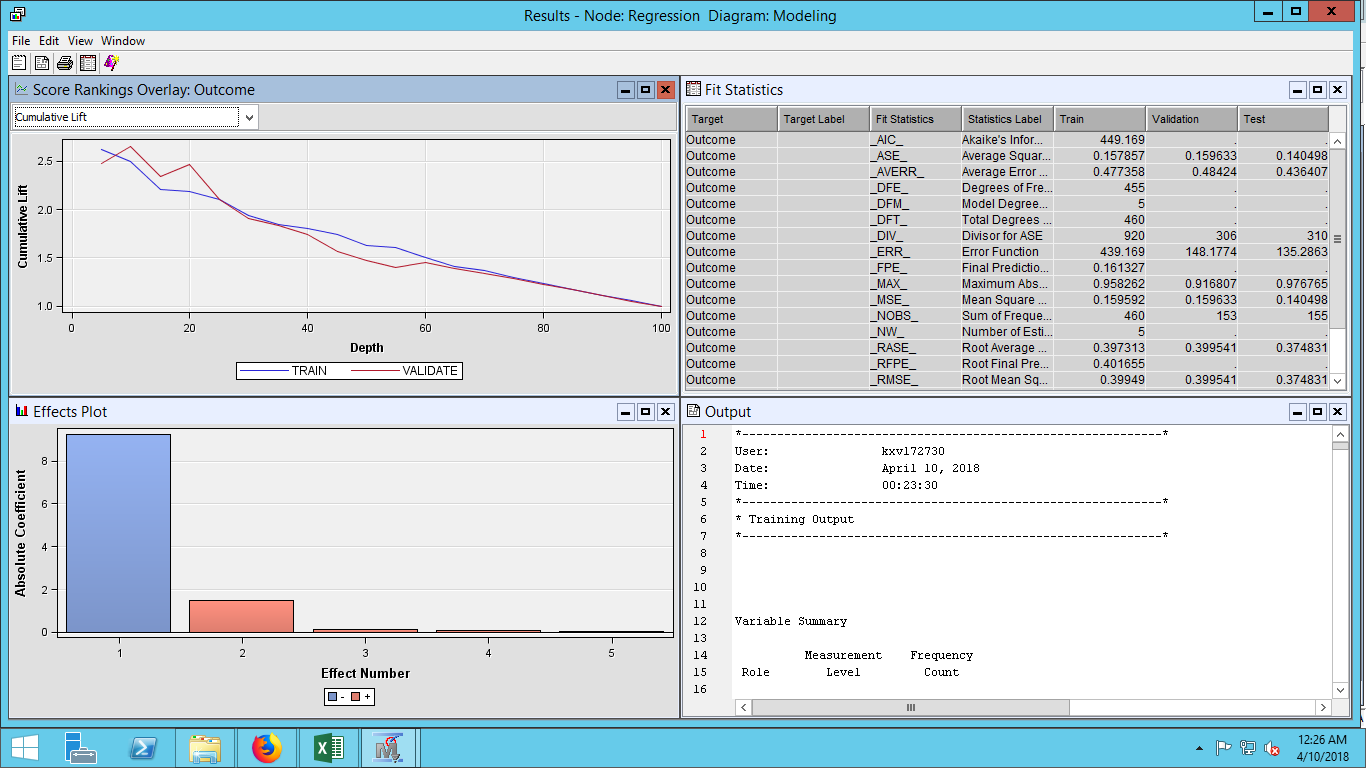
  


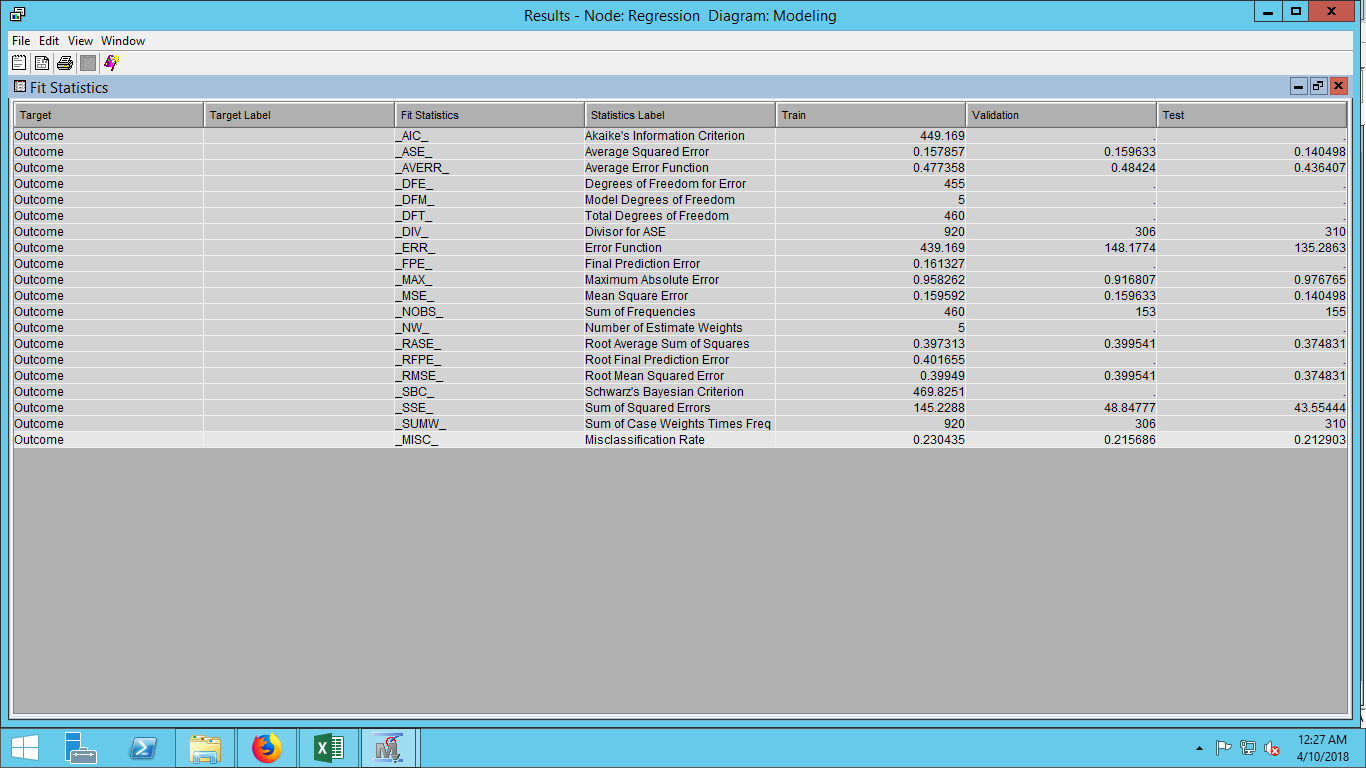
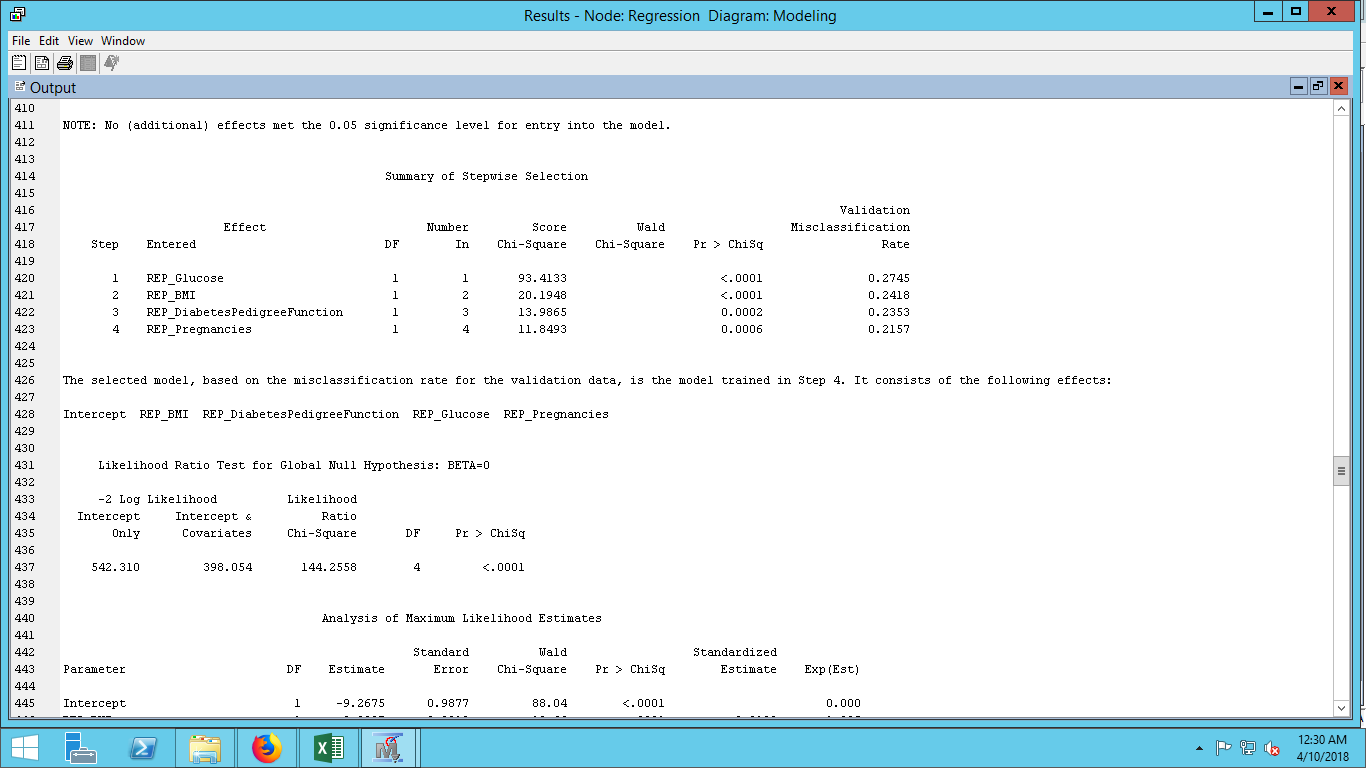
**NEURAL NETWORKS:**A class of parametric models that can accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable.   
The model selection criteria has been selected as misclassification rate and the standardization rate is set to yes.The optimization has default training technique and the maximum iteration is changed to 50. 

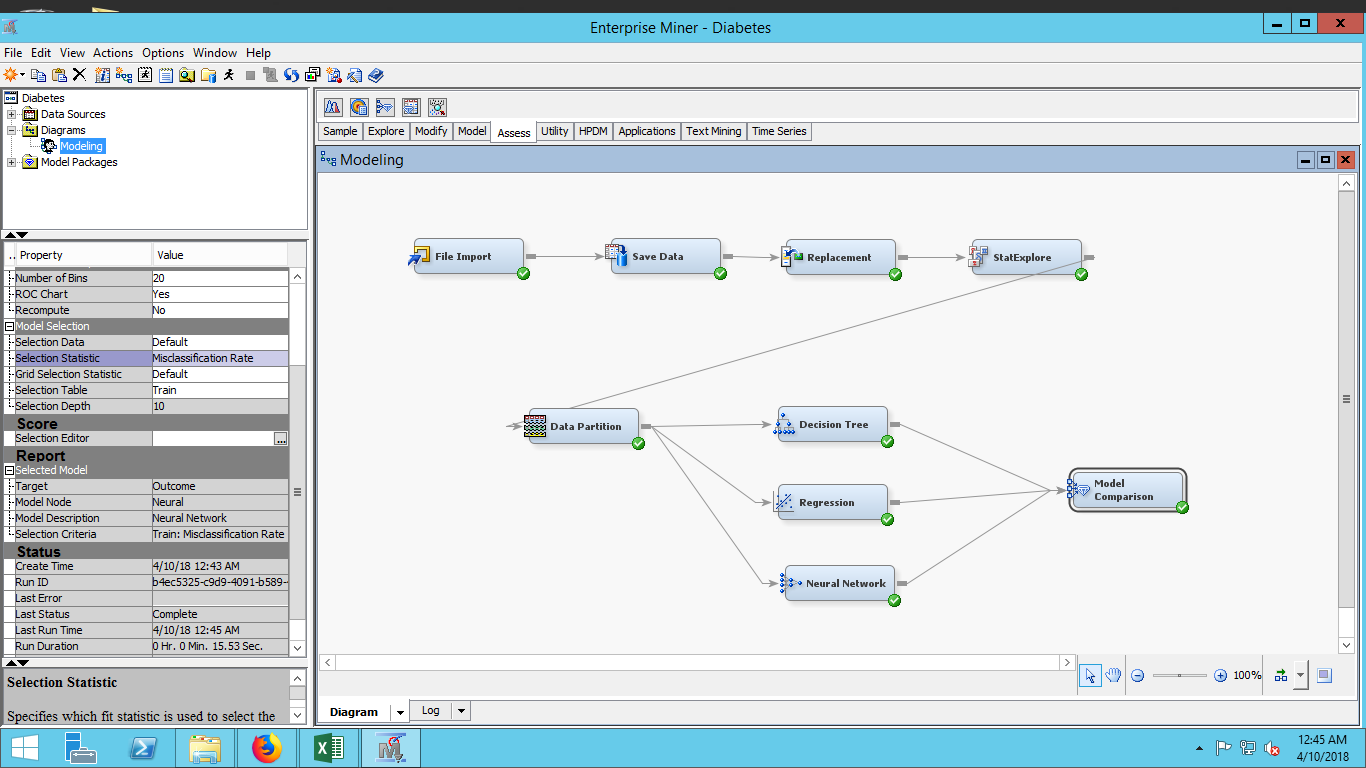
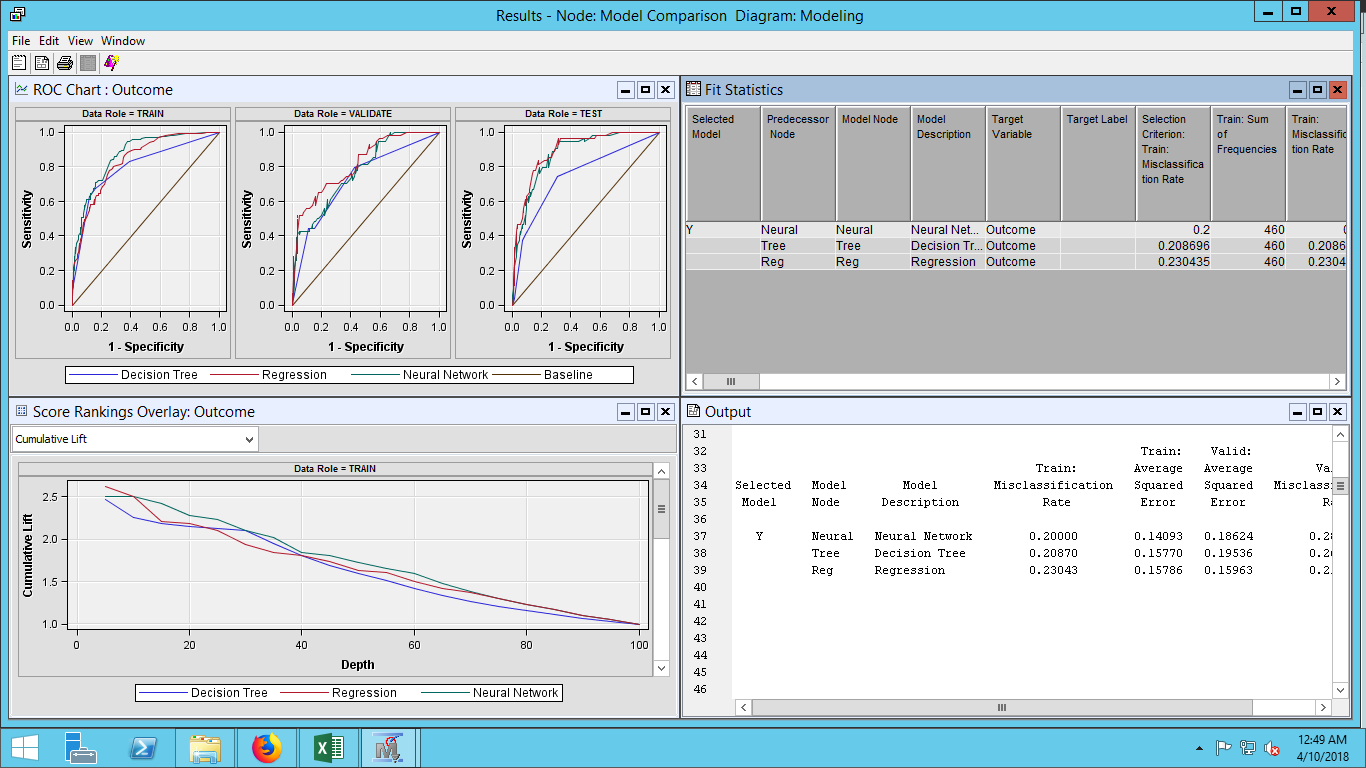
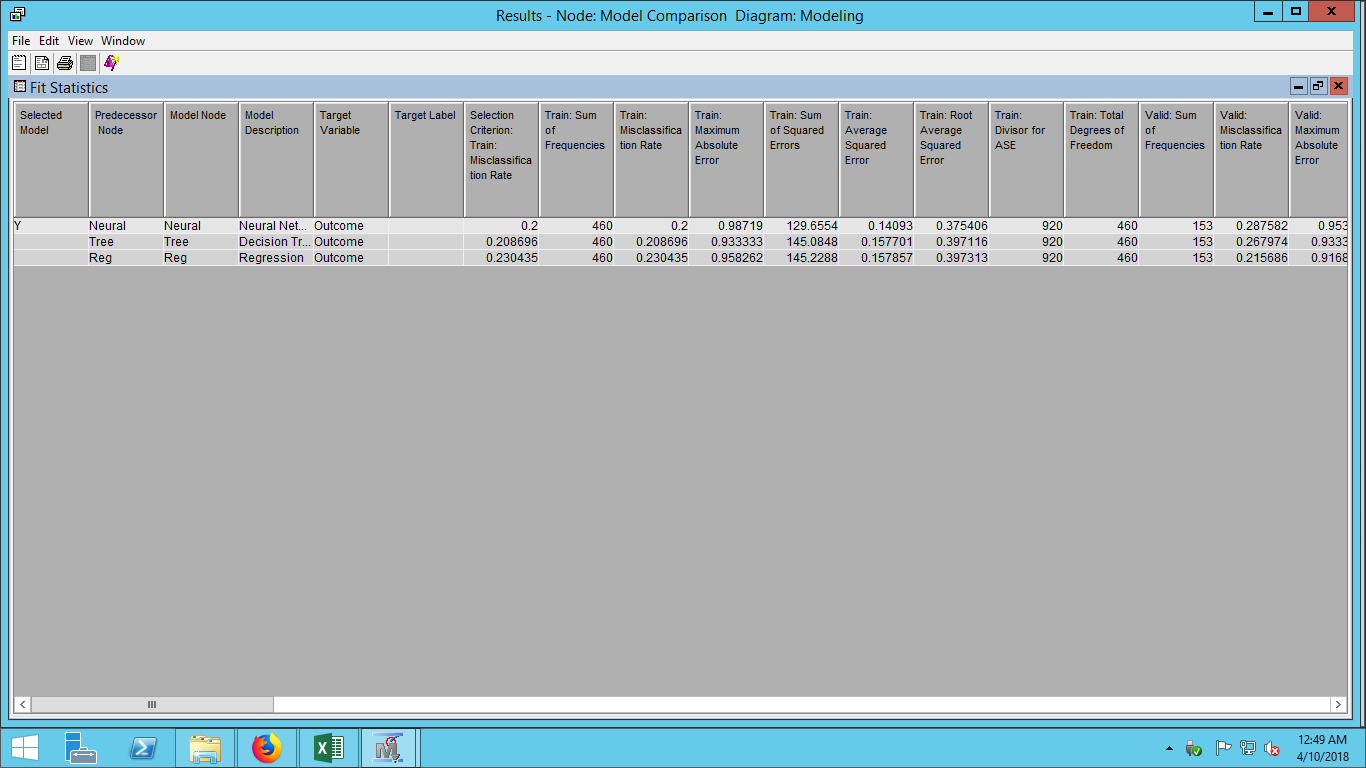
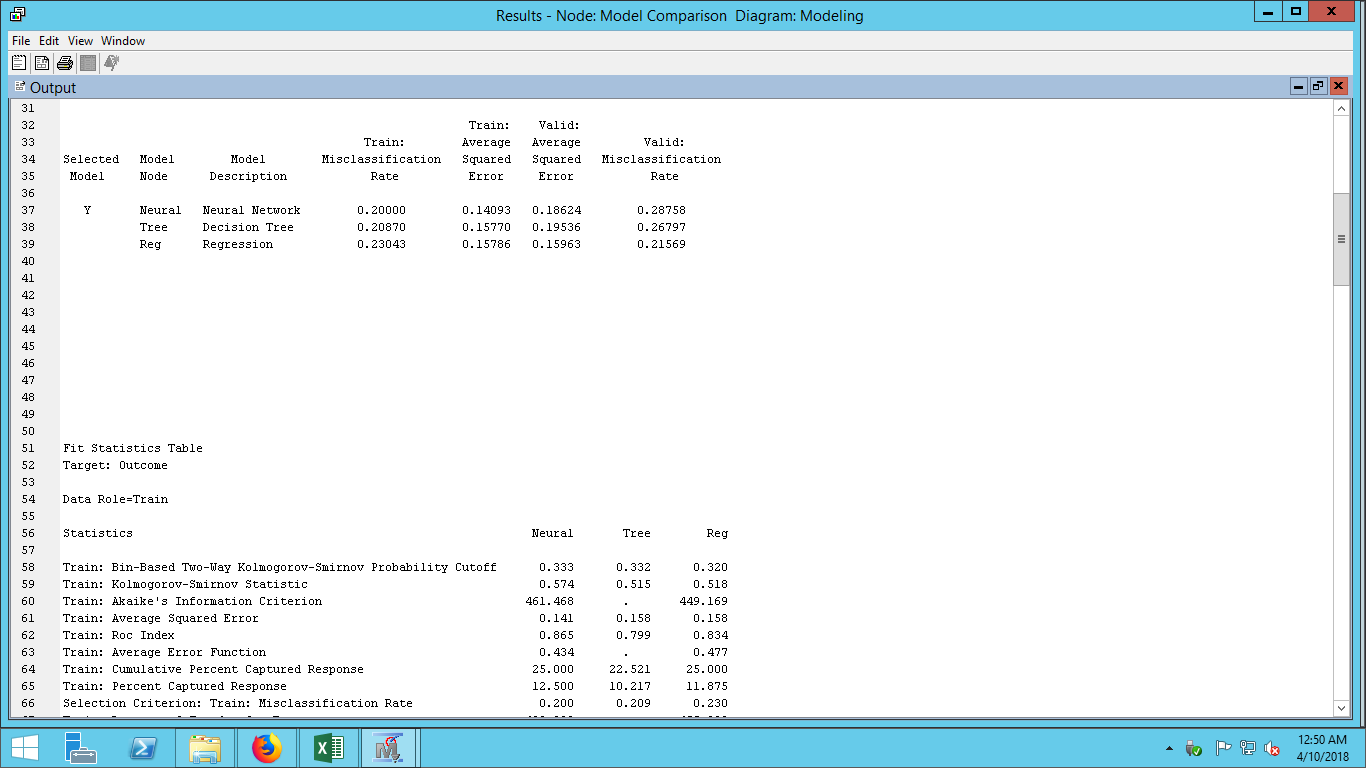
**REGRESSION**The regression type for the binary variable which is categorical has logistic regression. Model selection for the regression is stepwise and the selection criterion is validation misclassification.





**MODEL COMPARISON**

**CONCLUSION**

**REFERENCES**

Healthline. (2015). Type 1 and Type 2 diabetes: what is the difference? Retrieved from <https://www.healthline.com/health/difference-between-type-1-and-type-2-diabetes#causes>

Malik, R. (16 January, 2016). India is the diabetes capital of the world! Retrieved from <https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/India-is-the-diabetes-capital-of-the-world/articleshow/50753461.cms>