Jason Larosiliere

Professor Lamere

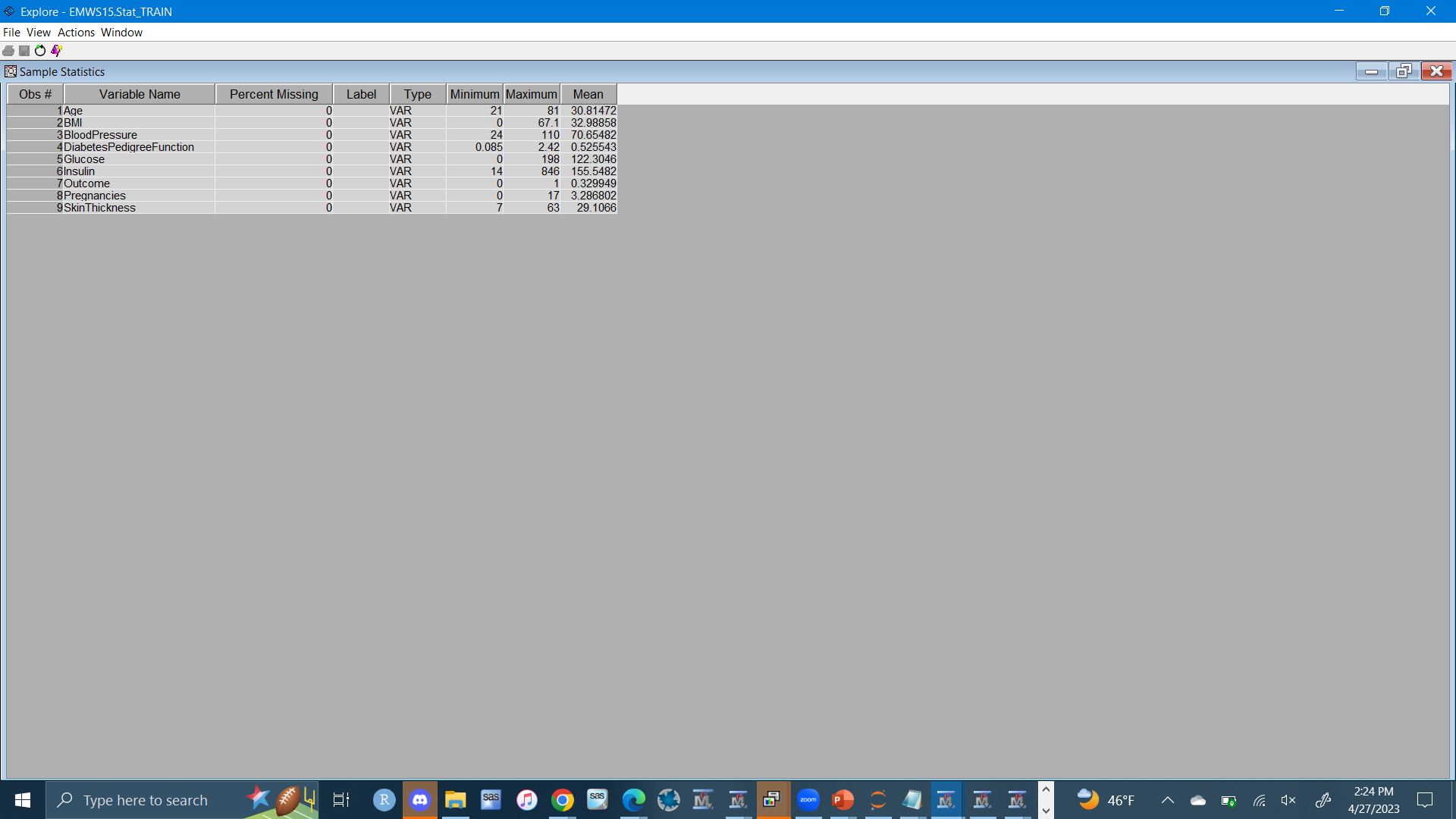
Applied Analytics Using SAS

10 May 2023

Applied Analytics Final Project

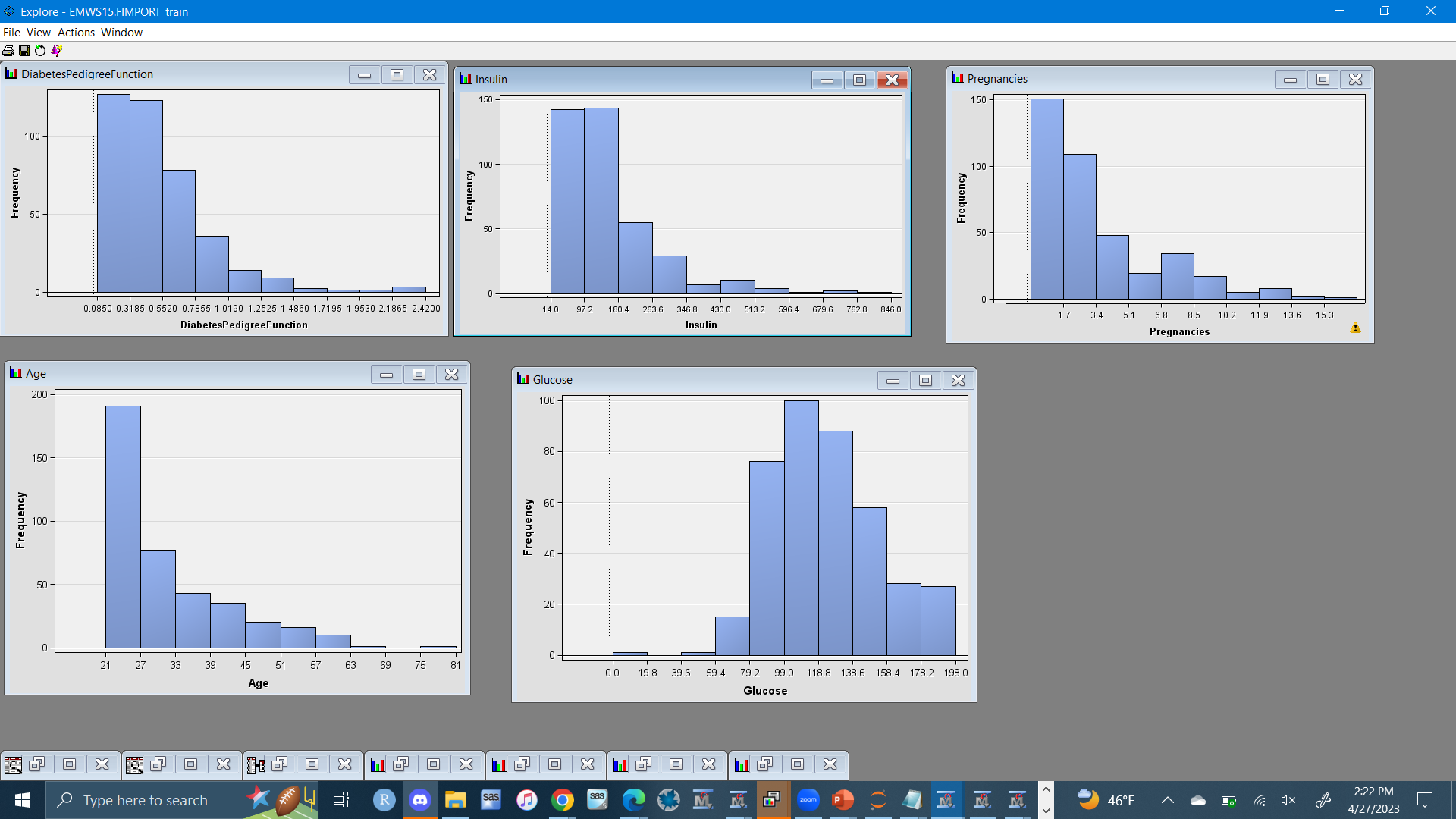
The goal of my project is to predict diabetes from several health measurements. My family has a long history of diabetes which motivates me to find insights into diabetes so I can learn more about it and the causes. I could potentially help my current family, myself and my family in the future from these causes. The dataset I have comes from Kaggle, a machine learning/data science community. It originally comes from the National Institute of Diabetes, Digestive and Kidney. The dataset consists of medical health indicators of female patients of Pima Indian heritage over 21 years old.

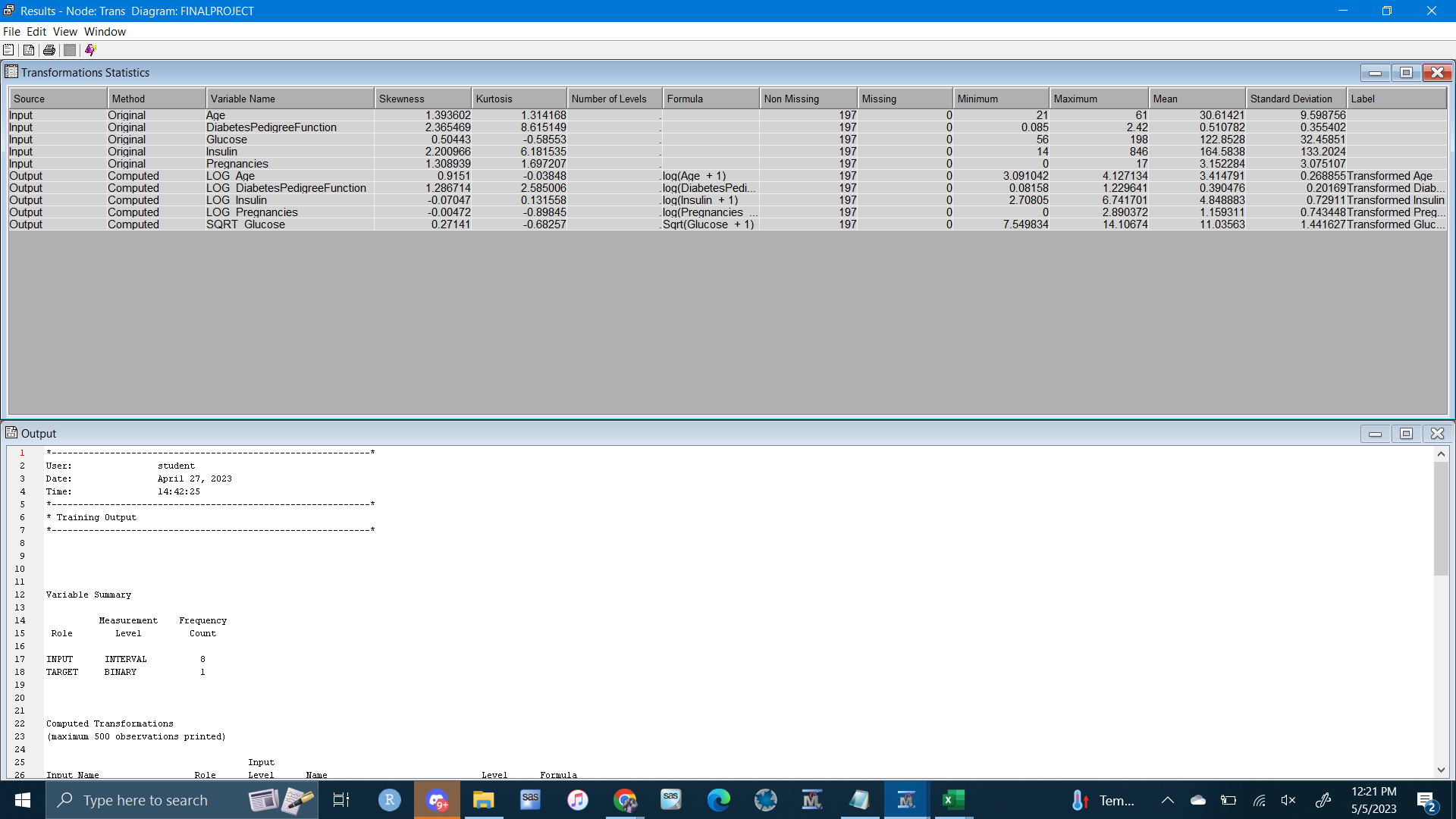
Before I start with the analysis, I must address missing data and transformations. I handled the missing data in Excel before bringing the data into SAS. Since I had a lot of observations to work with, I filtered out any zeros and missing values and deleted them.

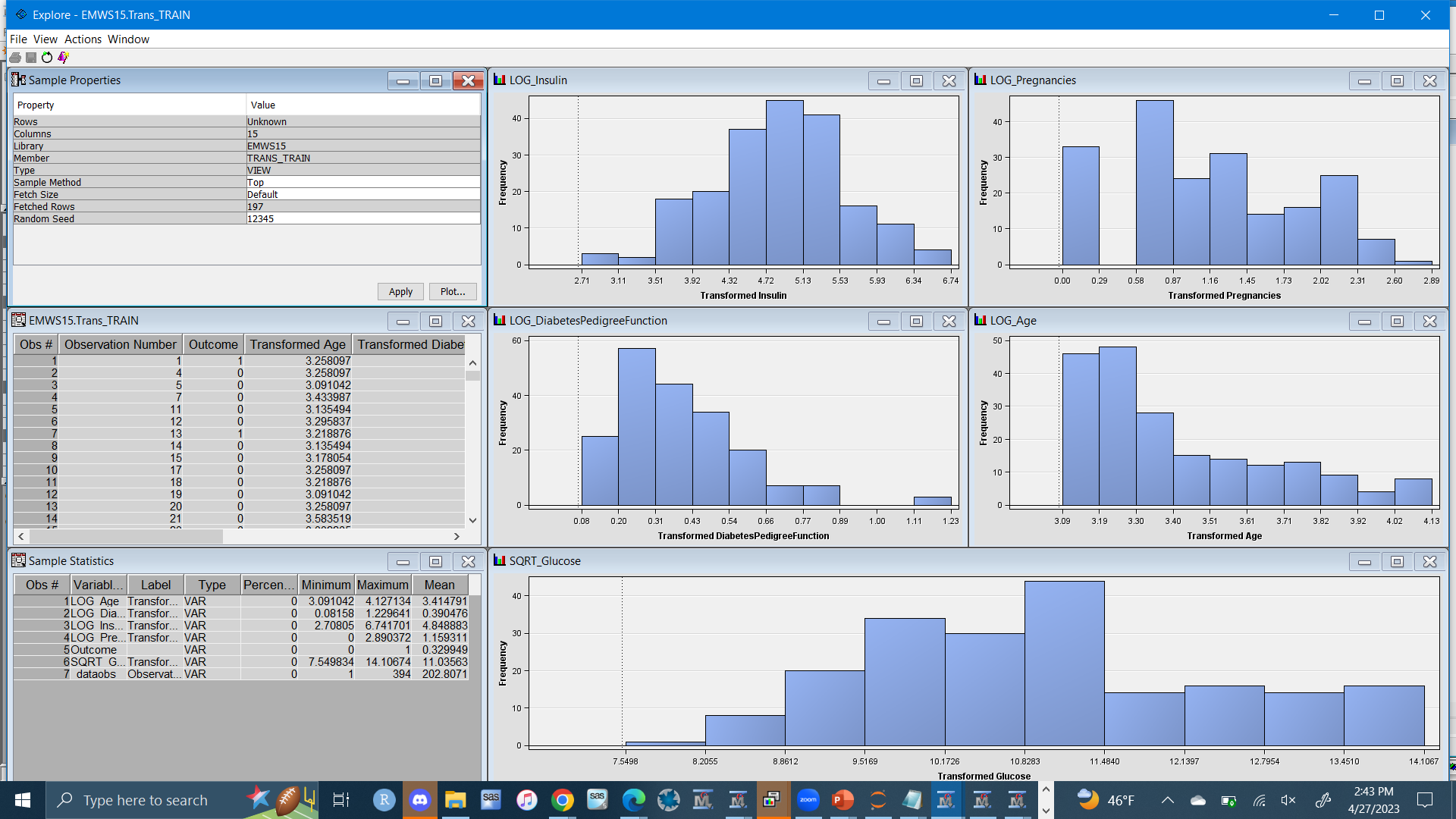


I ended up with 394 observations and the medical variables pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetespedigreefunction, age and outcome of diabetes as the target. Diabetes pedigree function represents the likelihood of diabetes based on family history. The variables that appear to have skewed distributions are diabetespedigreefunction, age, insulin, pregnancies and glucose. To resolve this, I will use a transform node and log transform diabetespedigreefunction, age, insulin and pregnancies. For glucose, I will use a square root transformation.

Before Transformation:

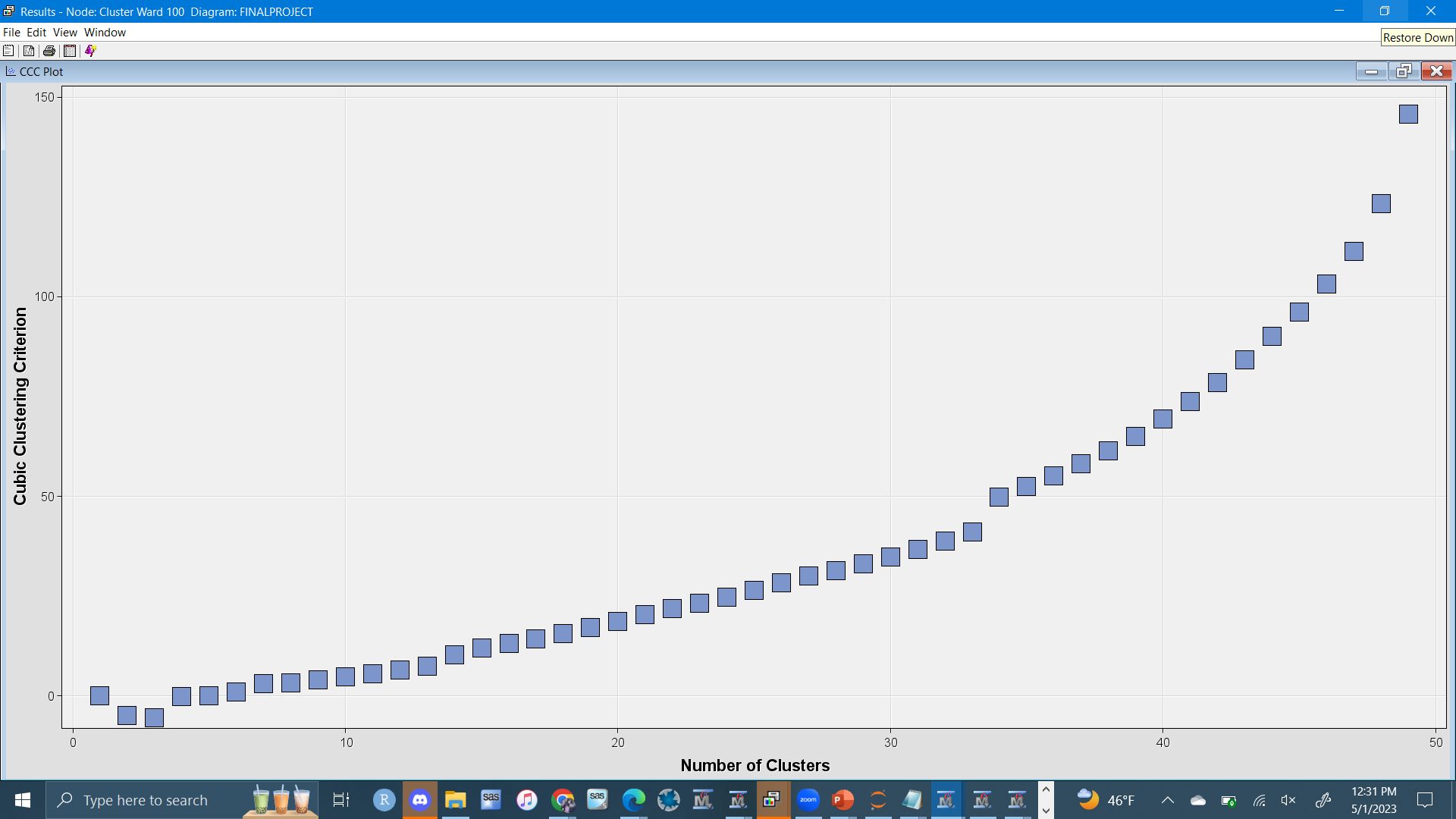
After Transformation:

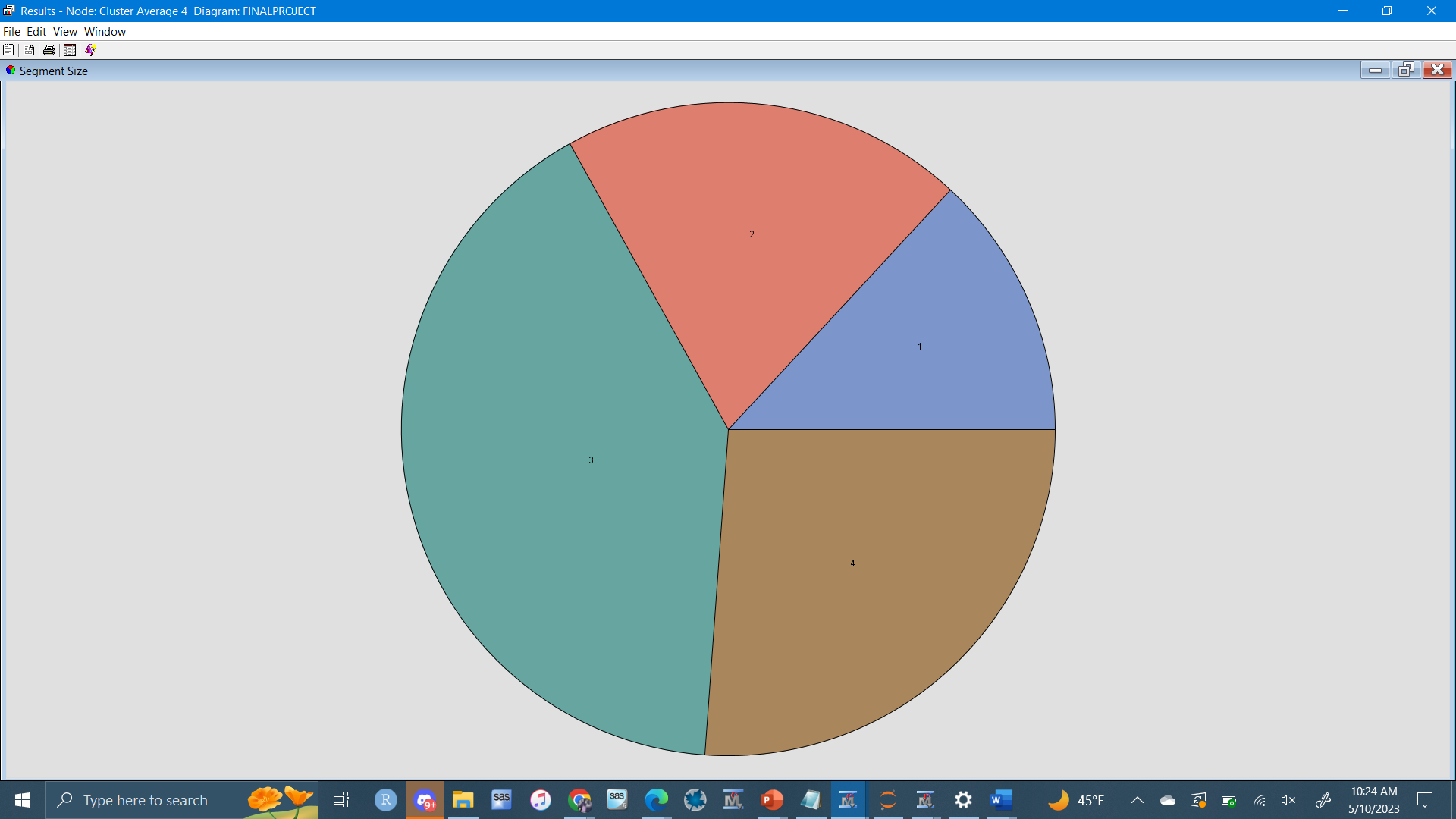


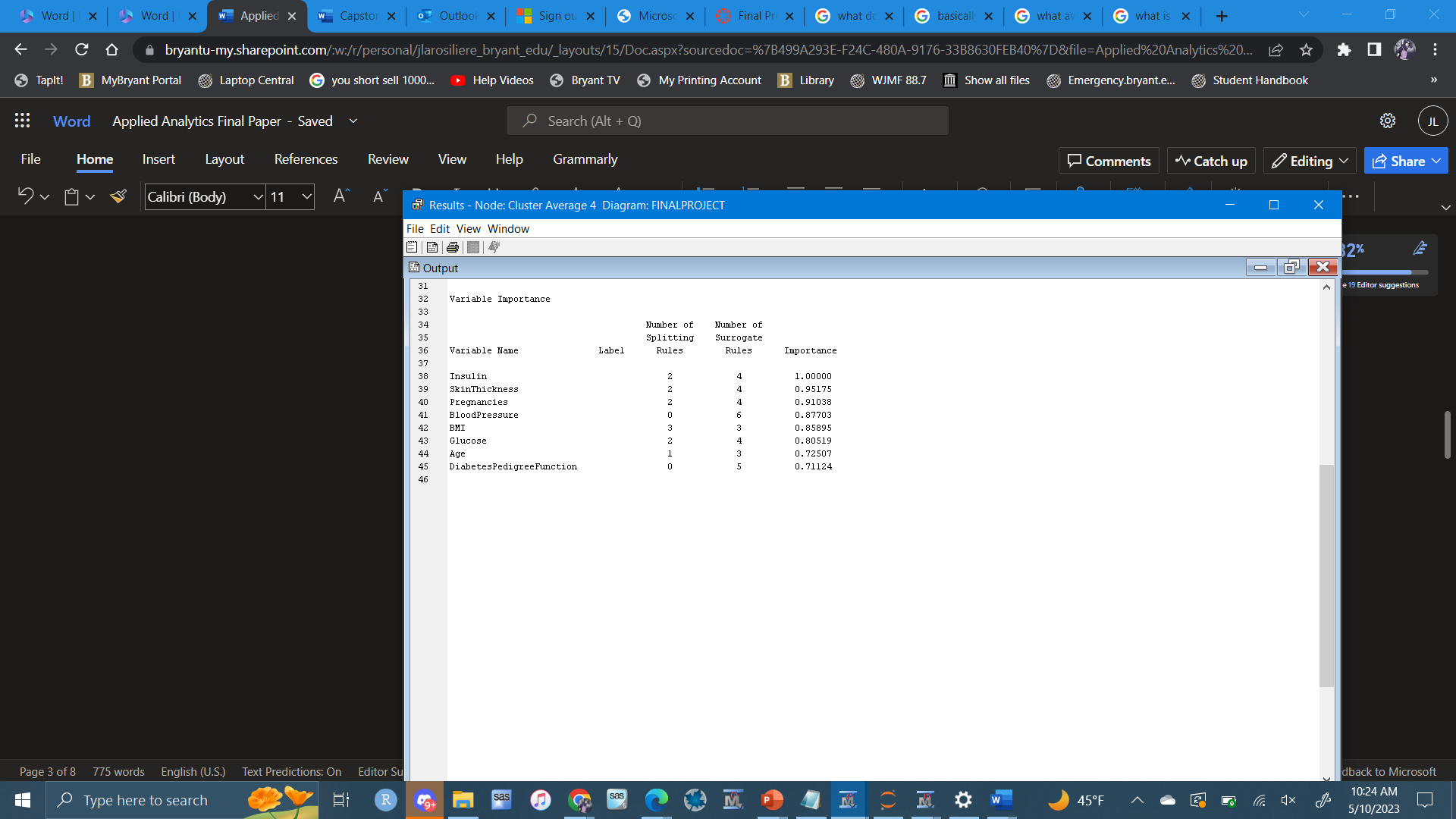
Unsupervised Analysis:

Clustering would create a specific number of cluster seeds and observations are assigned to the closest one based on their health conditions. Before running this, the data must be standardized so I'll use the range method. I went with the ward as the method. At first, I ran it with the preliminary max of 50 then 100 clusters. After taking a look at the CCC plot the local maximum was 4 clusters.

(analyze)





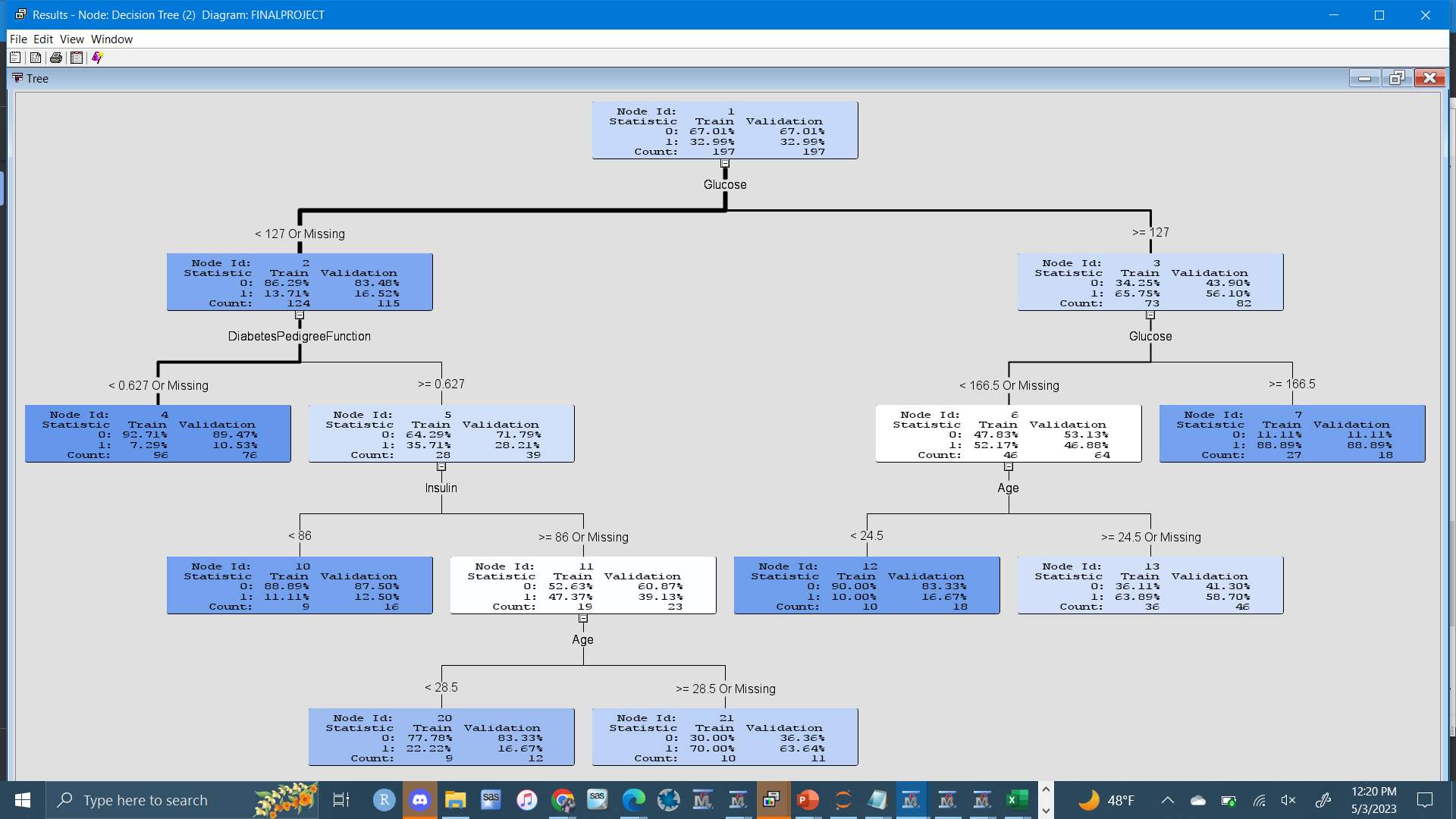


After analyzing the clusters there was not much separation. After exploring different relationships in X,Y and X,Y,Z scatterplots colored coded by segments, however I did not see any interesting observations.

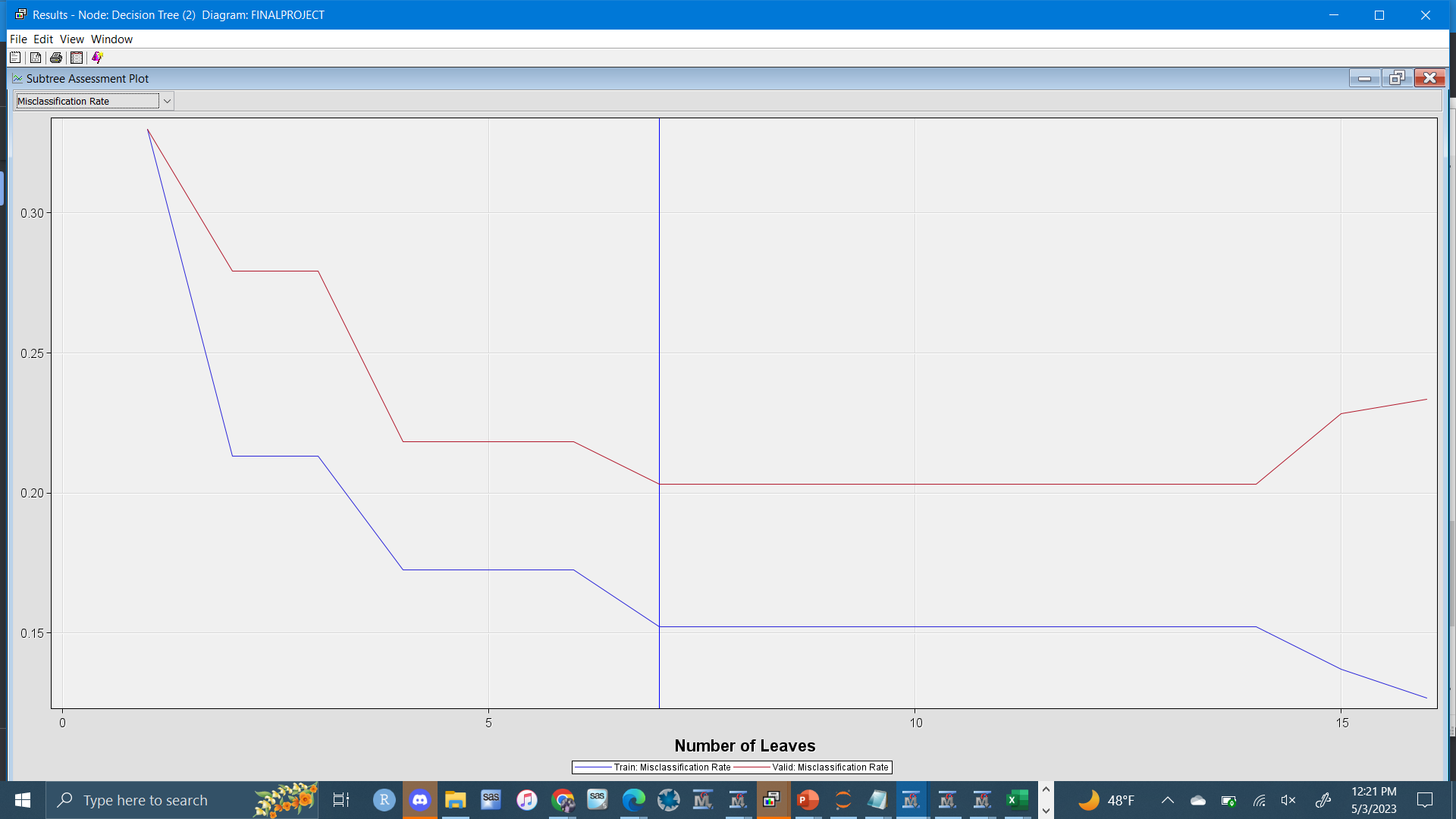
Supervised Analysis

Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A decision tree will serve as a method to predict the outcome of diabetes based on the health indicators. For the assessment measure I used decision, 8 for leaf size.

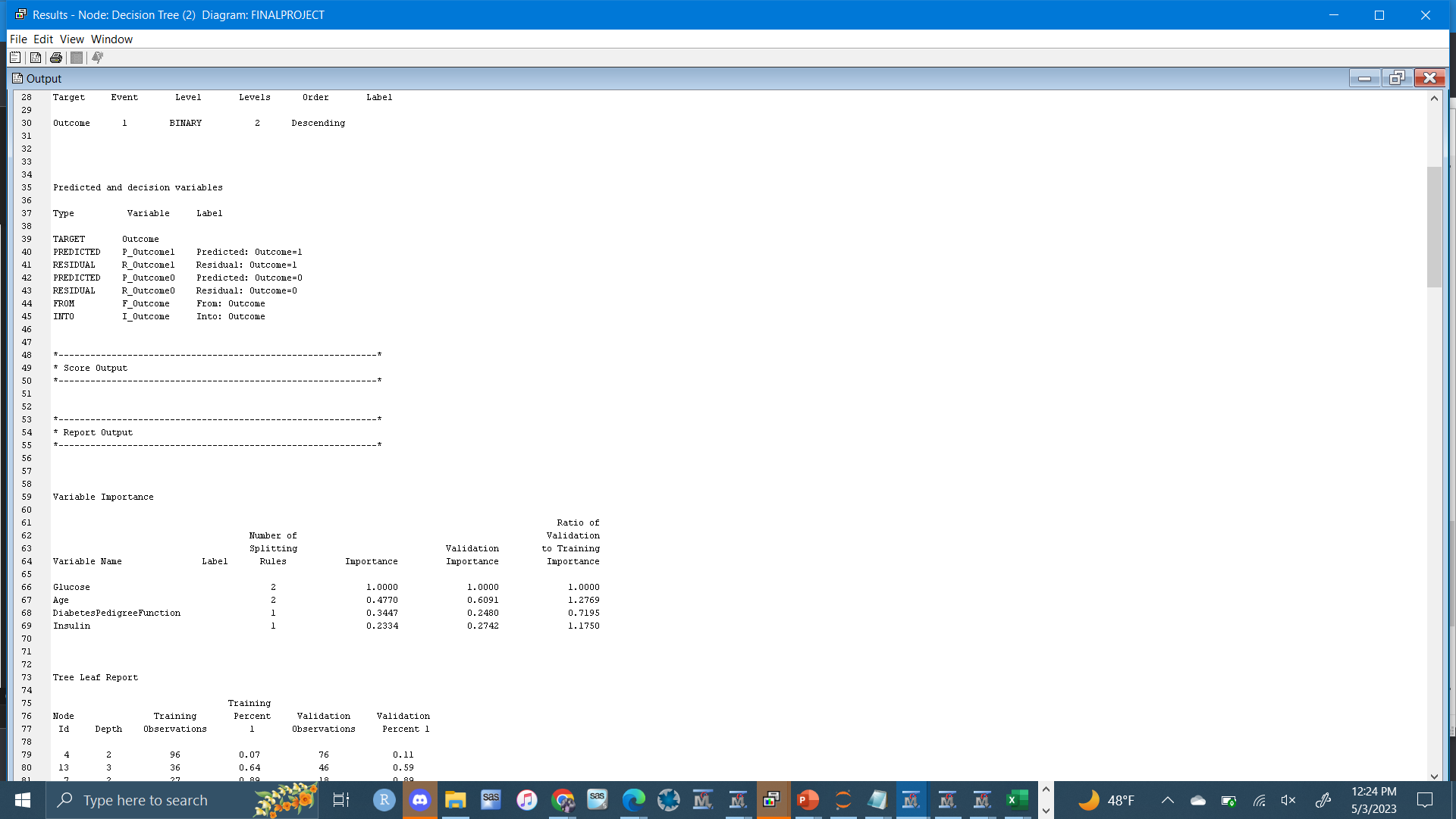
Tree



Subtree Assessment Plot

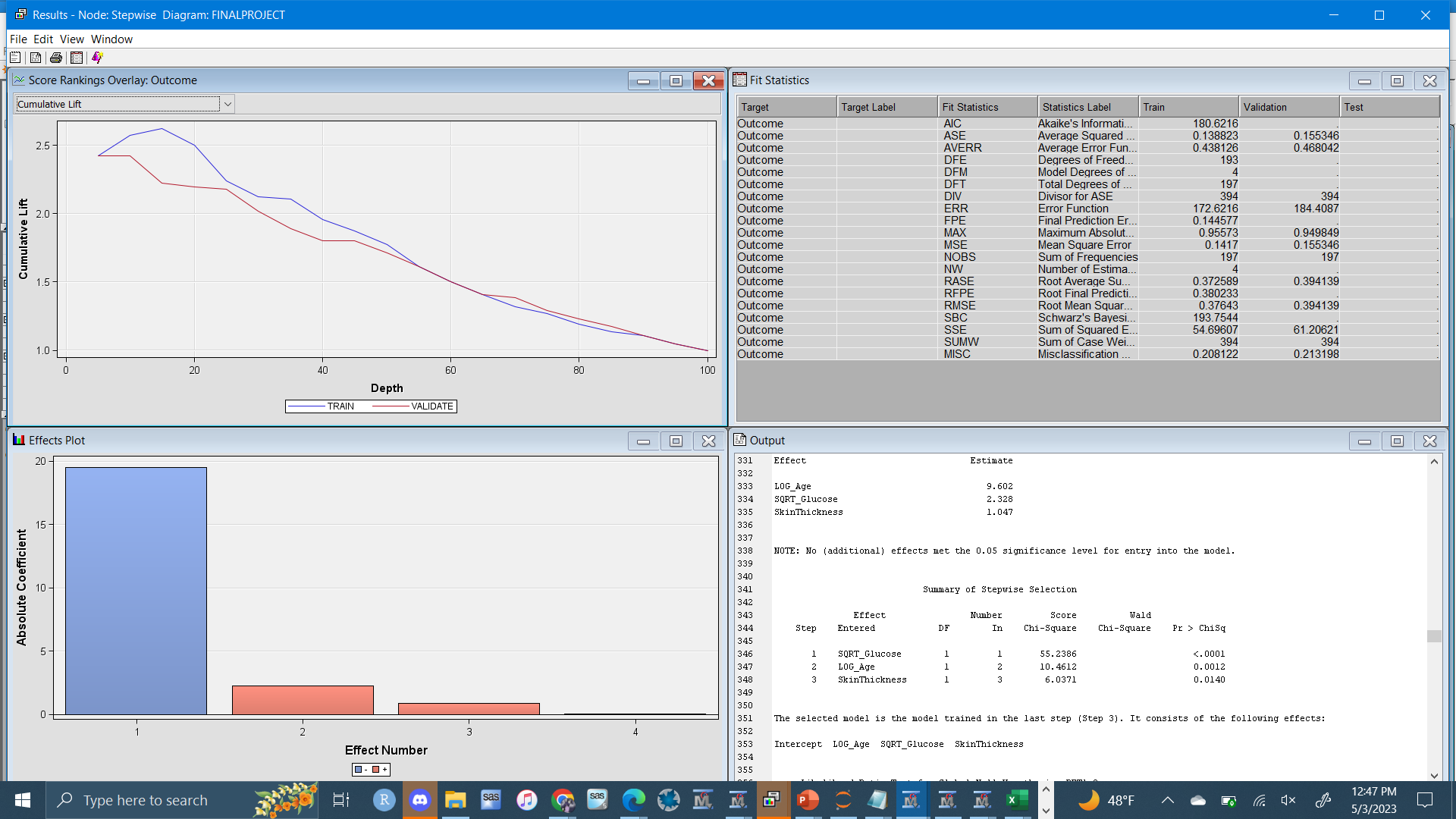


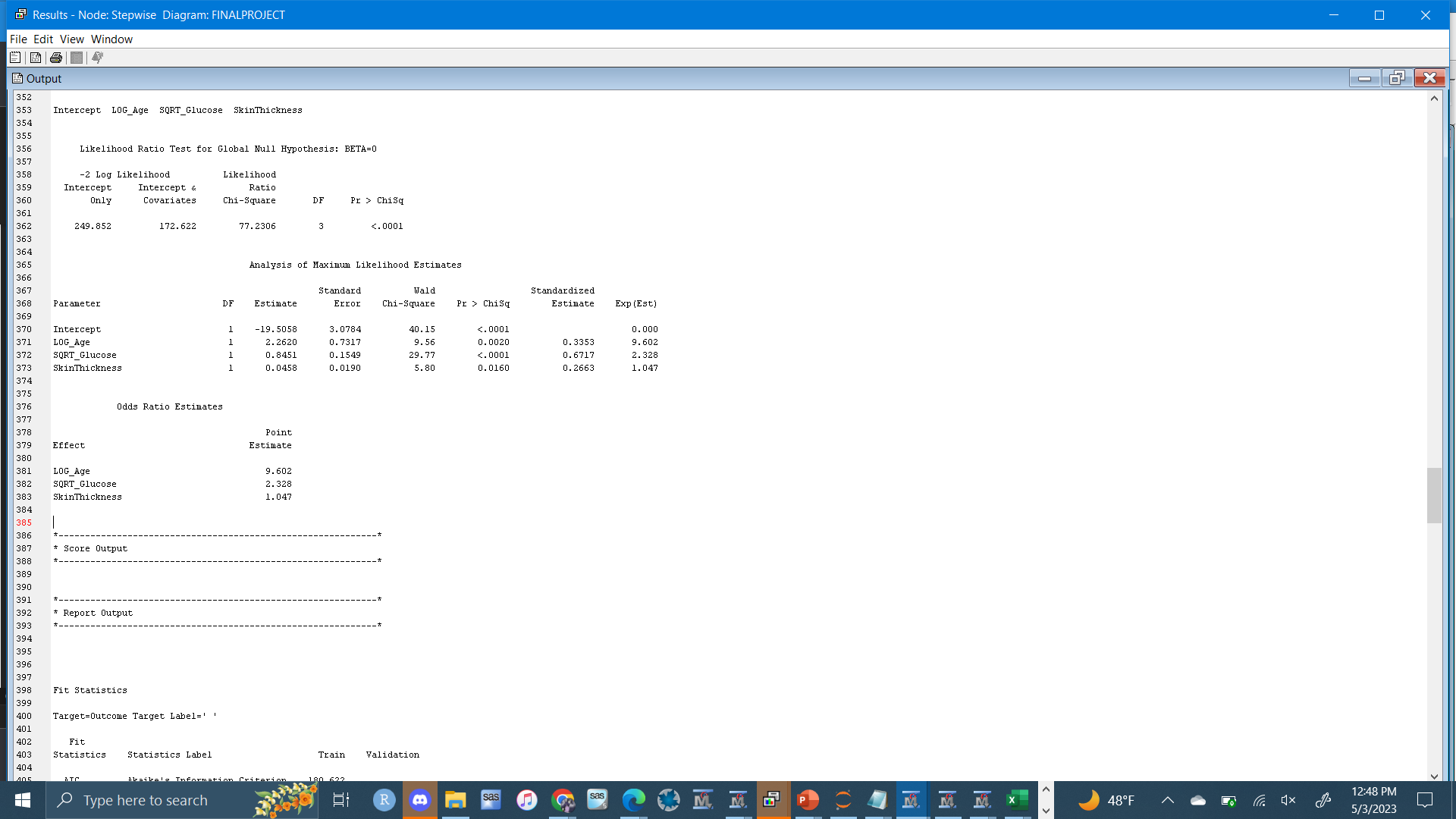
Variable Importance



The first thing we take from this is that the best chance at avoding diabetes is keeping glucose under 127 and diabetes pedigree function under .627. Also, there is 90% of getting diabetes if glucose hits above 166.5. Lastly high insulin strongly affects patients after the age of 29. Lastly the variables that came up as important from most to least is Glucose, Age, DiabetesPedigreeFunction and Insulin. The validation ASE is 14.67% which tells me the decision tree had a low error as a predictor.

Logistic Regression:



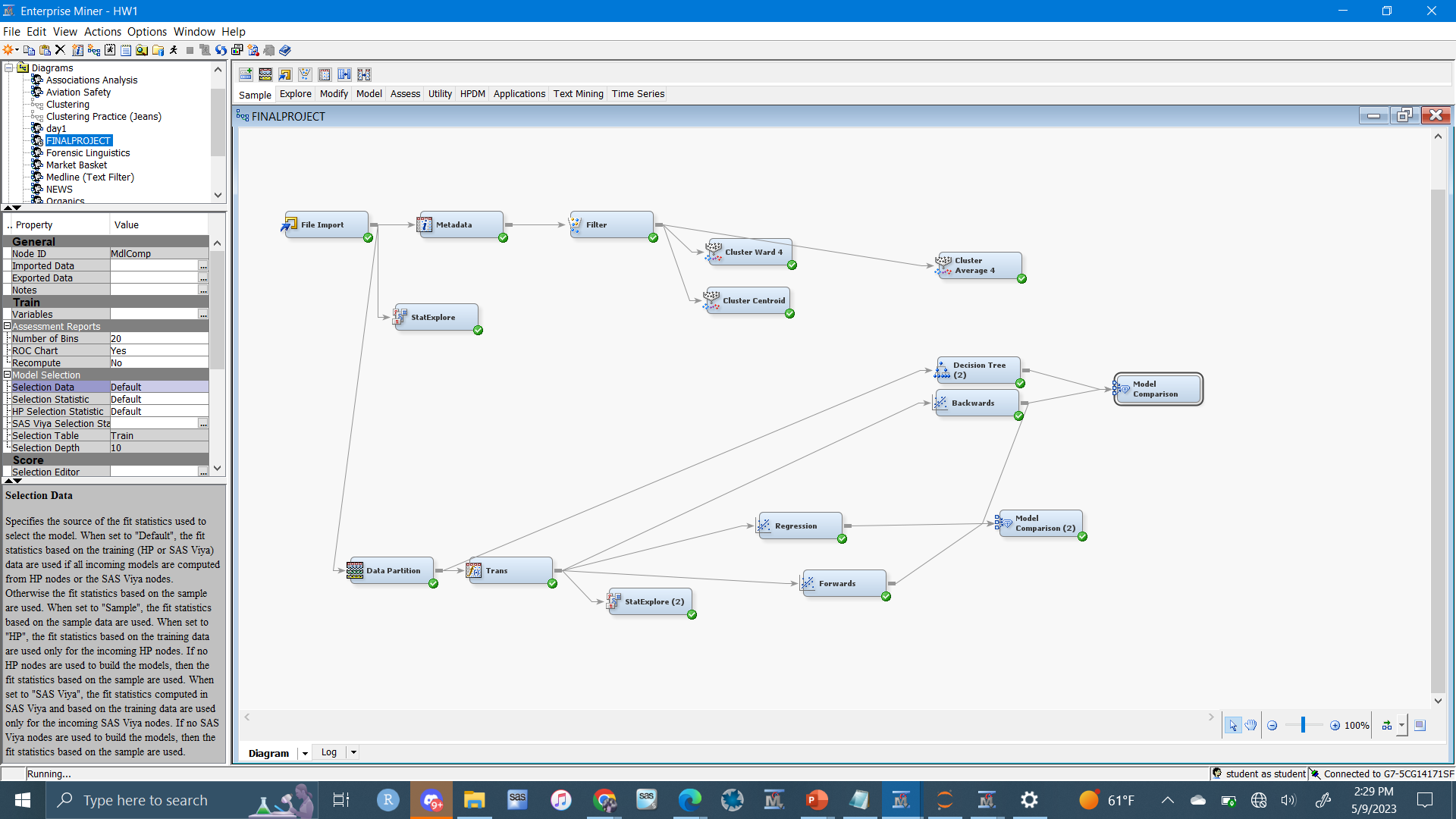


For the regression, I chose stepwise for the selection model. Since it is logistic regression we’ll take a look at the AIC as a replacement for R Squared. The AIC is 251.85. The variables entered into model were Glucose, Age and Skin Thickness.

After using the model comparison node, the model that performed the best was the decision tree. It was selected based on validation misclassification rate. This represents the % of the time the classifier is incorrect. Therefore, we want a lower percentage. The decision tree had a rate of 20.3% while the regression had a rate of 33.0%. So basically, the decision tree does a better job at predicting whether a patient has diabetes.

In conclusion, since the data only included women from India their way of living and the medical care could be totally different from the U.S. so applying this model to women here could potentially pose as a WMD. Another potential for WMD is there are some exceptions to the rule that may not be detected by my model. For instance, there are rare cases when people get diabetes from a genetic mutation not related to family history. There are also cases where medicine could disrupt the insulin leading to diabetes. The variables I have do not include all causes of diabetes. One thing I would do differently for a future study is include male and female patients from the U.S. Also, I would incorporate variables that involve habits such as exercising, smoking and diet to get a better understanding of the patient's everyday lifestyle and what their habits are.

Appendix:



The File Import Node brings the diabetes dataset into the diagram

The StatExplore Node displays variable statistics.

The MetaData node is used to modify the variable roles.

The Filter Node creates and applys filters to the dataset.

The Cluster nodes performs k-means cluster analysis to our data which creates segments. Each one uses a different clustering method between Ward, Centroid and Average.

The Data Partition Node splits the data into training validation and test datasets. Here I split the data 50% training and 50% validation.

The Decision Tree node uses the data and builds a decision tree. Here I used 8 leaves.

The Transformation node performs transformations on variables. Here I transform the variables that showed skewness in their histogram with the log values.

The Regression nodes ‘Regression’ ‘Backwards’ and ‘Forwards’ uses the data to perform a stepwise, backwards and forwards regression analysis.

The 2 Model Comparison nodes were used to compare the models connected to it based on validation misclassification rate.