# 

# 

MKTG 746 Group Assignment

* Data Prediction Modeling

Group 20

Ansley Ying Suet Liu 301138708

Tsao Chun Huang 301133343

Zhuocheng Zhang 301152869

# 

[**Executive Summary**](#_3ghf8ou30n5o) **3**

[**Introduction**](#_9oc3n89rzlz1) **3**

[**Data Wrangling**](#_vuomhr4my167) **3**

[**Decision Tree Modelling**](#_jjm8yv6kbqrn) **8**

[Maximal Tree](#_5lgaq38ul54v) 8

[Misclassification Tree](#_piahux4q7ifo) 9

[Average Squared Error Tree](#_6ot73fwkq9ng) 9

[Best Model](#_v1alm3jw17ej) 10

[**Logistic Regression**](#_bp942d66xvej) **11**

[Imputation and Replacement](#_5qomgvyh03ic) 11

[Full regression](#_d2nbojegrzqk) 15

[Forward Regression](#_7han9ucmledk) 16

[Backward Regression](#_sx0hx15p4tig) 16

[Stepwise Regression](#_km3xb02993aa) 18

[Best Model](#_n3774bamo19f) 18

[**Neural Network**](#_qdn6pi39n0lg) **19**

[Best Model](#_wbtz6a1bqbue) 21

[**Model Comparison**](#_r7zyjedr5tox) **22**

[**Conclusion**](#_hj84o7wvduxw) **27**

[**Appendix**](#_pltlb09kc2ew) **27**

# 

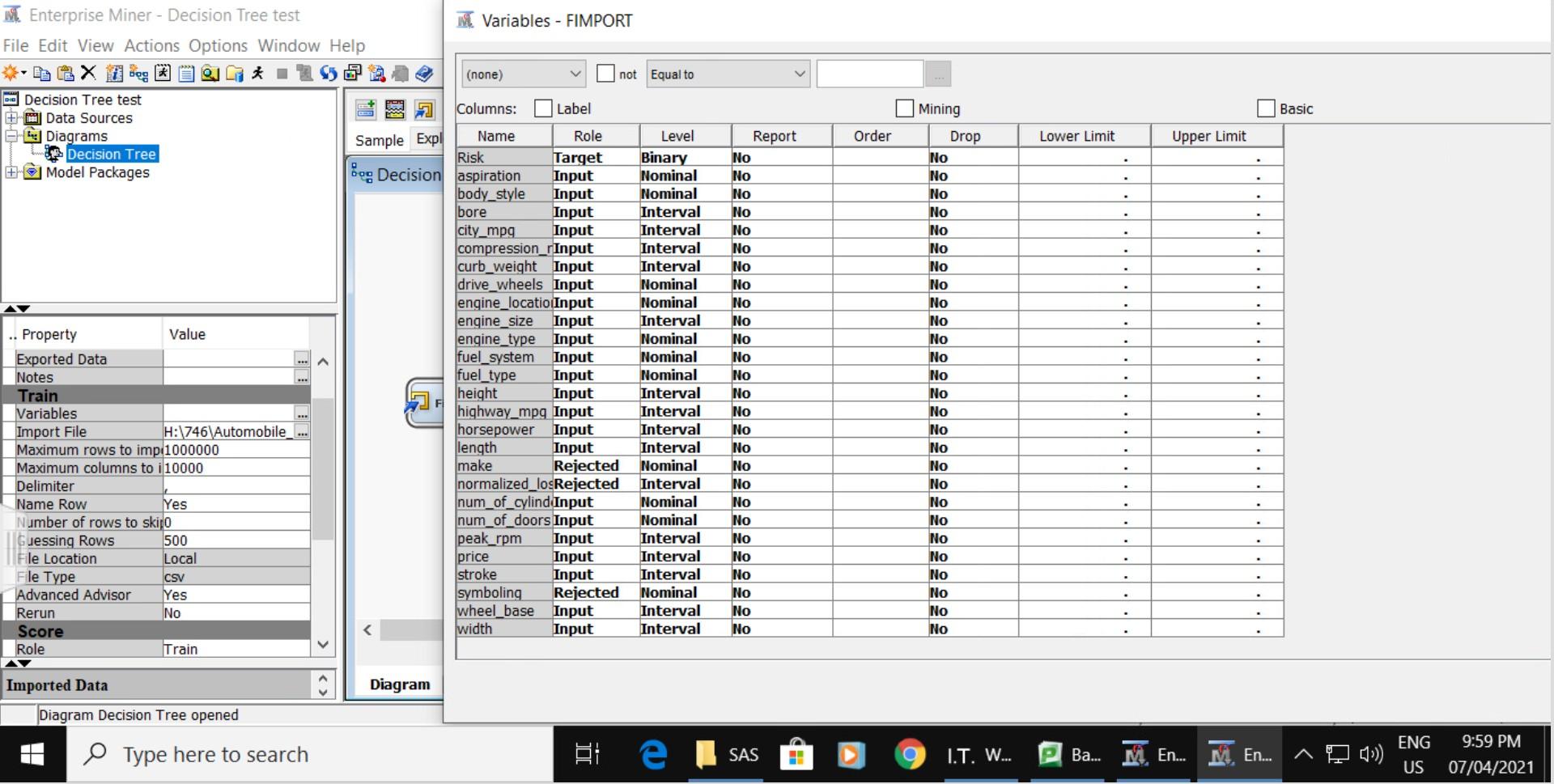
# Executive Summary

This research aims to utilize SAS Enterprise Miner to build and compare predictive modeling for car risks. In this study, Decision Trees, Logistic Regression, and Neural networks are constructed in SAS Enterprise Miner and compared. It turns out the number of doors, size of the car including height, width, wheelbase are the most important factors when determining the risk level of a car model.

# Introduction

Insurance companies all have their unique grading for car risks. So, what contributes to the risk grading of the car models? In this study, car models from 1985 will give you the answer. Given a few car model specifications when you consider when you purchase a car, the association with the risk will be converted into and estimated as a binary target.

# Data Wrangling

Variables explained:

The Automobile dataset contains 26 variables and 205 observations. The variables in the data set are shown below with the appropriate roles and levels:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| aspiration | Input | Nominal | Car aspiration type |
| body-style | Input | Nominal | Car body style |
| bore | Input | Interval | The size of a bore in millimeter |
| city-mpg | Input | Interval | Mile Per Gallon in city |
| compression-ratio | Input | Interval | The ratio of the volume of the cylinder and its headspace |
| curb-weight | Input | Interval | Weight of vehicle with standard equipment in-lbs |
| drive-wheels | Input | Nominal | Car drivetrain |
| engine-location | Input | Nominal | The location of the engine on a car |
| engine-size | Input | Interval | The volume of the engine in cubic inches |
| engine-type | Input | Nominal | The structure of the engine in the car |
| fuel-system | Input | Nominal | Method of injecting fuel to the engine |
| fuel-type | Input | Nominal | The type of fuel the car use |
| height | Input | Interval | Height of the car in inch |
| highway-mpg | Input | Interval | Mile Per Gallon in city |
| horsepower | Input | Interval | The power of the car in hp |
| length | Input | Interval | Length of the car in inch |
| make | Rejected | Nominal | Make or the car |
| normalized-losses | Rejected | Interval | Normalized losses in use as compared to other cars |
| num-of-cylinders | Input | Nominal | Number of cylinders in an engine |
| num-of-doors | Input | Nominal | Number of doors of the car |
| peak-rpm | Input | Interval | The peak round-per-minute the engine produce |
| price | Input | Interval | The retail price of the car |
| Risk | Target | Binary | 1if symboling is above zero, 0 if symboling is zero and below. |
| stroke | Input | Interval | Number of stroke of a cycle in an engine |
| symboling | Rejected | Nominal | The risk assigned by the automotive yearbook. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. |
| wheel-base | Input | Interval | Distance between the centers of the front and rear wheels in millimeter |
| width | Input | Interval | Width of the car in inch |

**Rejected variables:**

Make: Too many brands make this variable polygonal to interpret.

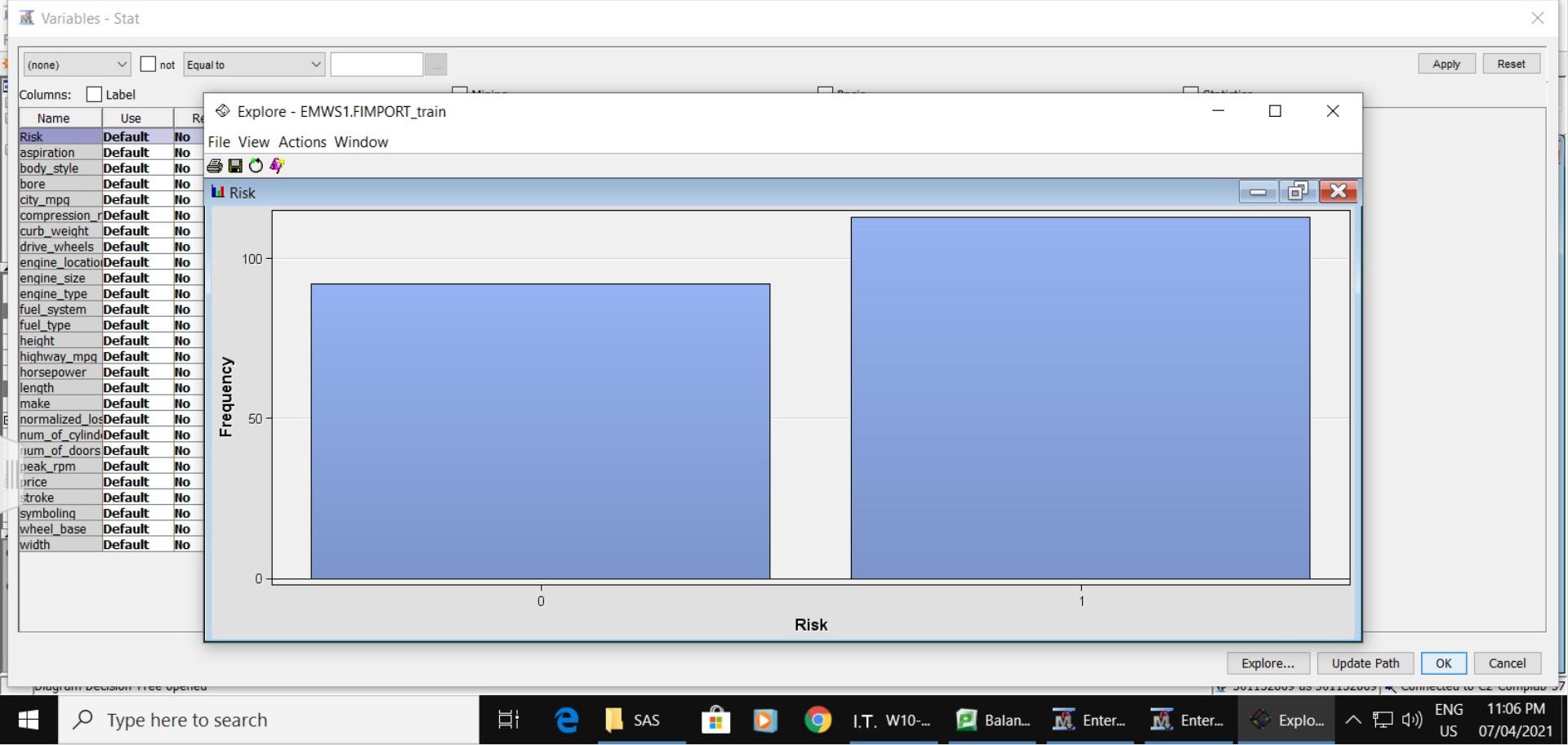
Normalized Losses: Directly associated with the target.

Symboling: This interval target is converted into a binary target for the purpose of this research.

**Variables added:**

Risk: Converted from symboling, where 1 represents symboling above zero and 0 if symboling is zero and below.

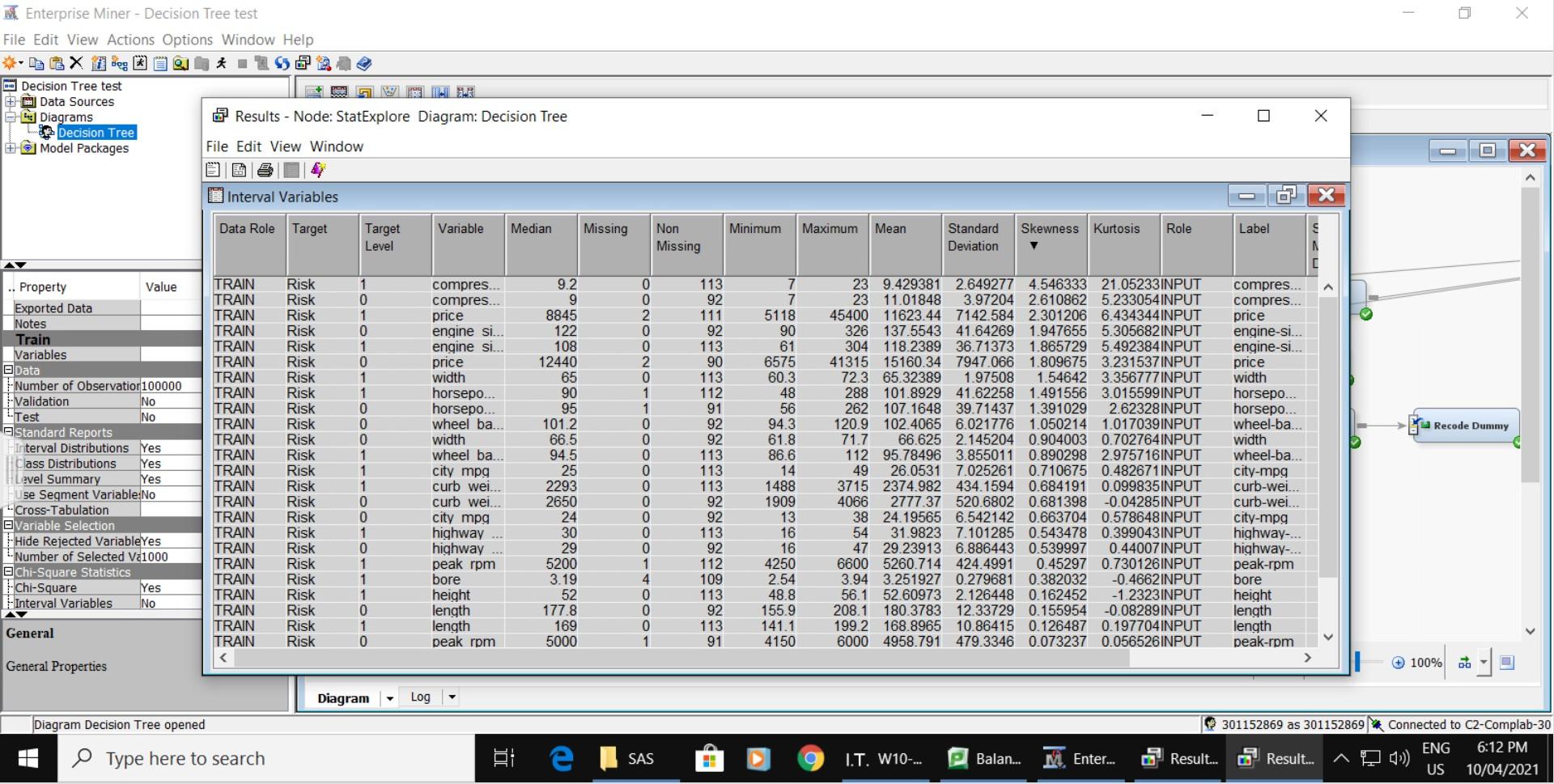
**Dataset overview:**



Out of 205 entities, 92 cars are considered safe and 113 cars are considered risky.

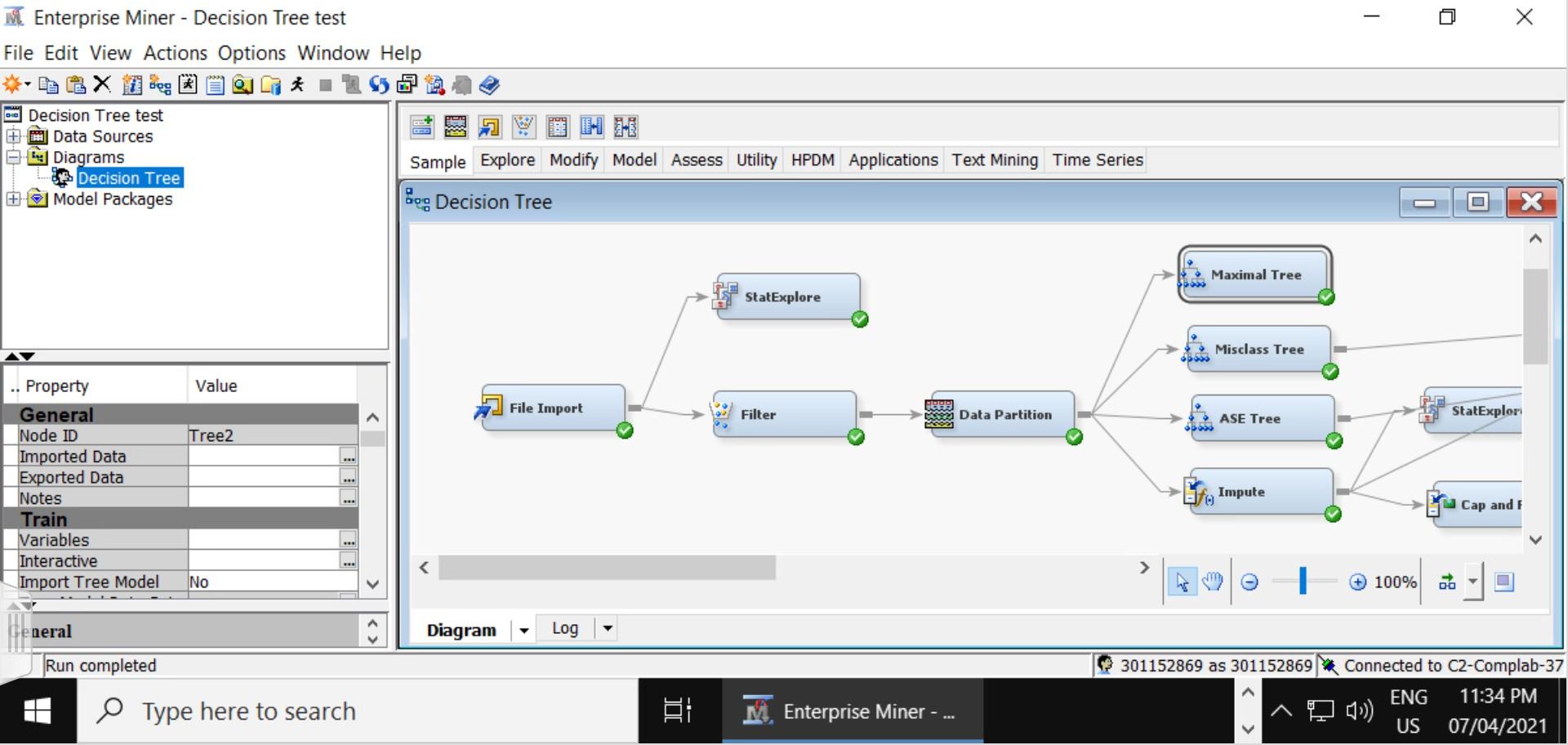


There are 5 missing variables: bore, stroke, price, peak rpm, and horsepower.



The compression ratio is highly skewed.

From the previous run, missing horsepower was included in the final model. But with only one entry of missing horsepower, it is not persuasive enough to prove the importance of a missing value. So, in the next step, this Renault will be filtered out.



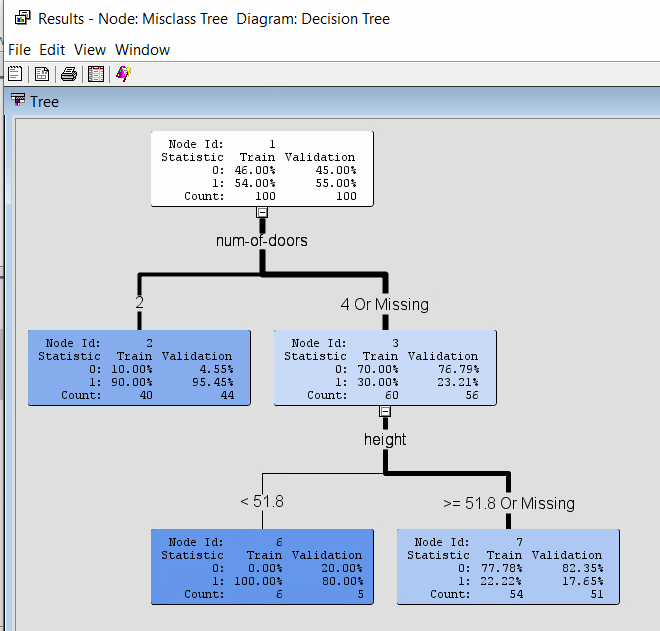
The data is split into 50 percent of training data and 50 percent of validation data. Three decision trees were created: Maximal Tree, Misclassification Tree, and Average Square Error Tree.

# Decision Tree Modelling

## Maximal Tree

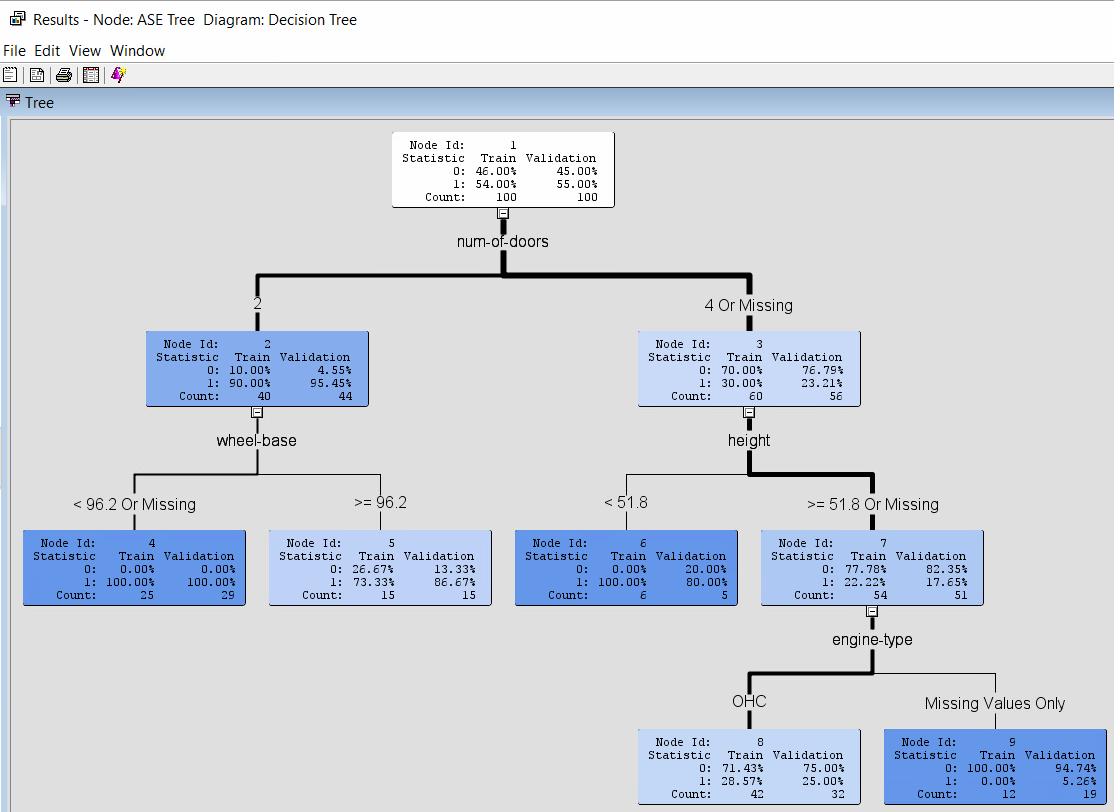
There are 4 splits and 6 leaves in the Maximal tree. The splits are on the number of doors, wheel-base, height, and engine type. The Misclassification Rate is 0.12 and the validation ASE of the maximal tree is 0.100408.

## Misclassification Tree



There are 2 splits and 3 leaves in the Misclassification Tree Model. The splits are on the number of doors, and height. The Misclassification rate is 0.12 and the validation ASE is 0.105585.

## Average Squared Error Tree



There are 4 splits and 6 leaves in the Average Square Error Tree Model. The splits are on the number of doors, and heights. The Misclassification Rate is 0.12 and the validation ASE is 0.100408.

## Best Model

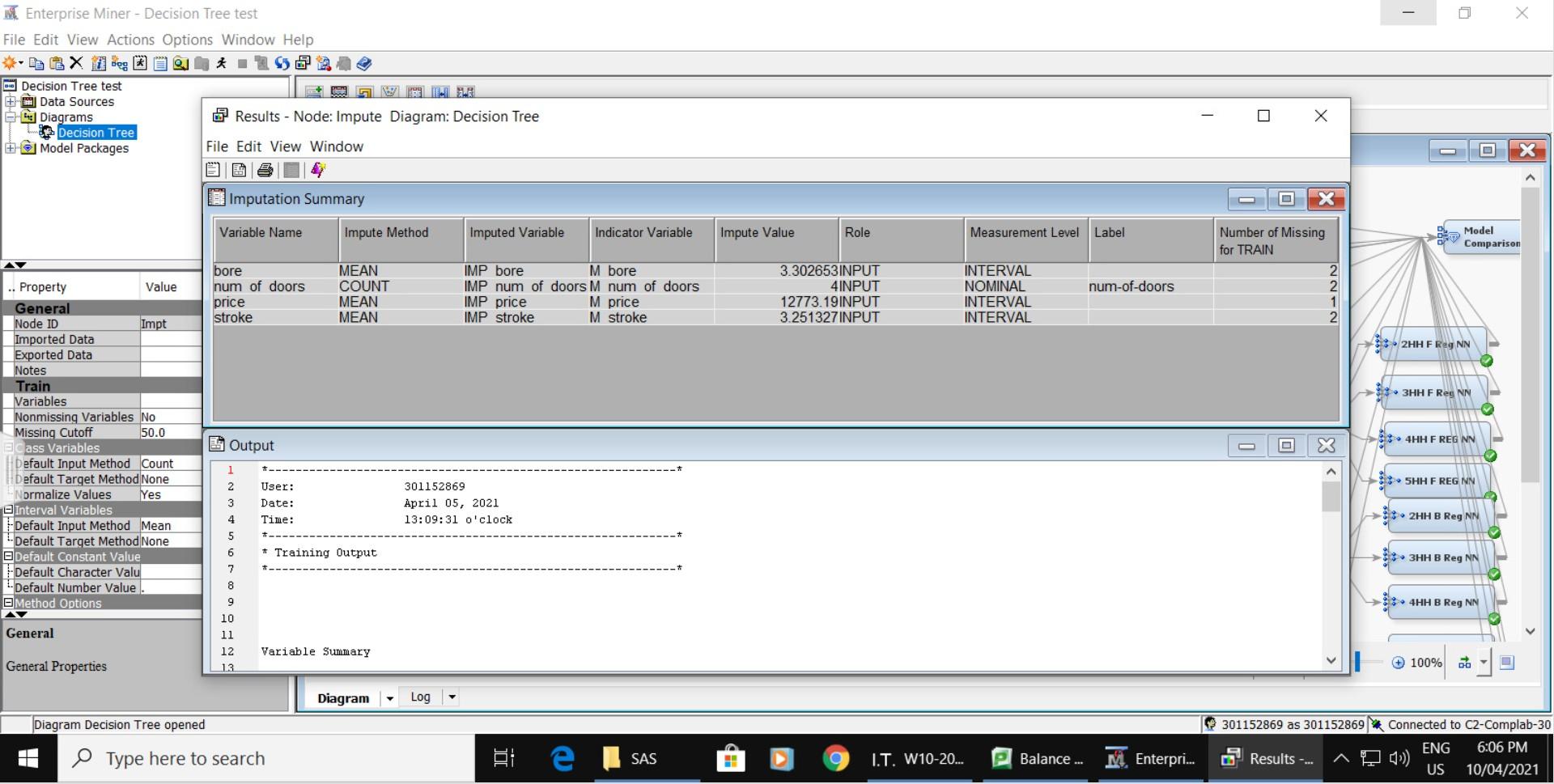
According to the model comparison, the Average Square Error Tree Model and the Maximal Tree are the same, which also have the lowest Misclassification and ASE.

* The first split in this model is on the number of doors.
* Two-door cars have a higher chance of being risky.
* Out of the two-door cars, those with a wheel-base smaller than 96.2 are riskier.
* Out of the four-door cars, those with a height lower than 51.8 inches are riskier.
* With those four-door cars higher than 51.8 inches or with a missing height, those with an Overhead Camshaft engine are riskier than those missing.

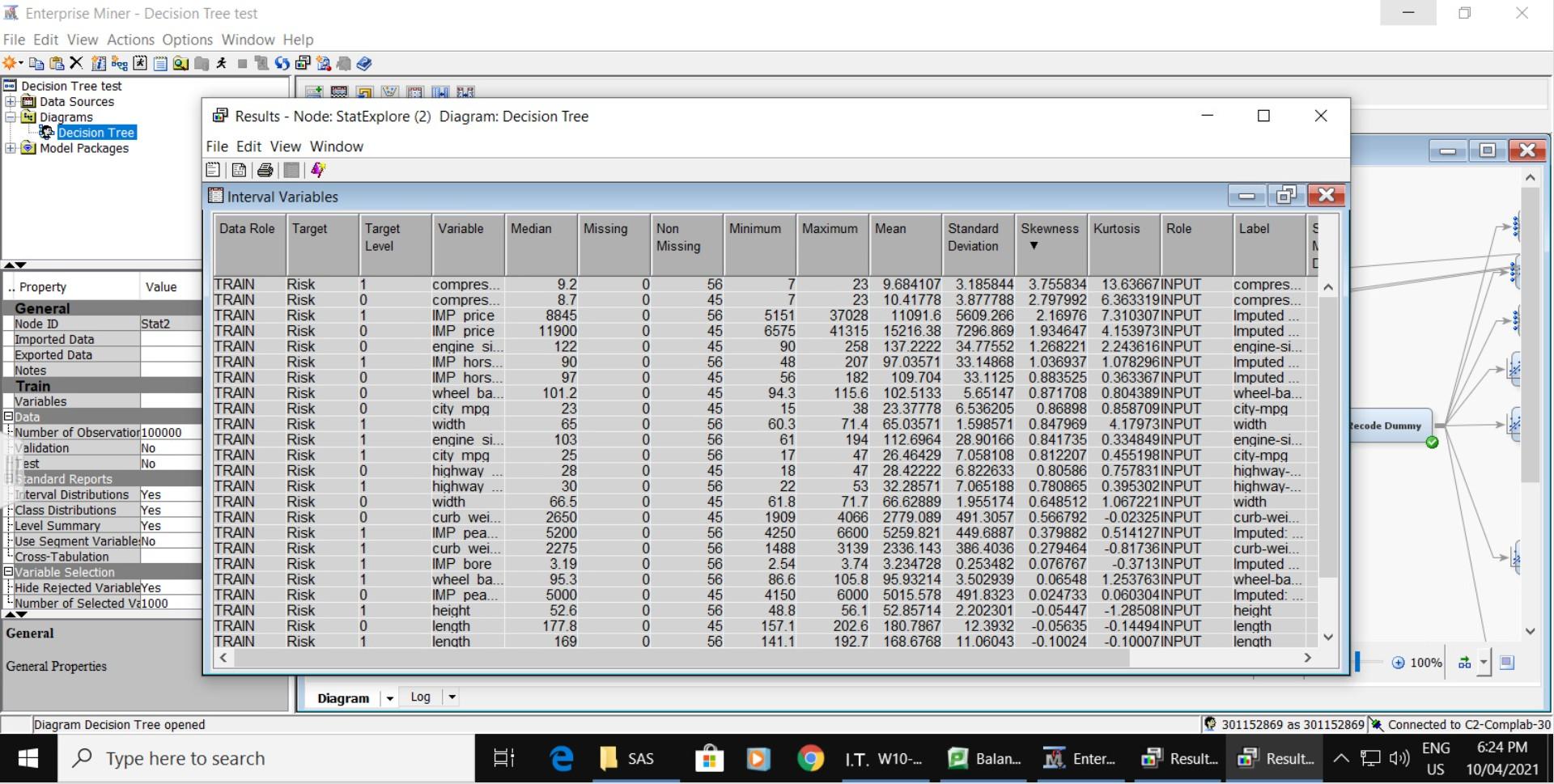
# Logistic Regression

## Imputation and Replacement

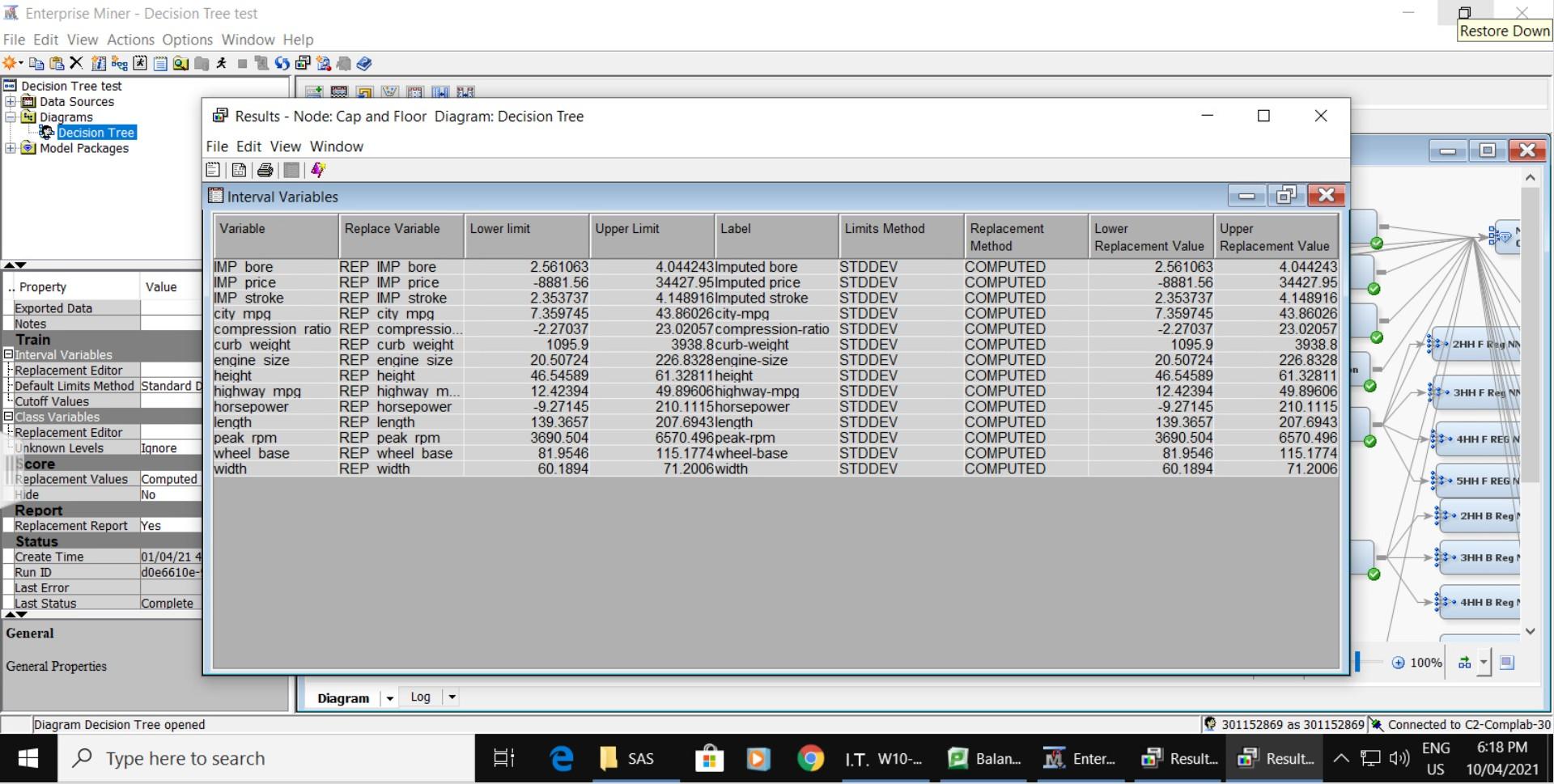
An impute node is added to the data partition to deal with the missing variables.



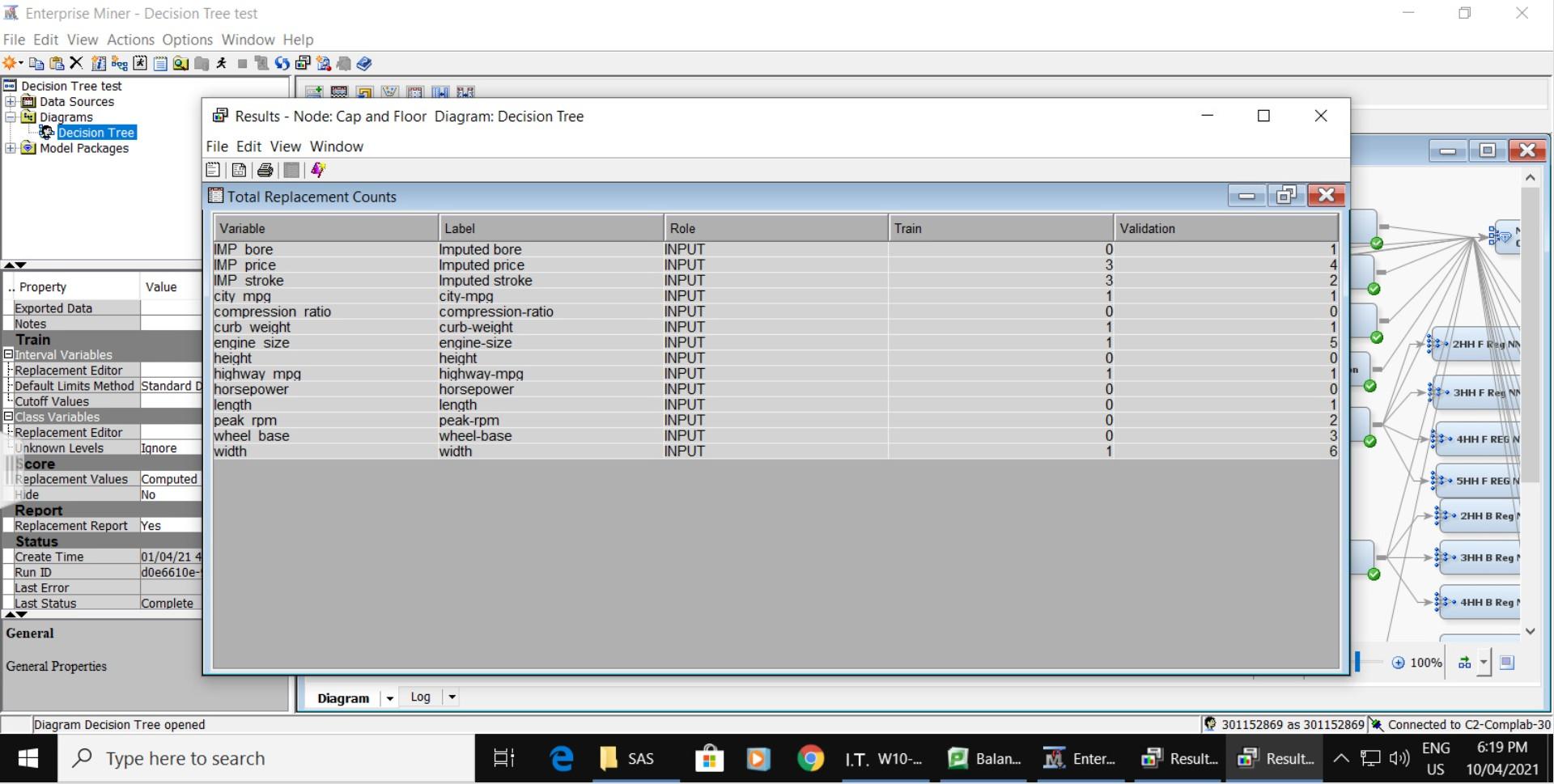
Four missing variables were imputed with the mean (bore, price, stroke) and count (number of doors)

.

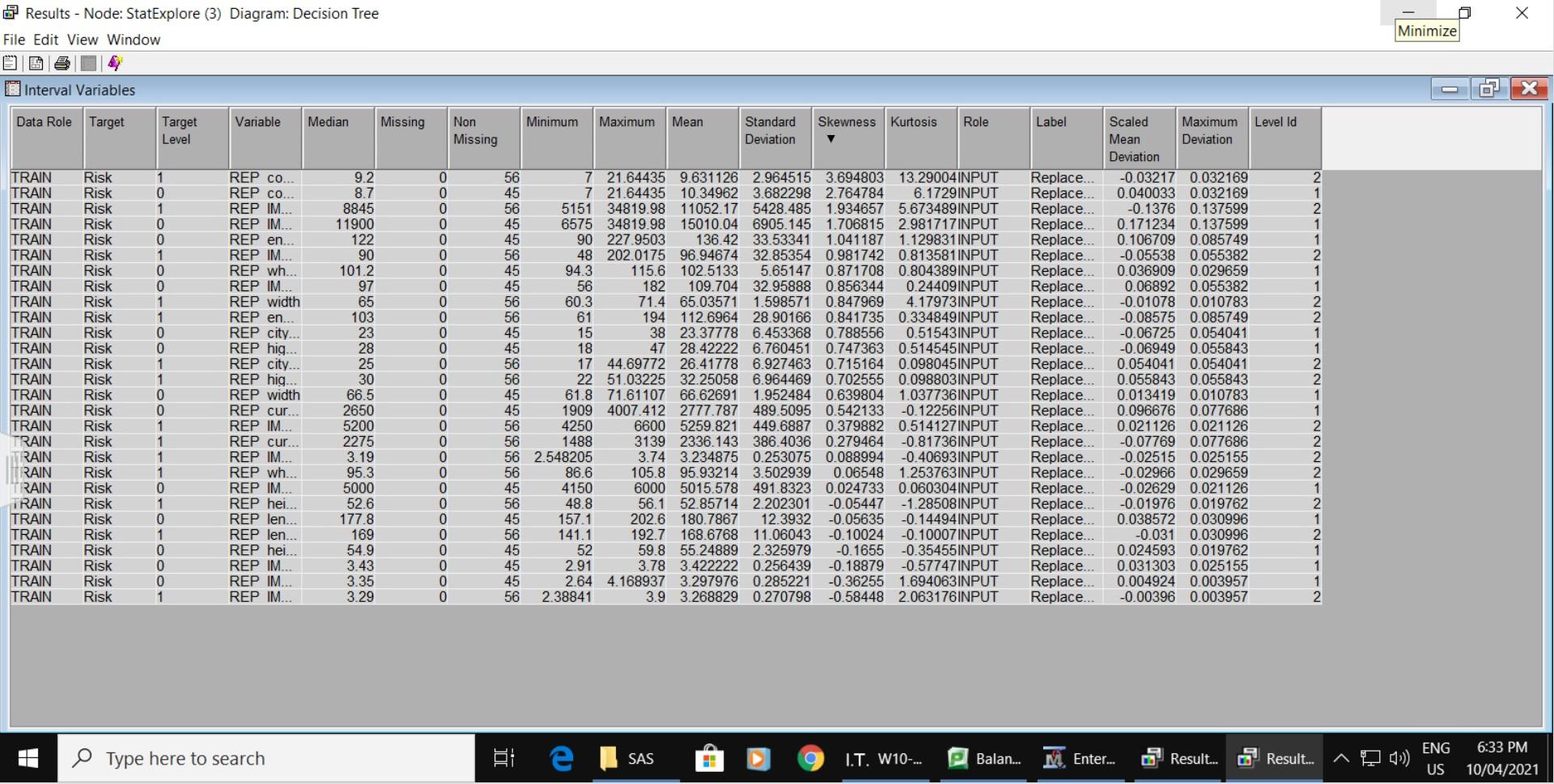
The skewness of compression ratio has decreased in imputation from 4.54 to 3.76. This indicates there are still extreme outliers in the compression ratio.



A replacement node is connected to the imputed node and named “cap and floor”. Lower and upper limits were automatically set when running.

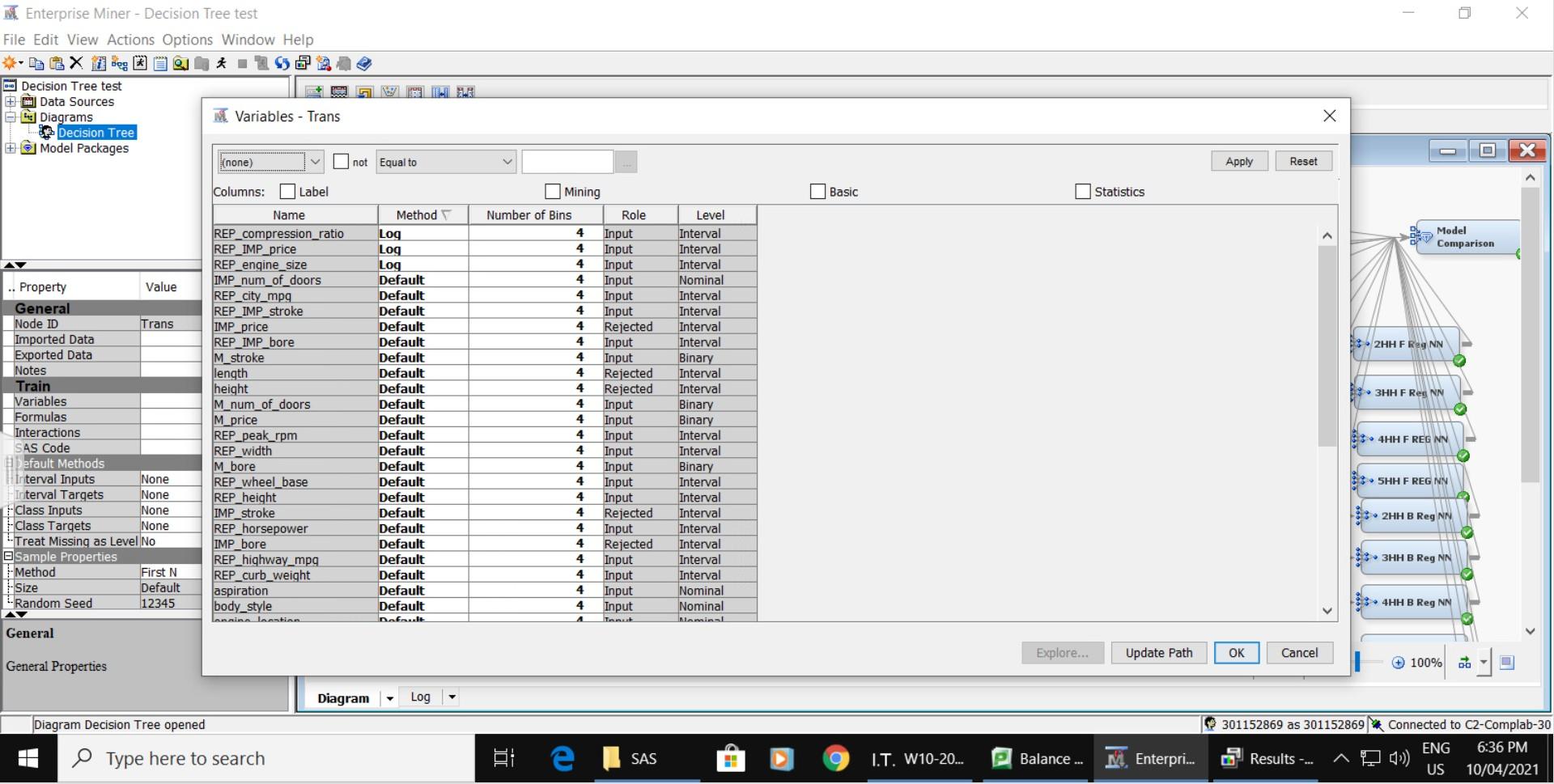


14 variables were replaced.

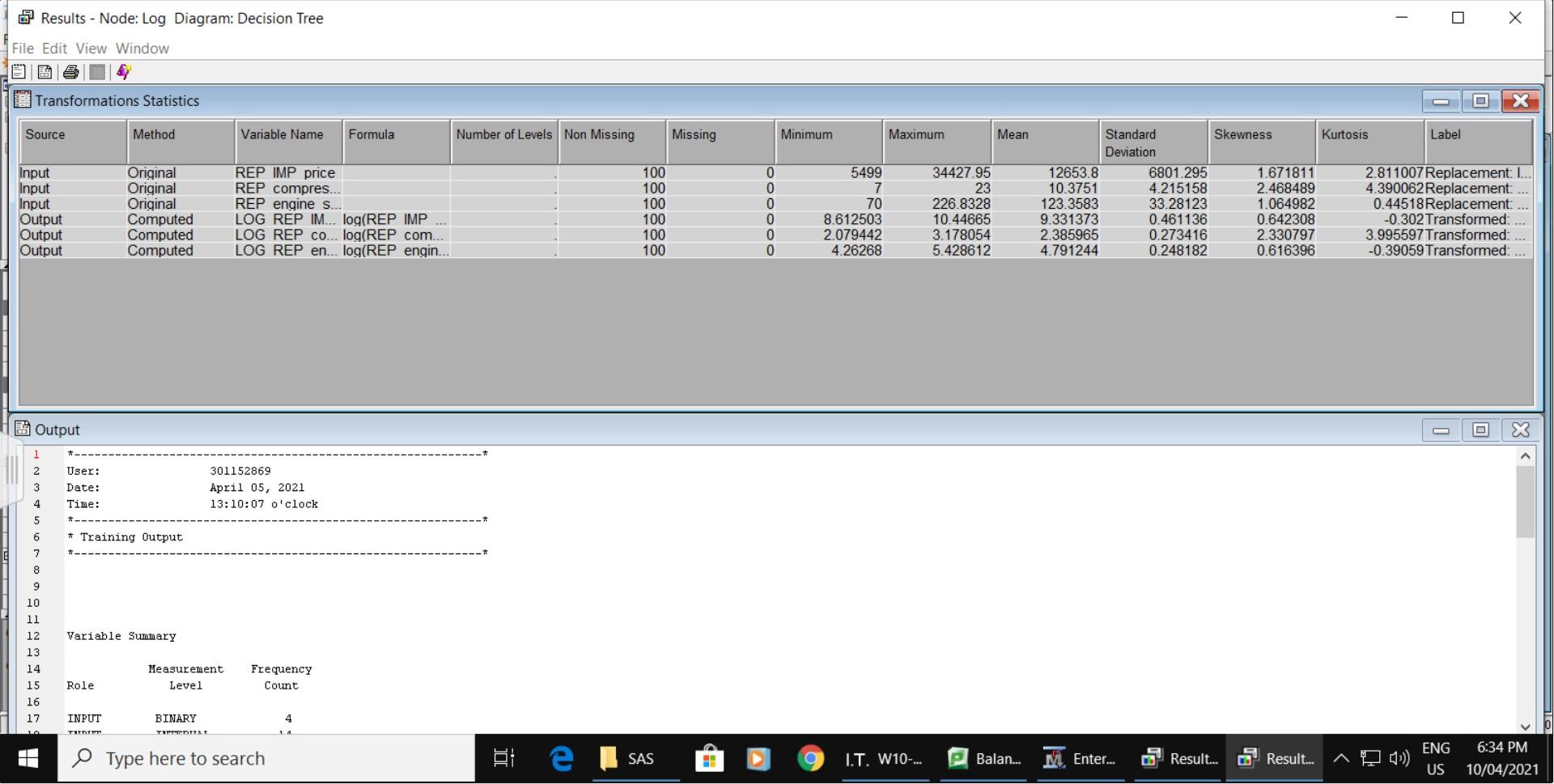


The skewness of compression ratio has decreased in Cap and Floor from 3.76 to 3.69.

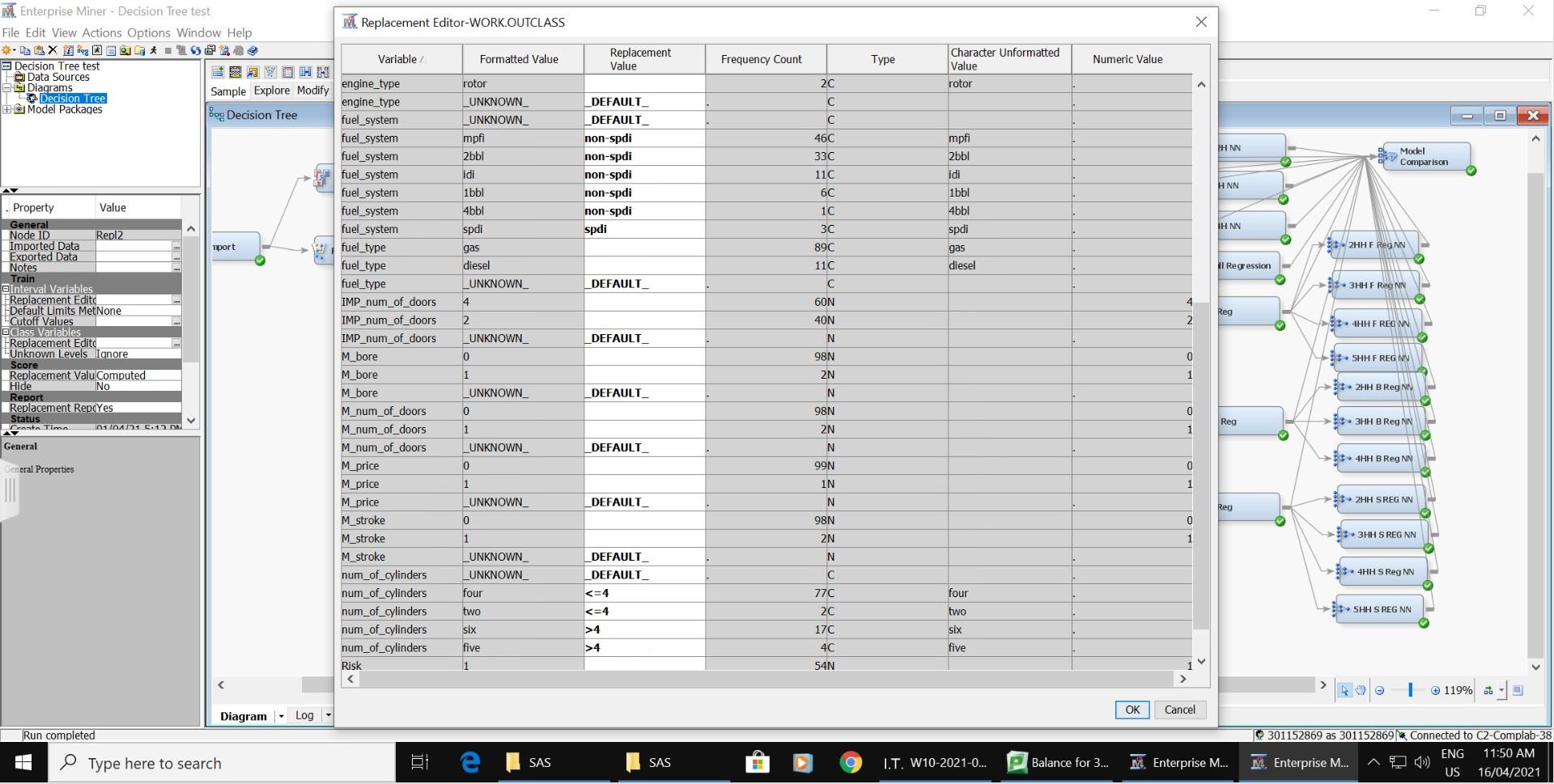
Log transformation worked the best among the results of all the transformation methods when reducing the skewness.



Log transformation was applied for three variables (compression, price, engine size).



Skewness for compression has decreased to 2.33 after Log Transformation.



In Recode Dummy, the number of cylinders was categorized into above 4 cylinders(>4) and 4 cylinders and below(<=4). Fuel type was categorized into “spdi” and “non-spdi” to reduce the curse of dimensionality, as spdi fuel type contributes to a higher risk level in the previous model.

# 

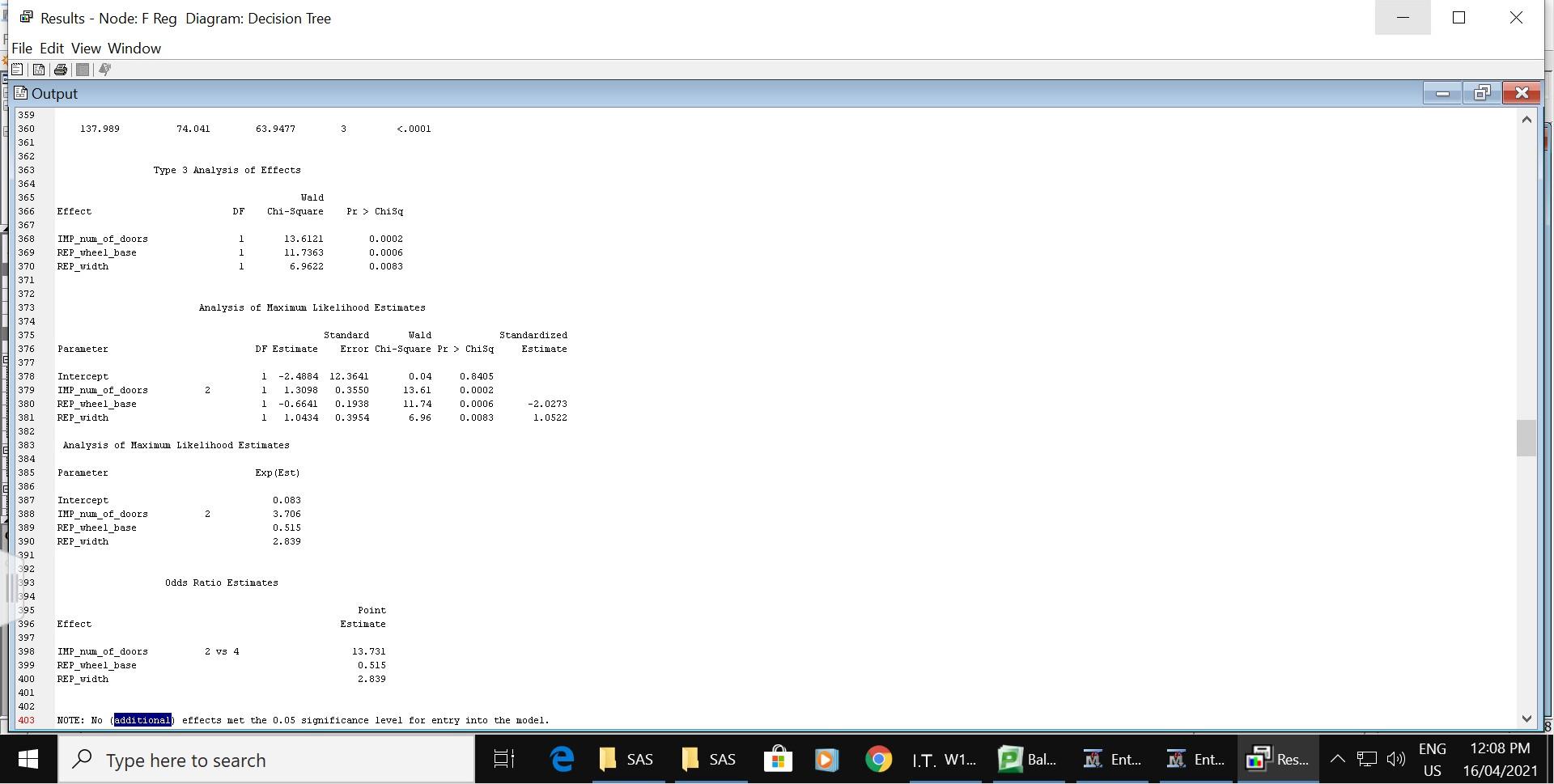
In the next step, four types of regression were added to the node.

## Full regression



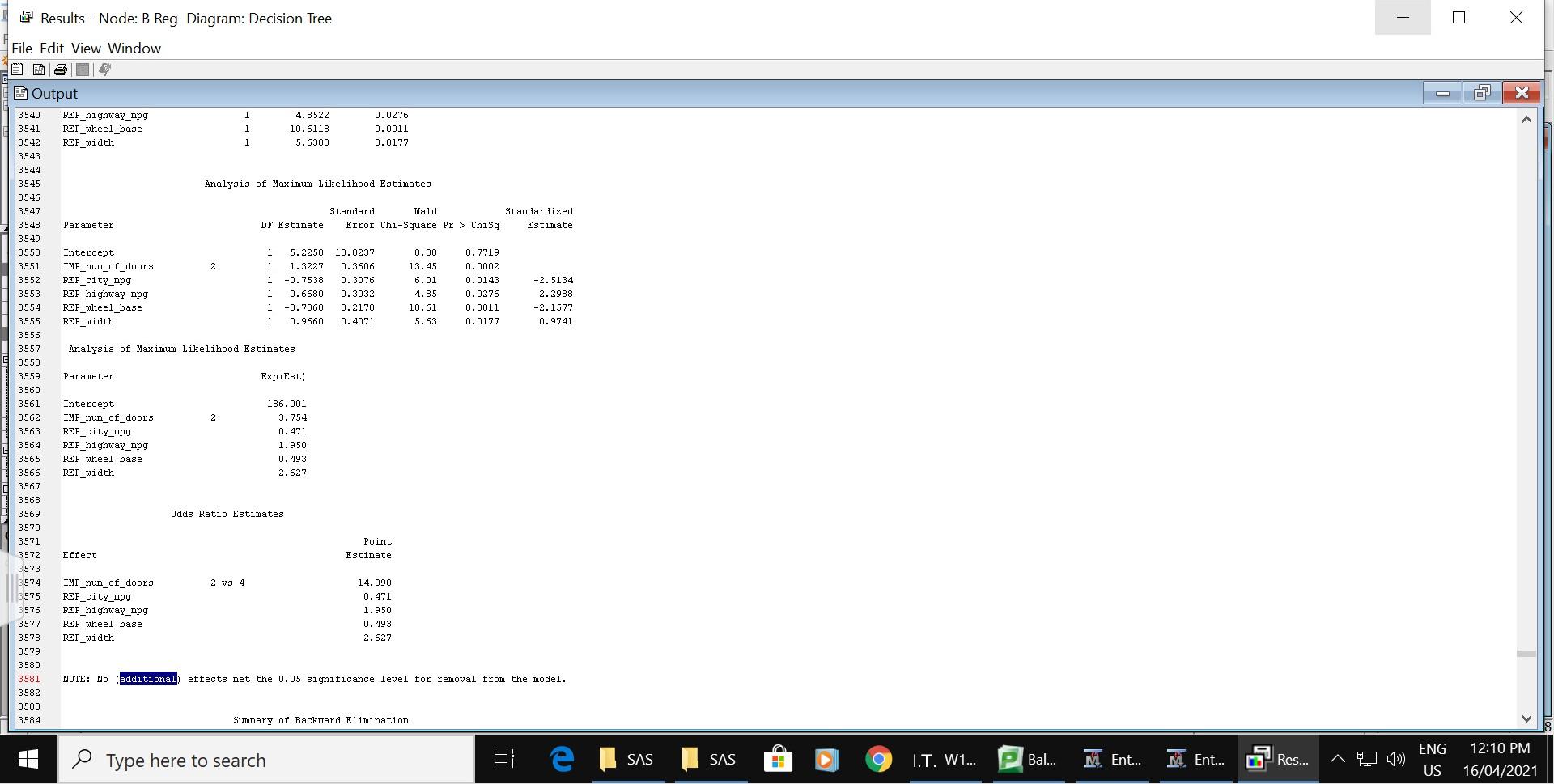
In the full regression, there are a lot of important variables. The Misclassification Rate is 0.17.

## Forward Regression



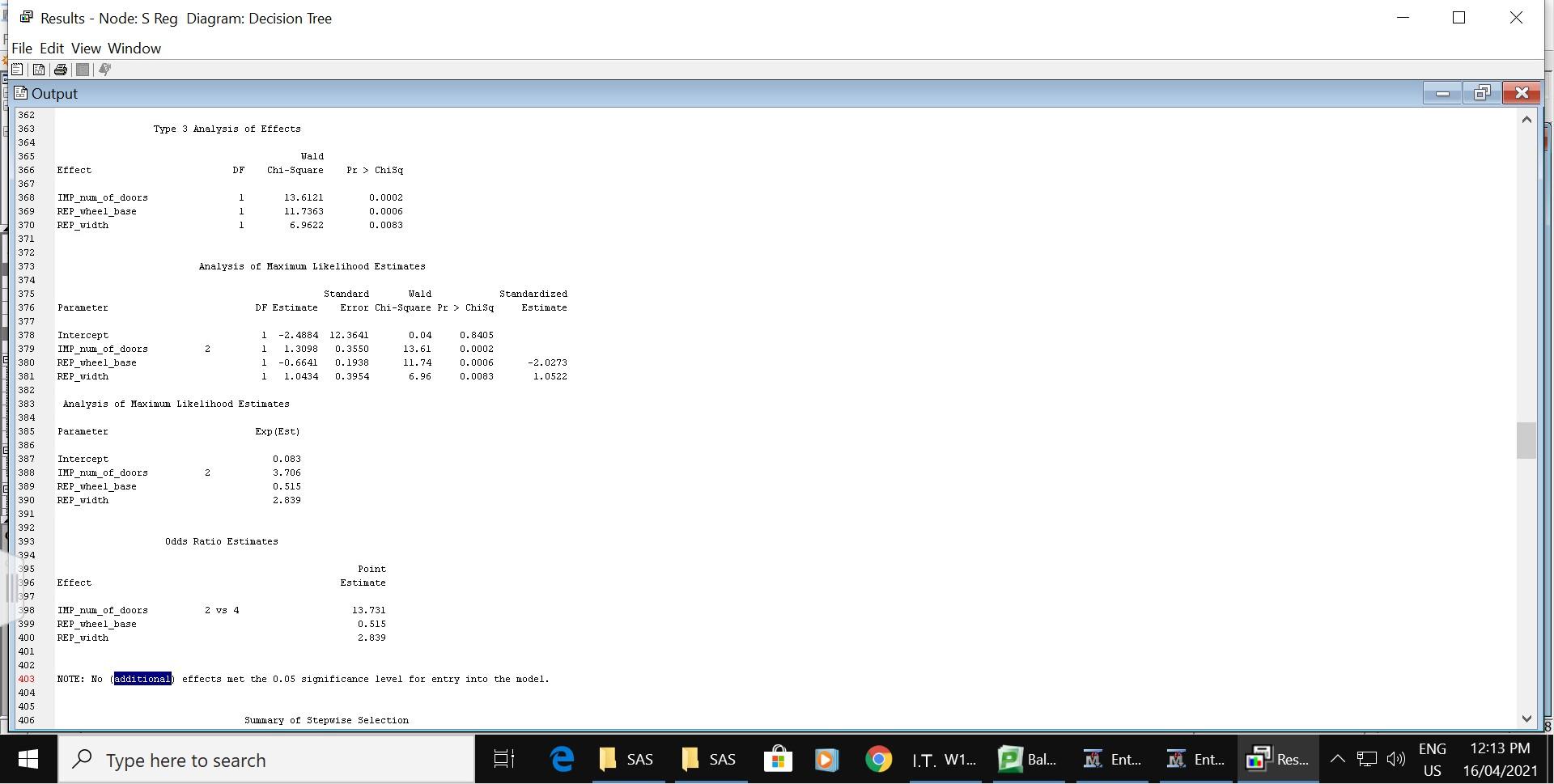
IMP\_num\_of\_doors, REP\_wheel\_base, REP\_width and fuel\_system are important since the p-value is less than 0.05 since the p-value is less than 0.05. The Misclassification Rate is 0.13.

## Backward Regression



IMP\_num\_of\_doors, LOG\_REP\_engine\_size, REP\_city\_mpg, REP\_height and REP\_highway\_mpg are important since the p-value is less than 0.05. The Misclassification Rate is 0.15.

## Stepwise Regression



IMP\_num\_of\_door, REP\_wheel\_base, REP\_width and Fuel\_system are important since the p-value is less than 0.05. The Misclassification Rate is 0.13.

## Best Model

The best model, in this case, is forward regression and stepwise regression. They have different procedures but end up with the same variables.

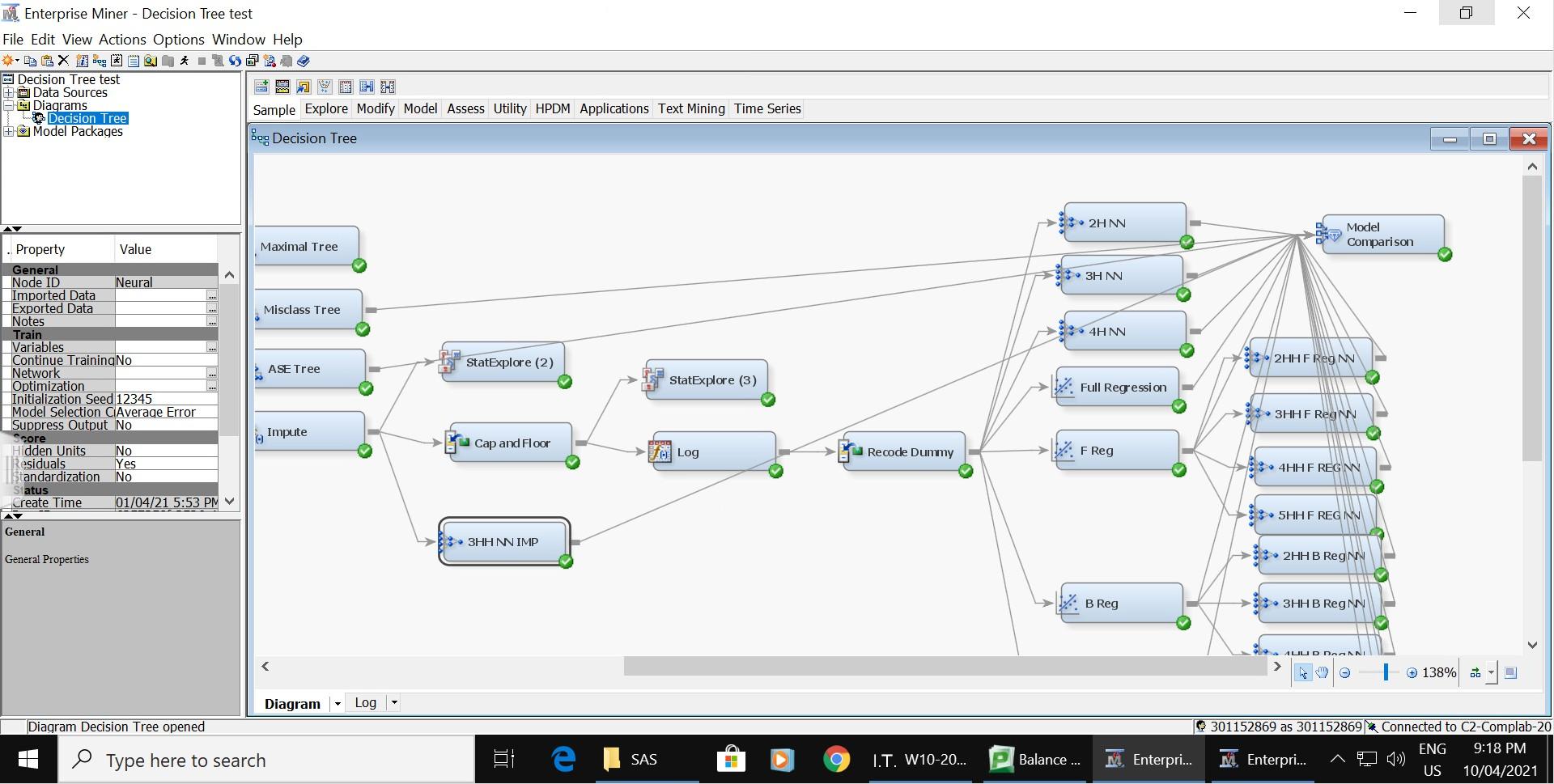
The number of doors is important, which indicates 2-door cars are 13.7 times as risky as 4-door cars.

Wheel-base is important, which indicates when the wheel-base increases by one unit, it is 48.5% less risky.

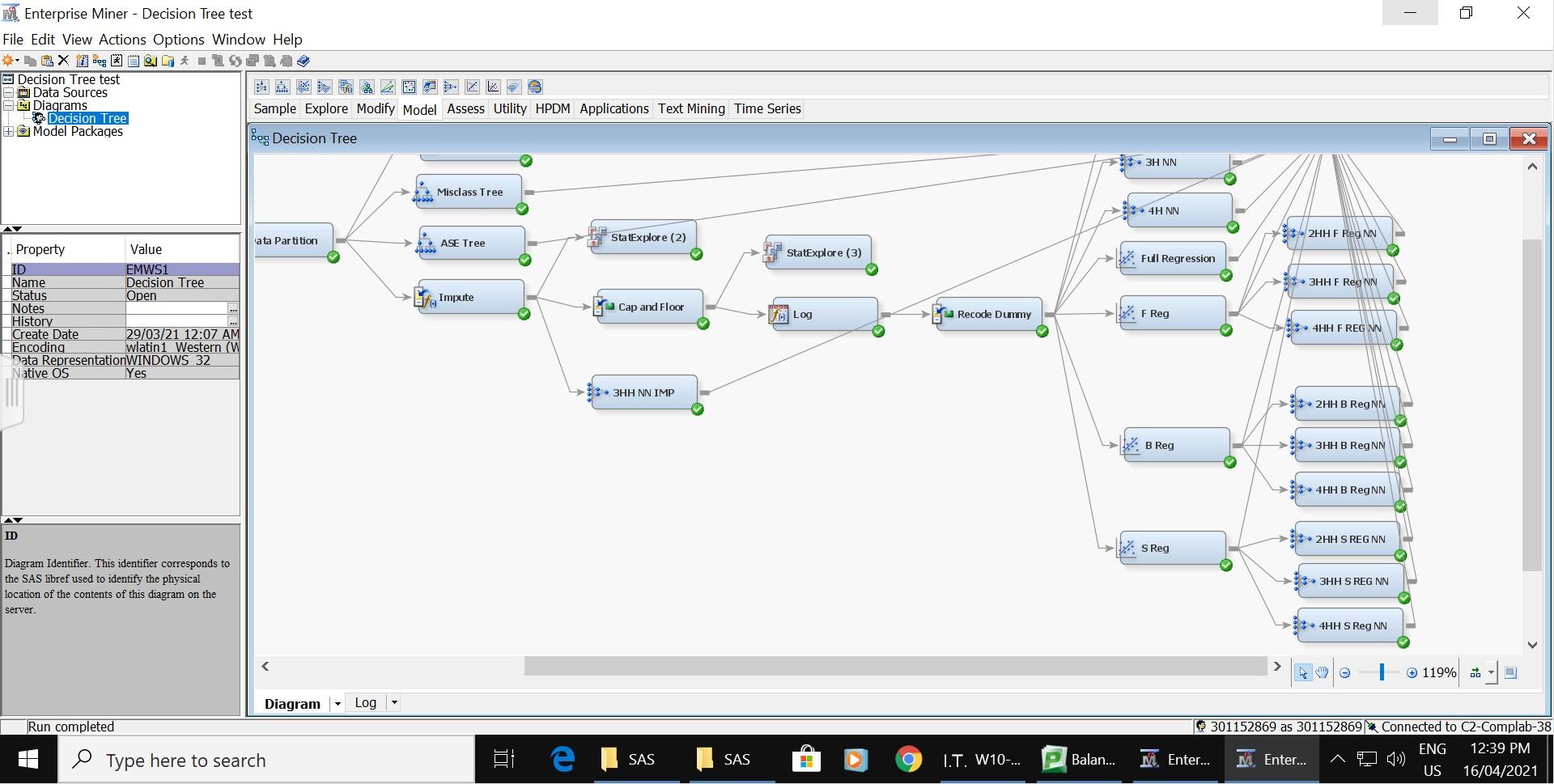
Width is important, indicating when the width increases by one unit, it is 1.8 times riskier.

# Neural Network

A 3 hidden unit neural network was added to the impute node.

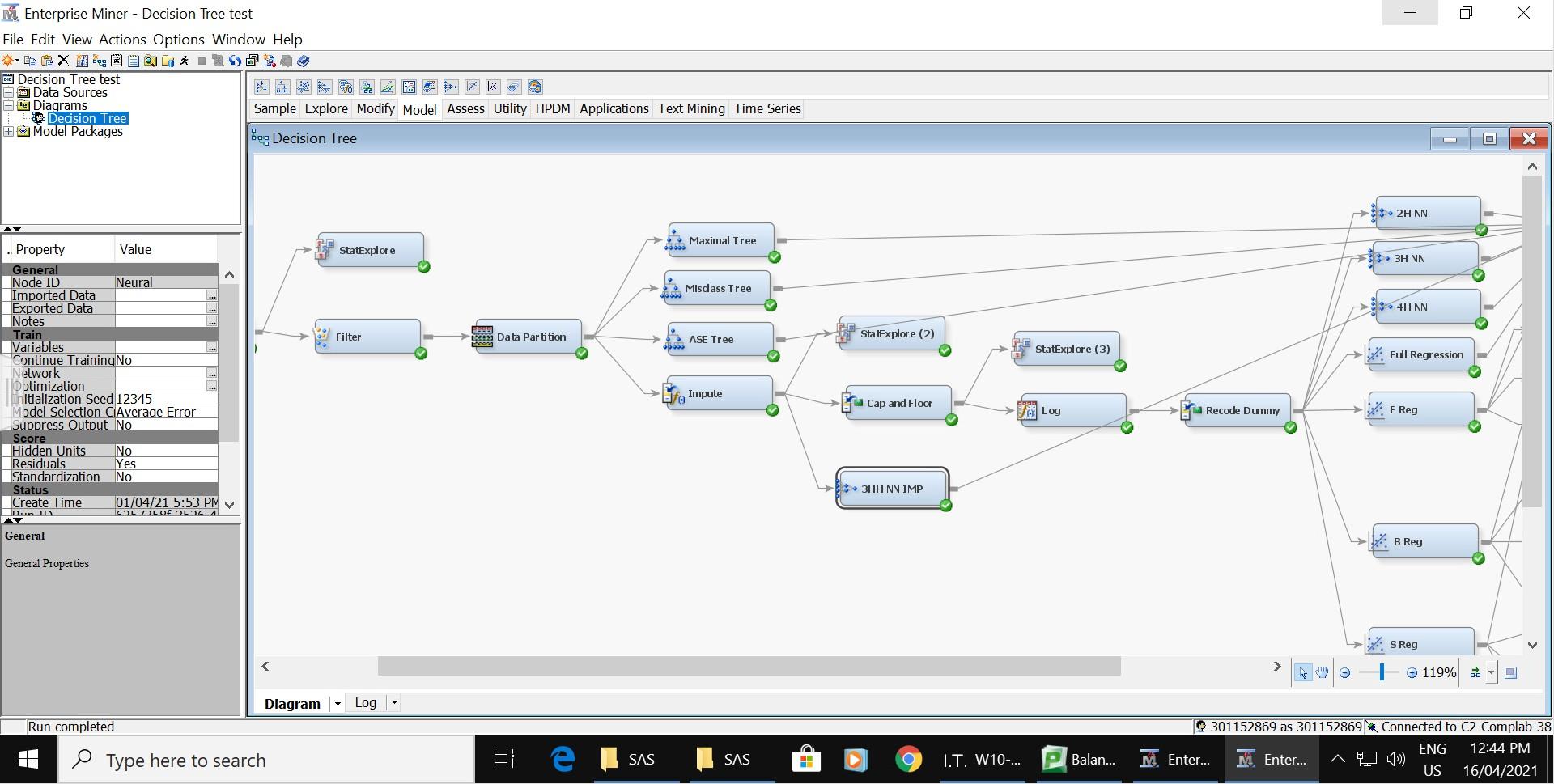


A two hidden unit neural network, a three hidden unit neural network, and a four hidden unit neural network were added to the recode dummy node. The two hidden unit network has the same Misclassification Rate as three hidden unit neural network, and better than four hidden unit neural network. It also has the lowest Validation Average Squared Error. In conclusion, two hidden unit neural network is good enough and no need to compare above four units.



A two hidden unit neural network, a three hidden unit neural network, a four hidden unit neural network were connected to the forward regression. They all have the same Misclassification Rate and two hidden unit have the lowest ASE. In conclusion, two hidden unit neural network is good enough and no need to compare above four units. The same applies to neural networks connected to stepwise regression.

A two hidden unit neural network, a three hidden unit neural network, a four hidden unit neural network were connected to the backward regression. It turns out two hidden unit neural network and four hidden unit neural network are both having the best Misclassification Rate. And two hidden unit have the lowest ASE. In conclusion, two hidden unit neural network is good enough and no need to compare above four units.



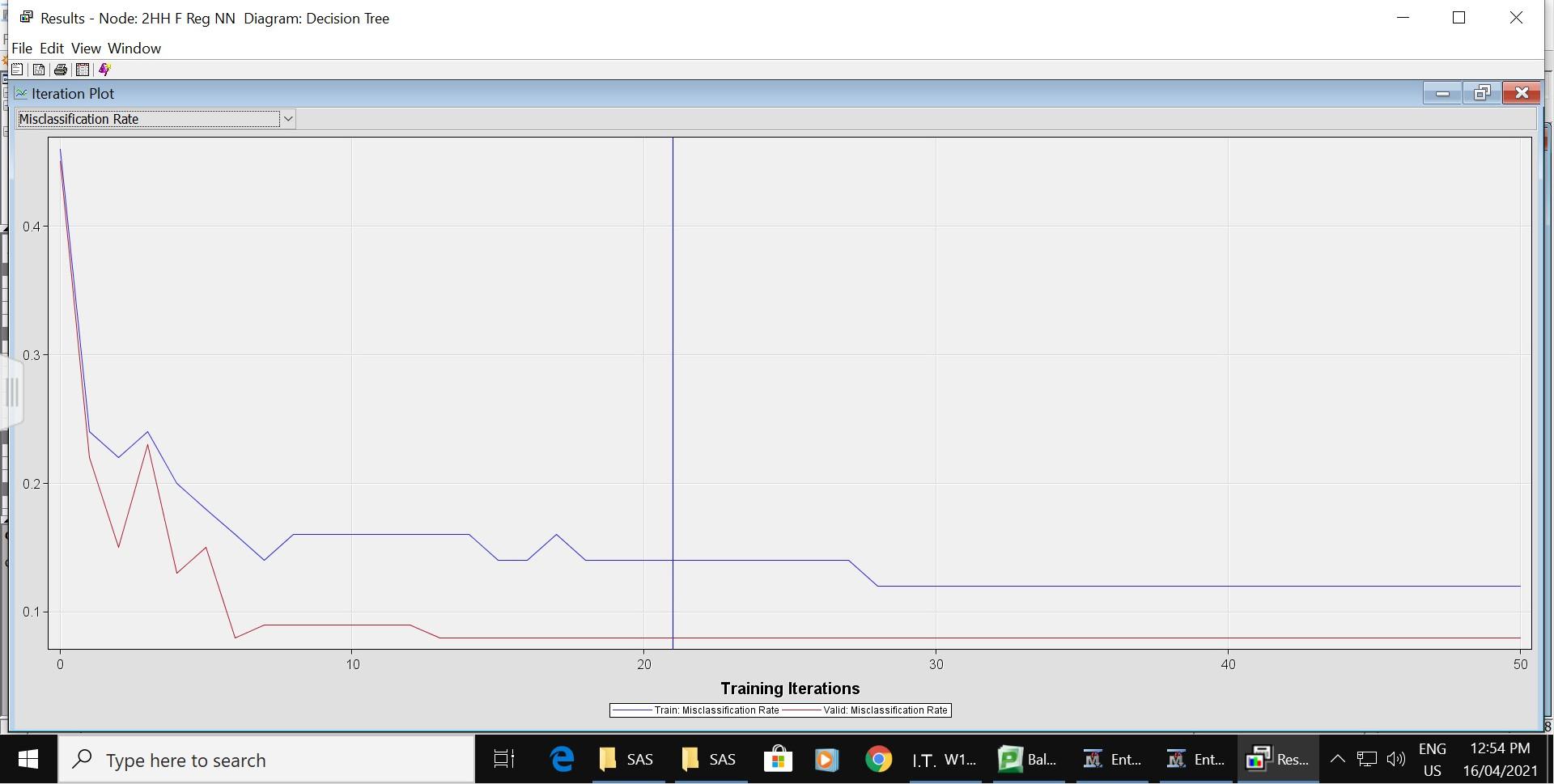
A three hidden unit neural network was connected to the impute node to test the effect of altering the model.

## Best Model

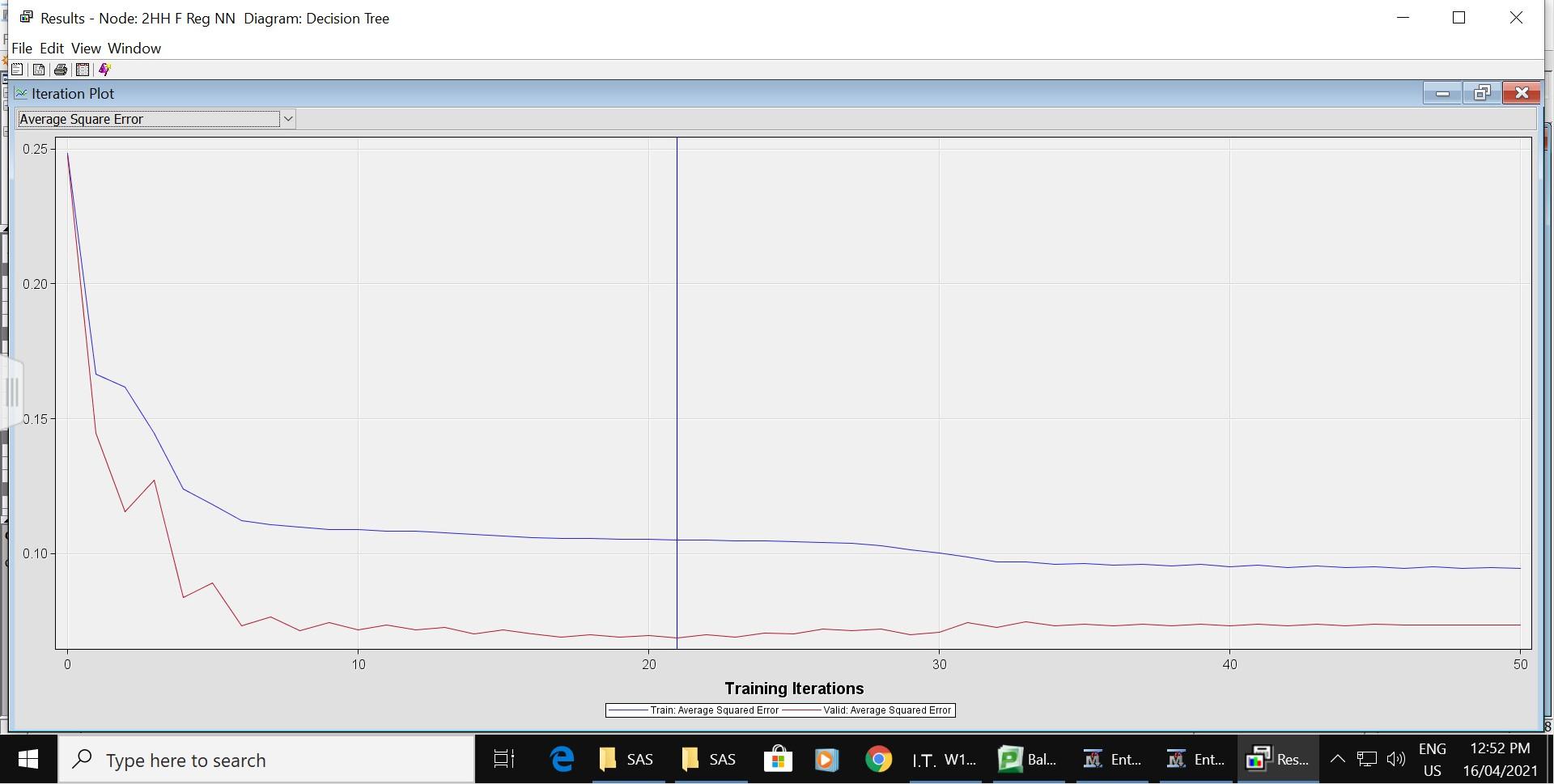
The best neural network is the two hidden unit neural network from forward/stepwise regression. It has the lowest Misclassification Rate of 0.08 and the Average Squared Error of 0.068825. The variables found important in forward/stepwise regression contribute to the best neural network.

**Iteration plot**

**Misclassification Rate**



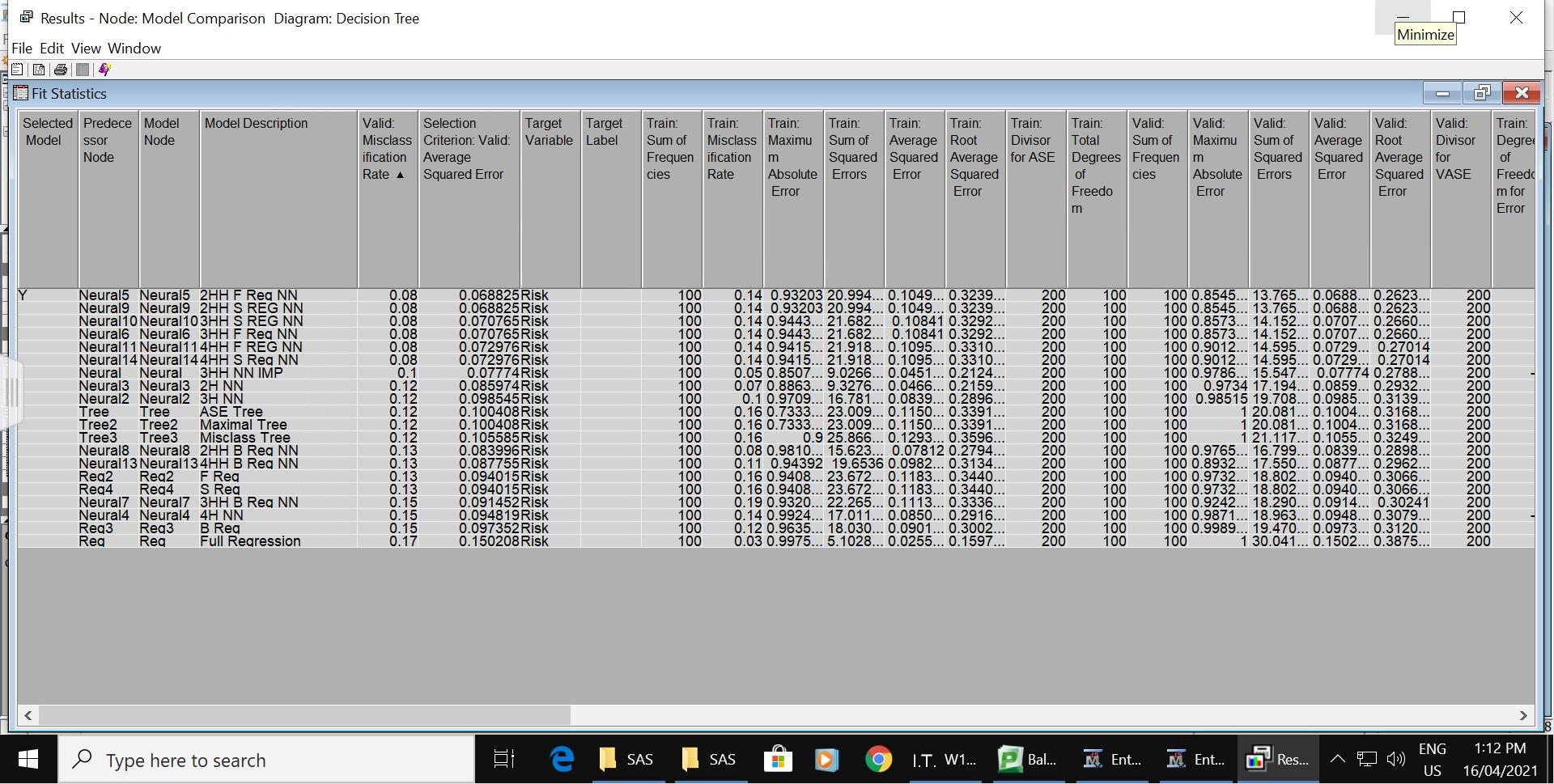
**Average Squared Error**



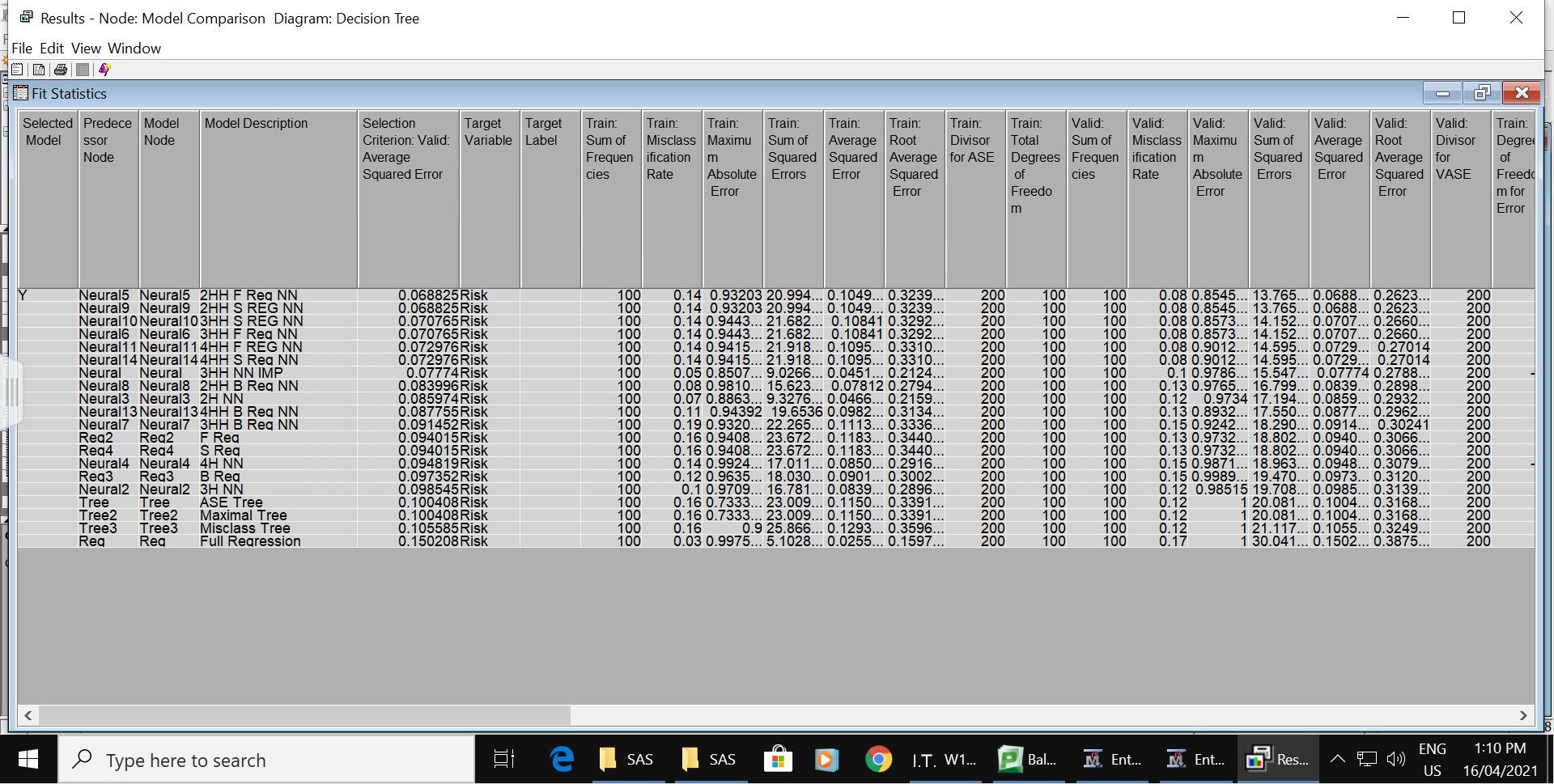
The iteration plot shows optimal validation Misclassification Rate and Average Squared Error occurring on iteration 21.

# Model Comparison

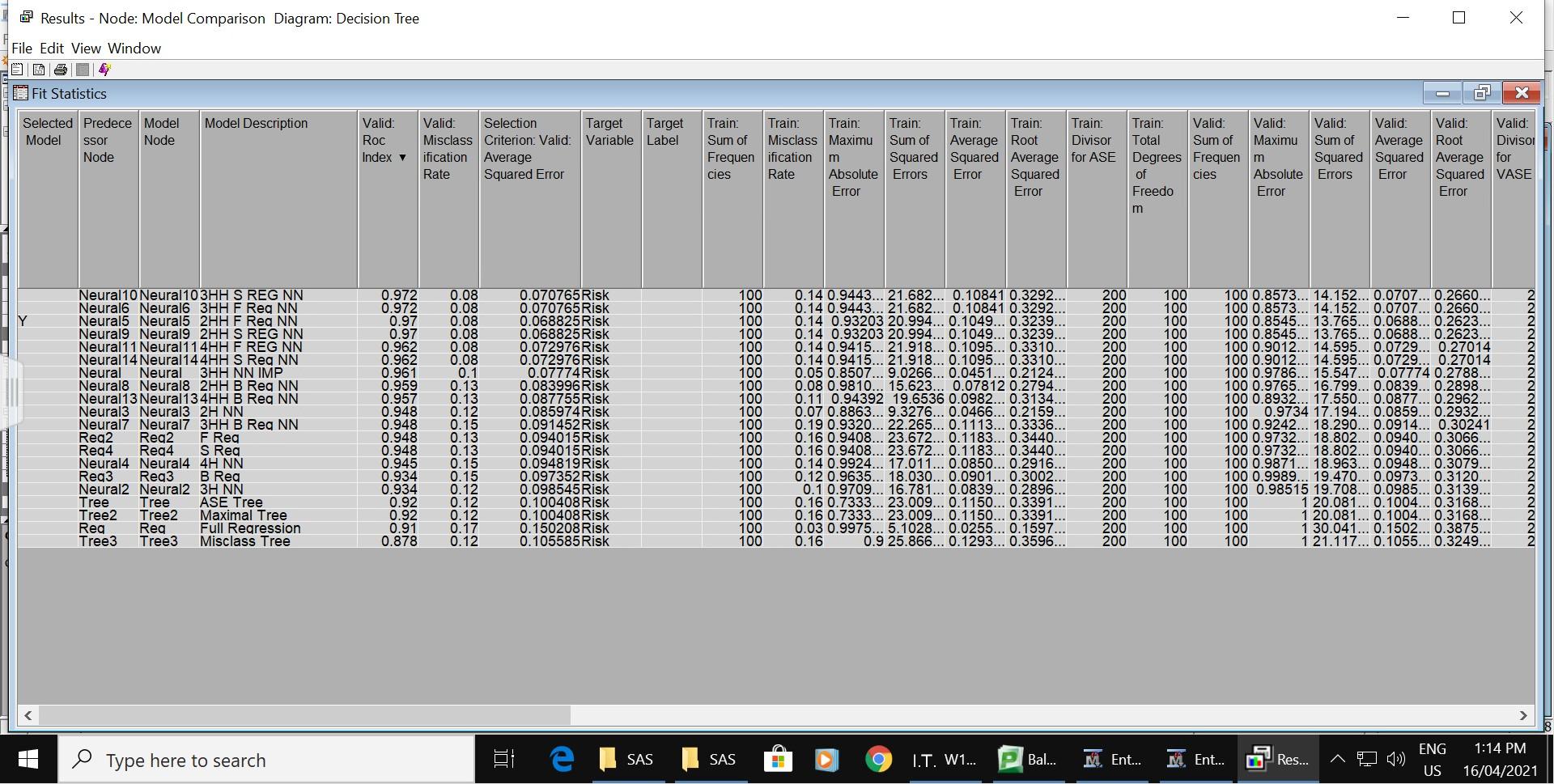
Best prediction model by Misclassification Rate:



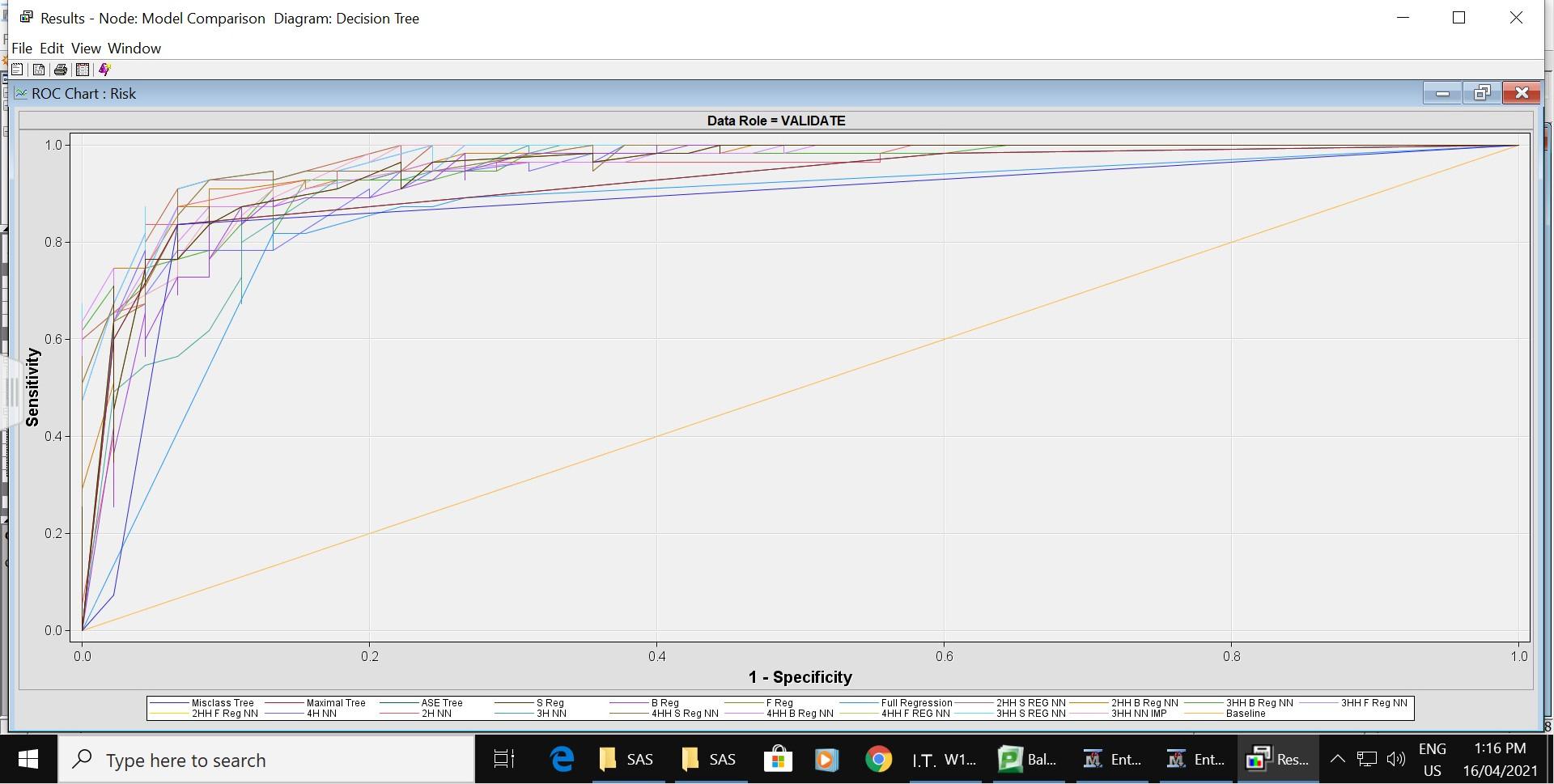
Best prediction model by Average Squared Error:



Best prediction model by ROC Index:



ROC Chart:



Based on the purpose of this predictive model, the prediction type is decisions.

|  |  |  |
| --- | --- | --- |
| Prediction Type | Validation Fit Statistic | Best Model |
| Decisions | Valid: Misclassification Rate | 1. **Neural Network:** Two Hidden Unit Neural Network from Forward/Stepwise Regression 2. **Decision Tree:** ASE/Maximal Tree |
| Estimates | Valid: Average Squared Error | 1. **Neural Network:** Two Hidden Unit Neural Network from Forward/Stepwise Regression 2. **Regression:** Forward/Stepwise Regression |
| Rankings | Valid: ROC Index | 1. **Neural Network:** Three Hidden Unit Imputed Neural Network 2. **Regression:** Forward/Stepwise Regression |

# Conclusion

In this case, the best model for determining whether a car model is risky or not is the Two Hidden Unit Neural Network from Forward/Stepwise Regression, as the most accurate model. A neural network is difficult to interpret for business decision-making, so the ASE/Maximal Tree will be used as the model for the explanation.

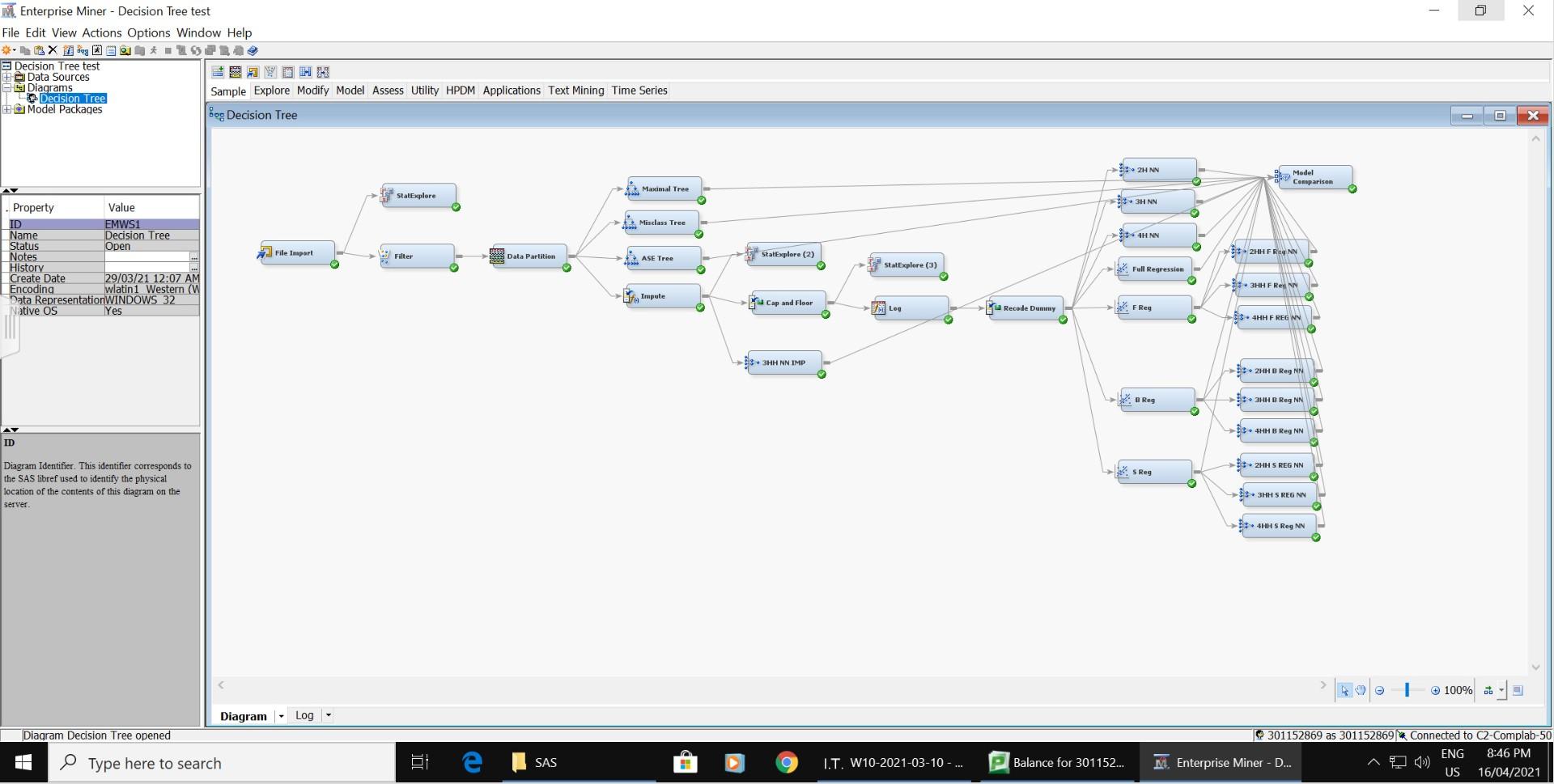
To determine whether a car model is risky or not, these variables are considered a lot:

* Two-door cars are always riskier than four-door cars
* A smaller wheelbase is riskier
* Lower ride height is riskier
* Wider cars are riskier.

# 

# Appendix

Appendix A - Full Diagram



Appendix B - Bibliography

Srinivasan,R. (2017) *Automobile Dataset.* Retrieved from https://www.kaggle.com/toramky/automobile-dataset