1. Executive Summary

The study utilizes data from the Behavioral Risk Factor Surveillance System (BRFSS) to predict heart disease based on multiple factors. Males are more susceptible than females to die from cardiovascular diseases (CVD), a significant global cause of death. The dataset includes nineteen variables covering lifestyle factors, categorized into continuous and categorical types. Data leakage points related to age group, height, weight, and BMI were identified and addressed. The study involves data exploration, sampling, partitioning, and predictive modeling using Neural Networks, Decision Trees, and Logistic Regression. Based on Average Squared Error, the Neural Network model with four hidden units and fifty iterations outperformed other models in predictive performance. According to the study results, several significant variables, such as arthritis, diabetes, sex, checkups, and alcohol consumption, primarily contribute to heart disease. Below is the pictorial view of the overall process flow of the project:

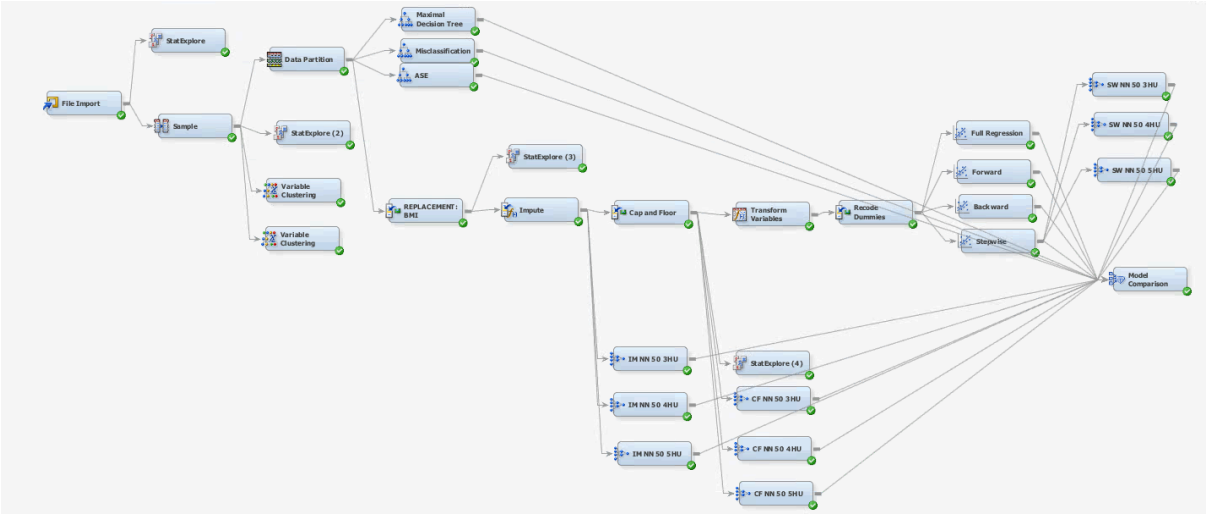


Figure: Overall Process Flow

2. Introduction

According to the World Heart Federation, Cardiovascular diseases (CVD) are known to be the leading cause of death worldwide and have accounted for millions of deaths. In 2019, 33% of the mortality was caused by Cardiovascular Diseases (World Heart Report 2023).

In most countries, it has been seen that Males have a higher tendency to death caused by CVDs than Females. There are several risk factors associated with CVDs. Some of the modifiable factors fall under behavioral, environmental, and metabolic categories, such as smoking and alcohol consumption in behavior, pollution in the environment, and high blood pressure and cholesterol in the metabolic type. Some unmodifiable factors are growing age, birth defects, and genetic mutations.

Many factors associated with the CVDs are interlinked with each other. Therefore, this study aims to find if there is any underlying relationship between these factors that can cause heart disease. And eventually, build the top-performing model to predict heart disease based on these factors.

2.1 Data Source

The dataset used for this predictive modeling was from the Behavioral Risk Factor Surveillance System (BRFSS). This nationally recognized health-related telephone survey system collects state information about U.S. residents' health-related risk behaviors, chronic health ailments, and prevention methods. It is a 2021 BRFSS Dataset from the CDC, titled "Cardiovascular Diseases Risk Prediction Dataset" and can be found on Kaggle - https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset.

The dataset contains information on various lifestyle factors of people that can eventually contribute as a risk factor for developing heart disease of any form.

2.2 Data Dictionary

The dataset has 19 variables, with 12 continuous ones and seven categorical ones.

Name	Level	Description
General Health	Nominal	Status of General Health: Excellent, Fair, Good, Poor and Very Poor
Checkup	Nominal	Time since the last routine checkup

Exercise	Nominal	Engage in any physical activity or not
Heart Disease	Nominal	Diagnosed with Heart Disease or not
Skin Cancer	Nominal	Diagnosed with Skin cancer or not
Other Cancer	Nominal	Diagnosed with any Other cancer or not
Depression	Nominal	Diagnosed with Depression or not
Arthritis	Nominal	Diagnosed with Arthritis or not
Sex	Nominal	Male or Female
Age_Category	Nominal	Falls into any of these age categories from 18-24 to 80+
Height (cm)	Interval	Height of the respondent in cm
Weight (kg)	Interval	Weight of the respondent in kg
BMI	Interval	The BMI Index of the respondent

Smoking History	Nominal	If smokes or not
Alcohol Consumption	Interval	Alcohol consumption in a month
Fruit Consumption	Interval	Fruit consumption in a month
Green Vegetable Consumption	Interval	Green vegetable consumption in a month
Fried Potato Consumption	Interval	Fried Potato consumption in a month

Table 2.2.1 - Data Dictionary

3. File Import

3.1. Import

A new project was created in SAS Enterprise Miner for this study. In this project, a new diagram named “Project” was created. A “**File Import**” Node was brought into this diagram. The CSV file of the cleaned data was imported through file “Import File” configuration on the Property Panel.

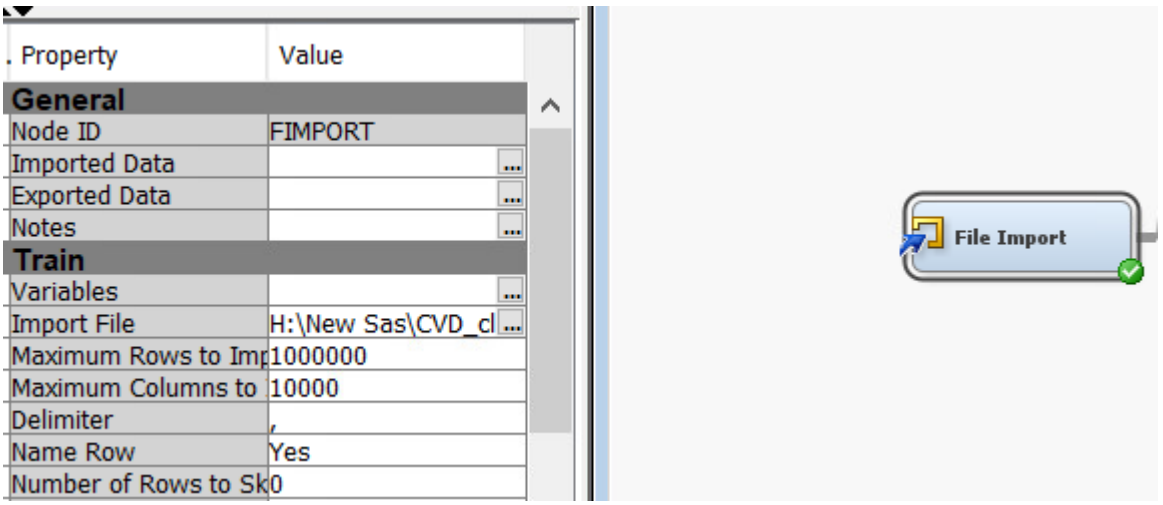


Figure 3.1.1. - "File Import" Node

3.2. Data Leakage (BMI | Height & Weight | Age Category)

From the variables, two data leakage points were identified:

- *BMI, Height & Weight*

The BMI variable was derived from the Height and Weight variable. Hence, there was a strong correlation between the BMI index and the weight and height of the individual. Thus to avoid the curse of dimensionality, we decided to reject the weight and height variables from the modelling.

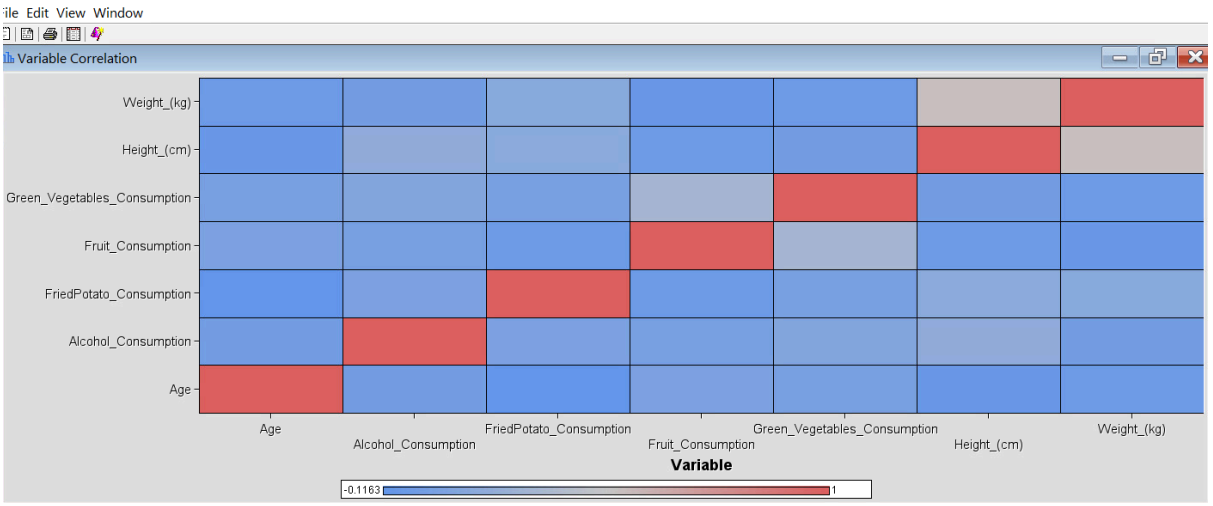


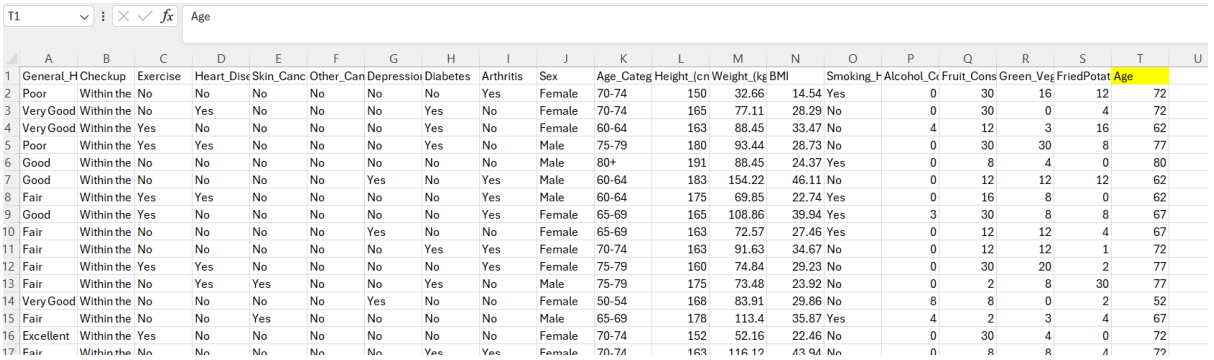
Figure 3.2.1 - Variable correlation showing a high correlation between height and weight variables. Results from Variable Clustering node

- *Age Category*

Another variable that we thought would cause an issue was the “Age category,” as it was nominal and was coming up as the most crucial variable in this data. Therefore, we created another variable named “Age,” which will be explained in the next section. It will be better to consider a continuous Age variable than a category Age variable in this study.

3.3. Creation of “Age” variable

We created another variable named “Age,” which was formed by taking the mean values of the Age category. For example, for the age category 50-54, the new Age would be 52. Thus, we rejected the “Age category” variable and kept the new “Age” variable as the input role and internal level.

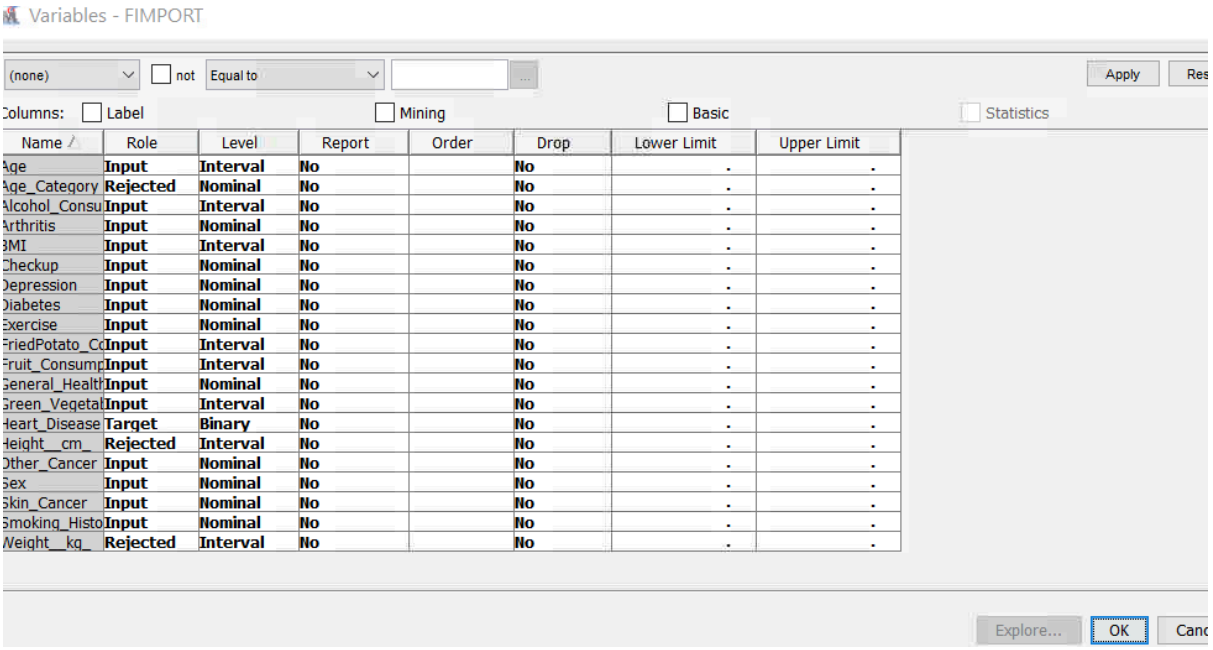


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	General_H	Checkup	Exercise	Heart_Dise	Skin_Canc	Other_Can	Depressio	Diabetes	Arthritis	Sex	Age_Categ	Height_(cn	Weight_(kg	BMI	Smoking_H	Alcohol_C	Fruit_Cons	Green_Veg	FriedPotat	Age	
2	Poor	Within the	No	No	No	No	No	No	Yes	Female	70-74	150	32.66	14.54	Yes	0	30	16	12	72	
3	Very Good	Within the	No	Yes	No	No	No	Yes	No	Female	70-74	165	77.11	28.29	No	0	30	0	4	72	
4	Very Good	Within the	Yes	No	No	No	No	Yes	No	Female	60-64	163	88.45	33.47	No	4	12	3	16	62	
5	Poor	Within the	Yes	Yes	No	No	No	Yes	No	Male	75-79	180	93.44	28.73	No	0	30	30	8	77	
6	Good	Within the	No	No	No	No	No	No	No	Male	80+	191	88.45	24.37	Yes	0	8	4	0	80	
7	Good	Within the	No	No	No	No	Yes	No	Yes	Male	60-64	183	154.22	46.11	No	0	12	12	12	62	
8	Fair	Within the	Yes	Yes	No	No	No	No	Yes	Male	60-64	175	69.85	22.74	Yes	0	16	8	0	62	
9	Good	Within the	Yes	No	No	No	No	No	Yes	Female	65-69	165	108.86	39.94	Yes	3	30	8	8	67	
10	Fair	Within the	No	No	No	No	Yes	No	No	Female	65-69	163	72.57	27.46	Yes	0	12	12	4	67	
11	Fair	Within the	No	No	No	No	No	Yes	Yes	Female	70-74	163	91.63	34.67	No	0	12	12	1	72	
12	Fair	Within the	Yes	Yes	No	No	No	No	Yes	Female	75-79	160	74.84	29.23	No	0	30	20	2	77	
13	Fair	Within the	No	Yes	Yes	No	No	Yes	No	Male	75-79	175	73.48	23.92	No	0	2	8	30	77	
14	Very Good	Within the	No	No	No	No	Yes	No	No	Female	50-54	168	83.91	29.86	No	8	8	0	2	52	
15	Fair	Within the	No	No	Yes	No	No	No	No	Male	65-69	178	113.4	35.87	Yes	4	2	3	4	67	
16	Excellent	Within the	Yes	No	No	No	No	No	No	Female	70-74	152	52.16	22.46	No	0	30	4	0	72	
17	Fair	Within the	No	No	No	No	No	Yes	Yes	Female	70-74	163	116.12	43.94	No	0	8	8	4	72	

Figure 3.3.1 - Variable correlation showing a high correlation between height and weight variables.

3.4. Roles and Variables

The role of the “Heart Disease” variable was set as the target, and its level was set as binary. As discussed above, the role of Height and Weight variables was rejected as it could cause a potential data leakage. Also, we created a new “Age” variable, whose role was set as input, and the variable “Age Category” was considered irrelevant and thus rejected. The rest of the variables were set to input and their respective levels as per Table 2.2.1 (Data dictionary).



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Age_Category	Rejected	Nominal	No		No	.	.
Alcohol_Consu	Input	Interval	No		No	.	.
Arthritis	Input	Nominal	No		No	.	.
BMI	Input	Interval	No		No	.	.
Checkup	Input	Nominal	No		No	.	.
Depression	Input	Nominal	No		No	.	.
Diabetes	Input	Nominal	No		No	.	.
Exercise	Input	Nominal	No		No	.	.
FriedPotato_Co	Input	Interval	No		No	.	.
Fruit_Consump	Input	Interval	No		No	.	.
General_Health	Input	Nominal	No		No	.	.
Green_Vegetal	Input	Interval	No		No	.	.
Heart_Disease	Target	Binary	No		No	.	.
Height__cm_	Rejected	Interval	No		No	.	.
Other_Cancer	Input	Nominal	No		No	.	.
Sex	Input	Nominal	No		No	.	.
Skin_Cancer	Input	Nominal	No		No	.	.
Smoking_Histo	Input	Nominal	No		No	.	.
Weight__kg_	Rejected	Interval	No		No	.	.

Figure 3.4.1- File Import Properties: Edit Variables

4. Data Wrangling

4.1. Data Exploration & Sampling

The imported data was initially explored using the StatExplore node to learn more about the data variables in detail. This node will help us identify information about the variables regarding their skewnesses, missing observations, etc. The StatExplore node was connected to the File Import node and was run.

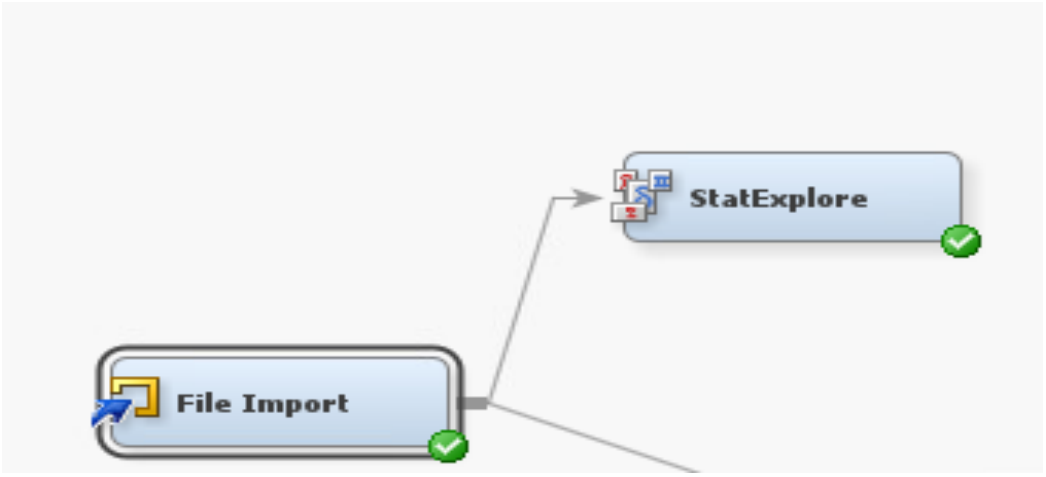
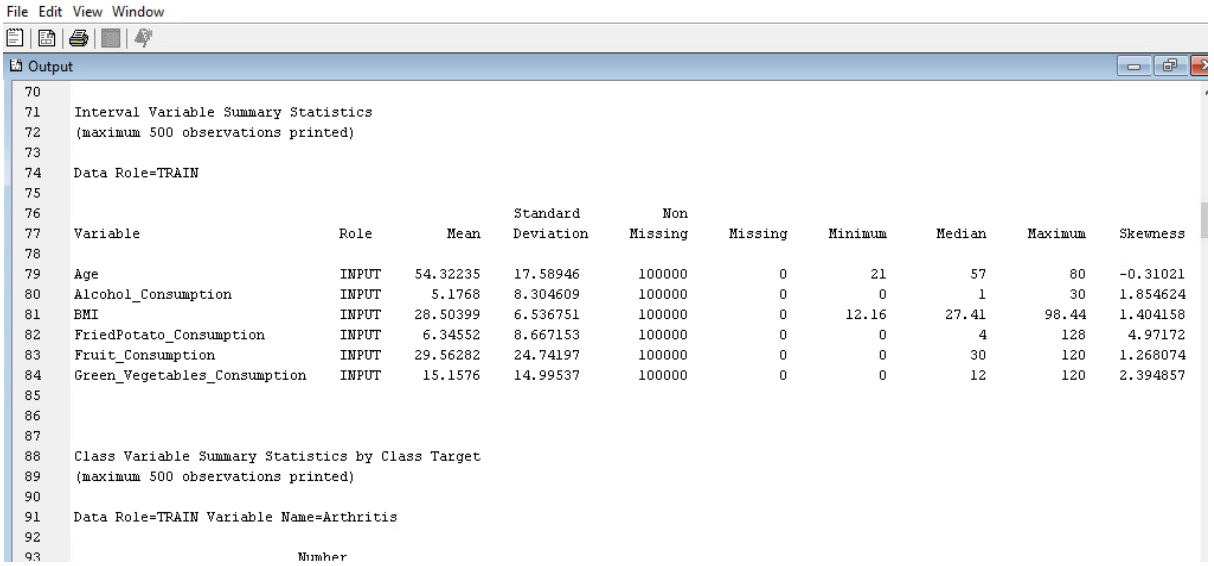


Figure 4.1.1 - "StatExplore" connected to "File Import" Node



File Edit View Window

Output

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
Age	INPUT	54.32235	17.58946	100000	0	21	57	80	-0.31021
Alcohol_Consumption	INPUT	5.1768	8.304609	100000	0	0	1	30	1.854624
BMI	INPUT	28.50399	6.536751	100000	0	12.16	27.41	98.44	1.404158
FriedPotato_Consumption	INPUT	6.34552	8.667153	100000	0	0	4	128	4.97172
Fruit_Consumption	INPUT	29.56282	24.74197	100000	0	0	30	120	1.268074
Green_Vegetables_Consumption	INPUT	15.1576	14.99537	100000	0	0	12	120	2.394857

Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Arthritis

Number

Figure 4.1.2. - Output Results of "StatExplore".

From the above results of the StatExplore connected to the File import, it can be observed that "Age" is the most critical variable in this analysis, and some other interval variables are skewed.

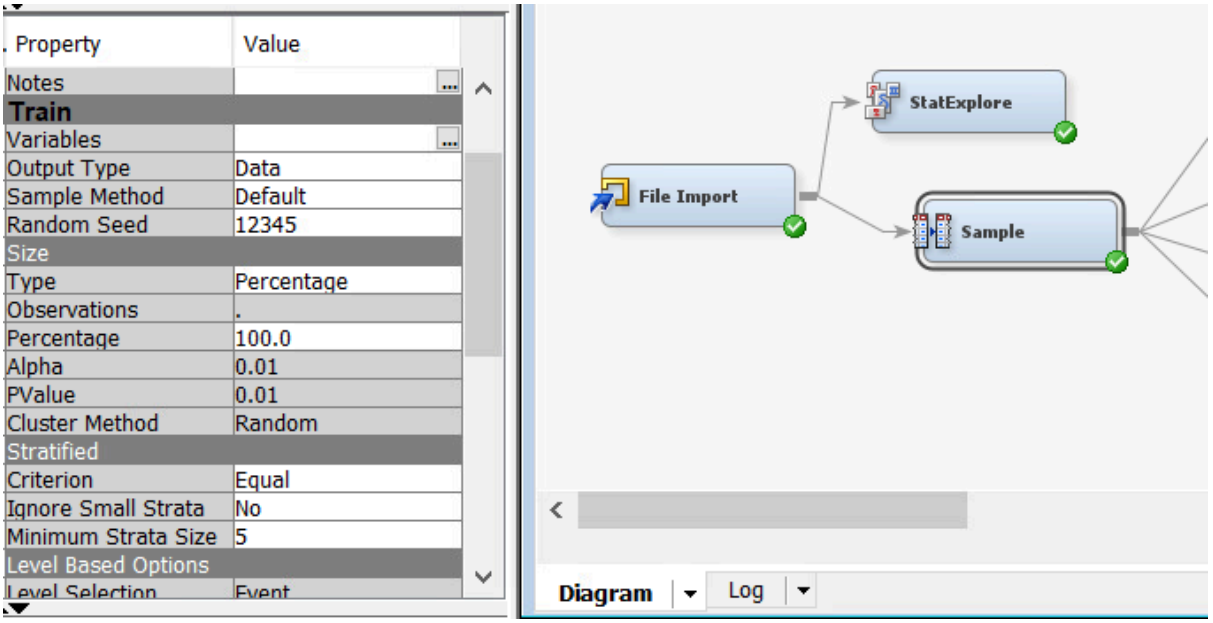


Figure 4.1.3. - "Sample" Node and its property panel on the left

We observed that our data is heavily imbalanced as our number of observations diagnosed with heart disease is way less than the number of observations having no heart disease. We brought the "Sample" node into the diagram to correct the imbalanced data and connected it with the "File Import" node.

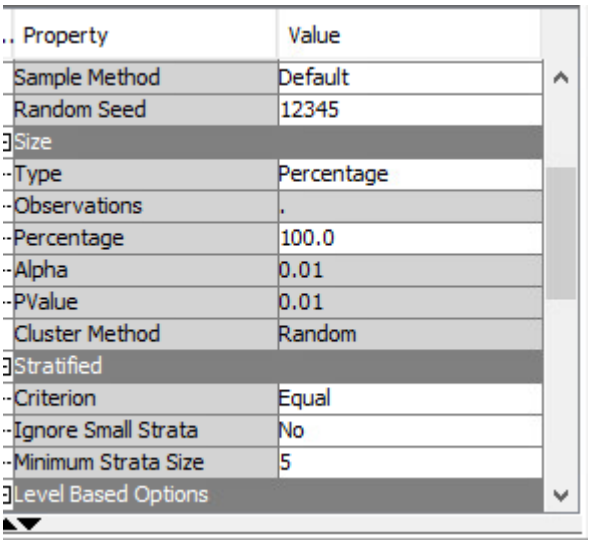


Figure 4.1.4. - "Sample" Node: property panel

We changed the property panel, as shown in the figure above, and ran it. To observe the changes made to the data, we connected another StatExplore node to the Sample node and ran it.

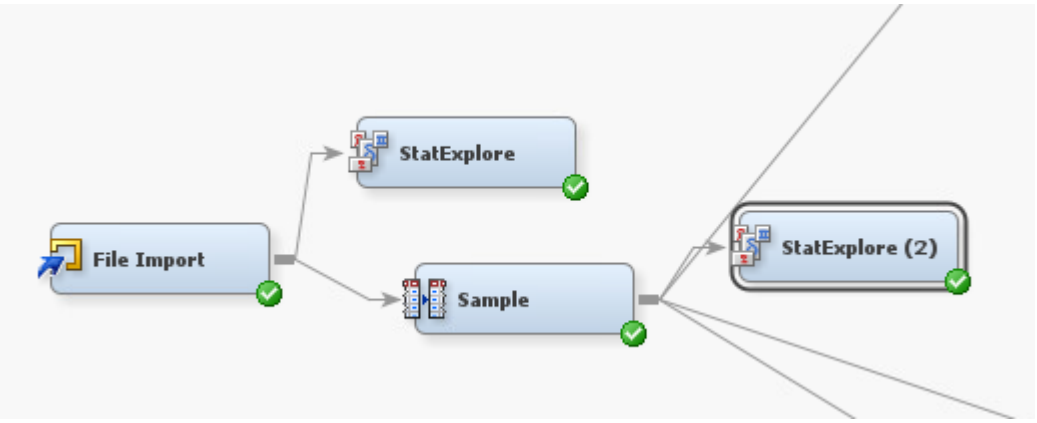


Figure 4.1.5. - The "StatExplore" node connected to the "Sample" Node

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Heart_Disease	.	No	24971	50	
Heart_Disease	.	Yes	24971	50	

Figure 4.1.6. - Output Results: "StatExplore" node connected to the "Sample" Node

It can be observed that the data was now balanced with 50% of each of the Yes & No values of the "Heart Disease" variable. All the rest of the data was excluded from the analysis.

Before moving forward, to understand any correlation between the variables, we brought the "Variable clustering" node into the diagram and connected it with the Sample node.

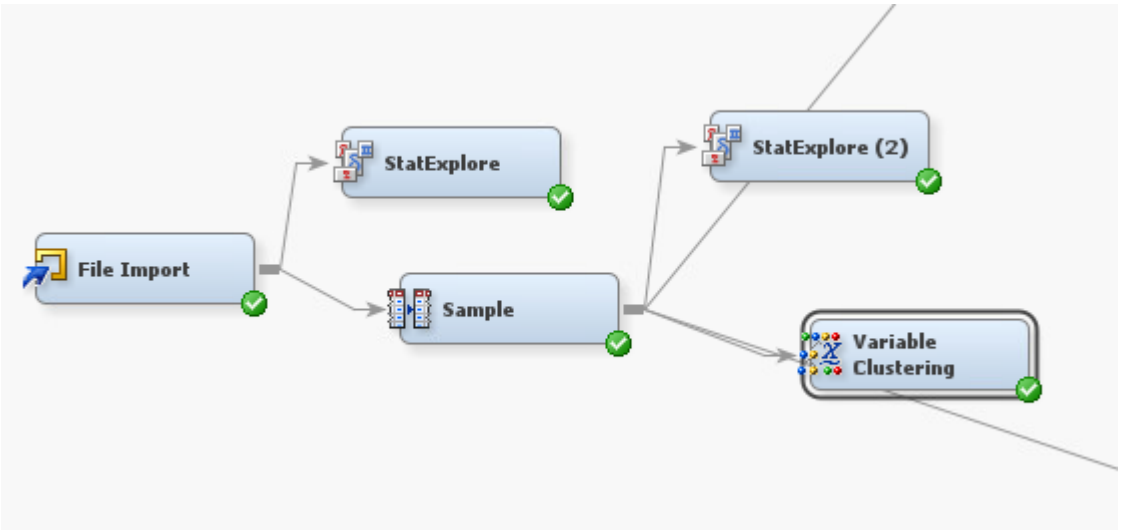


Figure 4.1.7. - The "Variable Clustering" node connected to the "Sample" Node

If we observe the results of the variable clustering node, we notice that now all the variables are uncorrelated. So, it was alright to move forward.

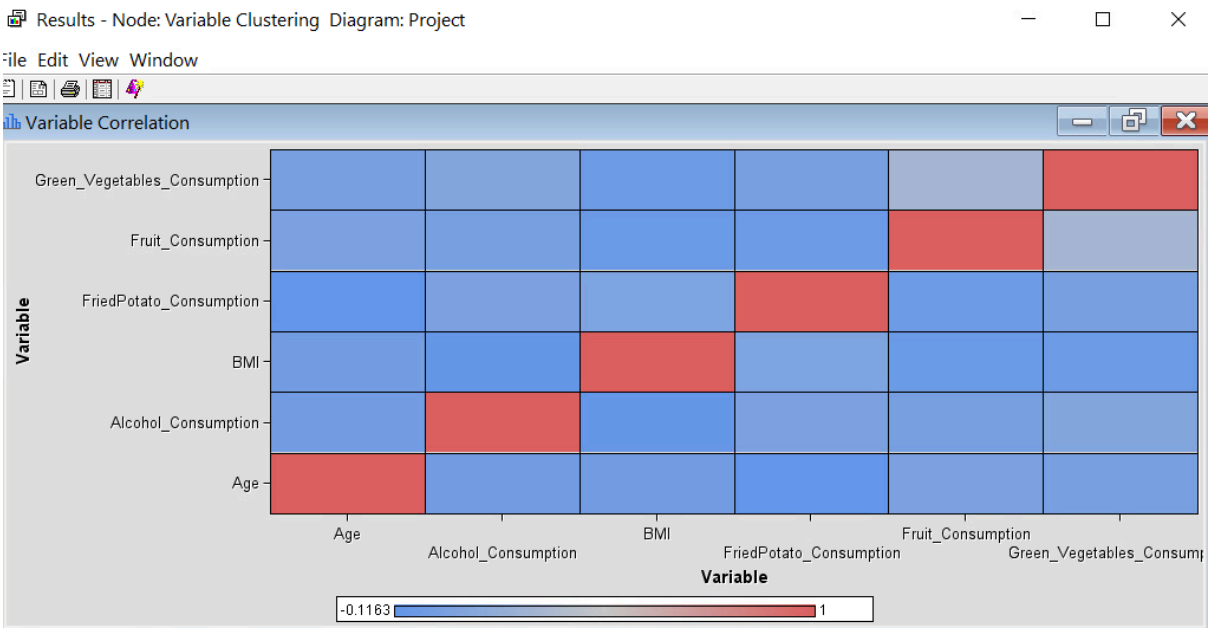


Figure 4.1.8. - The Variable Correlation matrix

4.2. Data Partition

The next step involved dividing the data into training and validation data sets at a ratio of 50:50. This was done to improve the accuracy of the analysis for this study and to get more generalized results.

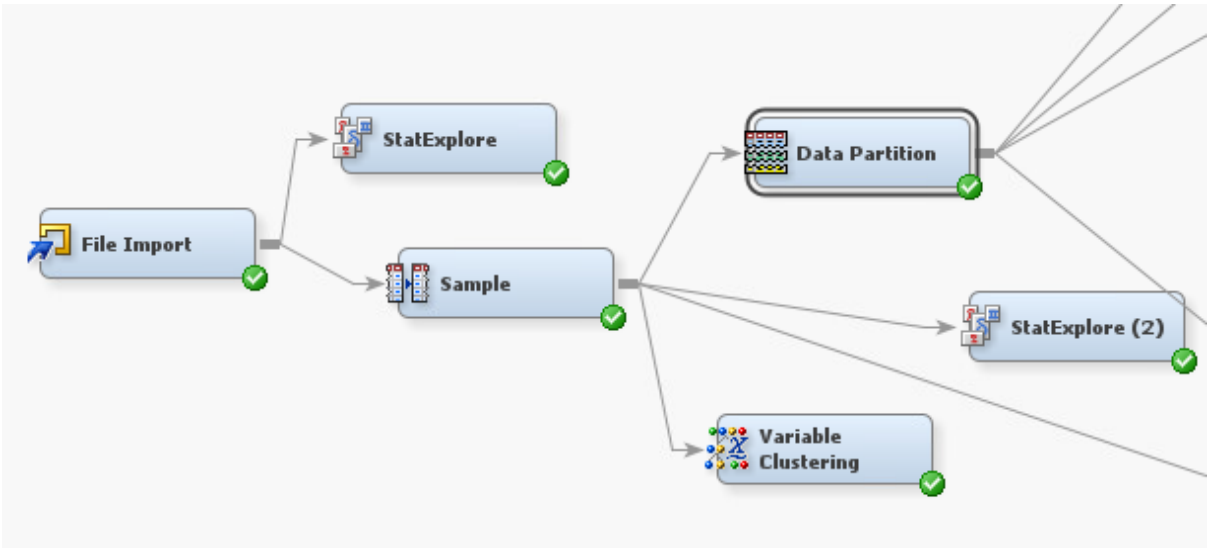


Figure 4.2.1. - The "Data Partition" node was connected to the "Sample" node

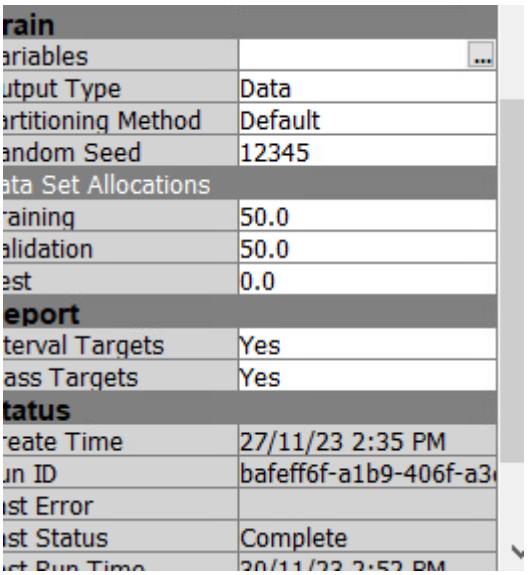


Figure 4.2.2. - The "Data Partition": Property Panel

5. Decision Tree

The Decision Tree was the first predictive modeling tool we used to study our data on the prediction of Heart Diseases. We made several decision trees to experiment with the data and observe which tree provided the best results. We initially performed a 2, 3, and 4-branch split maximal decision trees.

5.1. 2-Branch Split Maximal Tree

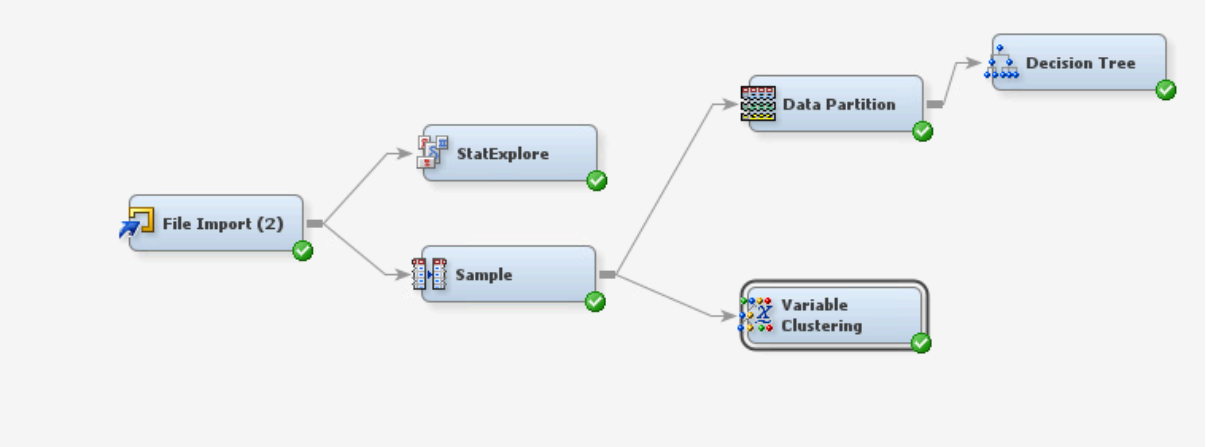


Figure 5.1.1 - 2-Branch Split Maximal Tree: The "Decision Tree" node was connected to the "Data Partition" node

We brought in the "Decision Tree" node from the Model tab, connected it to the Data partition node, and ran it.

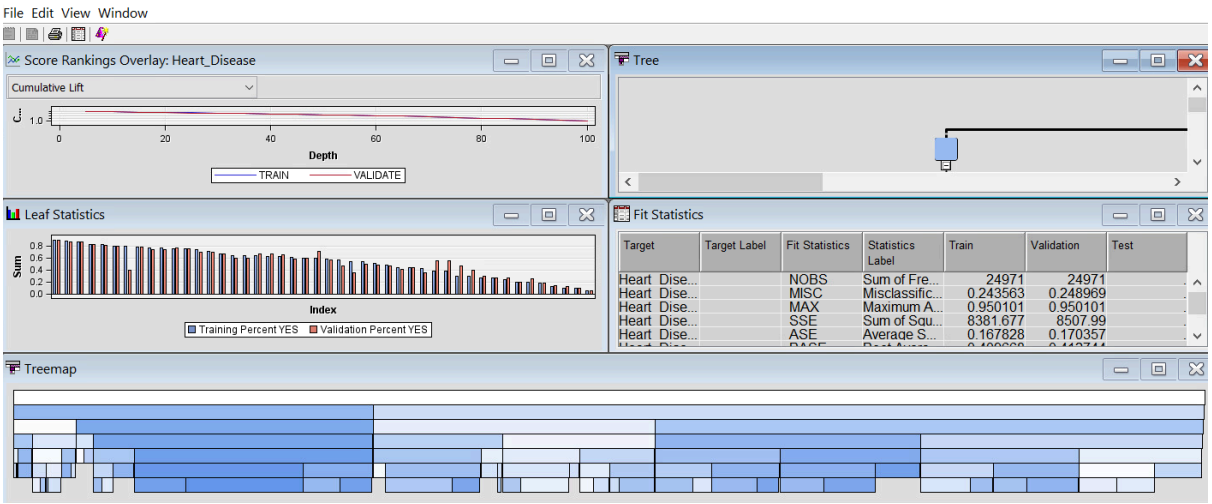


Figure 5.1.2 - 2-Branch Split Maximal Tree: Results

From the results, it can be observed that this 2-branch maximal tree had a validation **ASE of 0.170357.**

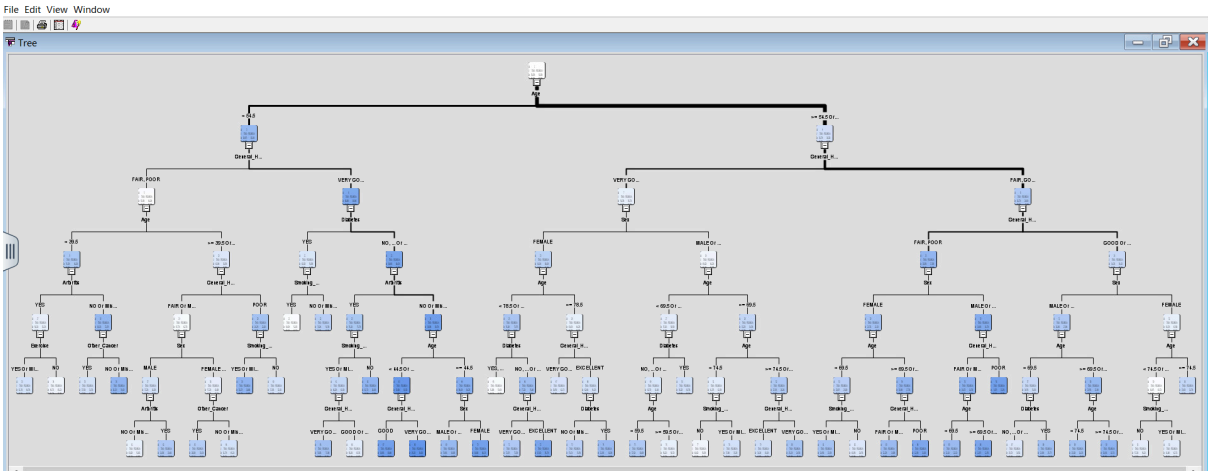


Figure 5.1.3 - 2-Branch Split Maximal Tree

5.2. 3-Branch Split Maximal Tree

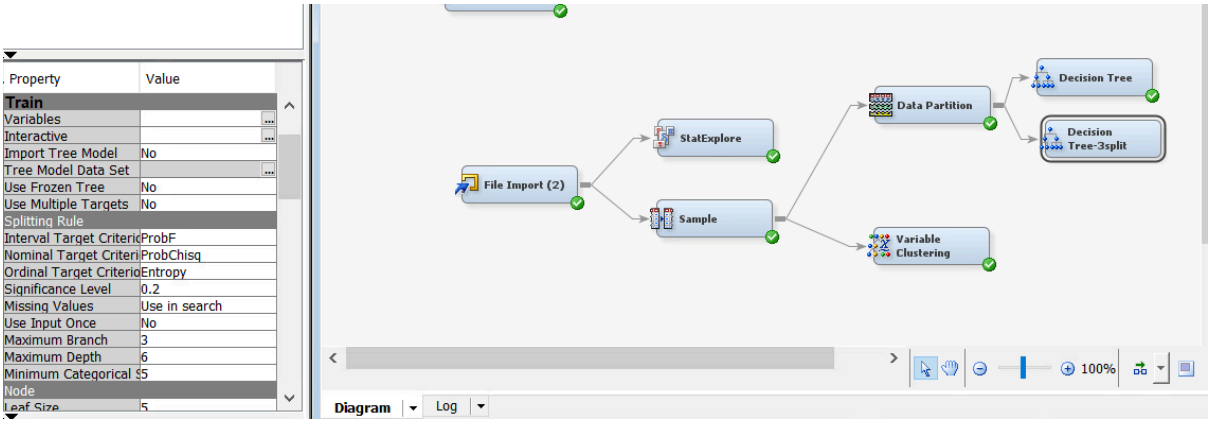


Figure 5.2.1 - 3-Branch Split Maximal Tree: The “Decision Tree” node was connected to the “Data Partition” node.

We brought in another “Decision Tree” node from the Model tab and connected it to the Data partition node. We changed the number of maximum branches from 2 to 3 and ran it.

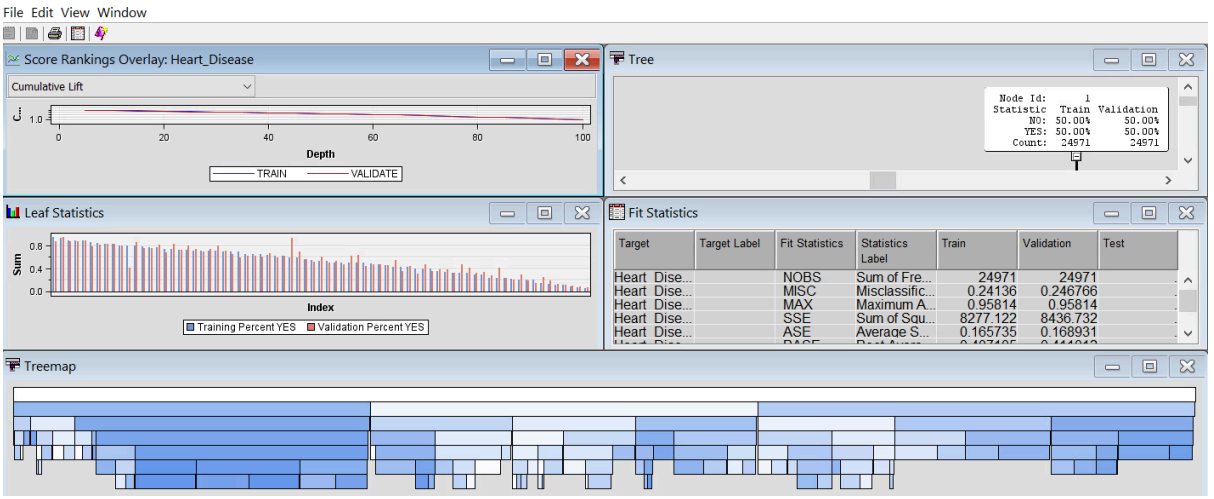


Figure 5.2.2 - 3-Branch Split Maximal Tree: Results

From the results, it can be observed that this 3-branch maximal tree had a validation **ASE of 0.168931.**

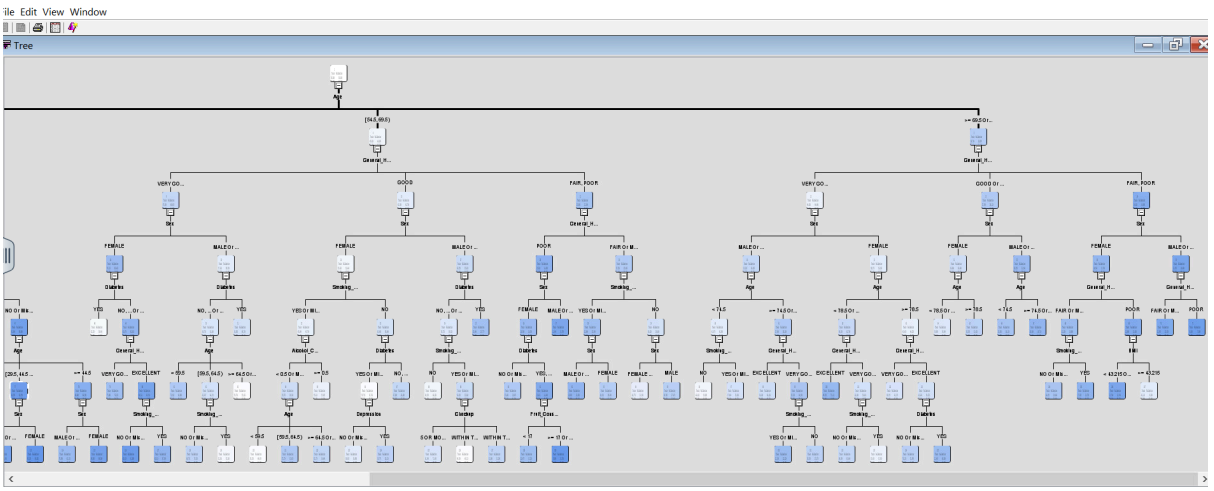


Figure 5.2.3 - 3-Branch Split Maximal Tree

5.3. 4-Branch Split Maximal Tree

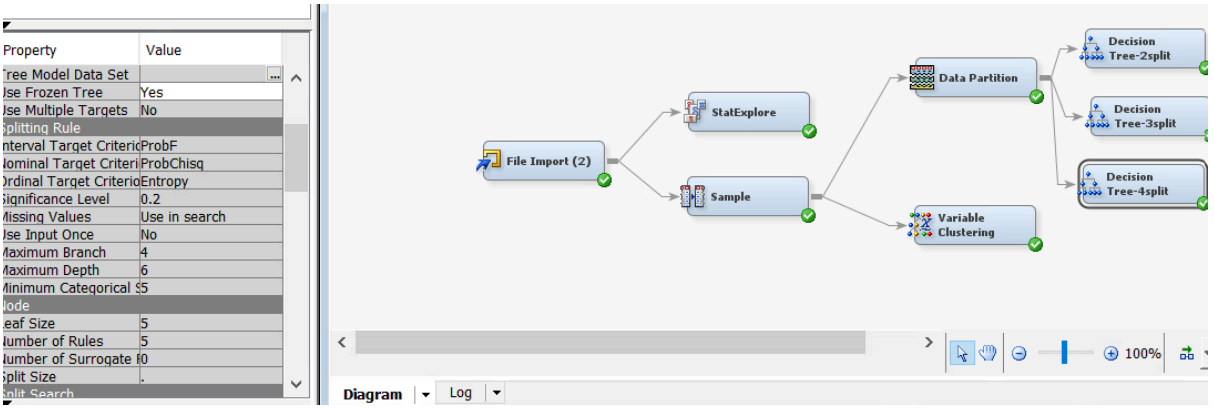


Figure 5.3.1 - 4-Branch Split Maximal Tree: The "Decision Tree" node was connected to the "Data Partition" node.

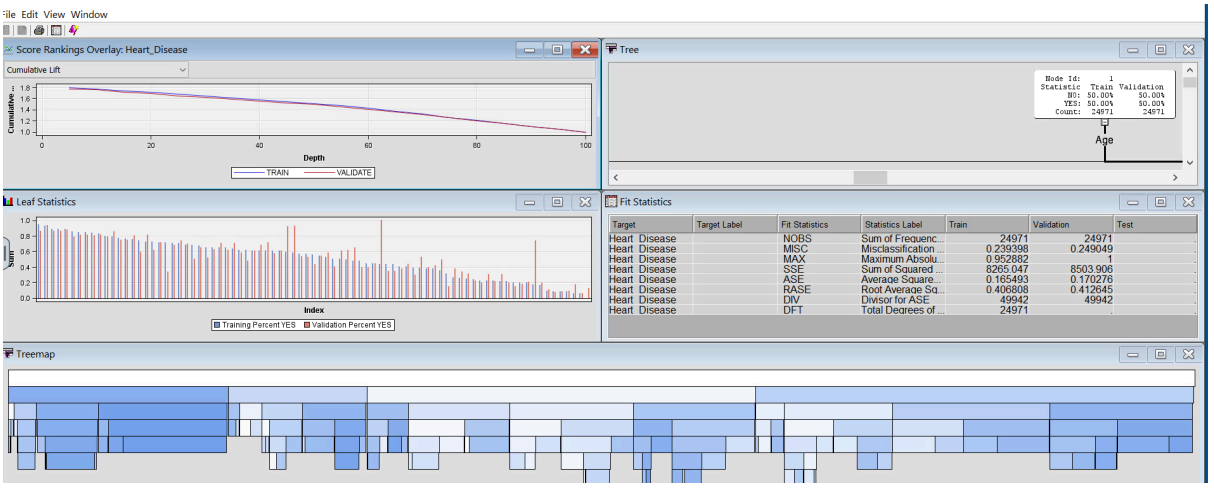


Figure 5.3.2 - 4-Branch Split Maximal Tree: Results

From the results, it can be observed that this 4-branch maximal tree had a validation **ASE of 0.170276.**

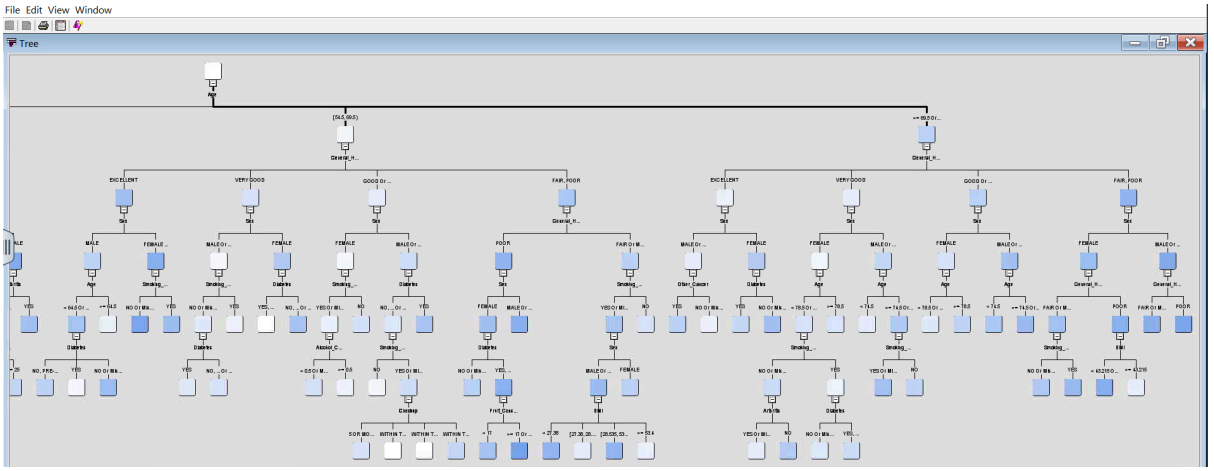
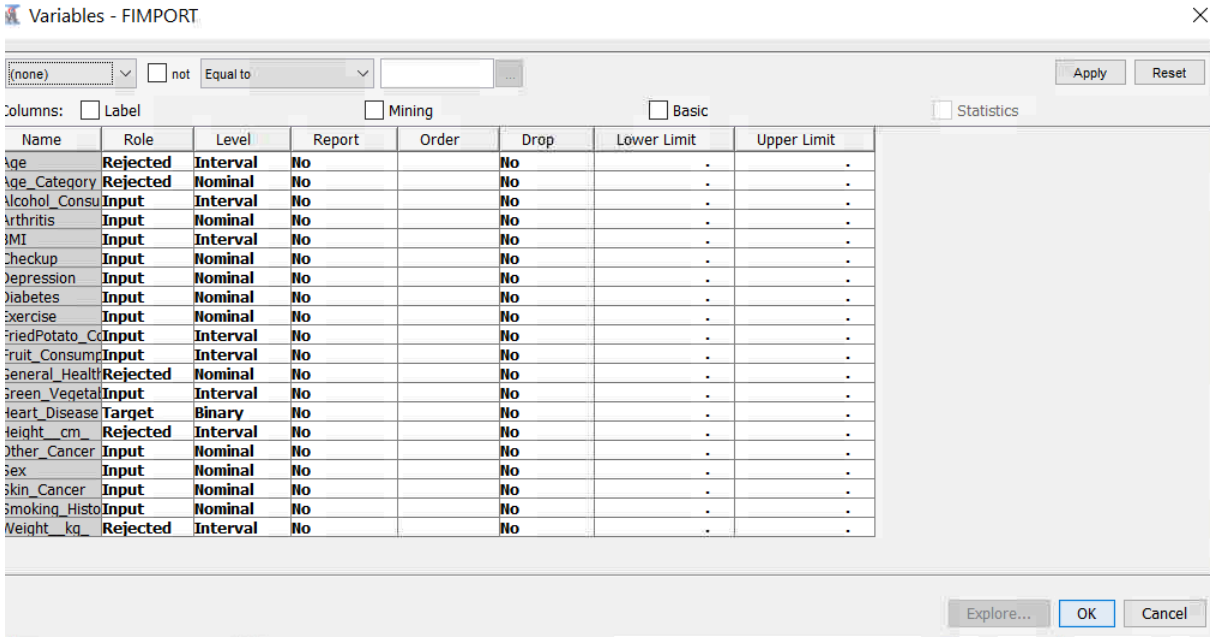


Figure 5.3.3 - 4-Branch Split Maximal Tree

5.4. Tackling the obvious variables

After this, we realized that the “Age” & “General Health” categories were obvious variables at the top of the tree. Growing Age and poor general health are apparent factors in deteriorating health and can eventually be top reasons for heart diseases.

However, because our study wanted to explore more unusual variables that contributed to heart diseases, we rejected the “Age” & “General Health” variables and re-ran the whole process flow.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Rejected	Interval	No		No	.	.
Age_Category	Rejected	Nominal	No		No	.	.
Alcohol_Consu	Input	Interval	No		No	.	.
Arthritis	Input	Nominal	No		No	.	.
BMI	Input	Interval	No		No	.	.
Checkup	Input	Nominal	No		No	.	.
Depression	Input	Nominal	No		No	.	.
Diabetes	Input	Nominal	No		No	.	.
Exercise	Input	Nominal	No		No	.	.
FriedPotato_Co	Input	Interval	No		No	.	.
Fruit_Consump	Input	Interval	No		No	.	.
General_Health	Rejected	Nominal	No		No	.	.
Green_Vegetal	Input	Interval	No		No	.	.
Heart_Disease	Target	Binary	No		No	.	.
Height__cm_	Rejected	Interval	No		No	.	.
Other_Cancer	Input	Nominal	No		No	.	.
Sex	Input	Nominal	No		No	.	.
Skin_Cancer	Input	Nominal	No		No	.	.
Smoking_Histo	Input	Nominal	No		No	.	.
Weight__kg_	Rejected	Interval	No		No	.	.

Figure 5.4.1 - Edit Variables window after rejecting Age and General Health variable.

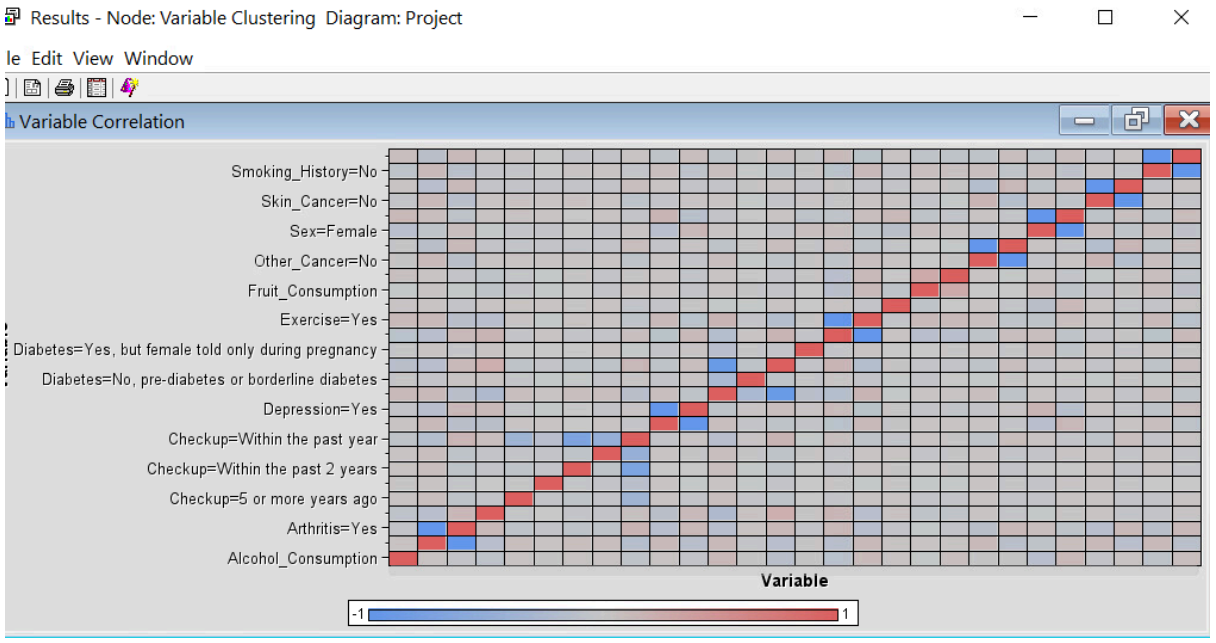


Figure 5.4.2 - Correlation matrix after rejecting Age and General Health variable.

5.5 Maximal Tree: Decision Tree Re-Ran

We reran three decision trees: maximal, misclassification, and ASE decision trees. We removed the initial three maximal decision trees, brought in another "Decision Tree" node from the Model tab, and connected it to the Data partition node. This time, we ran one maximal, one misclassification, and one ASE decision tree.

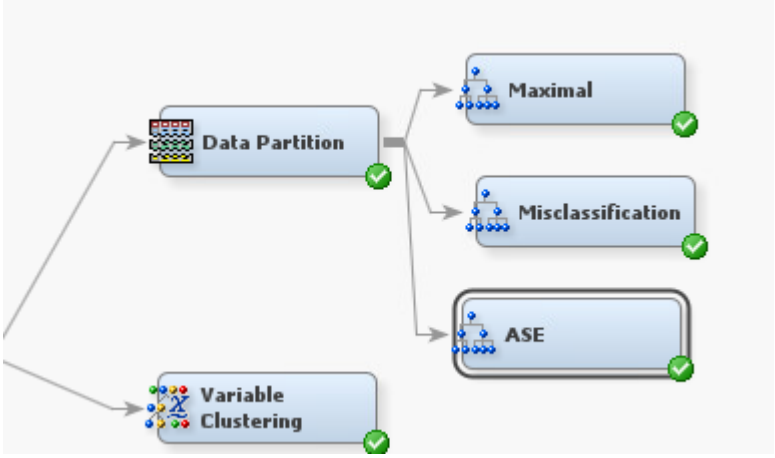


Figure 5.5.1 - Maximal Tree, Misclassification, and ASE: The "Decision Tree" nodes were connected to the "Data Partition" node.

Property	Value
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validati	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Imp	
Observation Based Imp	No
Number Single Var Im	5
P-Value Adjustment	
Bonferroni Adjustment	Yes

Figure 5.5.2 - Maximal Tree: Property Panel

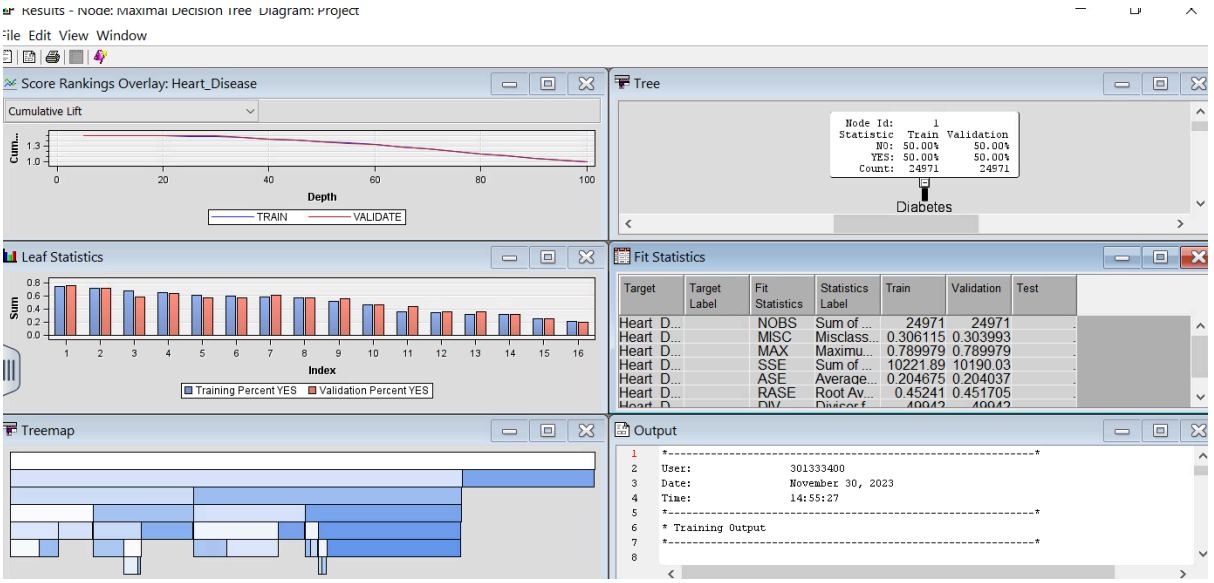


Figure 5.5.3 - Maximal Tree: Results

From the results, it can be observed that this maximal tree had a **validation ASE of 0.204037.**

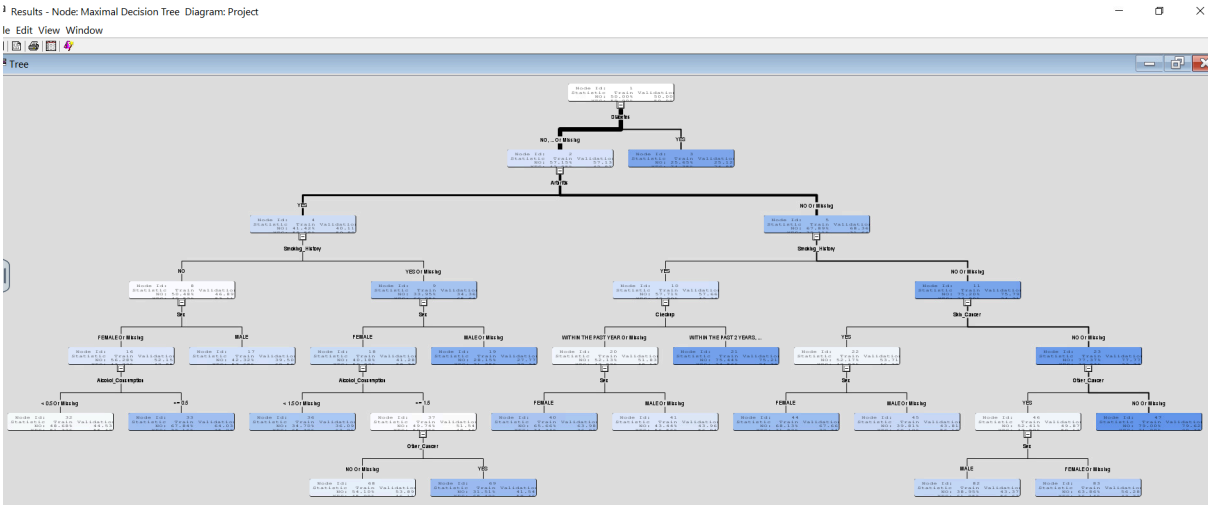


Figure 5.5.4 - Maximal Tree

5.6. Misclassification: Decision Tree

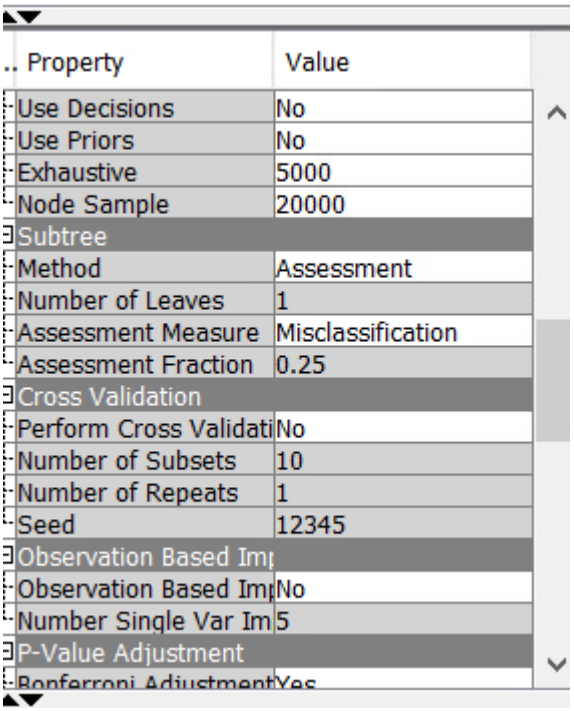


Figure 5.6.1 - Misclassification Tree: Property Panel

To achieve the misclassification tree, the assessment measure in the property panel was changed from decision to misclassification. We ran the decision tree.

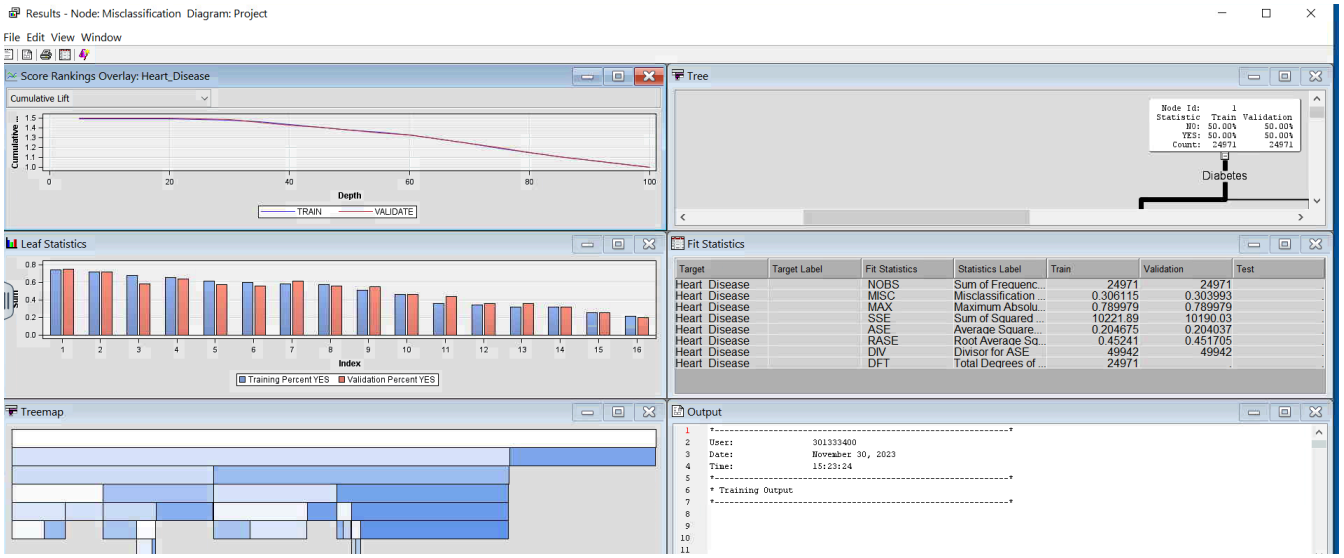


Figure 5.6.2- Misclassification Tree: Results

From the results, it can be observed that this maximal tree had a **validation ASE of 0.204037.**

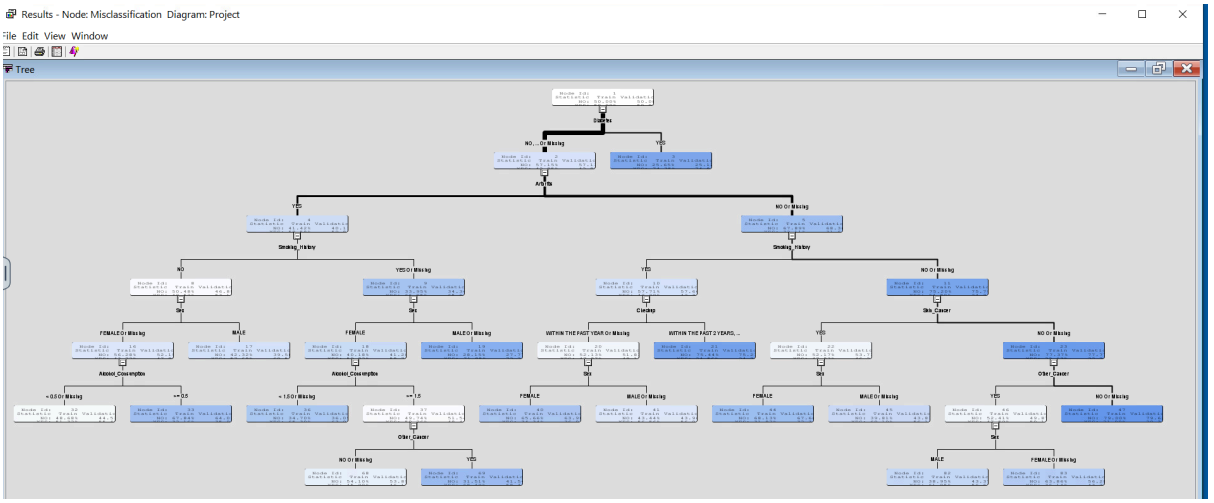


Figure 5.6.3 - MisclassificationTree

5.7. ASE: Decision Tree

Property	Value
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validati	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Im	
Observation Based Im	No
Number Singe Var Im	5

Figure 5.7.1 - ASE Tree: Property Panel

To achieve the ASE tree, the assessment measure in the property panel was changed from the decision to the Average Squared Error. We ran the decision tree.

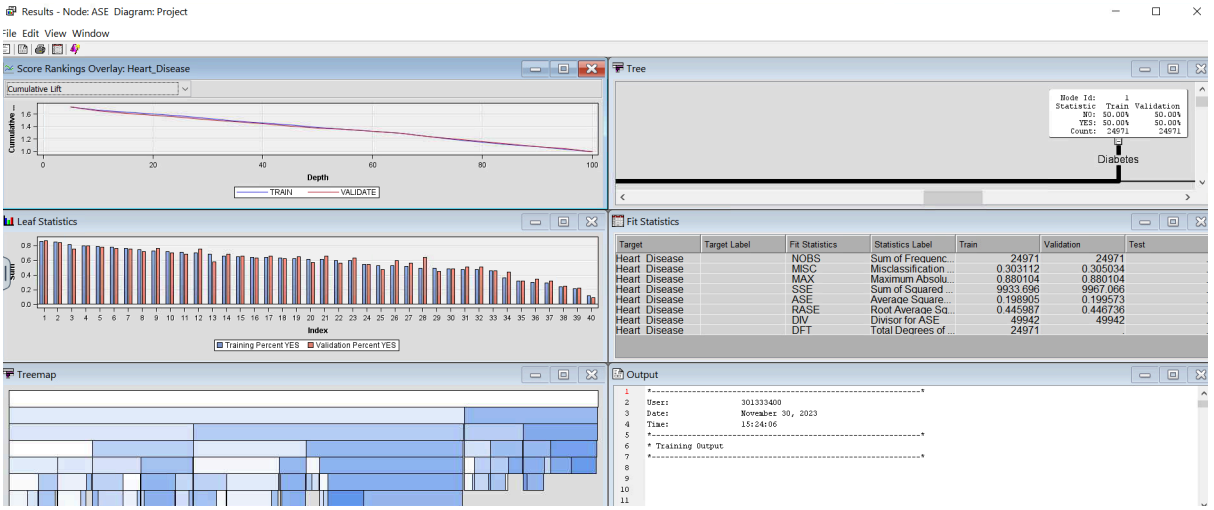


Figure 5.7.2 - ASE Tree: Results

From the results, it can be observed that this maximal tree had a **validation ASE of 0.199573.**

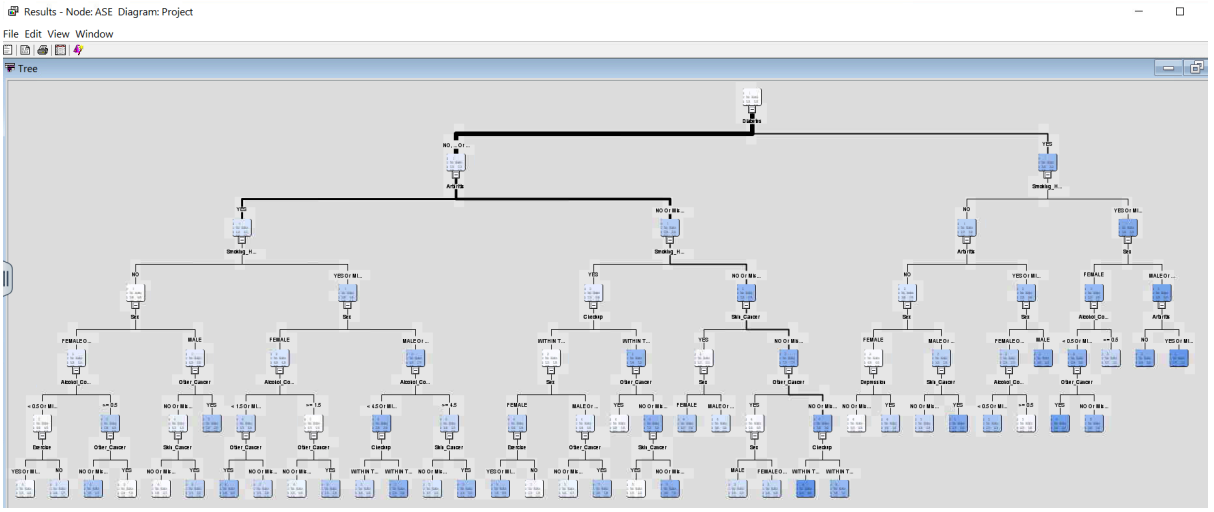


Figure 5.7.3 - ASE Tree

5.8. Decision Tree: Summary

A total of 3 decision trees were made. As the validation ASE for the **ASE: Decision Tree** was the lowest, we deduced that this tree was the most superior.

Decision Tree	ASE	# of leaves	First splitting rule: variable
Maximal Tree	**0.204037**	**15**	**Diabetes**
Misclassification Tree	**0.204037**	**15**	**Diabetes**
ASE Tree	**0.199573**	**40**	**Diabetes**

Table 5.8.1 - Summary Table of Decision Trees

6. Logistic Regression

6.1. Data Massaging

- Some BMI values are outliers, so they need to be filtered.
- As per initial expiration, we found alcohol consumption fried potato had outliers and high skewness with replacement and transformation.

6.1.1. Data Replacement

A data replacement node was added to the data partition node. The data replacement node was used to set the upper limit of the BMI index.

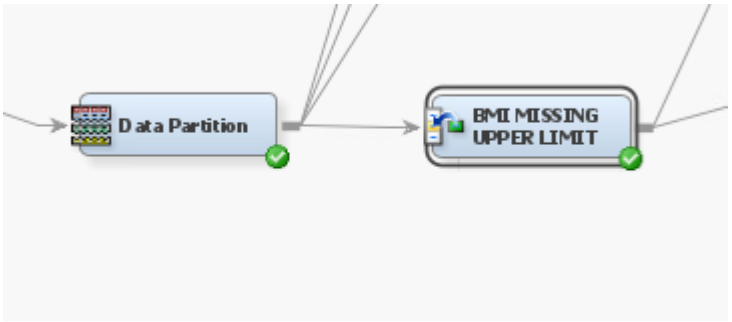


Figure 6.1.1- adding replacement node

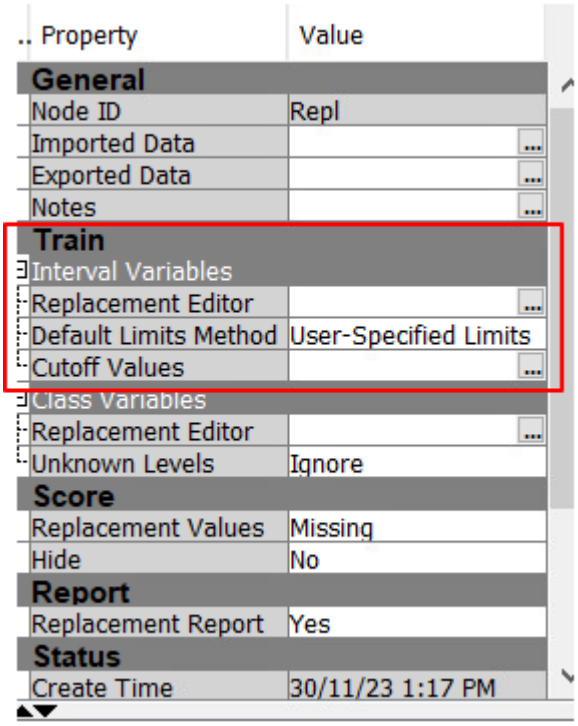


Figure 6.1.2- Property panel of Replacement node

We want to change the default limits method to user-specified limits to manually replace data of BMI that is above 45.

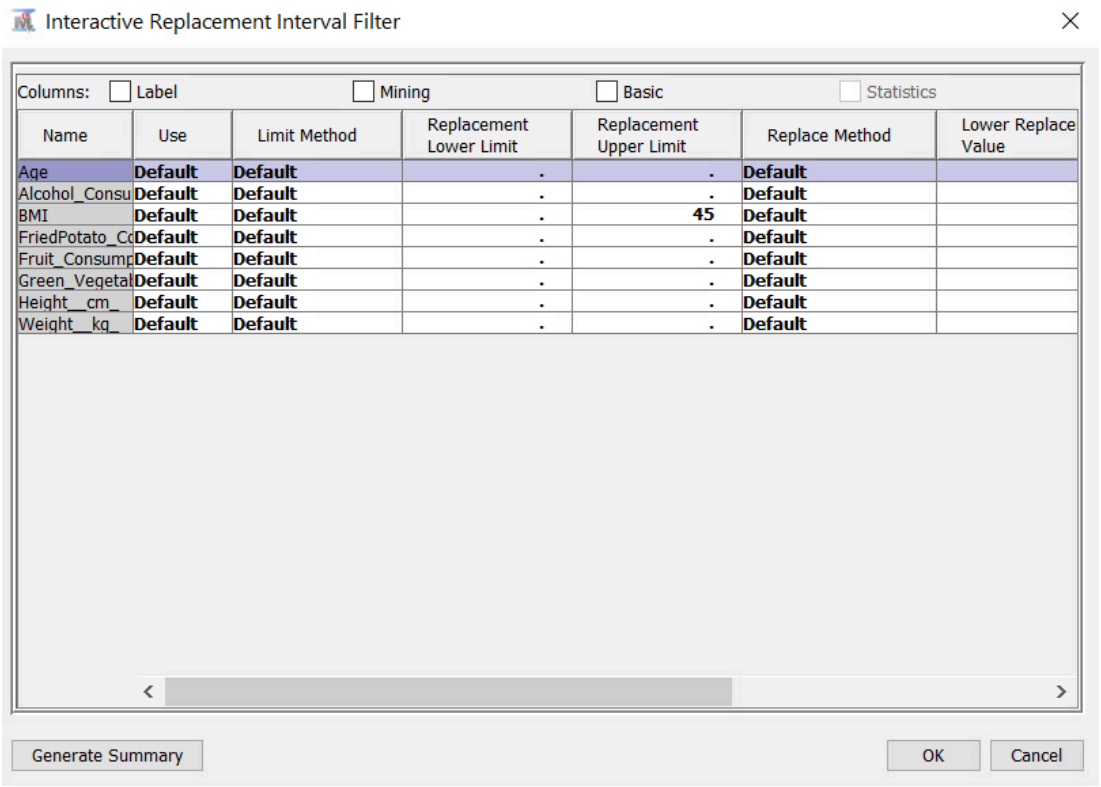


Figure 6.1.3- Replacement editor for interval variables.

The upper limit of BMI was set to 45 to filter out outliers.

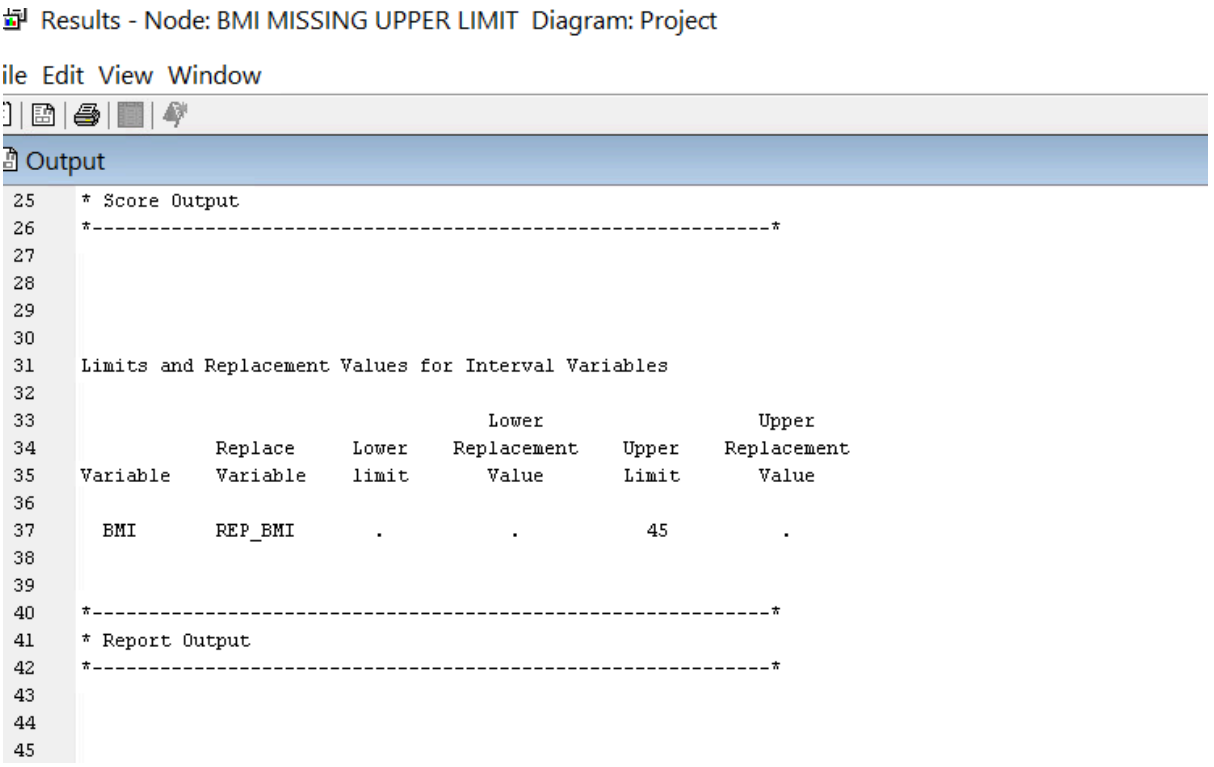


Figure 6.1.4- Results for data replacement

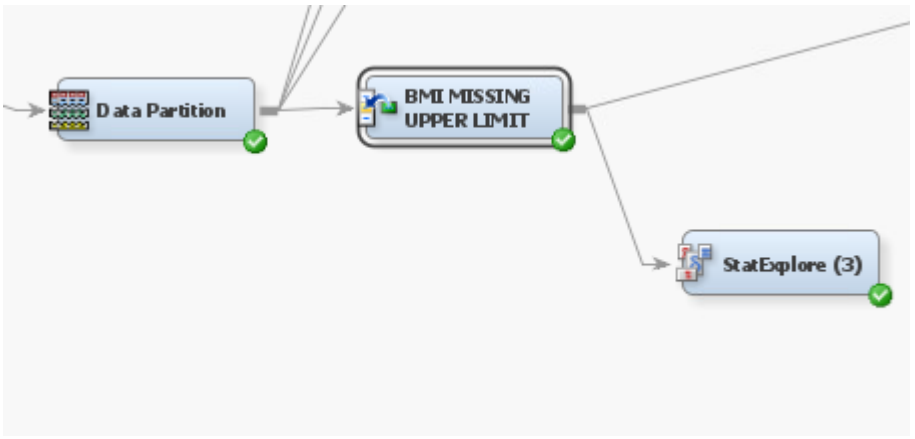


Figure 6.1.5 - Addition of "StatExplore" to the "Replacement" node

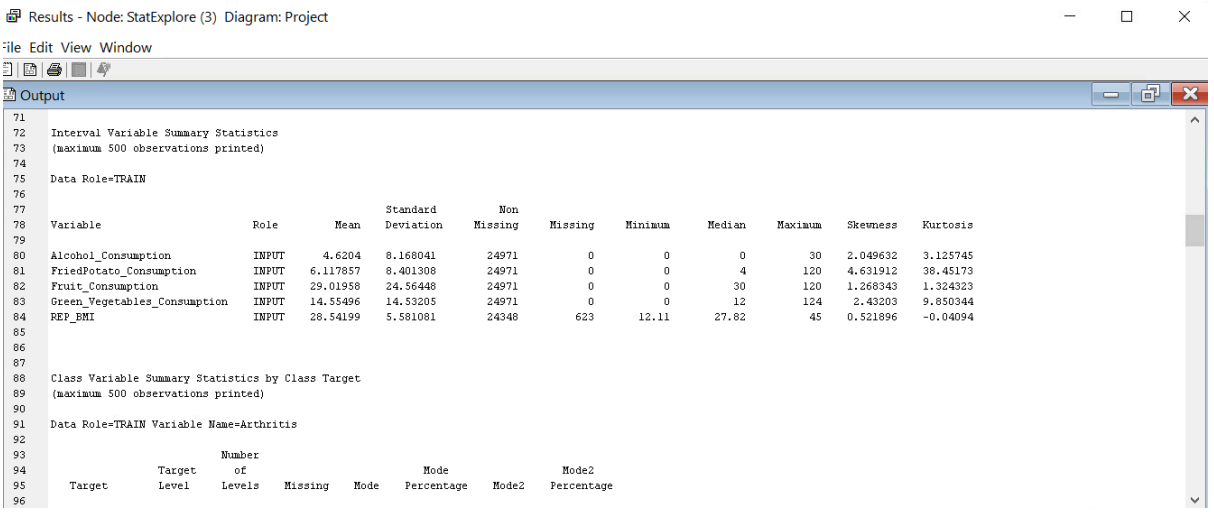


Figure 6.1.6 - "StatExplore": Results

The replaced BMI has much-reduced skewness, but the skewness didn't change much with other variables, so we added a cap and floor and checked for skewness again later.

6.1.2. Data Imputation

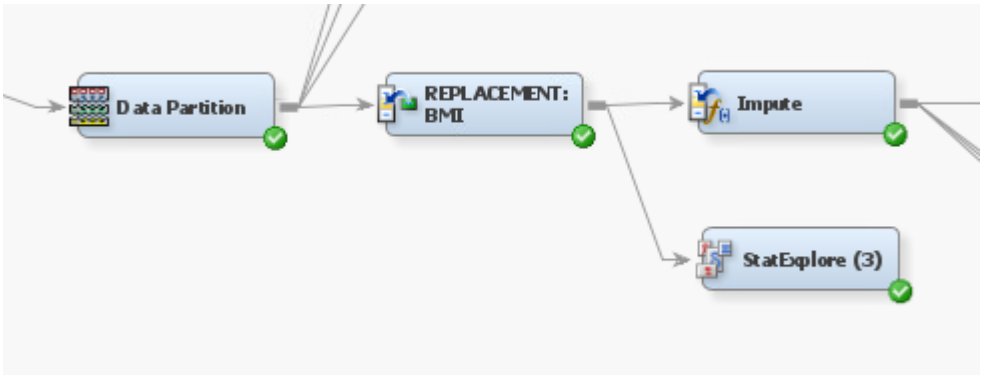


Figure 6.1.6 Addition of impute node to the replacement node

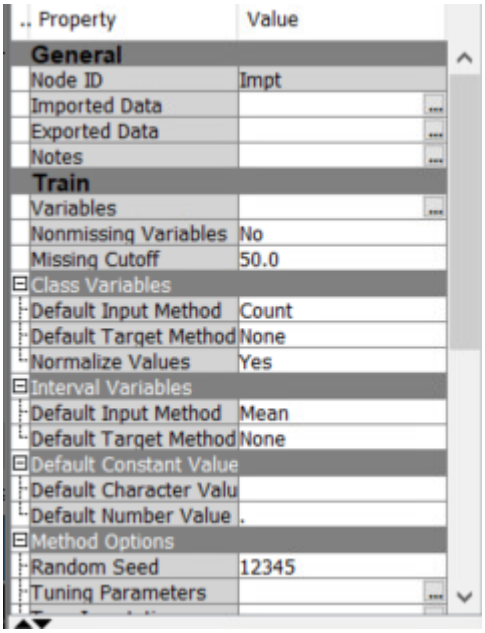


Figure 6.1.7 The Property panel of the Impute node

6.1.2 Cap and Floor

A replacement node was added to the impute node and renamed " Cap and floor." It is added to reduce the outliers.

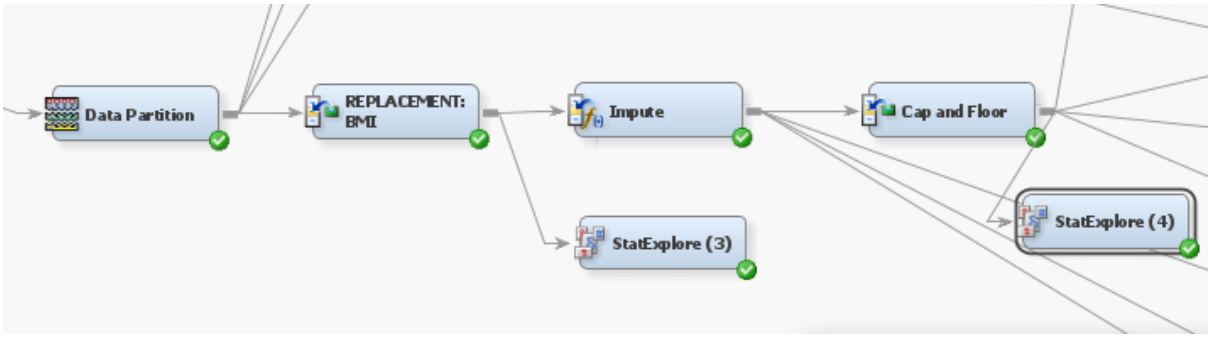


Figure 6.1.8 Cap and floor were added to the impute node.

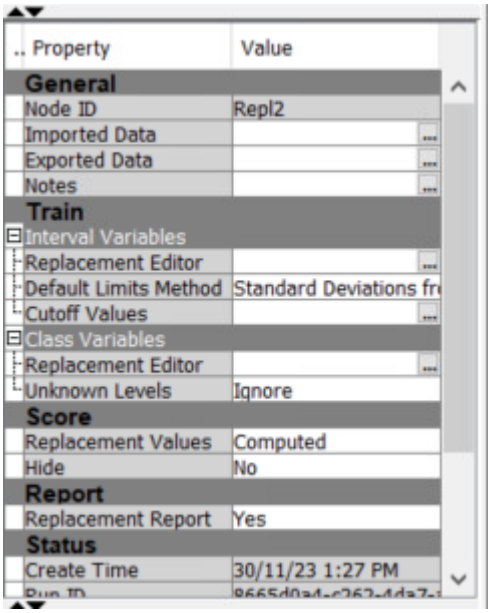
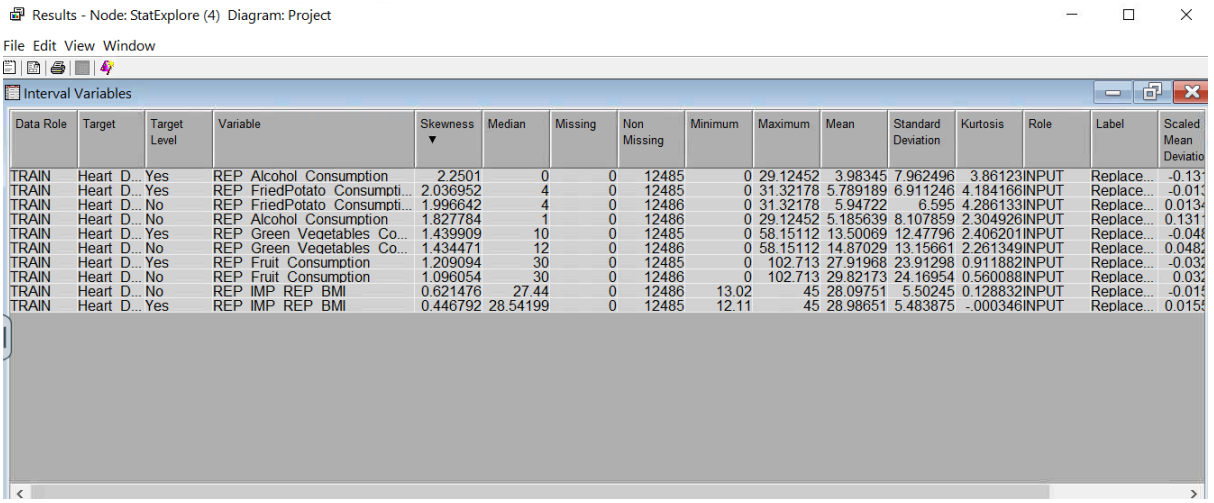


Figure 6.1.9 Property panel of cap and floor



Data Role	Target	Target Level	Variable	Skewness	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Kurtosis	Role	Label	Scaled Mean Deviatio
TRAIN	Heart_D...	Yes	REP_Alcohol_Consumption	2.2501	0	0	12485	0	29.12452	3.98345	7.962496	3.86123	INPUT	Replace...	-0.13
TRAIN	Heart_D...	Yes	REP_FriedPotato_Consumpti...	2.036952	4	0	12485	0	31.32178	5.789189	6.911246	4.184166	INPUT	Replace...	-0.01
TRAIN	Heart_D...	No	REP_FriedPotato_Consumpti...	1.996642	4	0	12486	0	31.32178	5.94722	6.595	4.286133	INPUT	Replace...	0.013
TRAIN	Heart_D...	No	REP_Alcohol_Consumption	1.827784	1	0	12486	0	29.12452	5.185639	8.107859	2.304926	INPUT	Replace...	0.131
TRAIN	Heart_D...	Yes	REP_Green_Vegetables_Co...	1.439909	10	0	12485	0	58.15112	13.50069	12.47796	2.406201	INPUT	Replace...	-0.04
TRAIN	Heart_D...	No	REP_Green_Vegetables_Co...	1.434471	12	0	12486	0	58.15112	14.87029	13.15661	2.261349	INPUT	Replace...	0.048
TRAIN	Heart_D...	Yes	REP_Fruit_Consumption	1.209094	30	0	12485	0	102.713	27.91968	23.91298	0.911882	INPUT	Replace...	-0.03
TRAIN	Heart_D...	No	REP_Fruit_Consumption	1.096054	30	0	12486	0	102.713	29.82173	24.16954	0.560088	INPUT	Replace...	0.03
TRAIN	Heart_D...	No	REP_IMP_REP_BMI	0.621476	27.44	0	12486	13.02	45	28.09751	5.50245	0.128832	INPUT	Replace...	-0.01
TRAIN	Heart_D...	Yes	REP_IMP_REP_BMI	0.446792	28.54199	0	12485	12.11	45	28.98651	5.483875	-.000346	INPUT	Replace...	0.015

Figure 6.1.10 Stat explorer results of cap and floor

We can see that after the cap and floor, the skewness has reduced drastically, except for two variables: alcohol consumption and fried potato consumption.

6.1.4 Data Transformation

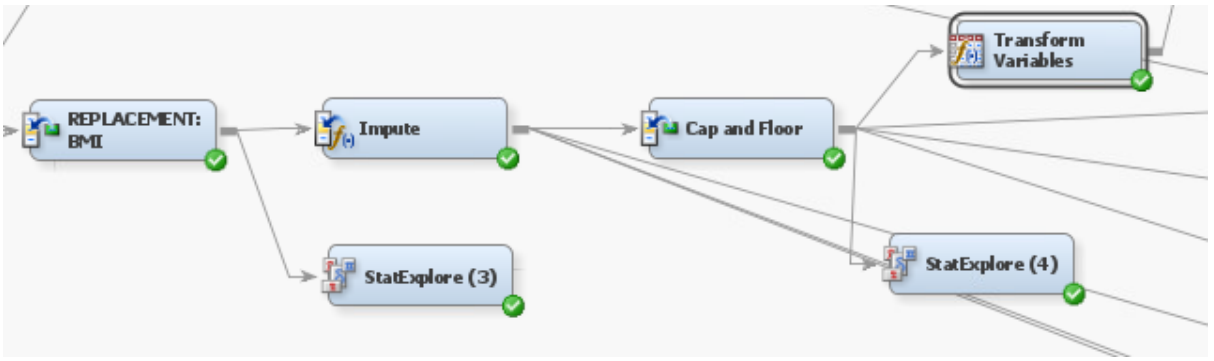
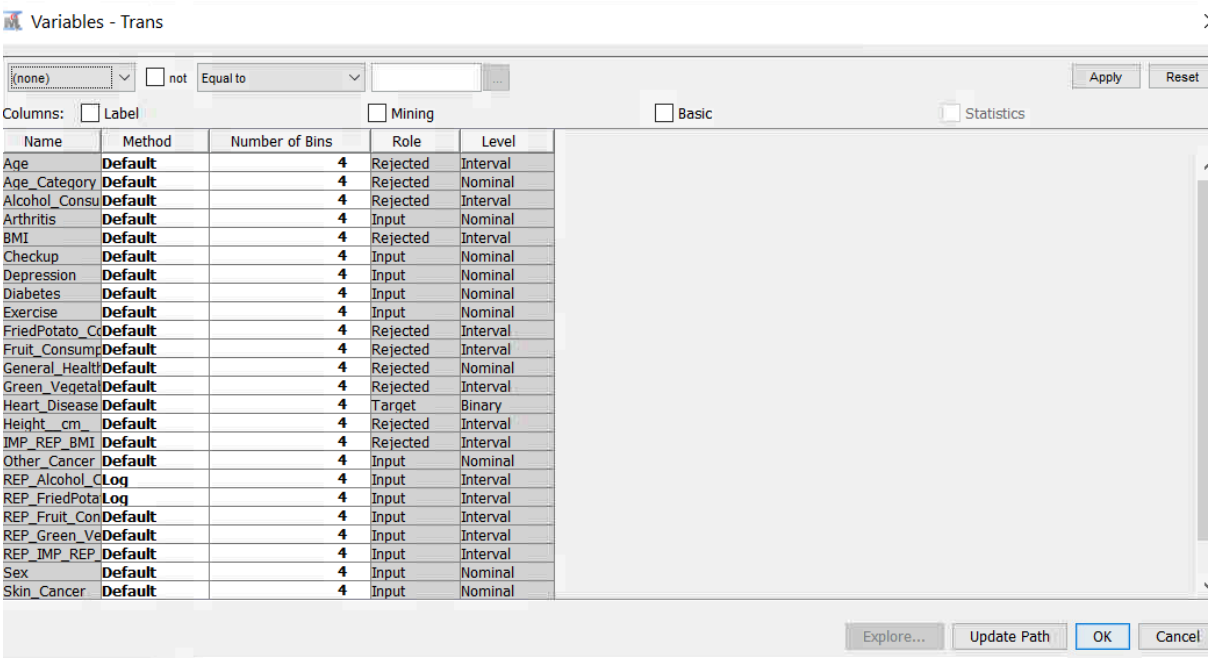


Figure 6.1.11 Addition of transformation node to cap and floor node



Name	Method	Number of Bins	Role	Level
Age	Default	4	Rejected	Interval
Age_Category	Default	4	Rejected	Nominal
Alcohol_Consu	Default	4	Rejected	Interval
Arthritis	Default	4	Input	Nominal
BMI	Default	4	Rejected	Interval
Checkup	Default	4	Input	Nominal
Depression	Default	4	Input	Nominal
Diabetes	Default	4	Input	Nominal
Exercise	Default	4	Input	Nominal
FriedPotato_Cc	Default	4	Rejected	Interval
Fruit_Consump	Default	4	Rejected	Interval
General_Health	Default	4	Rejected	Nominal
Green_Vegetal	Default	4	Rejected	Interval
Heart_Disease	Default	4	Target	Binary
Height__cm_	Default	4	Rejected	Interval
IMP_REP_BMI	Default	4	Rejected	Interval
Other_Cancer	Default	4	Input	Nominal
REP_Alcohol_C	Log	4	Input	Interval
REP_FriedPota	Log	4	Input	Interval
REP_Fruit_Con	Default	4	Input	Interval
REP_Green_Ve	Default	4	Input	Interval
REP_IMP_REP_	Default	4	Input	Interval
Sex	Default	4	Input	Nominal
Skin_Cancer	Default	4	Input	Nominal

Figure 6.1.12 Transformation variables

The two variables with skewness - alcohol consumption and fried potato consumption done a log transformation

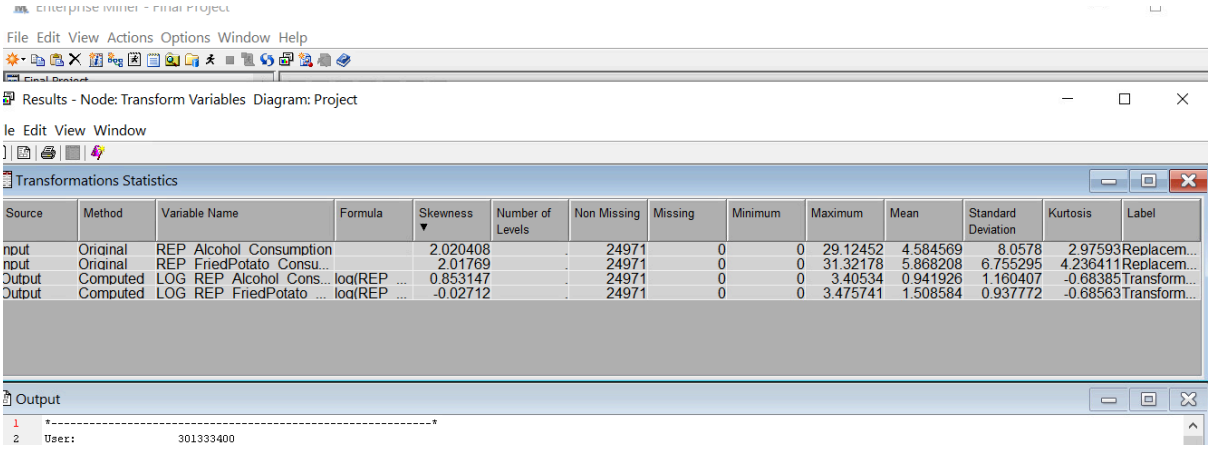


Figure 6.1.12 Results of transformation

6.1.5 Recode dummies

We use recode dummies to combine the categorical values into one variable and reduce the curse of dimension.

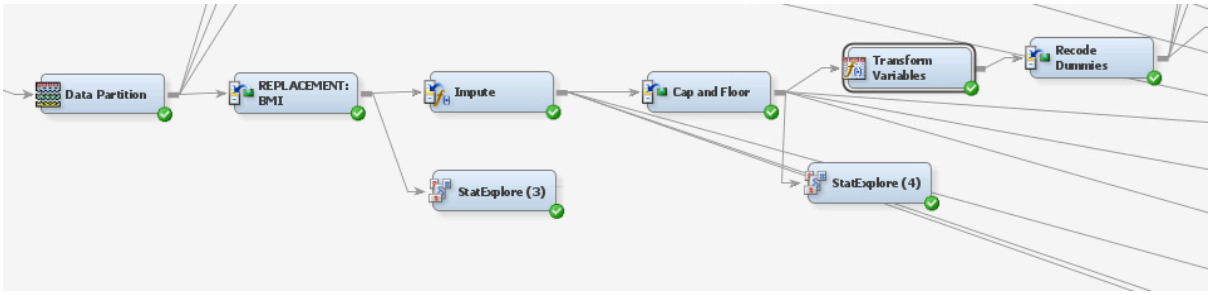


Figure 6.1.12 Addition of recode dummies

Property	Value
General	
Node ID	Repl3
Imported Data	
Exported Data	
Notes	
Train	
Interval Variables	
Replacement Editor	
Default Limits Method	Standard Deviations fror
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	04/12/23 4:38 PM
Run ID	eea9644c-4fc7-441e-b3(
Last Error	
Last Status	Complete
Last Run Time	04/12/23 4:43 PM
Run Duration	0 Hr. 0 Min. 20.01 Sec.
Grid Host	
User-Added Node	No

Figure 6.1.13 Property panel of recode dummies

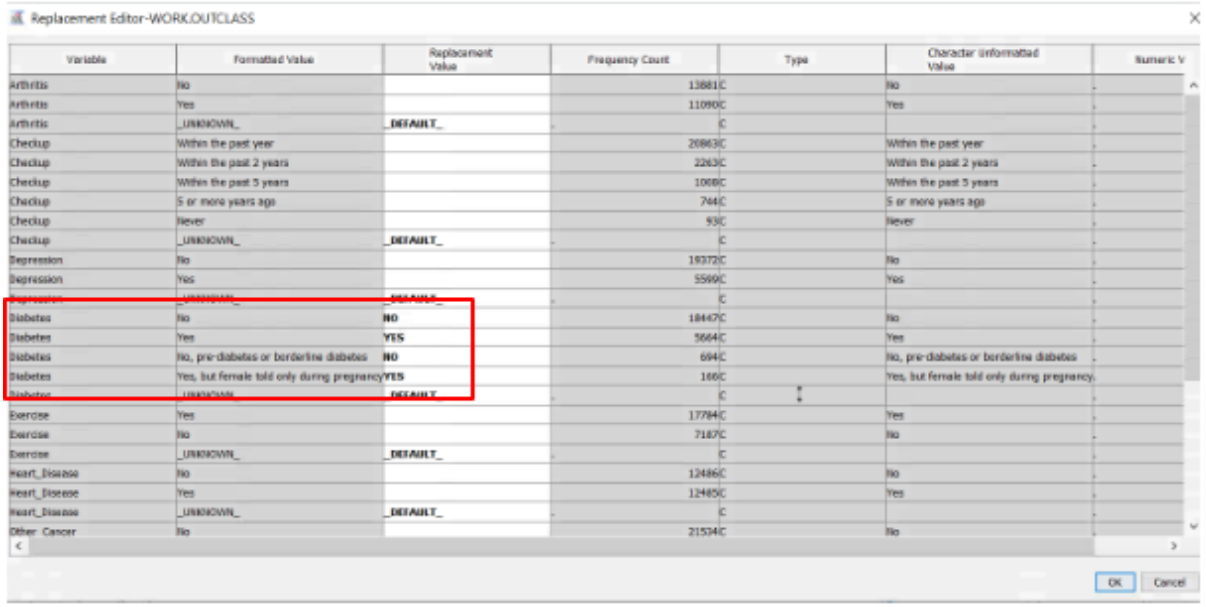


Figure 6.1.14 Replacement editor of recode dummies

we reduced four diabetes variables into two variables

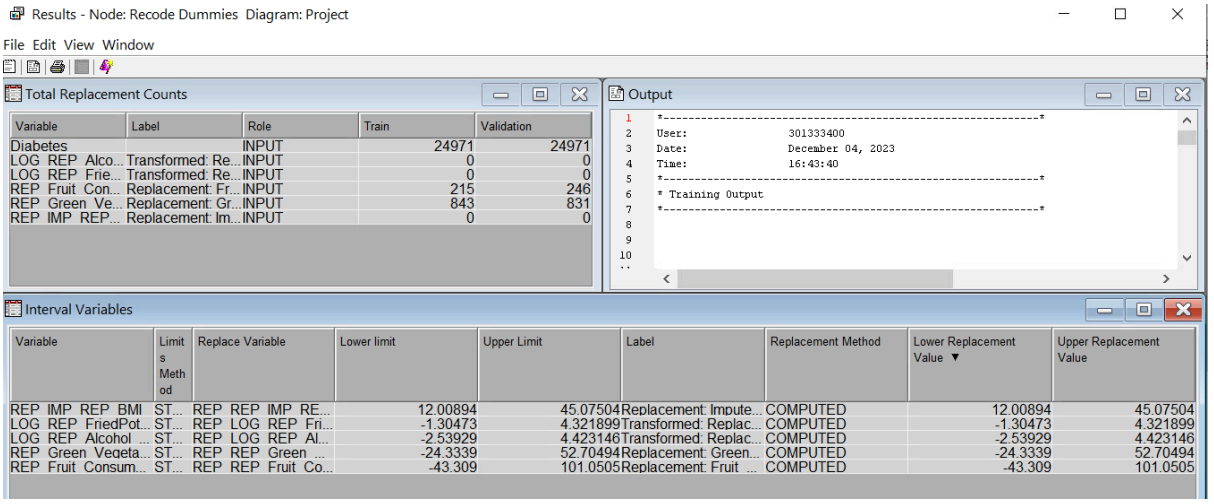


Figure 6.1.14 Results window of recode dummies

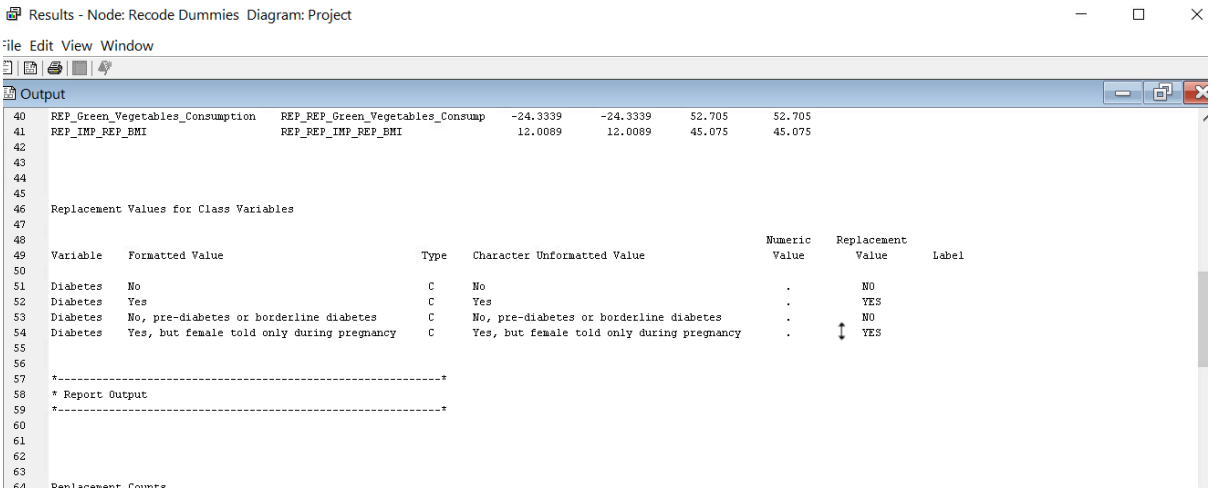


Figure 6.1.15 Output of recode dummies

We can see the categorical variables " diabetes" are grouped easily.

6.2. Full Regression

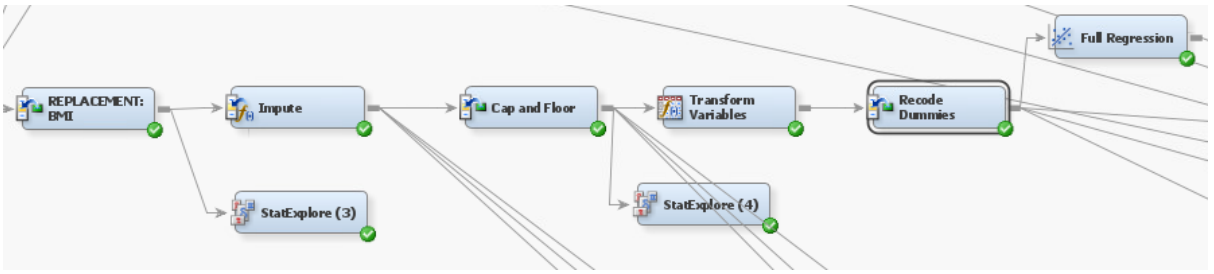
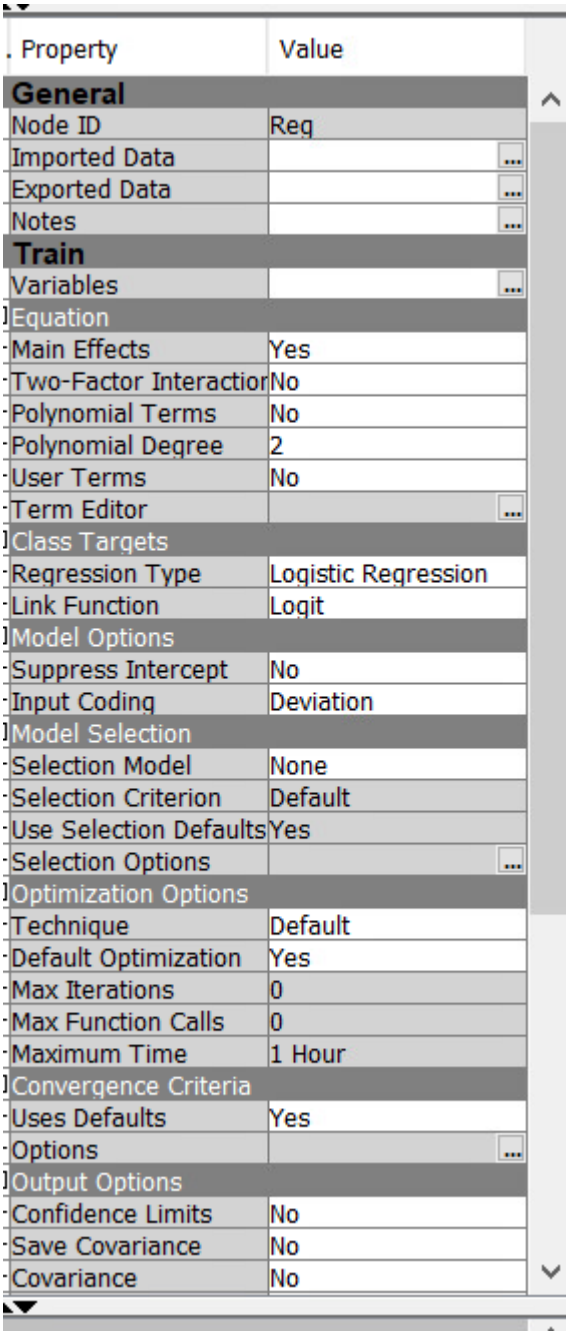
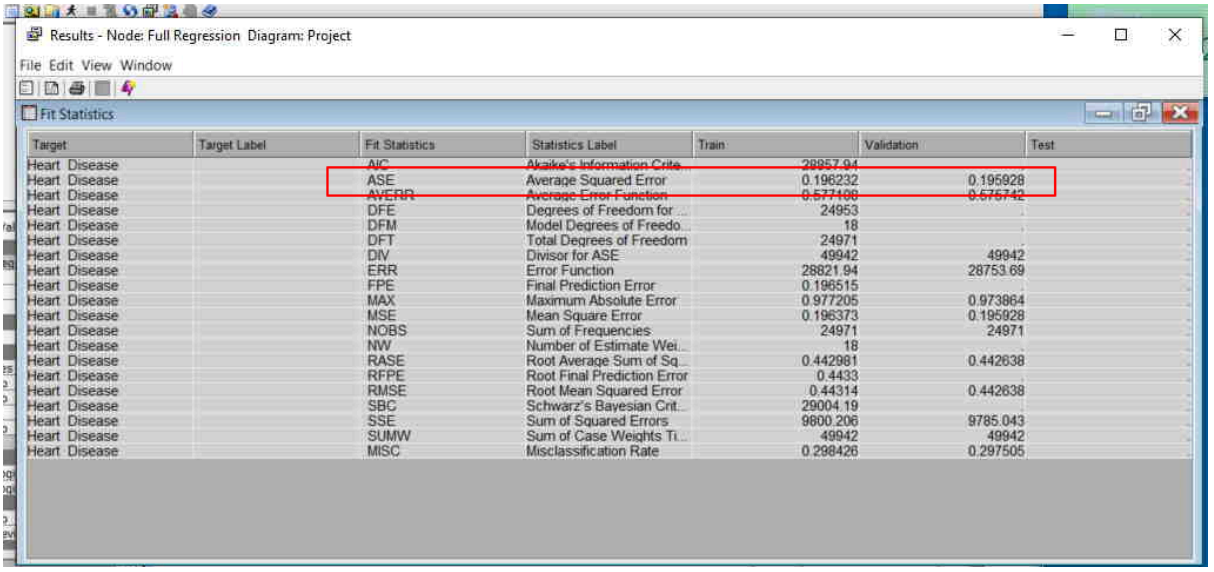


Figure 6.2.1 Full regression was added to recode dummies.



Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No

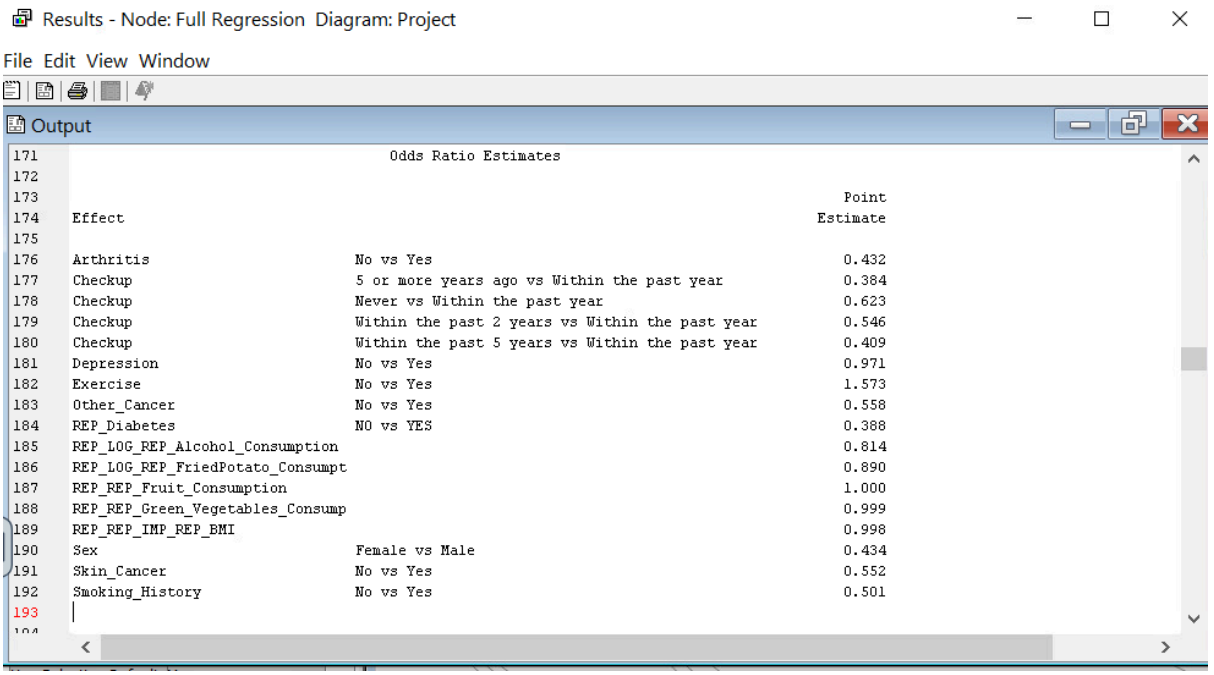
Figure 6.2.2 Property Panel of Full Regression Node



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Heart Disease		AIC	Akaike's Information Crite...	28857.94		
Heart Disease		ASE	Average Squared Error	0.196232	0.195928	
Heart Disease		AVERR	Average Error Function	0.577108	0.576742	
Heart Disease		DFE	Degrees of Freedom for ...	24953		
Heart Disease		DFM	Model Degrees of Freedo...	18		
Heart Disease		DFT	Total Degrees of Freedom	24971		
Heart Disease		DIV	Divisor for ASE	49942	49942	
Heart Disease		ERR	Error Function	28821.94	28753.69	
Heart Disease		FPE	Final Prediction Error	0.196515		
Heart Disease		MAX	Maximum Absolute Error	0.977205	0.973864	
Heart Disease		MSE	Mean Square Error	0.196373	0.195928	
Heart Disease		NOBS	Sum of Frequencies	24971	24971	
Heart Disease		NW	Number of Estimate Wei...	18		
Heart Disease		RASE	Root Average Sum of Sq...	0.442981	0.442638	
Heart Disease		RFPE	Root Final Prediction Error	0.4433		
Heart Disease		RMSE	Root Mean Squared Error	0.44314	0.442638	
Heart Disease		SBC	Schwarz's Bayesian Crit...	29004.19		
Heart Disease		SSE	Sum of Squared Errors	9800.206	9785.043	
Heart Disease		SUMW	Sum of Case Weights Ti...	49942	49942	
Heart Disease		MISC	Misclassification Rate	0.298426	0.297505	

Figure 6.2.3 Results of full Regression node

The ASE of complete regression is 0.195928



Odds Ratio Estimates

Effect		Point Estimate
Arthritis	No vs Yes	0.432
Checkup	5 or more years ago vs Within the past year	0.384
Checkup	Never vs Within the past year	0.623
Checkup	Within the past 2 years vs Within the past year	0.546
Checkup	Within the past 5 years vs Within the past year	0.409
Depression	No vs Yes	0.971
Exercise	No vs Yes	1.573
Other_Cancer	No vs Yes	0.558
REP_Diabetes	NO vs YES	0.388
REP_LOG_REP_Alcohol_Consumption		0.814
REP_LOG_REP_FriedPotato_Consumpt		0.890
REP_REP_Fruit_Consumption		1.000
REP_REP_Green_Vegetables_Consump		0.999
REP_REP_IMP_REP_BMI		0.998
Sex	Female vs Male	0.434
Skin_Cancer	No vs Yes	0.552
Smoking_History	No vs Yes	0.501

Figure 6.2.4 Output of Full regression node

From the odds ratio, we found that people who don't exercise are 57% more likely to get heart disease.

6.3. Forward Regression

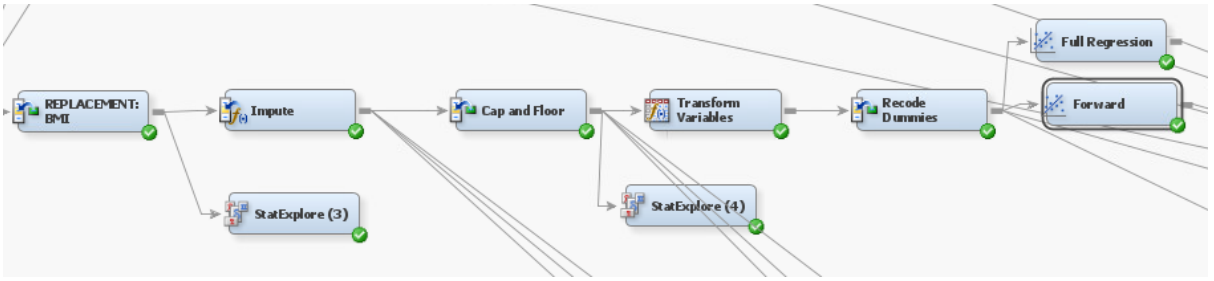


Figure 6.3.1 addition of forward regression

.. Property	Value
General	
Node ID	Reg2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0

Figure 6.3.2 Property panel of forward regression

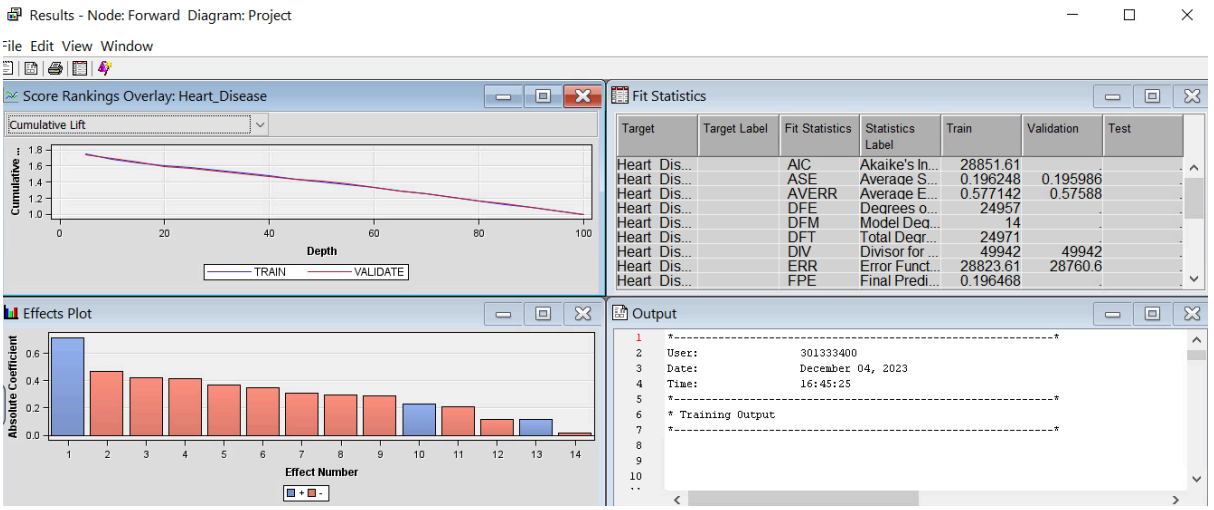


Figure 6.3.3 Results of Forward Regression

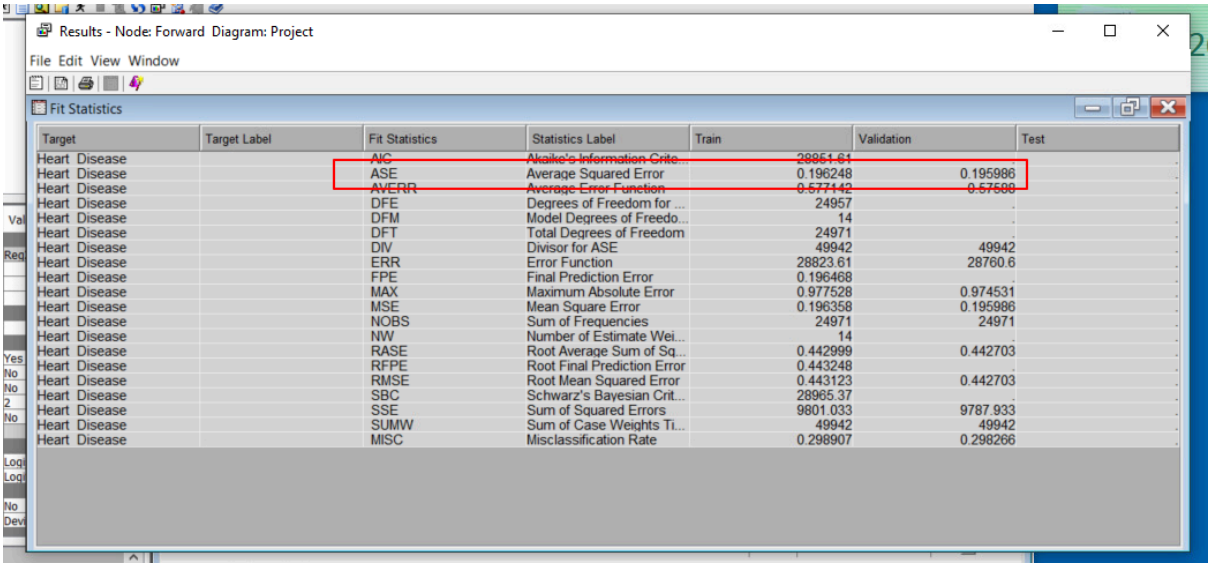


Figure 6..3.4 Fit statistics of forward regression

Validation ASE is 0.195986

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Arthritis	1	1	1657.3639	<.0001
2	REP_Diabetes	1	2	1313.9671	<.0001
3	Smoking_History	1	3	739.3747	<.0001
4	Sex	1	4	546.1231	<.0001
5	Checkup	4	5	444.7982	<.0001
6	REP_LOG_REP_Alcohol_Consumption	1	6	319.6800	<.0001
7	Other_Cancer	1	7	263.5192	<.0001
8	Skin_Cancer	1	8	192.4145	<.0001
9	Exercise	1	9	186.6458	<.0001
10	REP_LOG_REP_FriedPotato_Consumpt	1	10	56.4397	<.0001

ıe selected model is the model trained in the last step (Step 10). It consists of the following effects:

ıtercept Arthritis Checkup Exercise Other_Cancer REP_Diabetes REP_LOG_REP_Alcohol_Consumption REP_LOG_REP_FriedPotato_Consumpt Sex Skin_Cancer Smoking_Histo

Figure 6.3.5 Summary of Forward Regression

Significant variables are arthritis, REP diabetes, sex, checkup, skin cancer, REP alcohol consumption, and REP fried potato consumption.

Odds Ratio Estimates

Effect		Point Estimate
Arthritis	No vs Yes	0.432
Checkup	5 or more years ago vs Within the past year	0.385
Checkup	Never vs Within the past year	0.625
Checkup	Within the past 2 years vs Within the past year	0.547
Checkup	Within the past 5 years vs Within the past year	0.410
Exercise	No vs Yes	1.577
Other_Cancer	No vs Yes	0.557
REP_Diabetes	NO vs YES	0.390
REP_LOG_REP_Alcohol_Consumption		0.814
REP_LOG_REP_FriedPotato_Consumpt		0.890
Sex	Female vs Male	0.436
Skin_Cancer	No vs Yes	0.552
Smoking_History	No vs Yes	0.499

Figure 6.3.6 Odds ratio of forward regression

From the odds ratio, we found that people who don't exercise are 57% more likely to get heart disease.

6.4. Backward Regression

Backward regression is done to find if we can reduce the complexity of the analysis.

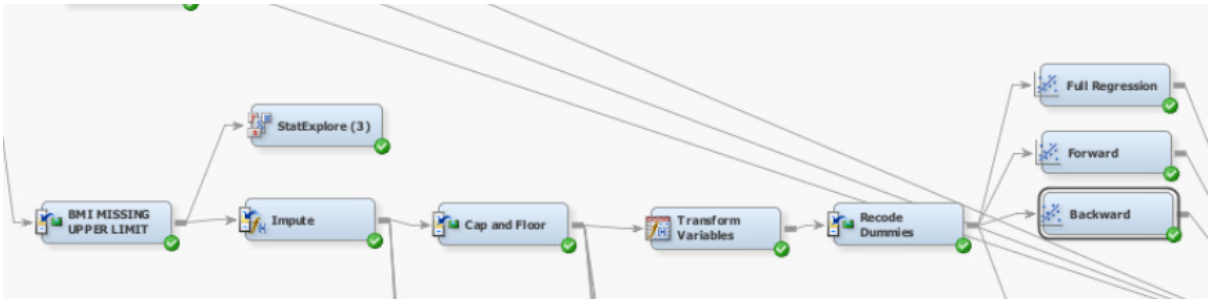


Figure 6.4.1 addition of forward regression

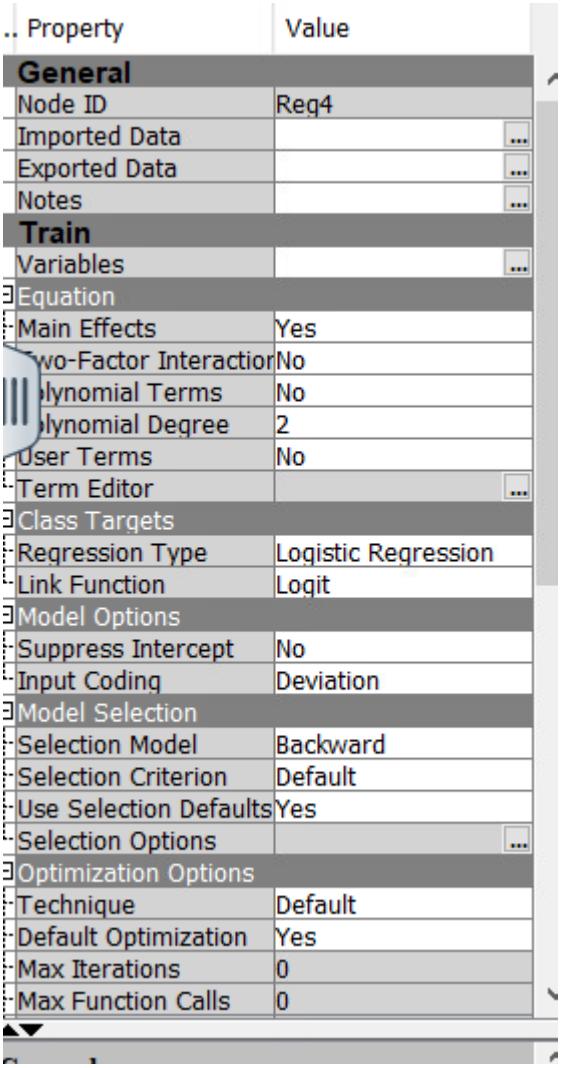


Figure 6.4.2 Property panel of forward regression

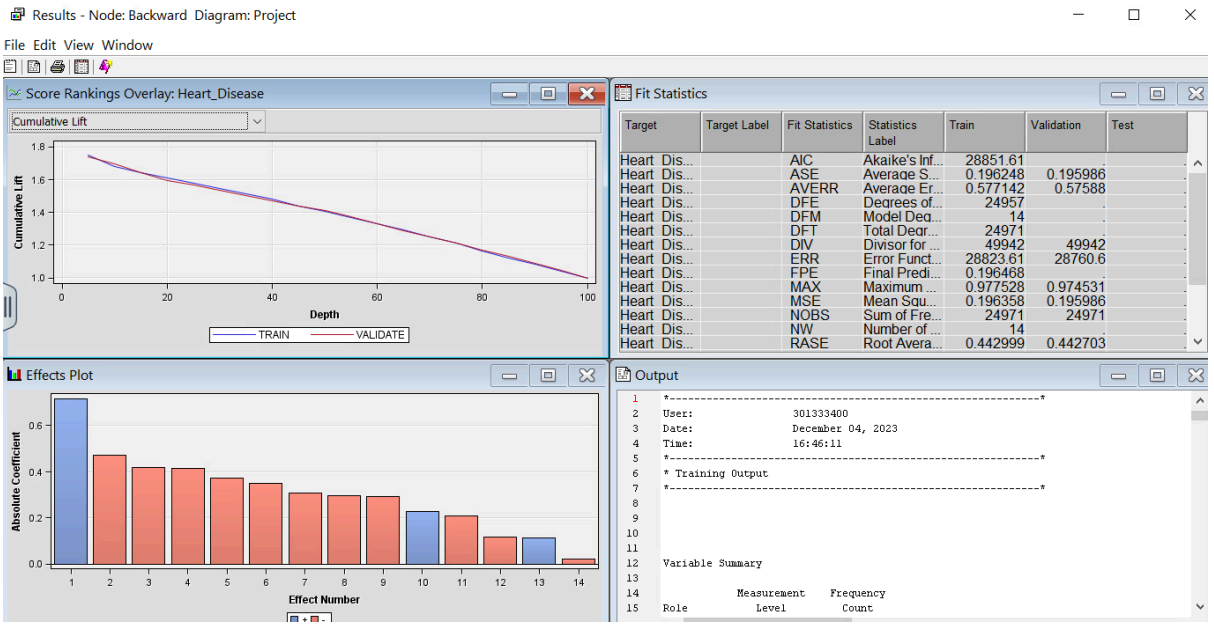
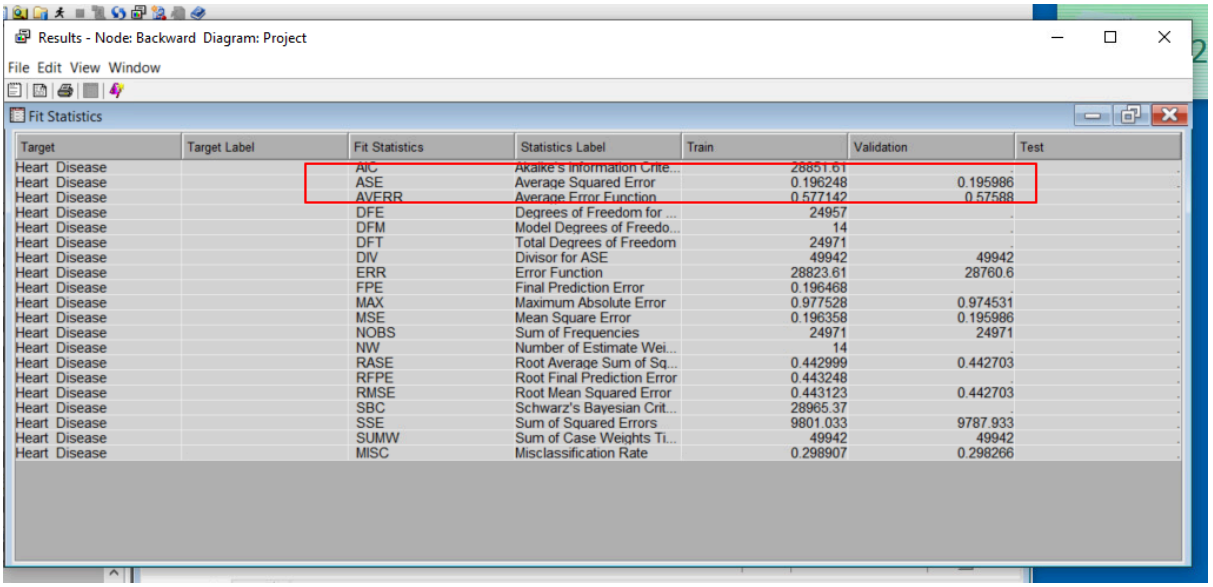


Figure 6.4.3 Results of forward regression



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Heart Disease		AIC	Akaike's Information Crite...	28851.61		
Heart Disease		ASE	Average Squared Error	0.196248	0.195986	
Heart Disease		AVERR	Average Error Function	0.577142	0.57588	
Heart Disease		DFE	Degrees of Freedom for ...	24957		
Heart Disease		DFM	Model Degrees of Freedo...	14		
Heart Disease		DFT	Total Degrees of Freedom	24971		
Heart Disease		DIV	Divisor for ASE	49942	49942	
Heart Disease		ERR	Error Function	28823.61	28760.6	
Heart Disease		FPE	Final Prediction Error	0.196468		
Heart Disease		MAX	Maximum Absolute Error	0.977528	0.974531	
Heart Disease		MSE	Mean Square Error	0.196358	0.195986	
Heart Disease		NOBS	Sum of Frequencies	24971	24971	
Heart Disease		NW	Number of Estimate Wei...	14		
Heart Disease		RASE	Root Average Sum of Sq...	0.442999	0.442703	
Heart Disease		RFPE	Root Final Prediction Error	0.443248		
Heart Disease		RMSE	Root Mean Squared Error	0.443123	0.442703	
Heart Disease		SBC	Schwarz's Bayesian Crit...	28965.37		
Heart Disease		SSE	Sum of Squared Errors	9801.033	9787.933	
Heart Disease		SUMW	Sum of Case Weights Ti...	49942	49942	
Heart Disease		MISC	Misclassification Rate	0.298907	0.298266	

Figure 6.4.4 Fit statistics of forward regression

Validation ASE is 0.195986

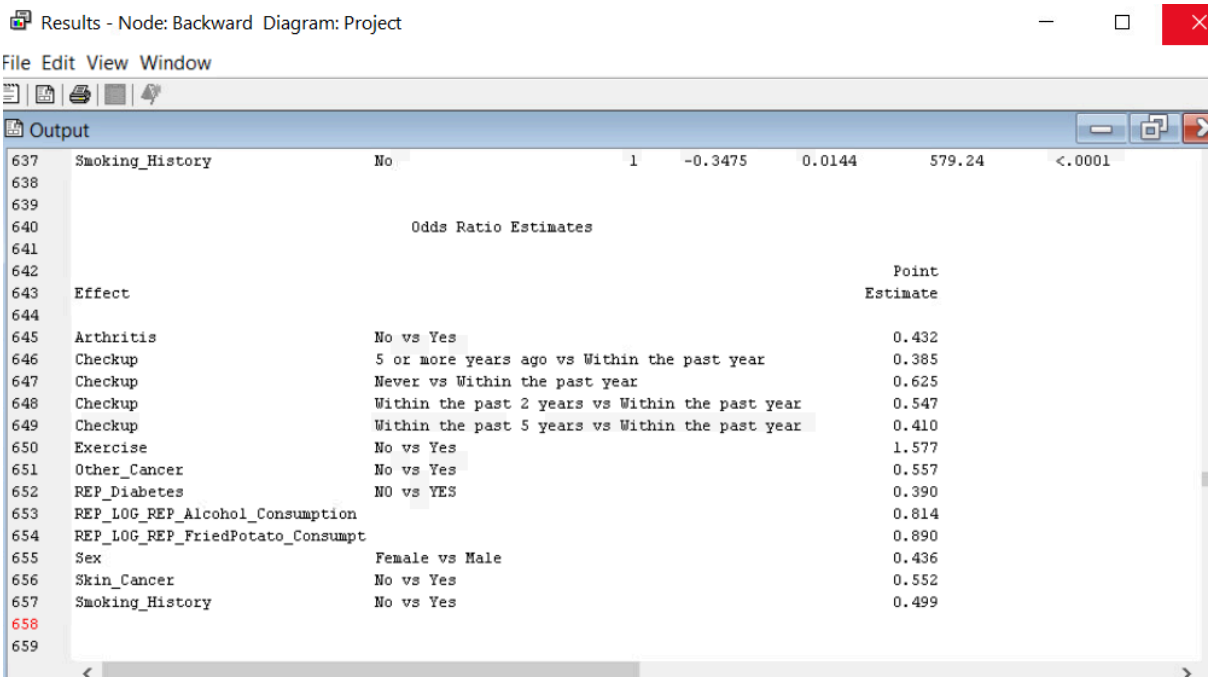
Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	REP_REP_Fruit_Consumption	1	13	0.0375	0.8464
2	REP_REP_Green_Vegetables_Consump	1	12	0.2676	0.6049
3	REP_REP_IMP_REP_BMI	1	11	0.7143	0.3980
4	Depression	1	10	0.6591	0.4169

The selected model is the model trained in the last step (Step 4). It consists of the following effects:

Intercept Arthritis Checkup Exercise Other_Cancer REP_Diabetes REP_LOG_REP_Alcohol_Consumption REP_LOG_REP_FriedPotato_Consumpt Sex Skin_Cancer Smoking_History

Figure 6.4.5 Summary of Forward Regression



637	Smoking_History	No	1	-0.3475	0.0144	579.24	<.0001

Odds Ratio Estimates

Effect		Point Estimate
Arthritis	No vs Yes	0.432
Checkup	5 or more years ago vs Within the past year	0.385
Checkup	Never vs Within the past year	0.625
Checkup	Within the past 2 years vs Within the past year	0.547
Checkup	Within the past 5 years vs Within the past year	0.410
Exercise	No vs Yes	1.577
Other_Cancer	No vs Yes	0.557
REP_Diabetes	NO vs YES	0.390
REP_LOG_REP_Alcohol_Consumption		0.814
REP_LOG_REP_FriedPotato_Consumpt		0.890
Sex	Female vs Male	0.436
Skin_Cancer	No vs Yes	0.552
Smoking_History	No vs Yes	0.499

Figure 6.4.6 Odds Ratio of forward regression

People who have fried potato less than 5 times a month are 11% less likely to have a heart disease

6.5. Stepwise Regression

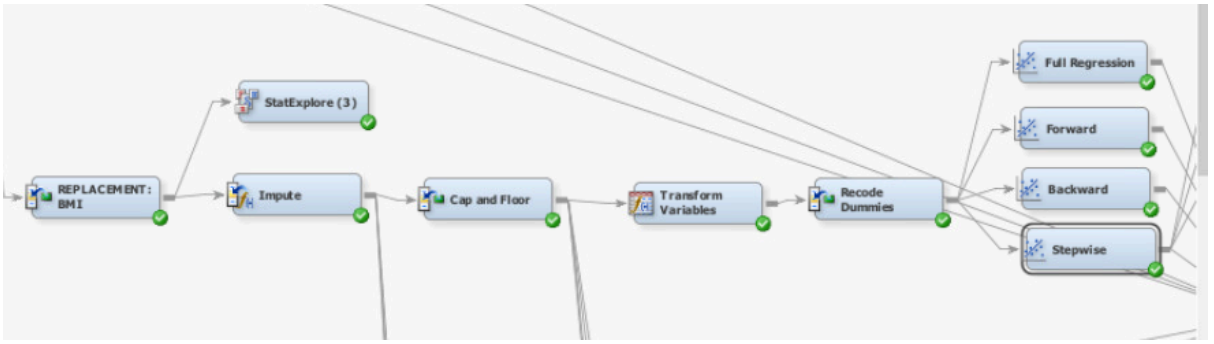


Figure 6.5.1 - Addition of stepwise regression node

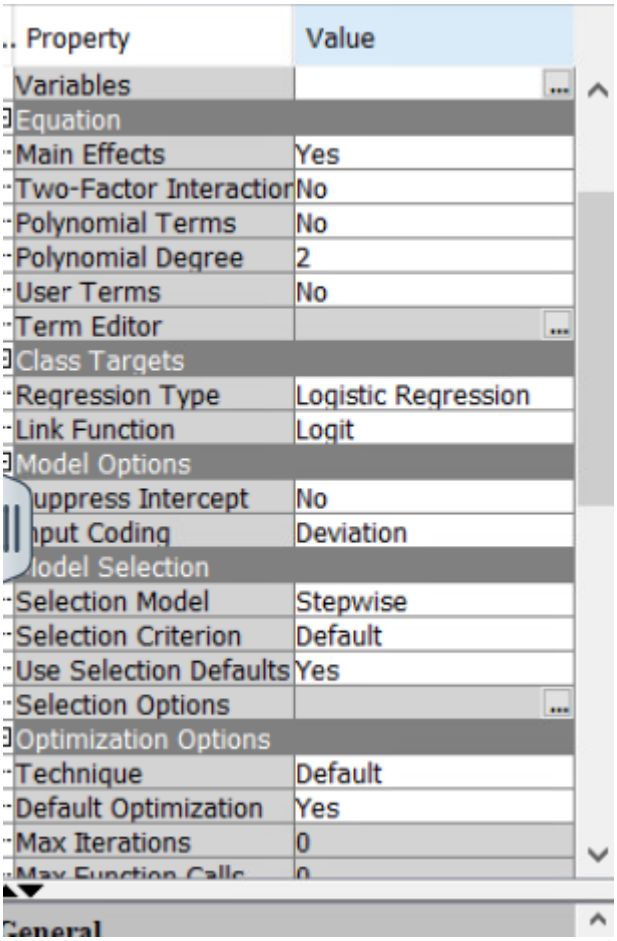


Figure 6.5.2 - Property Panel of stepwise regression node

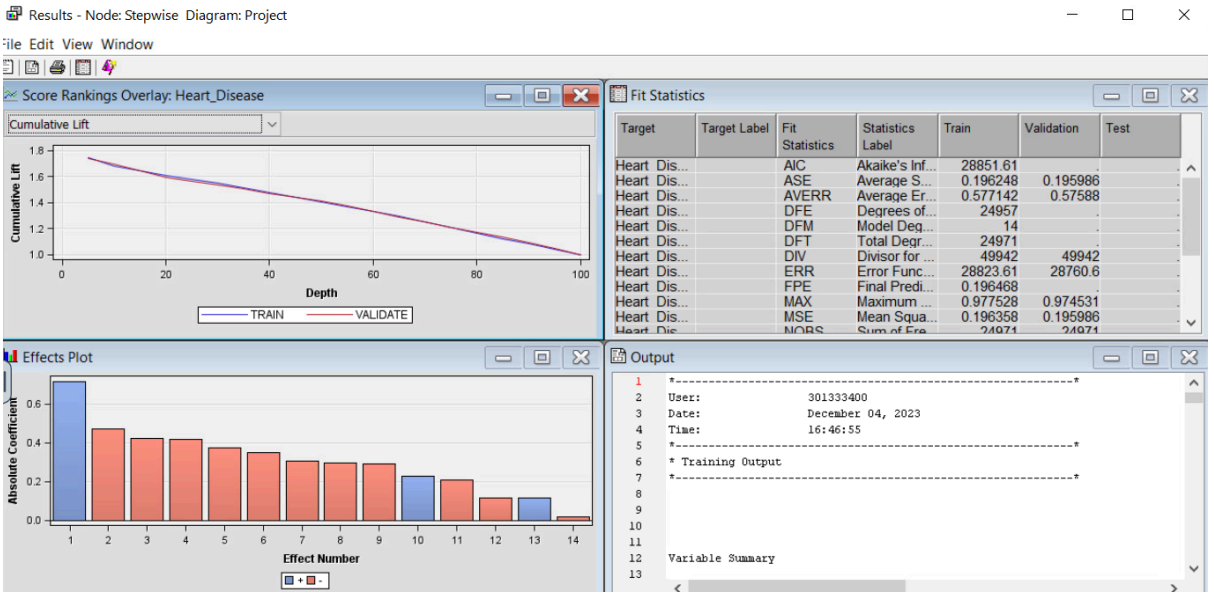
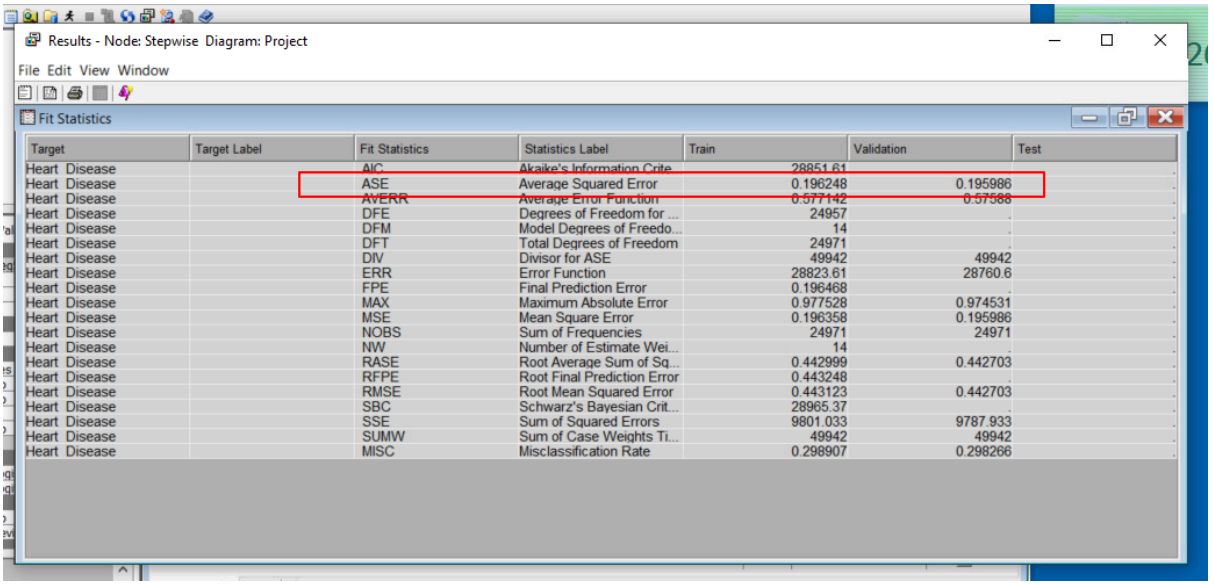


Figure 6.5.3 - Results of stepwise regression node



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Heart Disease		AIC	Akaike's Information Crite...	28851.61		
Heart Disease		ASE	Average Squared Error	0.196248	0.195986	
Heart Disease		AVERR	Average Error Function	0.577142	0.57588	
Heart Disease		DFE	Degrees of Freedom for ...	24957		
Heart Disease		DFM	Model Degrees of Freedo...	14		
Heart Disease		DFT	Total Degrees of Freedom	24971		
Heart Disease		DIV	Divisor for ASE	49942	49942	
Heart Disease		ERR	Error Function	28823.61	28760.6	
Heart Disease		FPE	Final Prediction Error	0.196468		
Heart Disease		MAX	Maximum Absolute Error	0.977528	0.974531	
Heart Disease		MSE	Mean Square Error	0.196358	0.195986	
Heart Disease		NOBS	Sum of Frequencies	24971	24971	
Heart Disease		NW	Number of Estimate Wei...	14		
Heart Disease		RASE	Root Average Sum of Sq...	0.442999	0.442703	
Heart Disease		RFPE	Root Final Prediction Error	0.443248		
Heart Disease		RMSE	Root Mean Squared Error	0.443123	0.442703	
Heart Disease		SBC	Schwarz's Bayesian Crit...	28965.37		
Heart Disease		SSE	Sum of Squared Errors	9801.033	9787.933	
Heart Disease		SUMW	Sum of Case Weights Ti...	49942	49942	
Heart Disease		MISC	Misclassification Rate	0.298907	0.298266	

Figure 6.5.4 - Fit statistics of stepwise regression node

Validation ASE is 0.195986

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Arthritis	1	1	1657.3639		<.0001
2	REP_Diabetes	1	2	1313.9671		<.0001
3	Smoking_History	1	3	739.3747		<.0001
4	Sex	1	4	546.1231		<.0001
5	Checkup	4	5	444.7982		<.0001
6	REP_LOG_REP_Alcohol_Consumption	1	6	319.6800		<.0001
7	Other_Cancer	1	7	263.5192		<.0001
8	Skin_Cancer	1	8	192.4145		<.0001
9	Exercise	1	9	186.6458		<.0001
10	REP_LOG_REP_FriedPotato_Consumpt	1	10	56.4397		<.0001

The selected model is the model trained in the last step (Step 10). It consists of the following effects:

Intercept Arthritis Checkup Exercise Other_Cancer REP_Diabetes REP_LOG_REP_Alcohol_Consumption REP_LOG_REP_FriedPotato_Consumpt Sex Skin_Cancer Smoking_History

Figure 6.5.5 - Summary of stepwise regression node

The significant variables are arthritis, REP diabetes, Sex, Checkup, REP LOG REP Alcohol consumption, Skin cancer, Exercise REP LOG REP Fried potato consumption.

Odds Ratio Estimates

Effect		Point Estimate
Arthritis	No vs Yes	0.432
Checkup	5 or more years ago vs Within the past year	0.385
Checkup	Never vs Within the past year	0.625
Checkup	Within the past 2 years vs Within the past year	0.547
Checkup	Within the past 5 years vs Within the past year	0.410
Exercise	No vs Yes	1.577
Other_Cancer	No vs Yes	0.557
REP_Diabetes	NO vs YES	0.390
REP_LOG_REP_Alcohol_Consumption		0.814
REP_LOG_REP_FriedPotato_Consumpt		0.890
Sex	Female vs Male	0.436
Skin_Cancer	No vs Yes	0.552
Smoking_History	No vs Yes	0.499

Figure 6.5.6 - Odds ratio of stepwise regression node

- People who don't have arthritis are 56% less likely to get heart disease
- People who don't have skin cancer are 45% less likely to get heart disease
- People who don't smoke are 51% less likely to get heart disease

6.6. Regression Summary

A Total of 4 regression nodes were done, and we found that all Forward, Backward, and Stepwise regression models had the same validation ASEs. We decided to move forward with Stepwise regression for further analysis.

Regression Model	ASE
Full Regression	0.195988
Forward Regression	0.195986
Backward Regression	0.195986

Stepwise Regression	0.195986

Table 6.6.1 - Regression Summary

7. Neural Network

A set of Neural Network nodes with three types of hidden units (3,4,5) and with iterations of (50) were used to analyze which model works the best.

One set of Neural Network nodes was attached to the impute node, the second set was connected to the cap and floor, and the third set was attached to stepwise regression to check which model works the best.

Average Error was the selection criterion used as a benchmark for all models to choose the best.

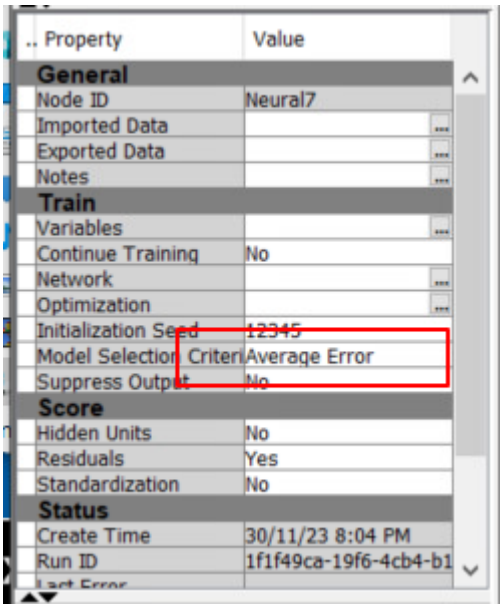


Figure 7.1 - Property panel of Neural Network Nodes

Since we have a large data set, we disabled the preliminary training under optimization configuration as we might need to train the neural network from scratch for greater effectiveness.

7.1 Neural Network from the Impute node

A set of 3 neural network nodes with varying hidden units (3,4,5) and iterations of 50 were attached to the impute node and tested.

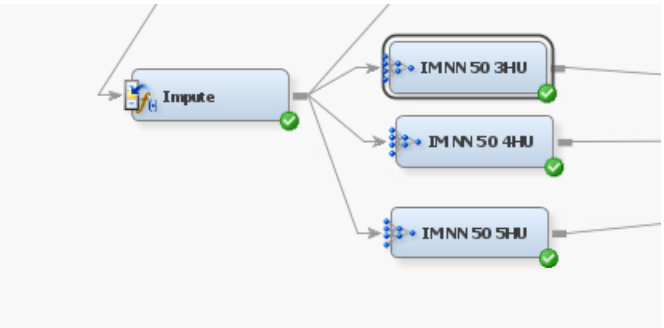


Figure 7.1.1 - Addition of Neural Networks to the Impute node

7.1.1 Impute Node - 3 Hidden Units,50 iterations

The first Neural Network node was added to the Impute node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 3. Under the Optimization configuration, the number of iterations was kept at 50 as shown in the following figures.

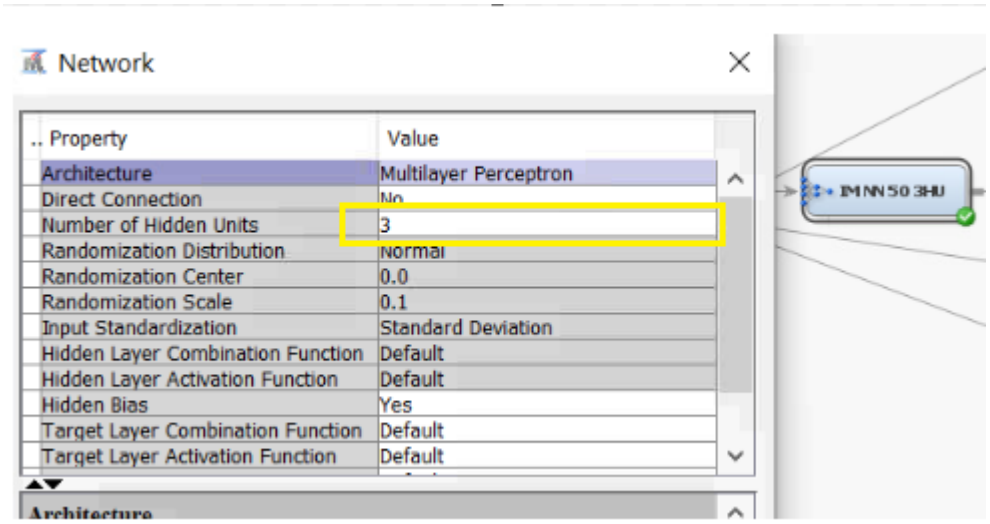


Figure 7.1.2 - Network configurations of IM NN 50 3HU

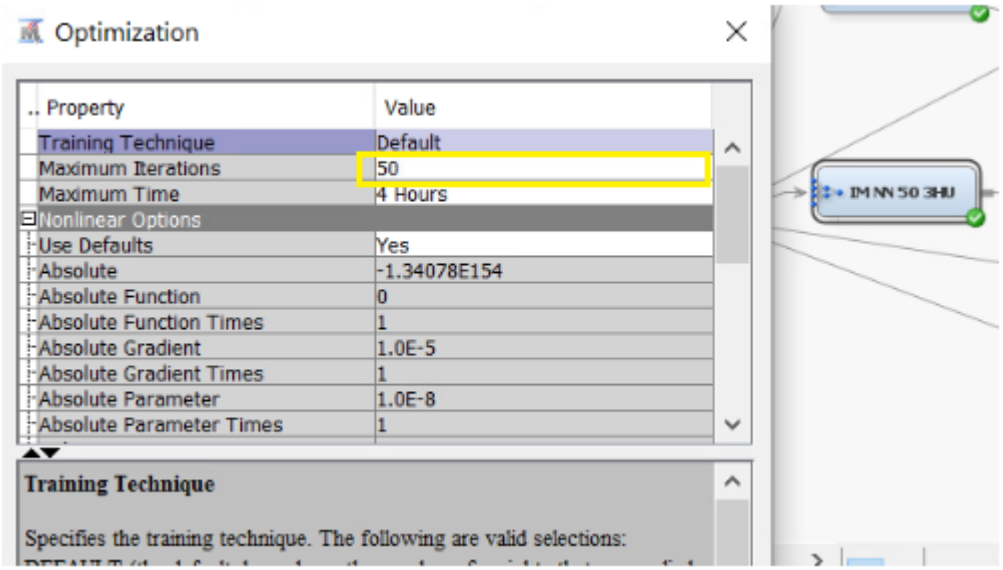


Figure 7.1.3 - Optimization configurations of IM NN 50 3HU

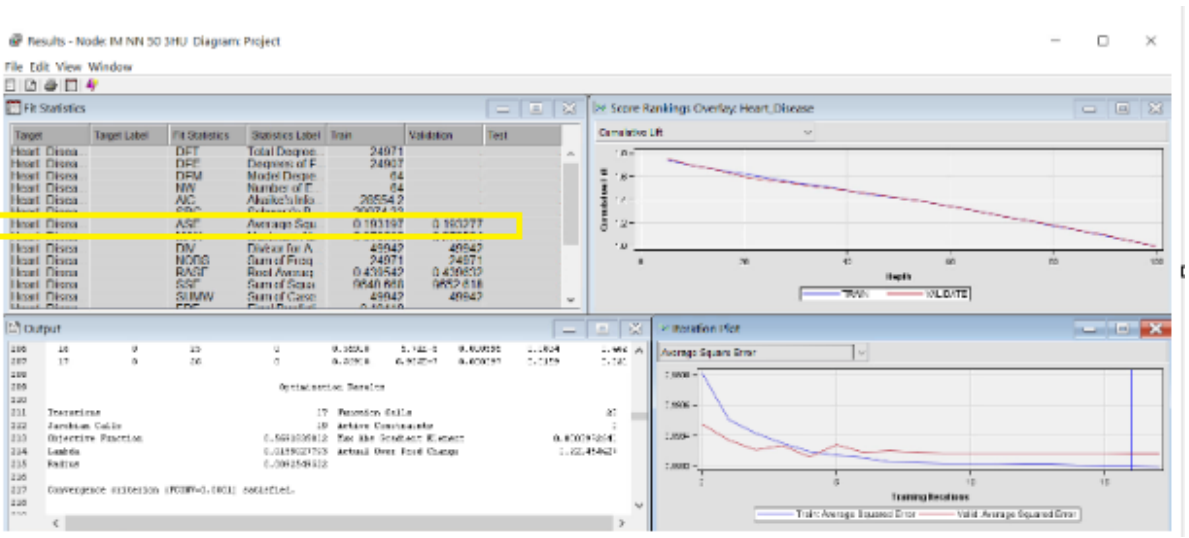


Figure 7.1.4 - Results for IM NN 50 3HU

This Neural Network had a validation **ASE of 0.193277.** As we wanted to check if we could get better results, we continued with four hidden units in the next step

7.1.2 Impute Node - 4 Hidden Units,50 iterations

The second Neural Network node was added to the Impute node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 4. Under the Optimization configuration, the number of iterations was kept at 50, as shown in the following figures.

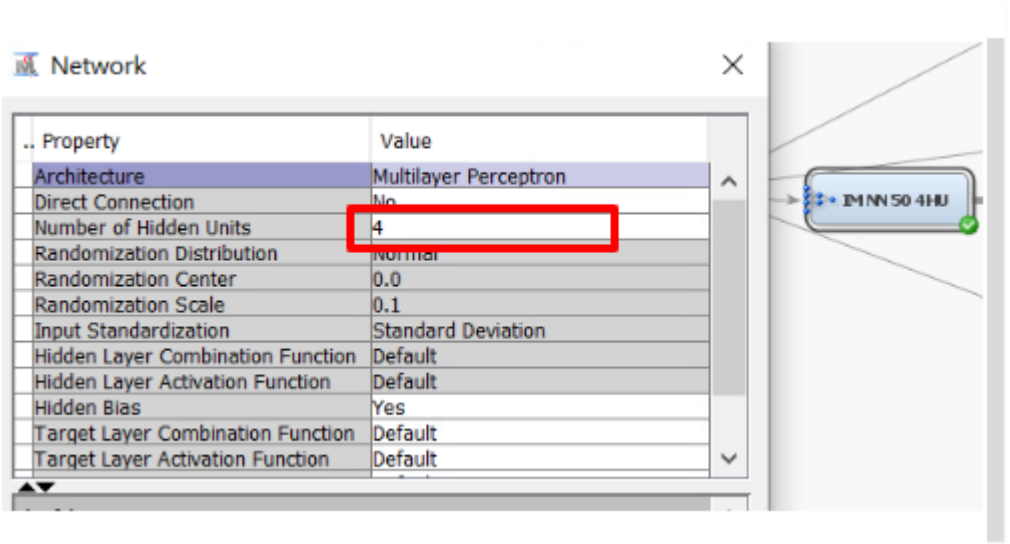


Figure 7.1.5 - Network properties for IM NN 50 4HU

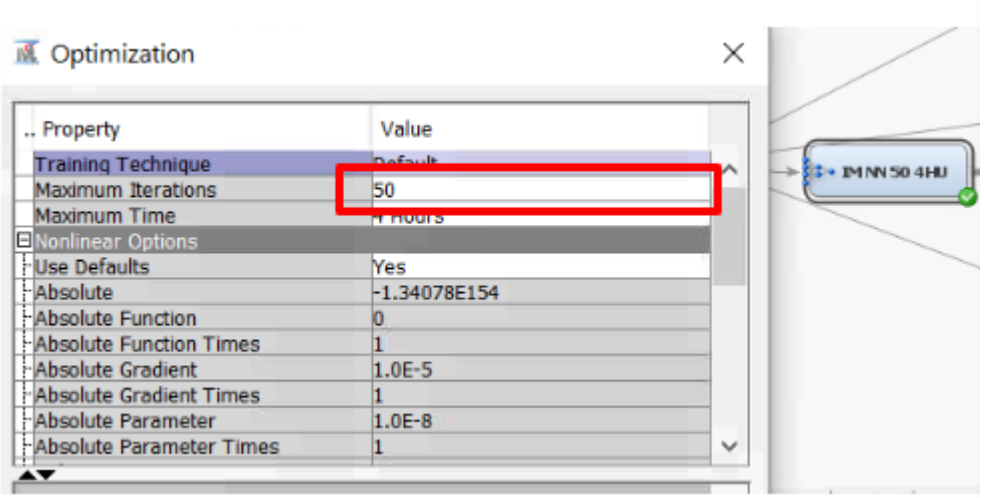


Figure 7.1.6 - Optimization properties for IM NN 50 4HU

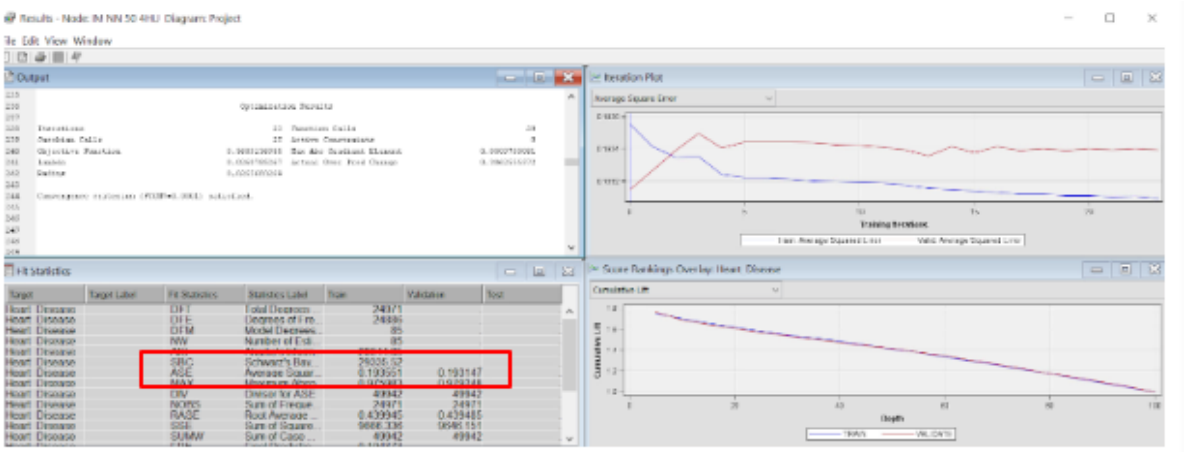


Figure 7.1.7 - Results for IM NN 50 4H This Neural Network had a validation **AS of 0.193147.** As we wanted to check for better results, we continued with five hidden units in the next step.

7.1.3 Impute Node - 5 Hidden Units,50 iterations

The third Neural Network node was added to the Impute node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 5. Under the Optimization configuration, the number of iterations was kept at 50 as shown in the following figures.

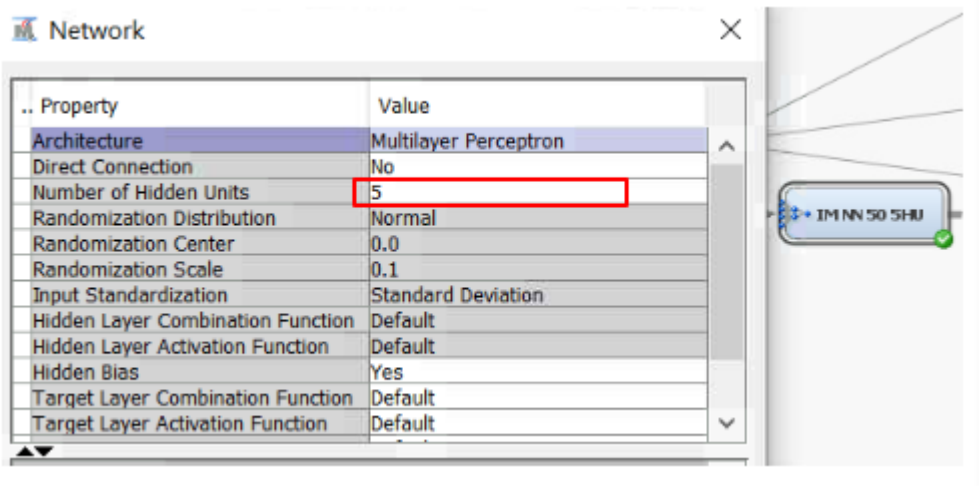


Figure 7.1.8 - Network properties for IM NN 50 5HU

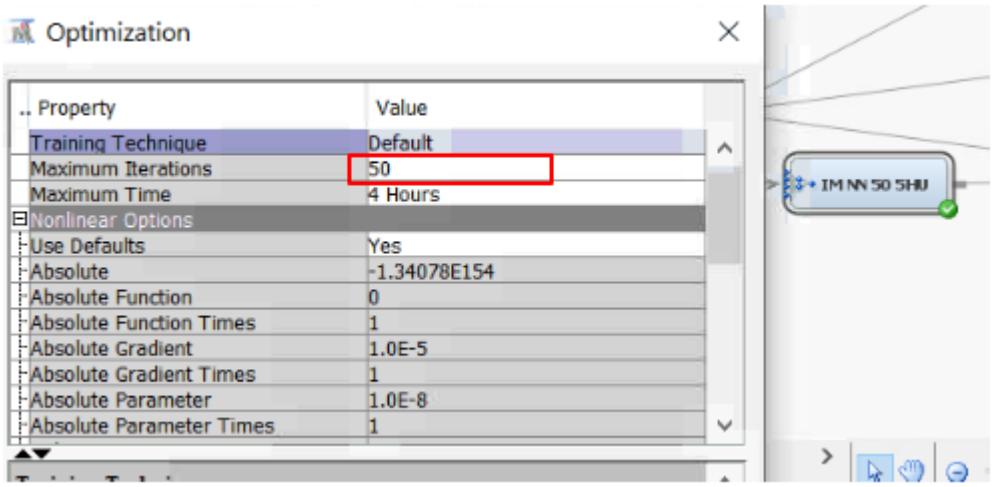


Figure 7.1.9 - Optimization properties for IM NN 50 5HU

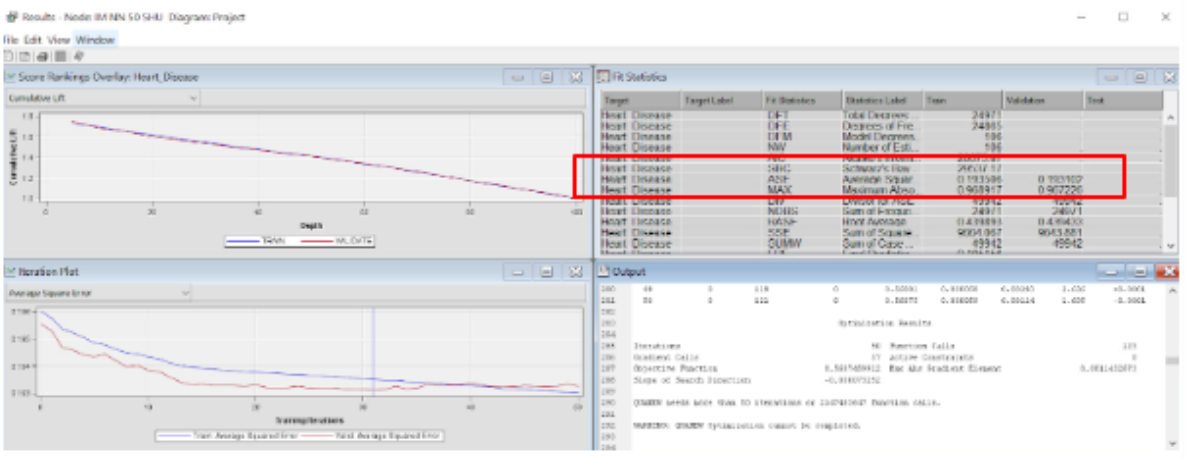


Figure 7.1.10- Results for IM NN 50 5HU

This Neural Network had a validation **ASE of 0.193102.** All the iterations were completed. There is no point in adding more hidden layers hence we stopped here.

7.2 Neural Network from the Cap and Floor Node

A total of 3 neural network nodes were connected to the cap and floor nodes and tested.

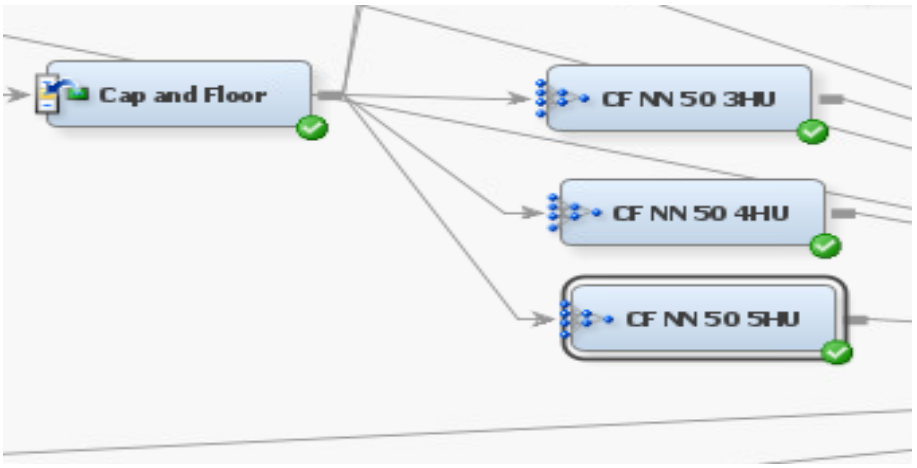


Figure 7.2.1 - Addition of Neural Network nodes to the cap and floor node

7.2.1 Cap and Floor Node - 3 Hidden Units,50 iterations

The first Neural Network node was added to the Cap and Floor node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 3. Under the Optimization configuration, the number of iterations was kept at 50 as shown in the following figures.

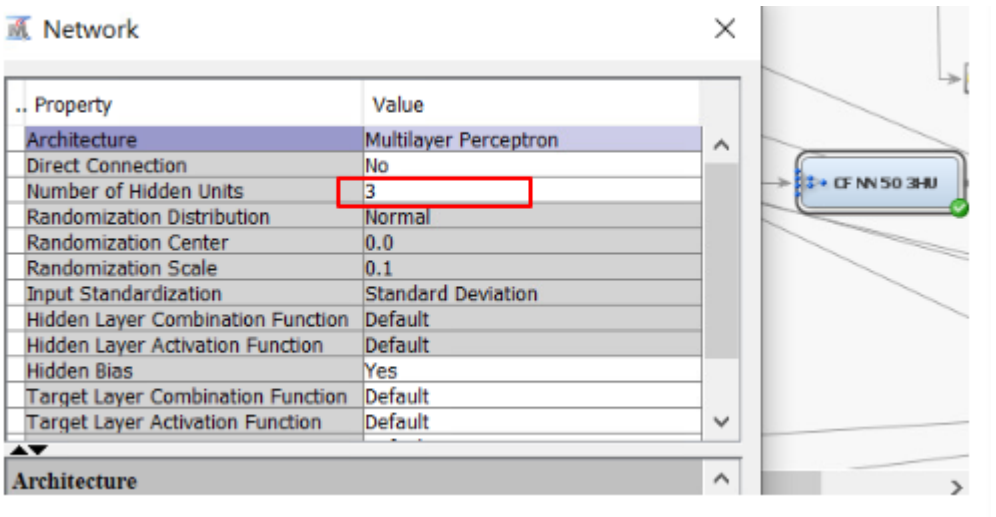


Figure 7.2.2 - Network settings of CF NN50 3HU

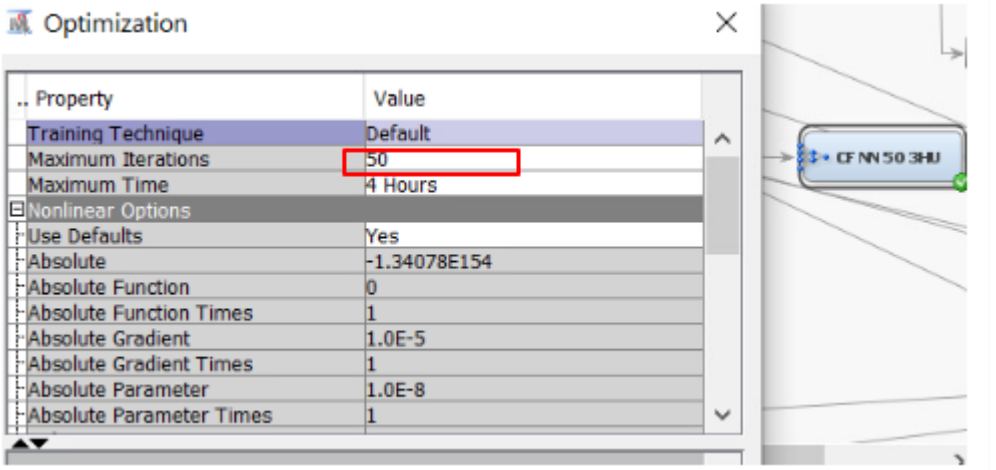


Figure 7.2.3 - Optimization settings of CF NN50 3HU

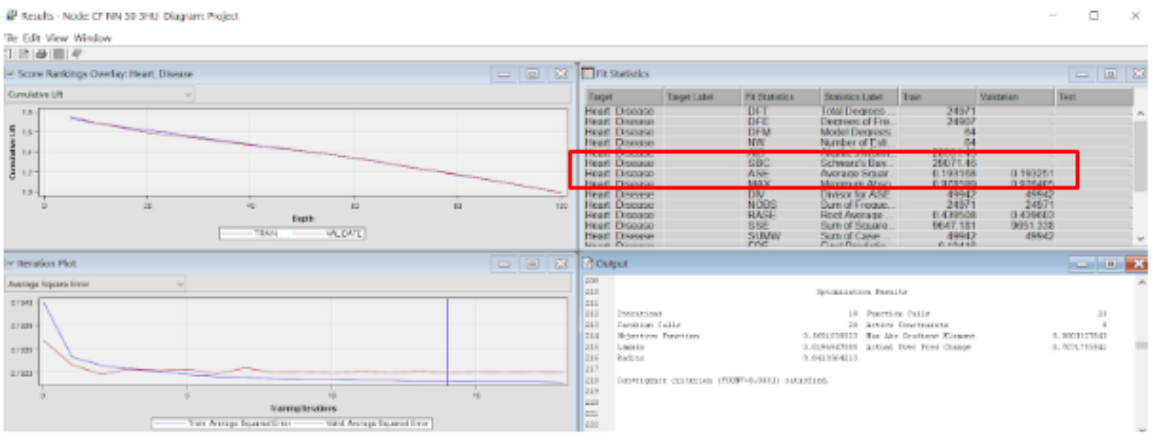


Figure 7.2.4 - Results of CF NN50 3HU

This Neural Network had a validation **ASE of 0.193251.** It was observed that the process was completed. So, we added a new node with four hidden layers.

7.2.2 Cap and Floor Node - 4 Hidden Units,50 iterations

The second Neural Network node was added to the Cap and Floor node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 4. Under the Optimization configuration, the number of iterations was kept at 50, as shown in the following figures.

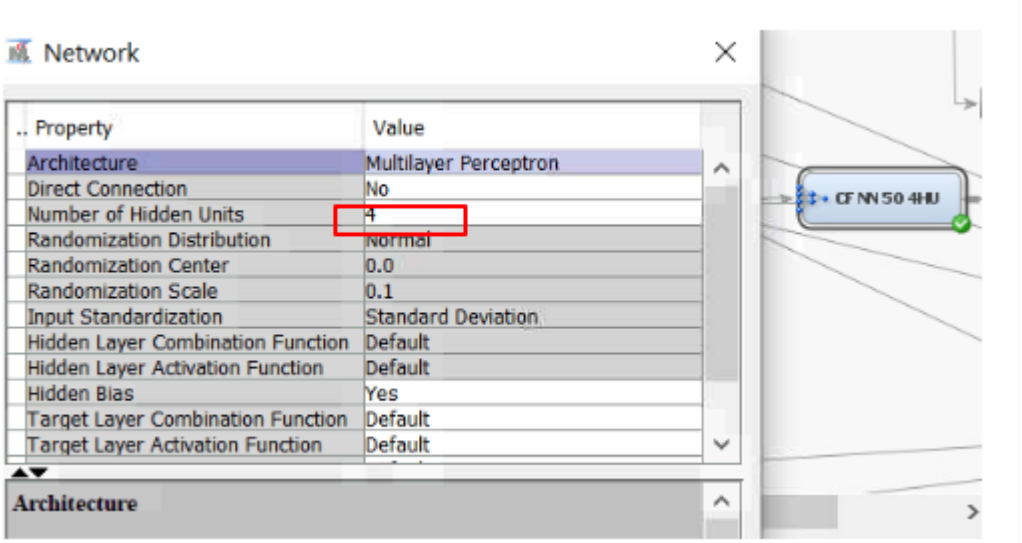


Figure 7.2.5 - Network settings of CF NN50 4HU

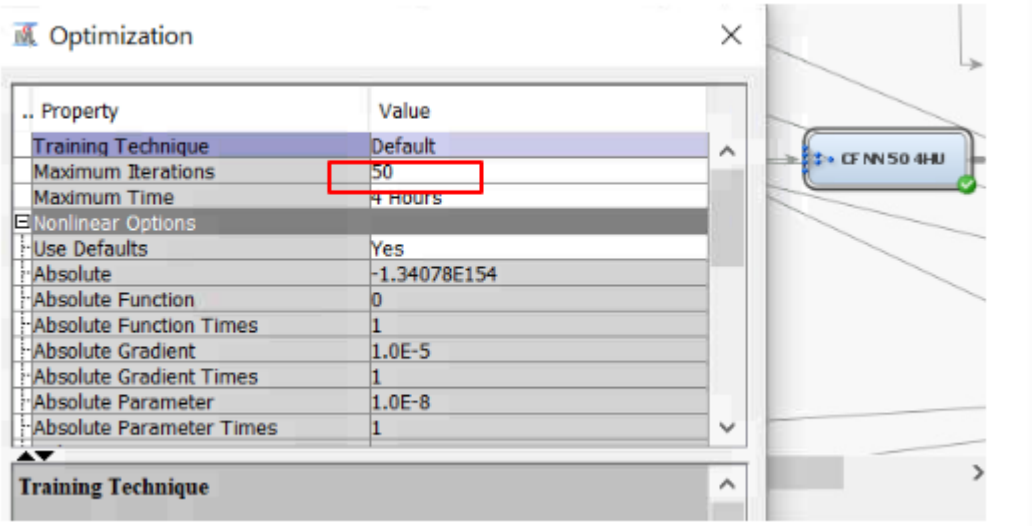


Figure 7.2.6 - Optimization settings of CF NN50 4HU

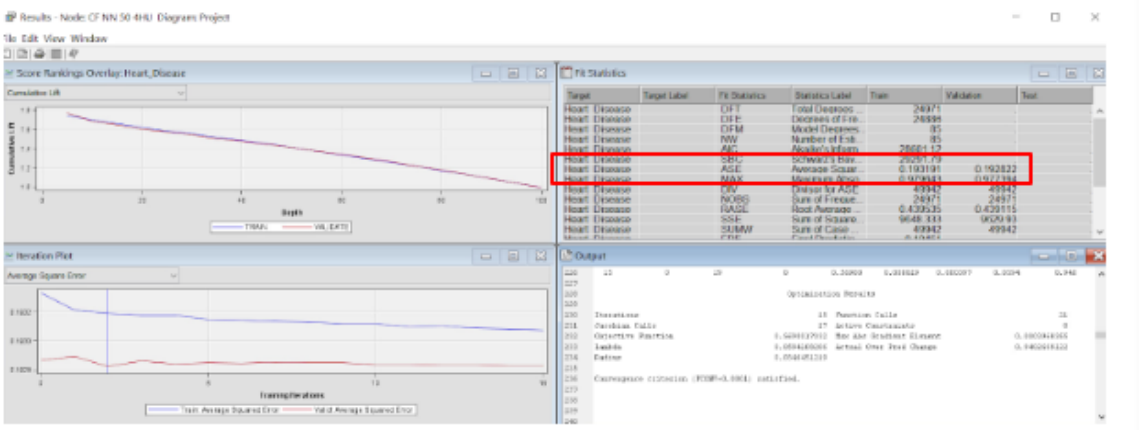


Figure 7.2.7 - Results of CF NN50 4HU

This Neural Network had a validation **ASE of 0.192822.** It was observed that the process was completed. So, we added a new node with five hidden layers.

7.2.3 Cap and Floor Node - 5 Hidden Units,50 iterations

The third Neural Network node was added to the Cap and Floor node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 5. Under optimization configuration, the number of iterations was kept at 50, as shown in the following figures.

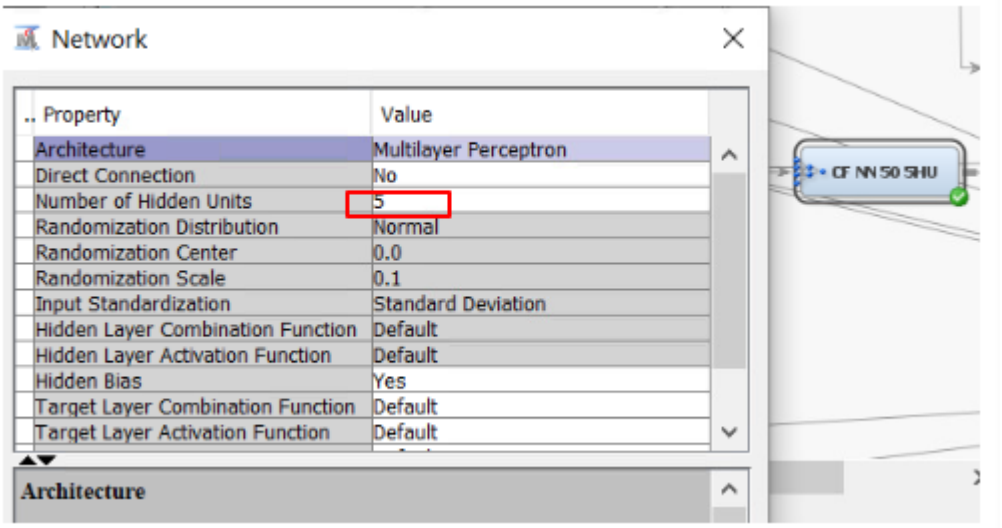


Figure 7.2.8 - Network settings of CF NN50 5HU

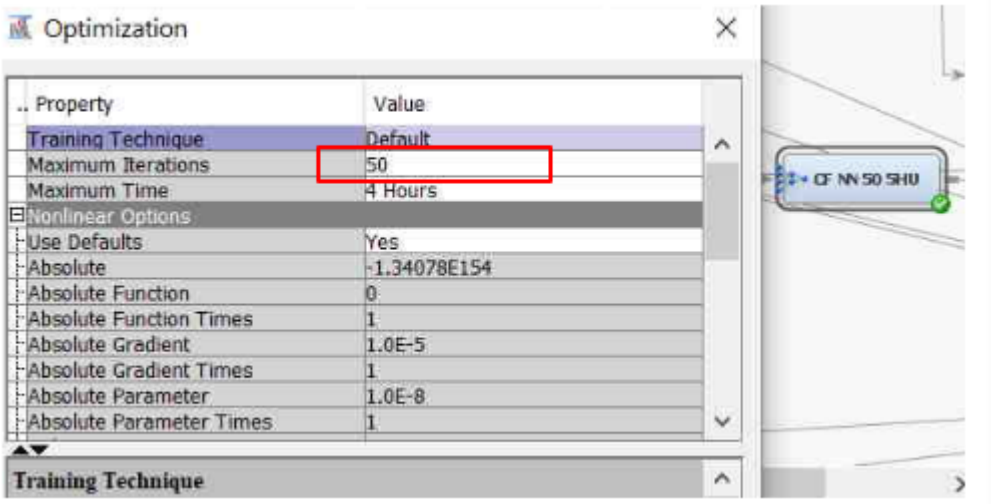


Figure 7.2.9 - Optimization settings of CF NN50 5HU

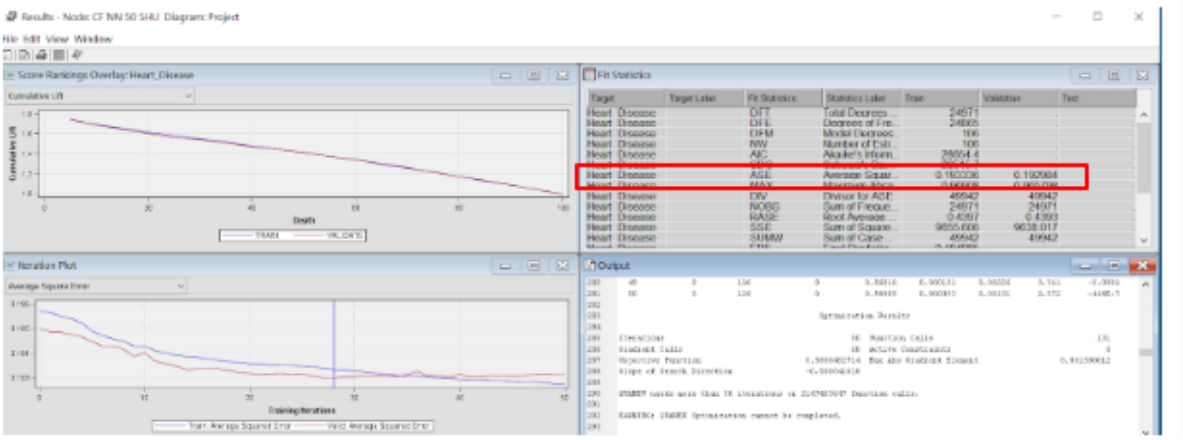


Figure 7.2.10 - Results of CF NN50 5HU

This Neural Network had a validation **ASE of 0.192984.** It was observed that the error was increasing, so we stopped here

7.3 Neural Network from the Best Regression Model

A total of 3 neural network nodes were connected to the Stepwise Regression node and tested.

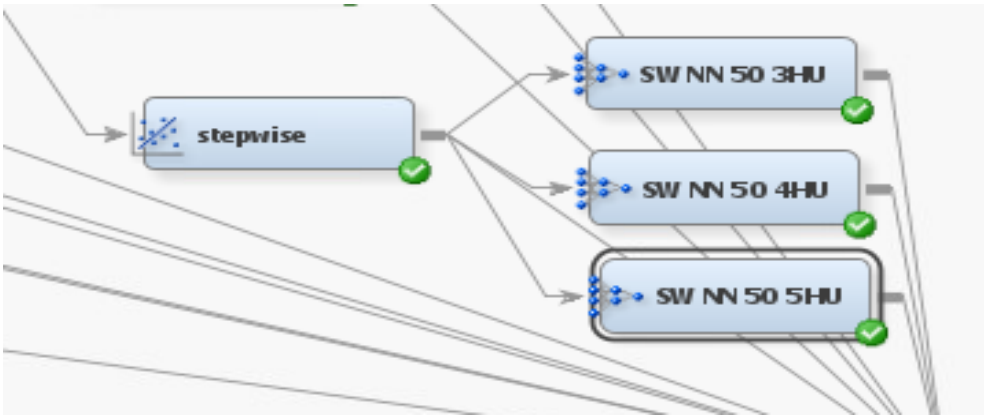


Figure 7.3.1 - Addition of Neural Network Nodes to the Stepwise Regression

7.3.1 Stepwise Regression Node - 3 Hidden Units,50 iterations

The first Neural Network node was added to the Stepwise Regression node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 3. Under the Optimization configuration, the number of iterations was kept at 50, as shown in the following figures.

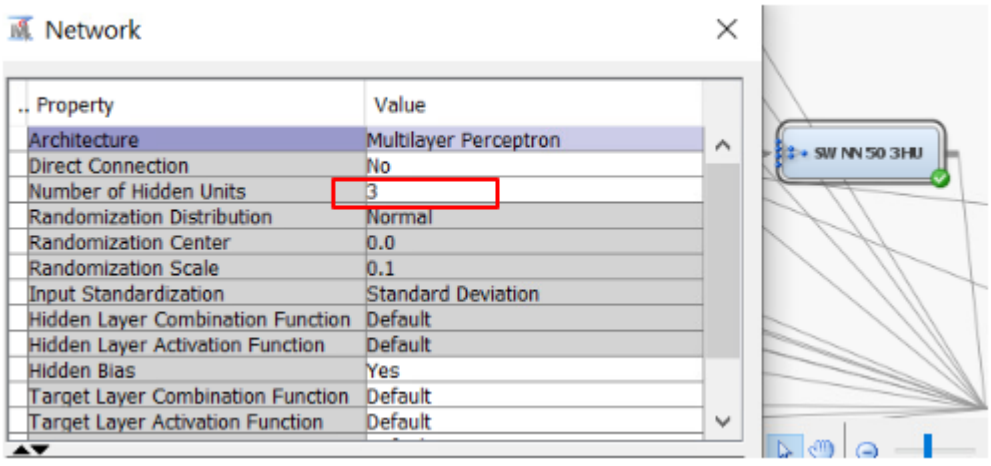


Figure 7.3.2 - Network settings of SW NN 50 3HU

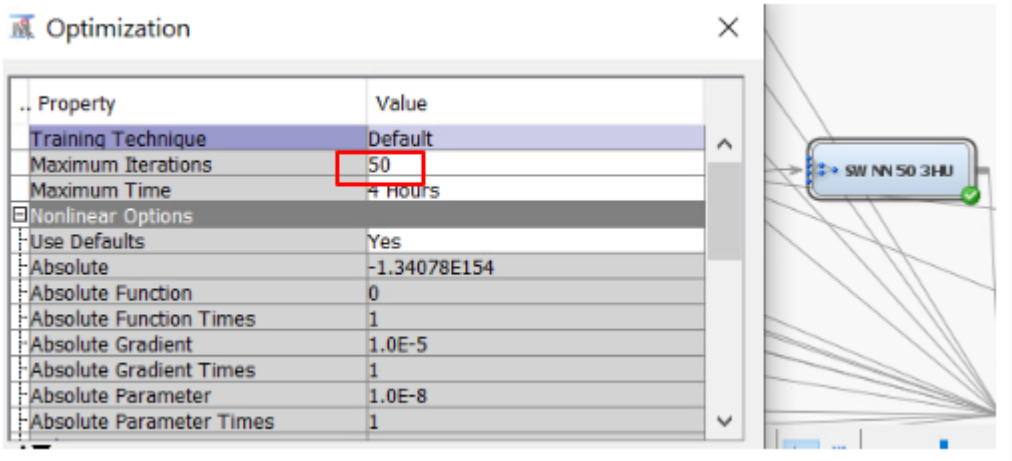


Figure 7.3.3 - Optimization settings of SW NN 50 3HU

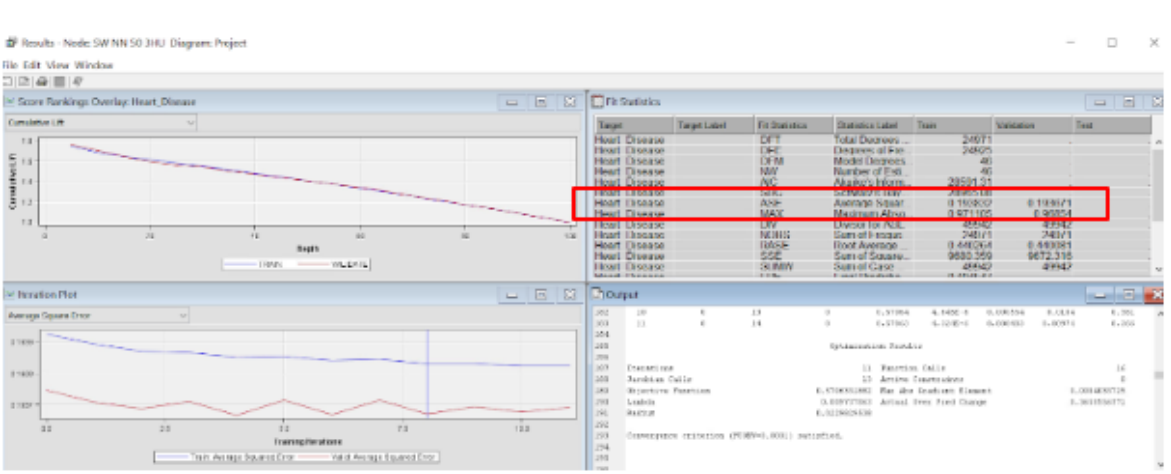


Figure 7.3.4 - Results of SW NN 50 3HU

This Neural Network had a validation **ASE of 0.193671.** It was observed that the process was completed. So, we added a new node with four hidden layers.

7.3.2 Stepwise Regression Node - 4 Hidden Units,50 iterations

The Second Neural Network node was added to the Stepwise Regression node. Under the Network configuration of the Property Panel, the number of hidden units

was kept at 4. Under the Optimization configuration, the number of iterations was marked as 50, as shown in the following figures.

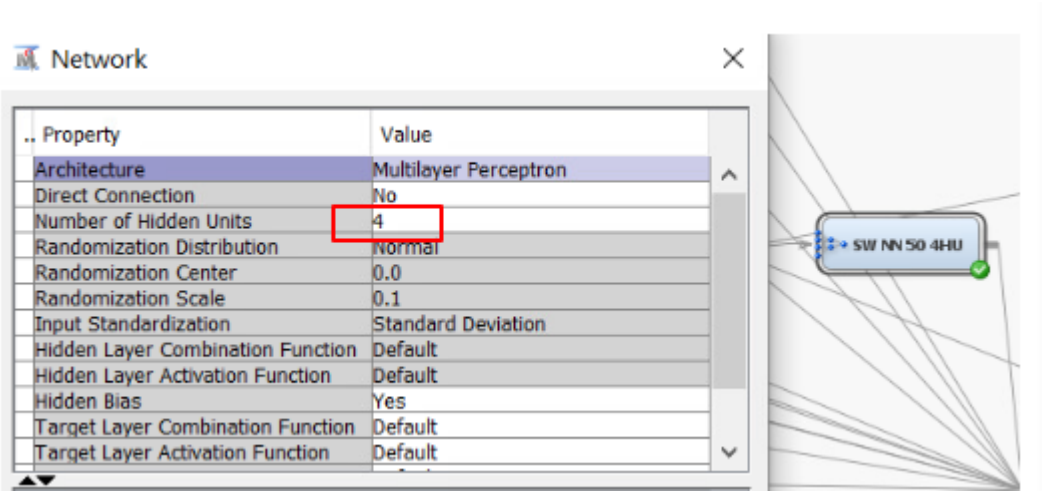


Figure 7.3.5 - Network settings of SW NN 50 4HU

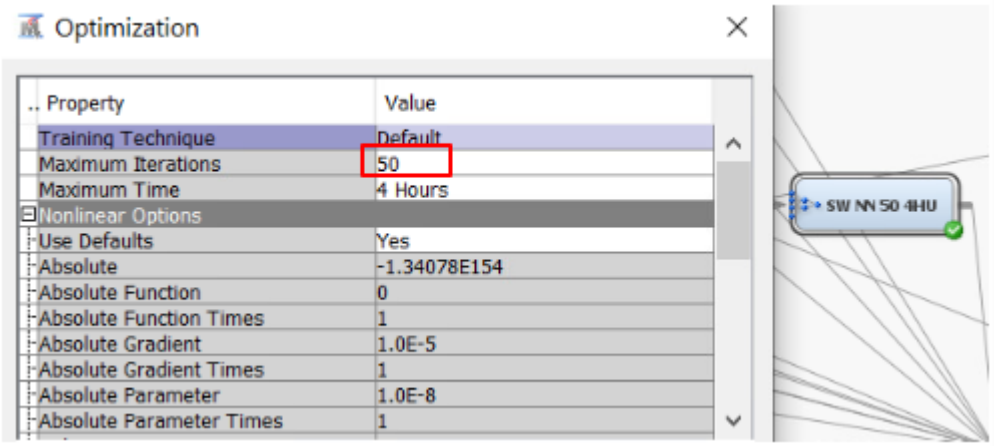


Figure 7.3.6 - Optimization settings of SW NN 50 4HU

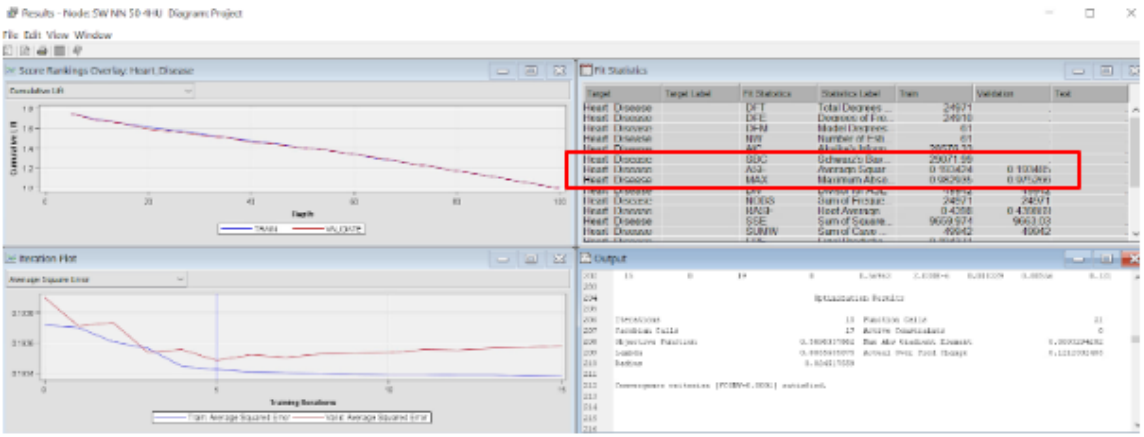


Figure 7.3.7 - Result of SW NN 50 4HU

This Neural Network had a validation **ASE of 0.193485.** It was observed that the process was completed. So, we added a new node with five hidden layers.

7.3.3 Stepwise Regression Node - 5 Hidden Units,50 iterations

The third Neural Network node was added to the Stepwise Regression node. Under the Network configuration of the Property Panel, the number of hidden units was kept at 5. Under the Optimization configuration, the number of iterations was observed as 50, as shown in the following figures.

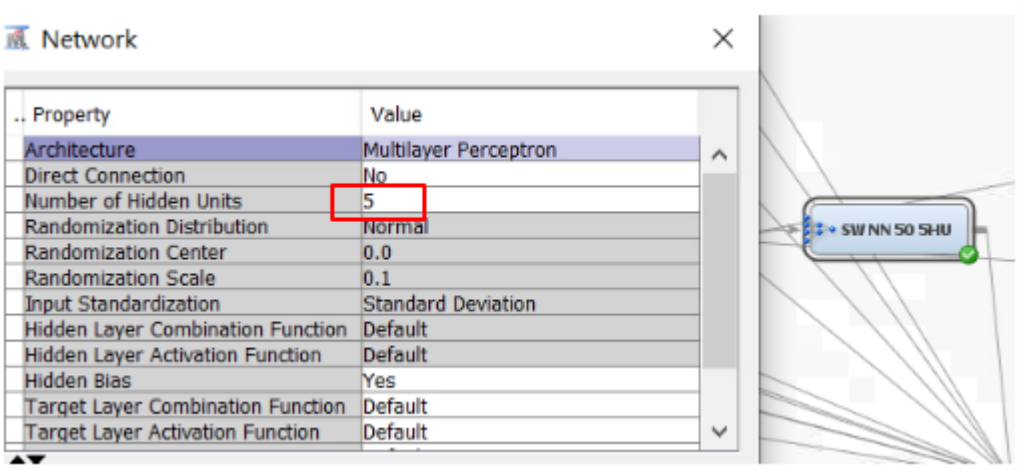


Figure 7.3.8 - Network settings of SW NN 50 5HU

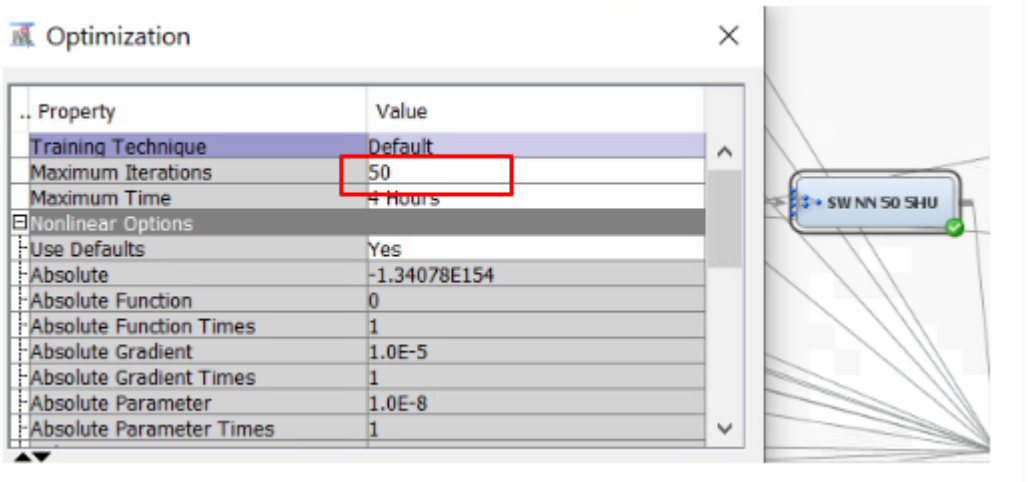


Figure 7.3.9 - Optimization settings of SW NN 50 5HU

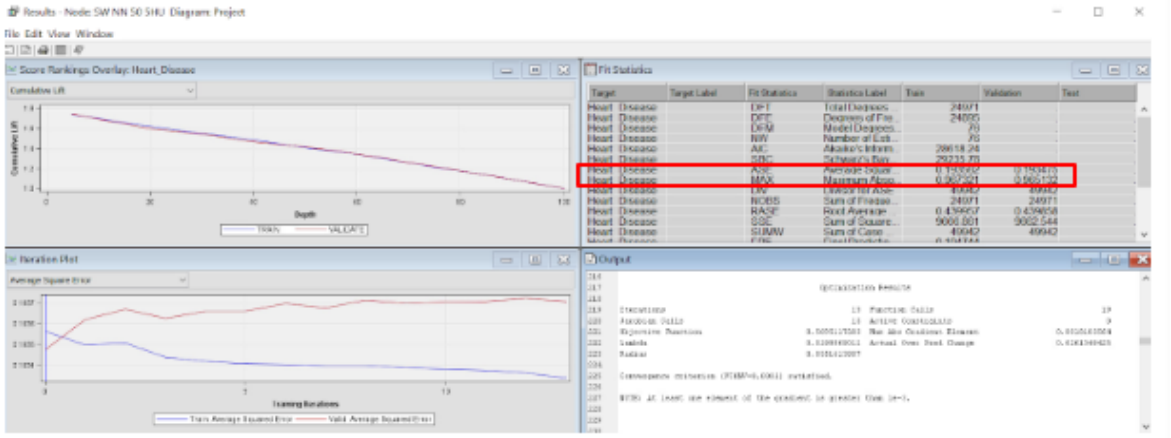


Figure 7.3.10 - Result of SW NN 50 5HU

This Neural Network had a validation **ASE of 0.193475.** Adding more hidden units will complicate the model, so we stop here.

7.4 Neural Network Summary

A total of 9 Neural Networks were run; the neural network with four hidden units and 50 iterations attached to cap and floor gave the lowest ASE.

Neural Network Model	Average Squared Error
IM NN 50 3HU	0.193277
IM NN 50 4HU	0.193147
IM NN 50 5HU	0.193102
CF NN 50 3HU	0.193251
CF NN 50 4HU	0.192822
CF NN 50 5HU	0.192984
SW NN 50 3HU	0.193671
SW NN 50 4HU	0.193485
SW NN 50 5HU	0.193475

Table 7.4.1 - Neural Network Summary

The exported validation report shows the first 15 people having the risk of heart disease with the first observation having a 95.08% chance of having heart disease.

EMWS1.Neural16_VALIDATE

	Predicted: Heart_Disease=Yes	Observation Number	General_Health	Checkup	Exercise	Heart_...	Skin_...	Other...	Depression	Diabetes	Arthritis	Sex	Age...	BMI	Smo...	Alcoh...	Frui...	Gree...	Frie...	Age	Height_(cm)	Weight_(kg)
1	0.9508909811213555	211080.0	Good	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	30-34	36.61	Yes	8	12	20	20	32	165.0	99.79
2	0.9458626646448226	88624.0	Good	Within the past year	No	Yes	Yes	No	No	Yes	Yes	Male	65-69	26.31	Yes	8	30	4	30	67	183.0	88.0
3	0.9369867051058 4	191365.0	Good	Within the past year	No	Yes	Yes	No	No	Yes	Yes	Male	75-79	29.53	Yes	1	90	30	60	77	188.0	104.33
4	0.9367798405456104	46579.0	Poor	Within the past year	Yes	Yes	Yes	No	Yes	Yes	Yes	Male	60-64	48.82	Yes	2	0	0	30	62	183.0	163.29
5	0.9355735810925062	232611.0	Poor	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	75-79	37.66	Yes	0	30	12	20	77	175.0	115.67
6	0.9349373119679547	41835.0	Very Good	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	80+	19.53	Yes	0	8	0	1	80	180.0	63.5
7	0.9331618886274275	185180.0	Fair	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	80+	27.26	Yes	1	5	8	0	80	183.0	91.17
8	0.9328131818414493	113037.0	Good	Within the past year	Yes	Yes	No	Yes	No	Yes	Yes	Male	70-74	32.78	Yes	8	30	60	60	72	180.0	106.59
9	0.9320130475271357	42608.0	Poor	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	75-79	31.32	Yes	3	12	1	0	77	170.0	90.72
10	0.931687618009481	237947.0	Fair	Within the past year	No	Yes	Yes	Yes	Yes	Yes	Yes	Male	50-54	27.39	Yes	0	4	28	0	52	193.0	102.06
11	0.9316765619419082	85839.0	Fair	Within the past year	No	Yes	Yes	Yes	No	Yes	Yes	Male	70-74	31.32	Yes	5	12	8	12	72	170.0	90.72
12	0.928952397240048	8646.0	Fair	Within the past year	No	Yes	Yes	No	Yes	Yes	Yes	Male	80+	26.61	Yes	0	4	2	2	80	173.0	79.38
13	0.9285349729323275	85288.0	Fair	Within the past year	No	Yes	Yes	No	No	Yes	Yes	Male	70-74	29.24	Yes	0	2	0	0	72	175.0	89.81
14	0.9274544425215513	230987.0	Fair	Within the past year	No	Yes	No	Yes	No	Yes	Yes	Male	75-79	32.61	Yes	3	0	2	60	77	163.0	86.18
15	0.9271960739046574	141938.0	Fair	Within the past year	No	Yes	Yes	No	No	Yes	Yes	Male	80+	28.25	Yes	0	30	4	0	80	168.0	79.38
16	0.9270236441072781	22433.0	Good	Within the past year	No	Yes	Yes	No	Yes	Yes	Yes	Male	70-74	28.5	Yes	0	8	4	2	72	175.0	87.54
17	0.9269782753982737	136682.0	Poor	Within the past year	No	Yes	Yes	No	No	Yes	Yes	Male	80+	31.02	Yes	0	1	1	0	80	173.0	92.53
18	0.9263703177292084	87316.0	Poor	Within the past year	No	Yes	Yes	Yes	No	Yes	Yes	Male	70-74	27.12	Yes	0	3	2	12	72	183.0	90.72
19	0.9258388764396066	100240.0	Fair	Within the past 2 years	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Male	65-69	34.87	Yes	0	20	0	30	67	180.0	113.4
20	0.92575699267913	194755.0	Good	Within the past year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Male	80+	28.89	Yes	5	5	10	8	80	173.0	86.18
21	0.9255633249011324	171563.0	Poor	Within the past year	No	Yes	Yes	No	Yes	Yes	Yes	Male	80+	31.79	Yes	0	30	30	0	80	170.0	92.08

Figure 7.4.1 - Exported Validation Data for CF NN 50 4HU

8. Model Comparison

All decision trees, neural networks, and regression nodes were attached to the Model Comparison node to find the best model. The model that has the lowest error is considered the best model.

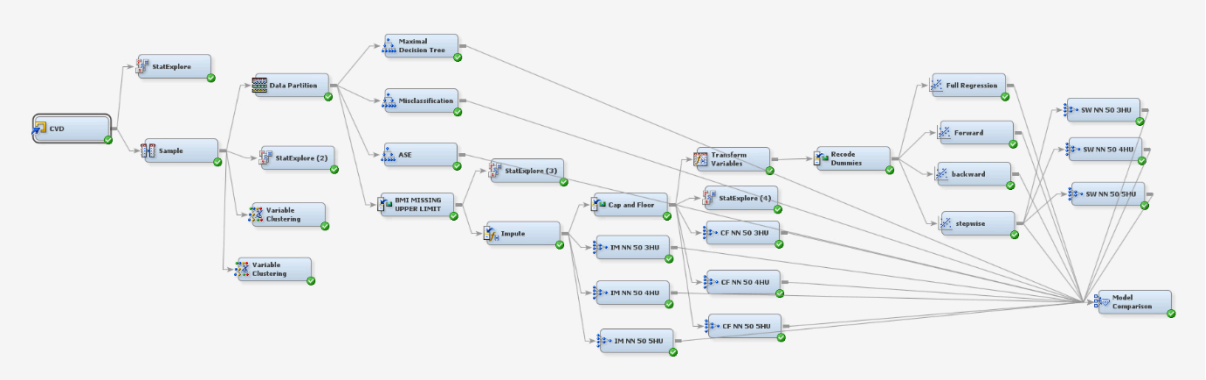


Figure 8.1 - Addition to Model Comparison

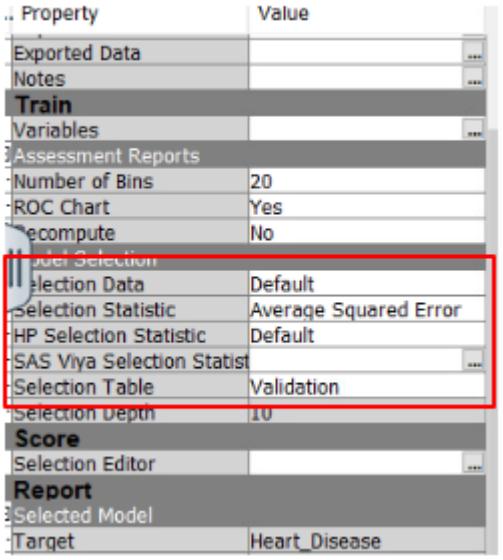


Figure 8.2 - Property panel of Model Comparison

As all models were based on ASE, it was used as the selection statistic and validation data was used in the selection table as it is more important than the training data.

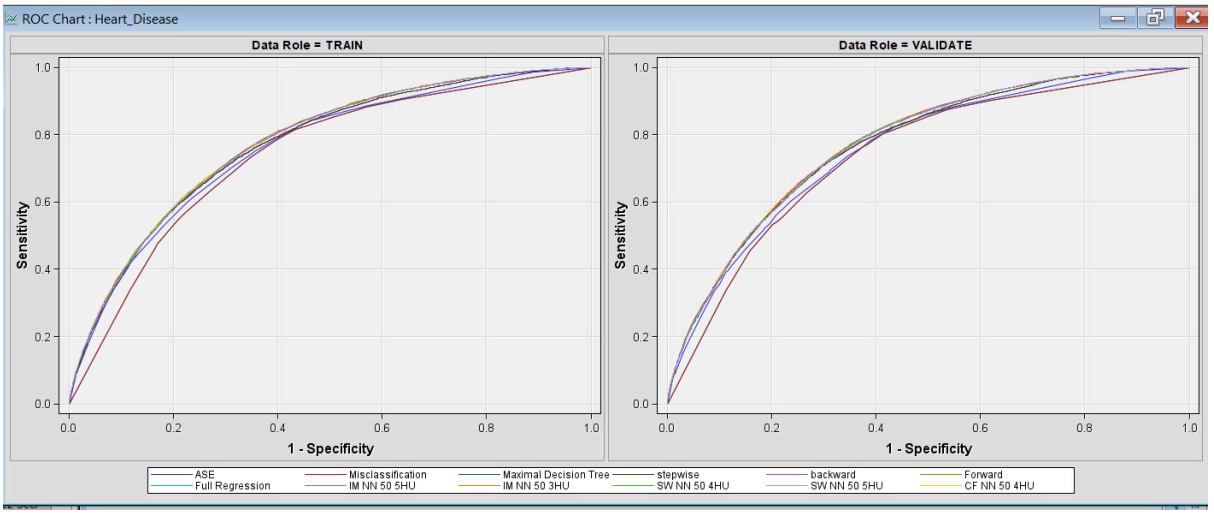
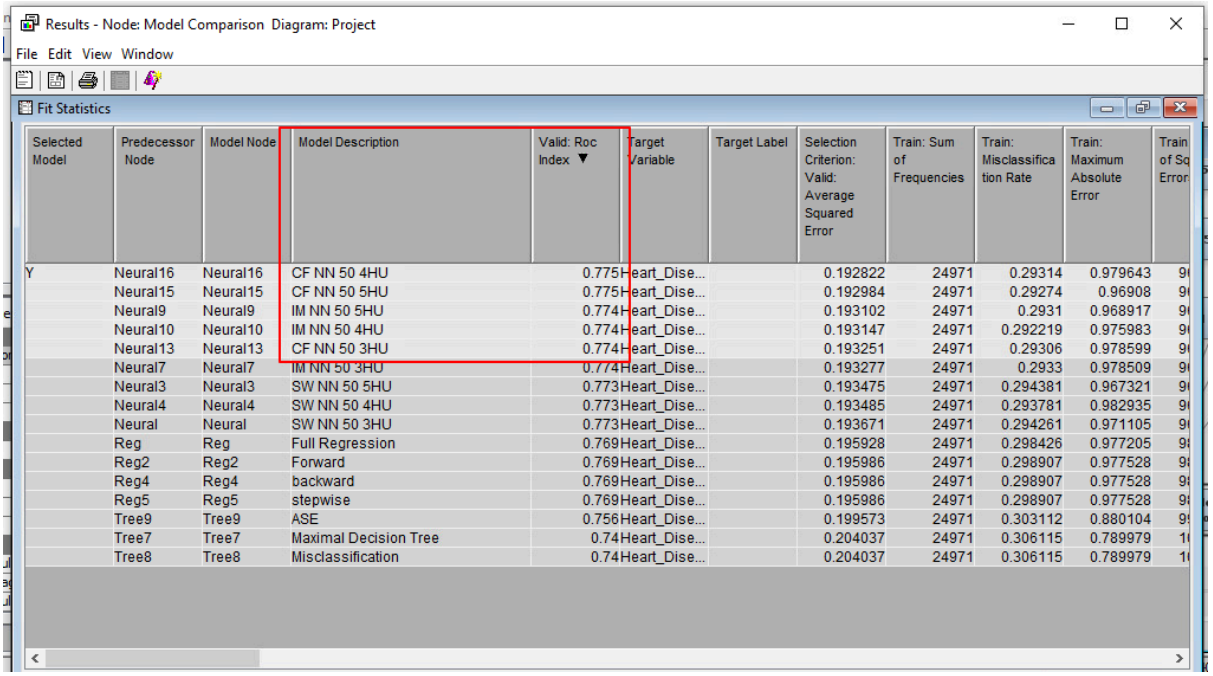


Figure 8.3 - ROC chart of Model Comparison

All models showed a high ROC index; among them CF NN 50 4HU, CF NN 50 5HU, IM NN 50 5HU, IM NN 50 4HU and CF NN 50 3HU showed an index of 0.774 and more



Selected Model	Predecessor Node	Model Node	Model Description	Valid: Roc Index ▼	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train of Sq Error
Y	Neural16	Neural16	CF NN 50 4HU	0.775	Heart_Dise...		0.192822	24971	0.29314	0.979643	9(
	Neural15	Neural15	CF NN 50 5HU	0.775	Heart_Dise...		0.192984	24971	0.29274	0.96908	9(
	Neural9	Neural9	IM NN 50 5HU	0.774	Heart_Dise...		0.193102	24971	0.2931	0.968917	9(
	Neural10	Neural10	IM NN 50 4HU	0.774	Heart_Dise...		0.193147	24971	0.292219	0.975983	9(
	Neural13	Neural13	CF NN 50 3HU	0.774	Heart_Dise...		0.193251	24971	0.29306	0.978599	9(
	Neural7	Neural7	IM NN 50 3HU	0.774	Heart_Dise...		0.193277	24971	0.2933	0.978509	9(
	Neural3	Neural3	SW NN 50 5HU	0.773	Heart_Dise...		0.193475	24971	0.294381	0.967321	9(
	Neural4	Neural4	SW NN 50 4HU	0.773	Heart_Dise...		0.193485	24971	0.293781	0.982935	9(
	Neural	Neural	SW NN 50 3HU	0.773	Heart_Dise...		0.193671	24971	0.294261	0.971105	9(
	Reg	Reg	Full Regression	0.769	Heart_Dise...		0.195928	24971	0.298426	0.977205	9(
	Reg2	Reg2	Forward	0.769	Heart_Dise...		0.195986	24971	0.298907	0.977528	9(
	Reg4	Reg4	backward	0.769	Heart_Dise...		0.195986	24971	0.298907	0.977528	9(
	Reg5	Reg5	stepwise	0.769	Heart_Dise...		0.195986	24971	0.298907	0.977528	9(
	Tree9	Tree9	ASE	0.756	Heart_Dise...		0.199573	24971	0.303112	0.880104	9(
	Tree7	Tree7	Maximal Decision Tree	0.74	Heart_Dise...		0.204037	24971	0.306115	0.789979	1(
	Tree8	Tree8	Misclassification	0.74	Heart_Dise...		0.204037	24971	0.306115	0.789979	1(

Figure 8.4- Fit statistics - ROC index of Model Comparison

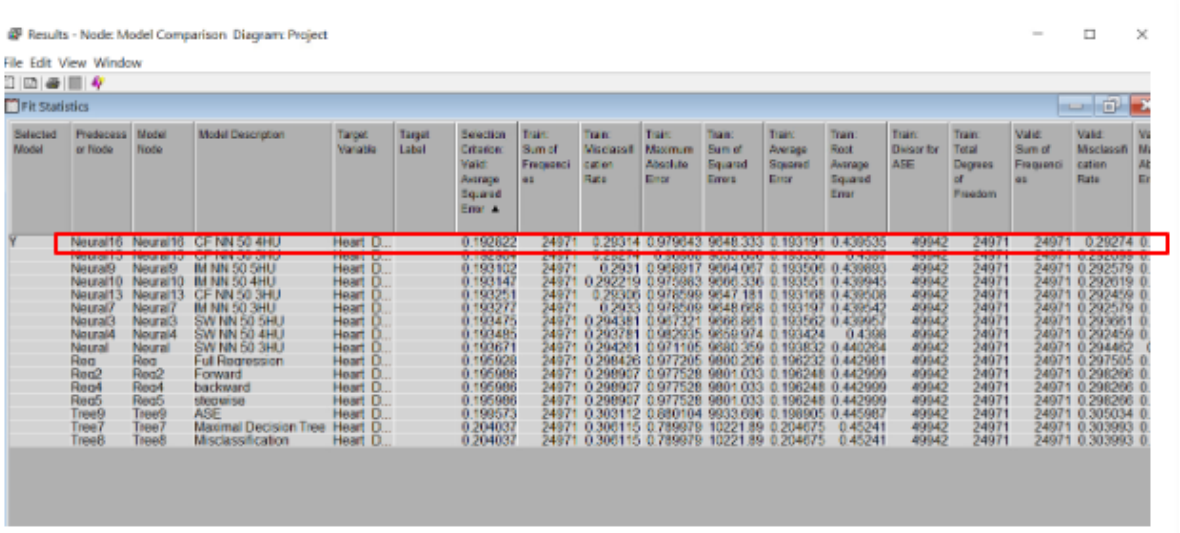


Figure 8.5- Fit Statistics of Model Comparison

The CF NN 50 4HU model has the lowest ASE, so it is considered the best model.

8. T-Test Results

As we rejected the “Age” variable due to its high correlation with other variables, we wanted to explore more on how it is associated with the other variables through t-tests using the SAS Enterprise Guide.

t Test

The TTEST Procedure

Variable: Age

General_Health=Excellent Checkup=5 or more years ago Exercise=No Skin_Cancer=No Other_Cancer=No Depression=No Diabetes=No Arthritis=No Sex=Female Smoking_History=No

N	Mean	Std Dev	Std Err	Minimum	Maximum
79	45.0886	17.2821	1.9444	21.0000	80.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
45.0886	41.2176	48.9596	17.2821	14.9441	20.4943

DF	t Value	Pr > |t|
78	23.19	<.0001

Figure 8.1- T-test results 1

From the above Figure 8.1., the p-value is less than 0.05 significance and therefore can be termed as statistically significant. Hence the null hypothesis was rejected and we were able to say that we had enough evidence to conclude that there is a correlation between age and other variables.

9. Conclusion

Summary

This study explored the various factors leading to heart disease using predictive modeling techniques. Several valuable insights were uncovered as a result of this analysis. We found out the significant variables contributing to the chances of heart disease are Arthritis, REP diabetes, Sex, Checkup, REP LOG REP Alcohol consumption, Skin cancer, Exercise REP LOG REP Fried potato consumption.

Unlike other modeling tools, neural networks performed best in grasping the complex relationships between the variables and the target. The CF NN 50 4HU model performed the best with the lowest validation ASE value. Using neural networks and stepwise regression, these models provided the most valid predictions of heart diseases.

Recommendations

Following are the recommendations for preventing heart diseases:

1. To be more physically active to avoid ailments like Arthritis.

2. Try to eat a more healthy and balanced diet. Avoid sugar and fatty food to reduce the chances of getting diabetes.
3. Reduce the consumption of Alcohol.

4. Visit the doctor for regular checkups once in 6 months or a year.
5. Avoid using items with carcinogens to avoid cancer risk.

10. References

1. World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation. 2023. https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf

2. Cardiovascular Diseases Risk Prediction Dataset.A 2021 BRFSS Dataset from CDC. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

11. Appendix

Data Source:

https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset.

Team Responsibilities

Kenil Gondaliya	SAS Procedures (Data Preparation, Decision Tree, Regression, Neural Network, Model Comparison) Report writing and review

Dhinesh Durairaju	SAS Procedures, Regression, Neural Network, Model Comparison Editing & Formatting Review
Rahul Lad	SAS Procedures (Data Preparation, Decision Tree, Regression, Neural Network, Model Comparison) Executive Summary, Review
Keerthana Sandhya	Data Preparation & Cleaning SAS Procedures, Report Writing, formatting, & review Introduction, Data Wrangling, File Import, Decision Tree Appendix
Neha Sharma	SAS Procedures, Report Writing, ,Logistics Regression,Data massaging, Full Regression & Review