BA 706 Applied Analytic Modeling
Group Project F22 – Group 8

Revised Version

Submitted to: David Parent

Ka Yee Lam, Angel 301251452

Chik Hung Tang, Ricky 301243794

Chun Yin Lee, Michael 301243243

July 12, 2023

Table of Contents

Introduction

Diabetes is a chronic disease in which the human's body does not produce enough insulin to ensure blood sugar is properly transformed as energy. With diabetes, one is potentially prone to multiple health problems such as heart disease, stroke, eye disease, etc (CDC, 2022). According to Centers for Disease Control and Prevention (CDC), statistics show that there is approximately 10% of women in the United States are diagnosed with diabetes during their pregnancies and there are chances that they would not be recovered from diabetes even after deliveries. There are varieties of measurements, for instance, glucose level, blood pressure, skin thickness, BMI, etc., being used to predict whether one has diabetes.

In the project, the dataset containing various measurements of females who are at least 21 years old is identified. Data modeling and assessment will be conducted using SAS Enterprise Miner to predict which indicators most contribute to the likeliness of being diagnosed with diabetes for female population whose age is over 21 years old. The project aims to provide clinicians visions identifying the relationships between measurement variables and the probability of diabetes, excluding glucose. The predictive model could be used by clinicians to pre-screen their patients in a more efficient way.

Questions:

1. Which independent variables are statically significant to diabetes?
2. How are the variables correlated with diabetes? What are their relationships?
3. Which type of clients have higher probability of being diagnosed with diabetes?

Data Setup and Exploration

The first time to begin with the project is to explore the dataset and understand each variable's modeling role and its appropriate measurement level. For the dataset, the following highlighted in blue are the edits made regarding the modeling role and measurement level.

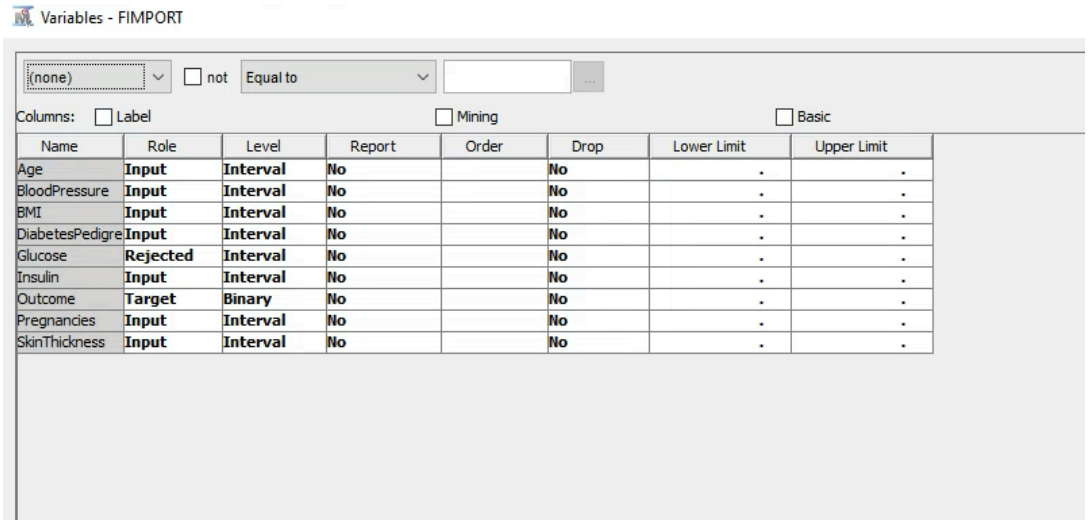
Data Table

Name	***Model Role***	***Measurement Level***	***Description***
Age	Input	Interval	Age (years)
BloodPressure	Input	Interval	Diastolic blood pressure (mm Hg)
BMI	Input	Interval	Body Mass Index (weight in kg/ (height in m)^2)
DiabetesPedigreeFunction	Input	Interval	The function which scores likelihood of diabetes based on family history
Glucose	*Rejected*	Interval	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Insulin	Input	Interval	2-hour serum insulin (mu U/ml)
Outcome	*Target*	*Binary*	Class variable 0 or 1, with 1='tested positive for diabetes'
Pregnancies	Input	Interval	Number of times pregnant
SkinThickness	Input	Interval	Triceps skin fold thickness (mm)

Notes:

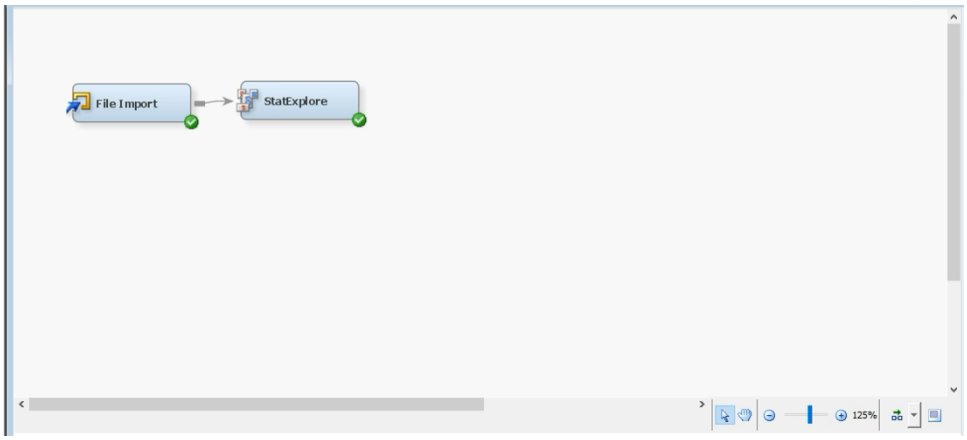
- Glucose is rejected since the variable is too correlated with the outcome (diabetes).
- "Outcome" is set as the binary target variable.
- Variables except "Outcome" remains as "Interval" as they are the numeric data type.
- No nominal data type was found in the dataset.

In SAS Enterprise Miner, variables are edited according to the table above as shown in below screenshot:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
BloodPressure	Input	Interval	No		No	.	.
BMI	Input	Interval	No		No	.	.
DiabetesPedigre	Input	Interval	No		No	.	.
Glucose	Rejected	Interval	No		No	.	.
Insulin	Input	Interval	No		No	.	.
Outcome	Target	Binary	No		No	.	.
Pregnancies	Input	Interval	No		No	.	.
SkinThickness	Input	Interval	No		No	.	.

It is important to first understand your data before performing types of models. To start with, add the "**Stats Explore**" node in SAS Enterprise Miner as shown in below screenshot.



Results are shown as follows:

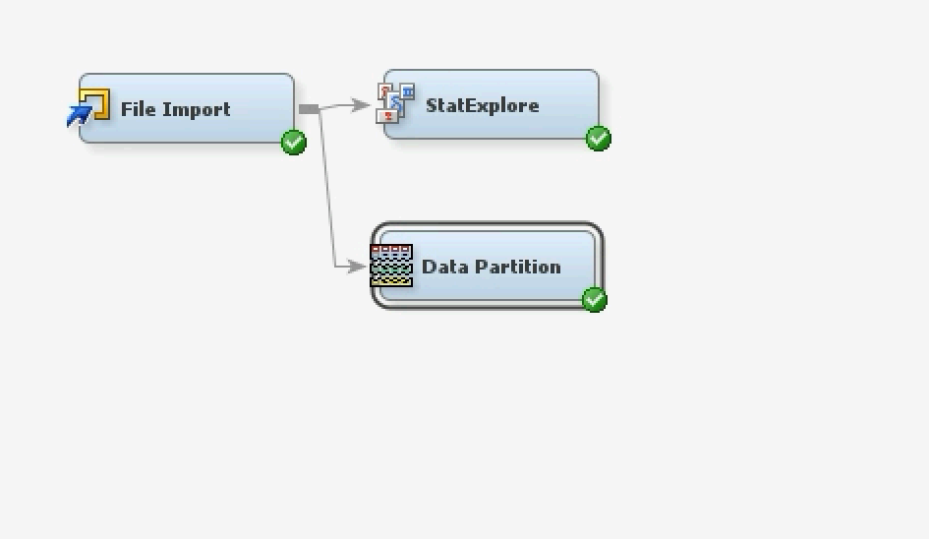
Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation
TRAIN	Outcome	0	Pregnancies	2	0	500	0	13	3.298	3.017185	1.114105	0.687268	INPUT	Pregnancies	-0.14227
TRAIN	Outcome	1	Pregnancies	4	0	268	0	17	4.865672	3.741239	0.503749	-0.44164	INPUT	Pregnancies	0.265437
TRAIN	Outcome	0	Insulin	38	0	500	0	744	68.792	98.86529	2.498741	9.457788	INPUT	Insulin	-0.13794
TRAIN	Outcome	1	Insulin	0	0	268	0	846	100.3358	138.6891	1.843831	4.360493	INPUT	Insulin	0.257349
TRAIN	Outcome	0	DiabetesPe...	0.336	0	500	0.078	2.329	0.429734	0.299085	2.006242	6.120934	INPUT	DiabetesPe...	-0.08931
TRAIN	Outcome	1	DiabetesPe...	0.447	0	268	0.088	2.42	0.5505	0.372354	1.722373	4.559083	INPUT	DiabetesPe...	0.166619
TRAIN	Outcome	0	Age	27	0	500	21	81	31.19	11.66765	1.571609	1.963428	INPUT	Age	-0.0617
TRAIN	Outcome	1	Age	36	0	268	21	70	37.06716	10.96825	0.581646	-0.34794	INPUT	Age	0.115108
TRAIN	Outcome	0	BMI	30	0	500	0	57.3	30.3042	7.689855	-0.6659	3.058286	INPUT	BMI	-0.05277
TRAIN	Outcome	1	BMI	34.2	0	268	0	67.1	35.14254	7.262967	.0005968	4.763822	INPUT	BMI	0.098459
TRAIN	Outcome	0	SkinThickn...	21	0	500	0	60	19.664	14.88995	0.031155	-1.03821	INPUT	SkinThickn...	-0.04248
TRAIN	Outcome	1	SkinThickn...	27	0	268	0	99	22.16418	17.67971	0.11591	-0.20928	INPUT	SkinThickn...	0.07926
TRAIN	Outcome	0	BloodPress...	70	0	500	0	122	68.184	18.06308	-1.80982	5.686977	INPUT	BloodPress...	-0.01333
TRAIN	Outcome	1	BloodPress...	74	0	268	0	114	70.82463	21.49181	-1.94363	4.699552	INPUT	BloodPress...	0.024877

From the StatsExplore, we have understood that there is no missing values in the analysis data set. While all variables have an appropriate range, the skewness is one thing that we need to be aware of and to be addressed in later stages.

Data Partition

To prevent generalization in prediction, it is pivotal to have competing models. As per SAS Enterprise Miner screenshot below, we use the **Data Partition** tab to allocate 50% of the data to training, and another 50% to validation from our raw analysis data. As such, we have cases to build the model and to adjust and compare models.



Outcome has shown in the following indicating how data is partitioned:

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Outcome	0	0	500	65.1042	
Outcome	1	1	268	34.8958	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Outcome	0	0	250	65.2742	
Outcome	1	1	133	34.7258	

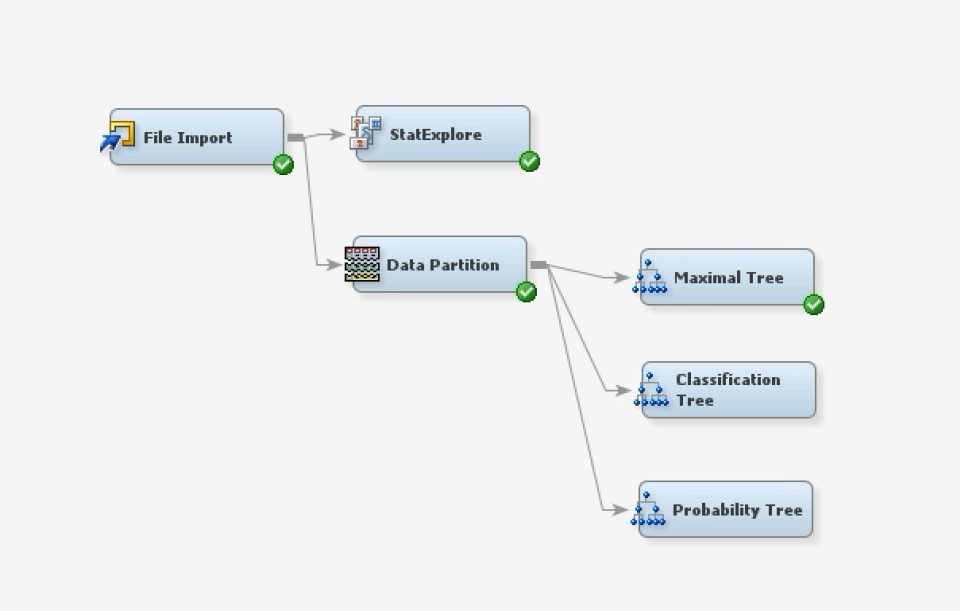
Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Outcome	0	0	250	64.9351	
Outcome	1	1	135	35.0649	

Now that the data has been partitioned, predictive models are to be explored in the following stages.

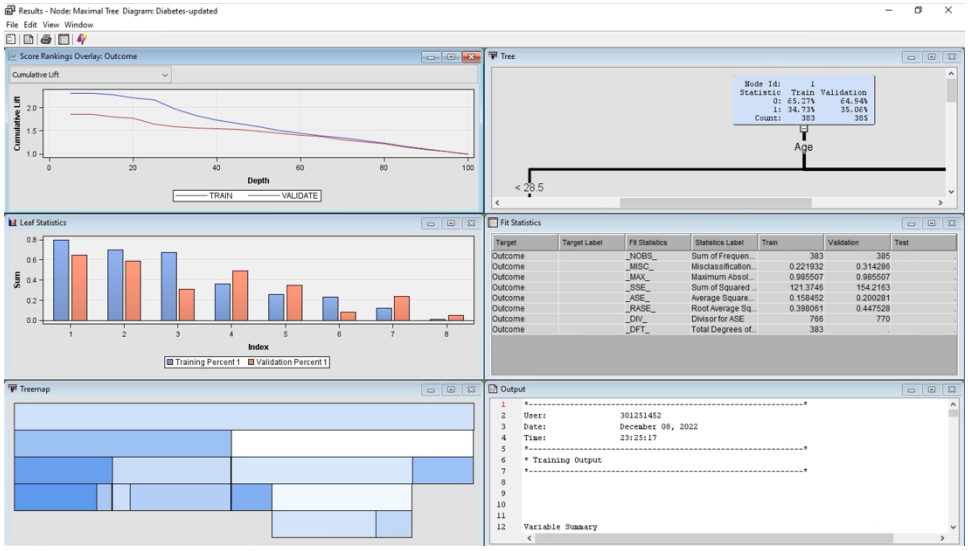
Decision Trees

In SAS Enterprise Miner, we have connected the **Data Partition** node to the **Decision Tree** (i.e., Maximal Tree, Classification Tree, Probability Tree) nodes to construct the tree models.



Maximal Tree

To create a Maximal Tree, we have selected 'Largest' for method and 'Decision' for assessment measure in the subtree section. Below screenshot shows the results, including Fit Statistics, Output, and Tree.



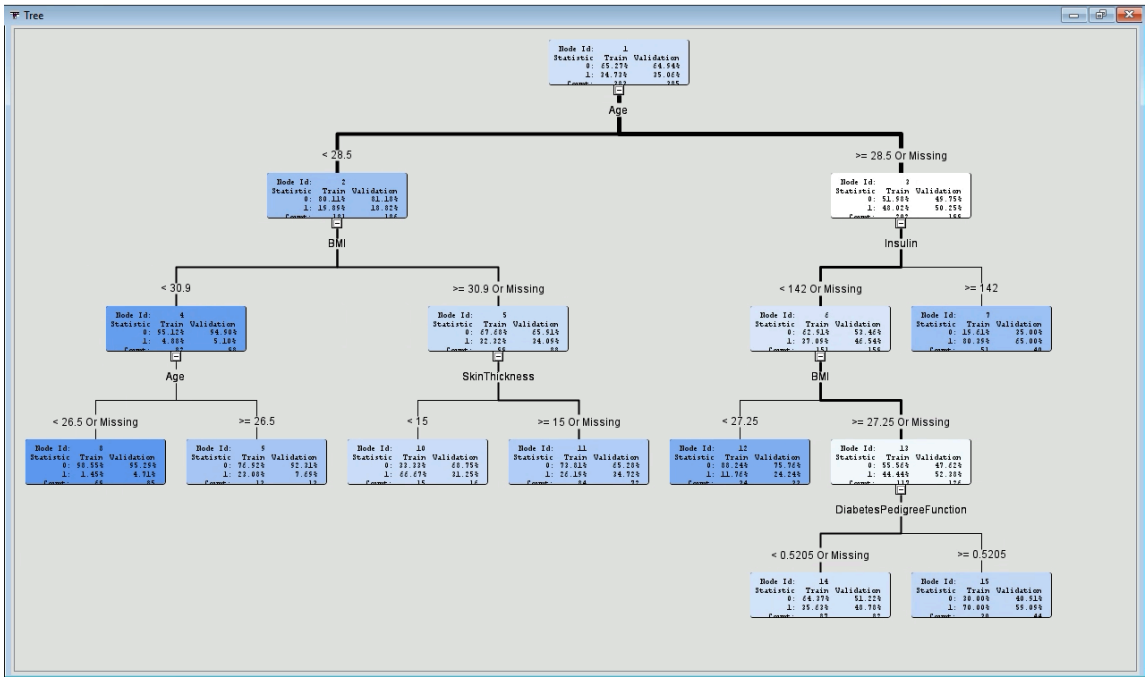
On Fit Statistics, SAS Miner has calculated the ASE and the misclassification rate for maximal tree are 0.200281 and 0.314286 respectively.

Fit Statistics

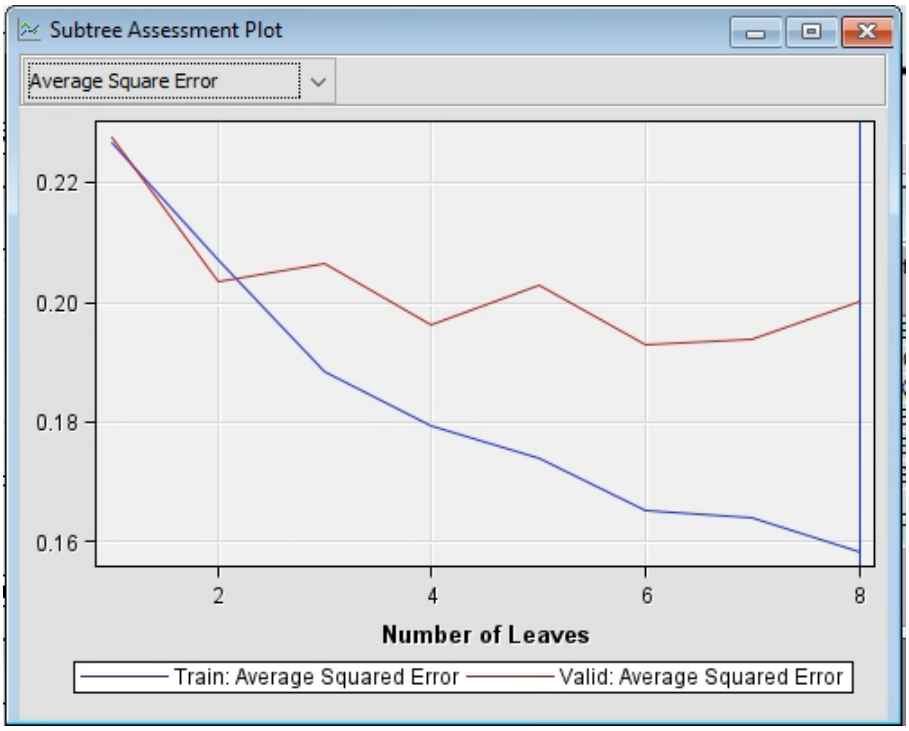
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_NOBS_	Sum of Frequencies	383	385	
Outcome		_MISC_	Misclassification Rate	0.221932	0.314286	
Outcome		_MAX_	Maximum Absolute Error	0.985507	0.985507	
Outcome		_SSE_	Sum of Squared Errors	121.3746	154.2163	
Outcome		_ASE_	Average Squared Error	0.158452	0.200281	
Outcome		_RASE_	Root Average Squared Error	0.398061	0.447528	
Outcome		_DIV_	Divisor for ASE	766	770	
Outcome		_DFT_	Total Degrees of Freedom	383		

For the tree diagram, below are our observations:

Starting from the root node, the maximal tree has eight leaves and indicates that people who are 28.5 years old have 18.82% chances of having diabetes. The probability is even smaller to be diagnosed with diabetes if their BMI index is less than 30.9, which is only 5.1%. The most optimal node would be the bottom left box, only 4.71% chance of getting diabetes when people are younger than 26.5 years old and less than 30.9 BMI. This tree model shows that age and BMI are negatively correlated with diabetes, the younger and lower the BMI, the less likely the chance of getting diabetes.

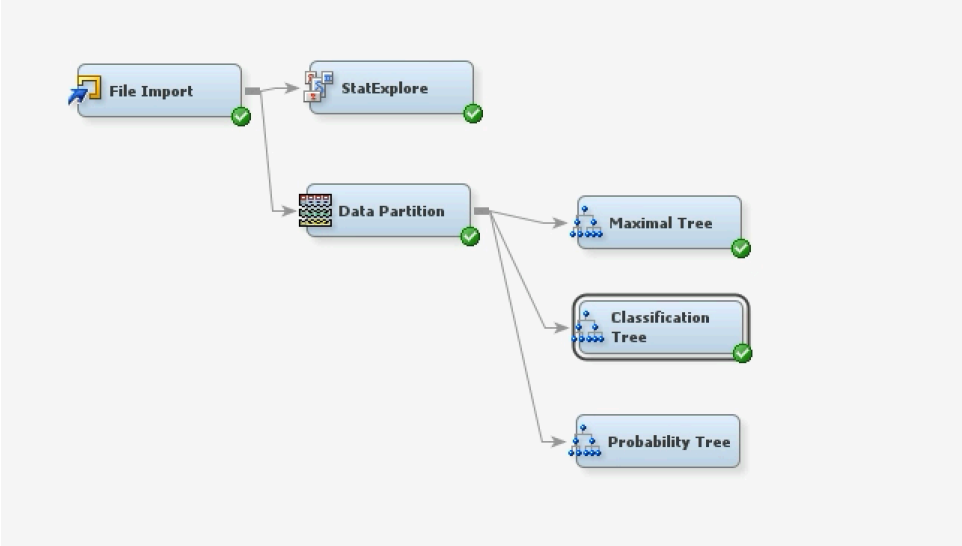


And the following screenshot shows the Average Square Error graph:

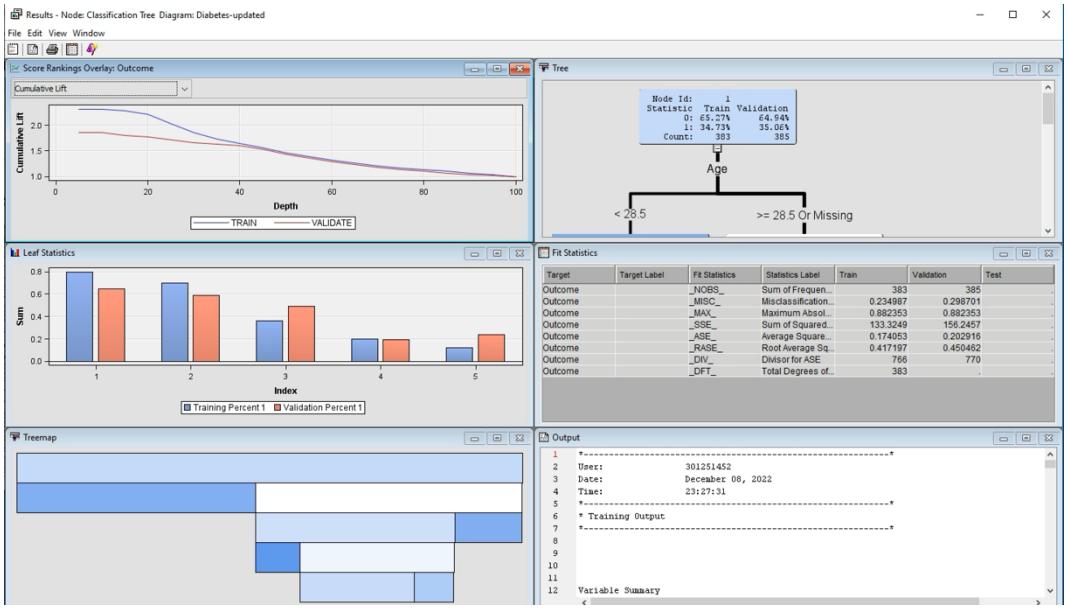


Classification Tree

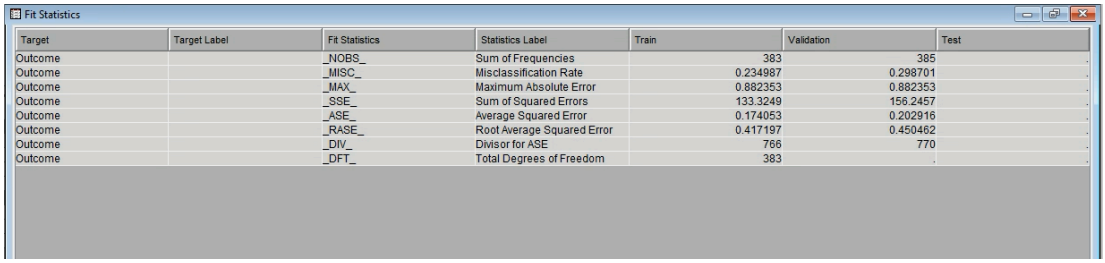
To construct a classification tree, we have selected 'Assessment' for method and 'Misclassification' for assessment measure in the subtree section.



Similar with Maximal Tree, below screenshot shows an overview of the results:

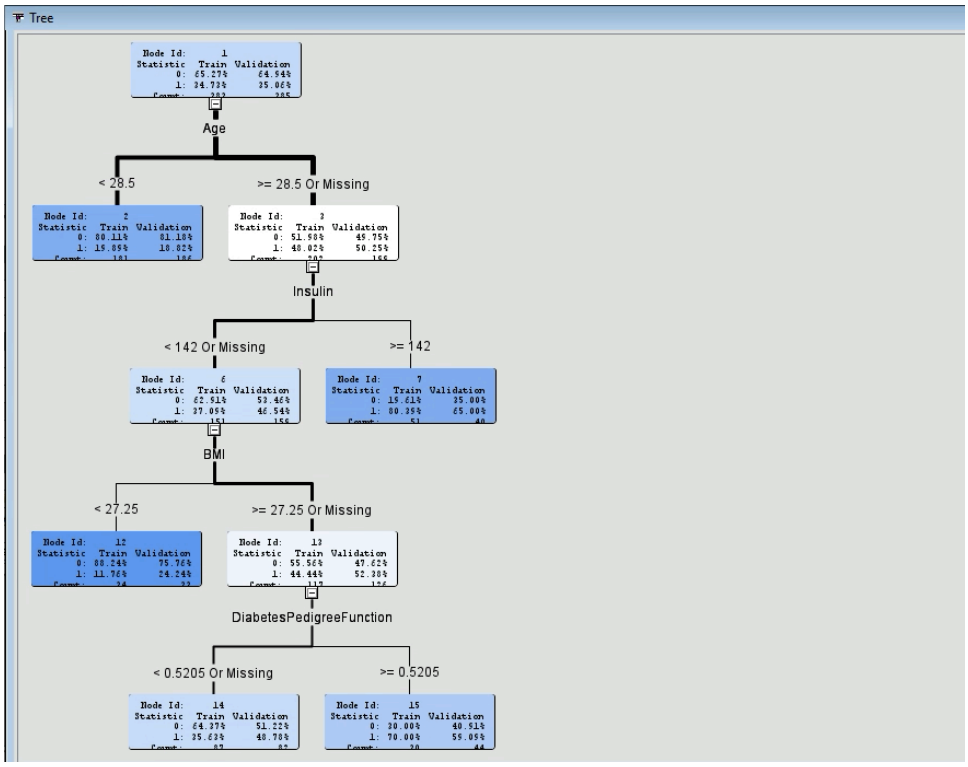


On Fit Statistics, SAS Miner has calculated the ASE and the misclassification rate for classification tree are 0.202916 and 0.298701 respectively.

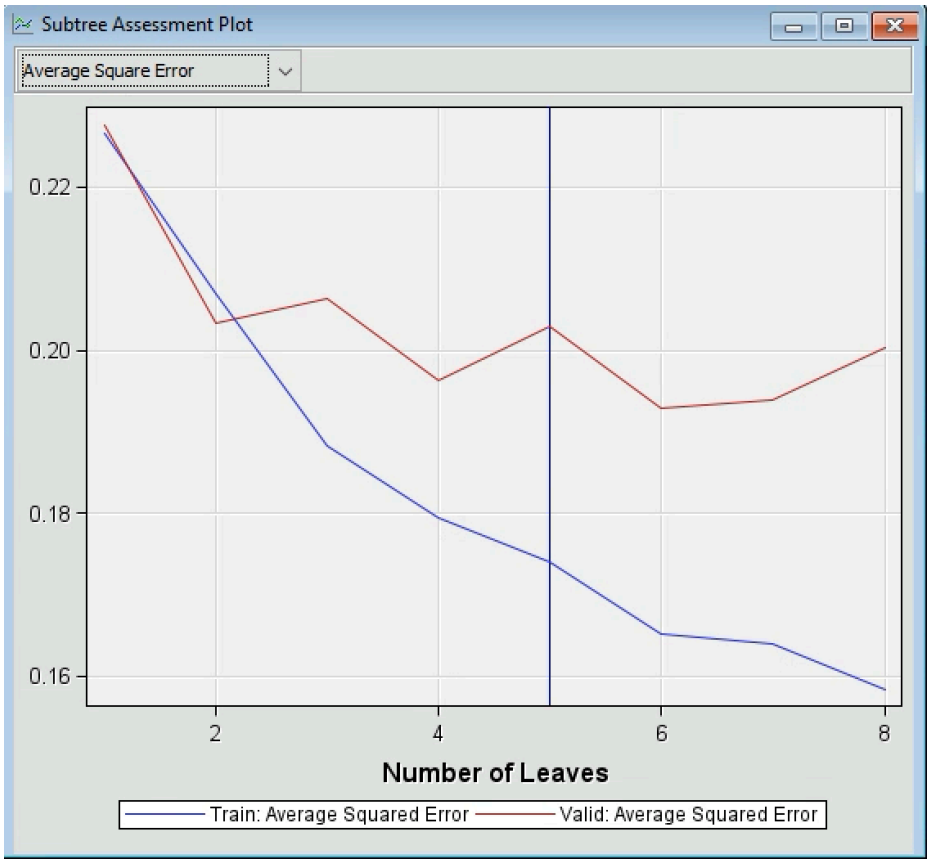


For the tree diagram, we have the following interpretations:

The classification tree model has five leaves and explains the negative correlations between age, insulin level, BMI, diabetes pedigree function, and the likelihood of having diabetes. All these independent variables indicate the lower numbers represent less chance of getting diabetes, such as younger age, lower insulin level, lighter weight, and fewer diabetes symptoms in family history, etc. Top three splits indicate only 24.24% chance of getting diabetes if people have less than 27.25 BMI, fewer than 142 insulin levels, and younger than 28.5 years old.

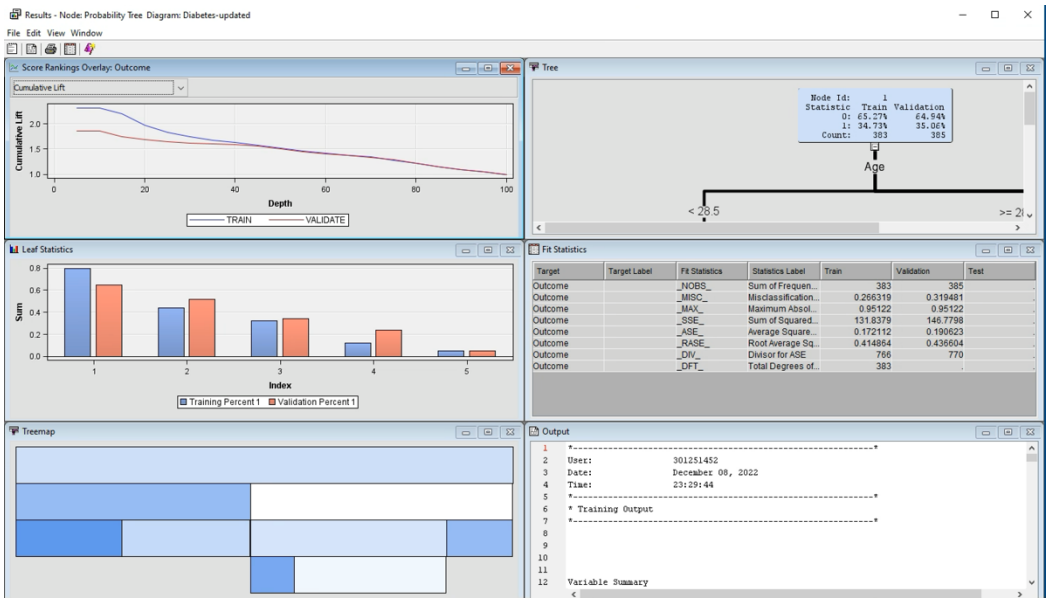


The following graph shows the average square error for the classification tree:

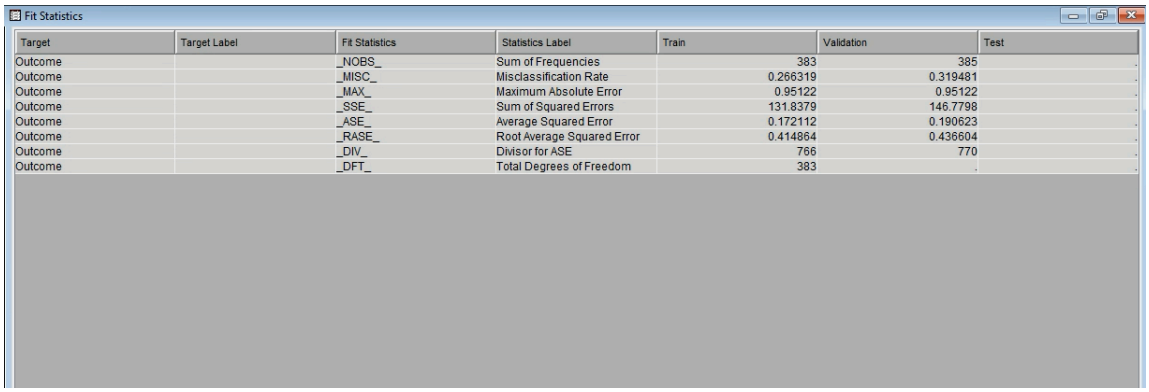


Probability Tree

To construct a probability tree, we have selected 'Assessment' for method and 'Average Square Error' for assessment measure in the subtree section. The following screenshot shows an overview of the results.



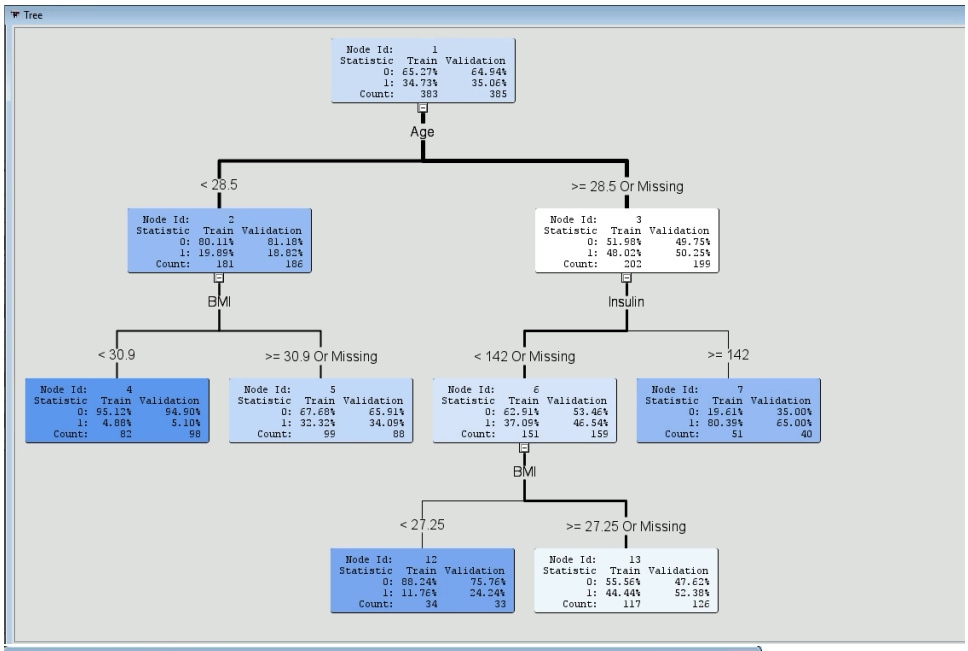
On Fit Statistics, SAS Miner has calculated the ASE and the misclassification rate for classification tree are 0.190623 and 0.319481 respectively.

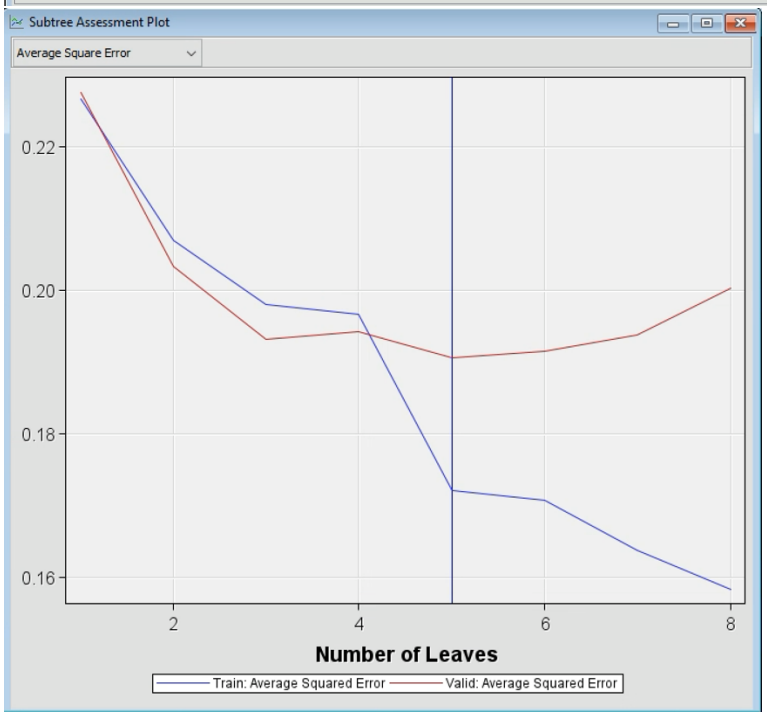


For the tree diagram, below is our observations:

This decision tree model also has five leaves and represents the reverse relationship between all independent variables and the likelihood of having diabetes. The story behind the tree nodes does make lots of sense because generally the lower numbers of different variables

mean a healthier lifestyle or physique. The healthier you get also indicates the less chance of getting diabetes symptoms. The most optimal node would be the middle dark blue box, indicating only a 5.10% chance of being diagnosed with diabetes if people have less than 30.9 BMI and younger than 28.5 years old.





Summary

Types of Decision Tree	Average Square Error
Maximal Tree	0.200281
Classification Tree	0.202916
Probability Tree (**Optimal Tree)	0.190623 (**lowest)

Out of the three types of decision trees, Probability Tree is our optimal tree in which it has the lowest average square error. This implies that it has the lowest error between the train and validation dataset.

Impute

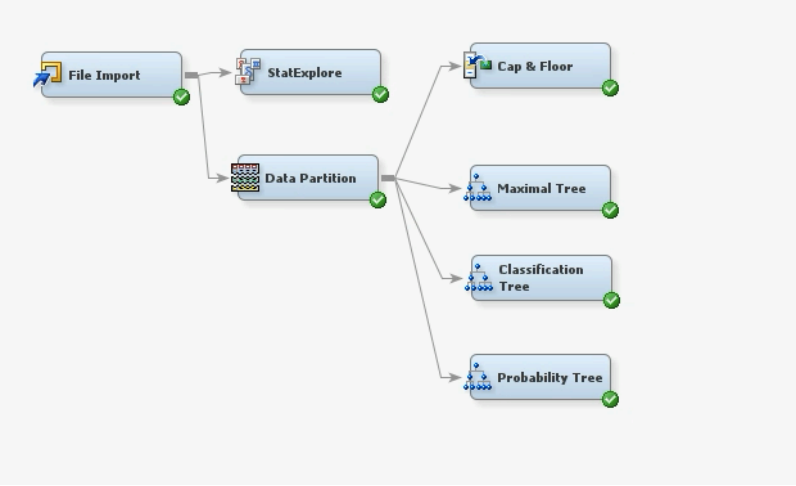
Per previously discussed, no missing values were found in the analysis data. Hence, no imputation is required.

Adjust Outliers

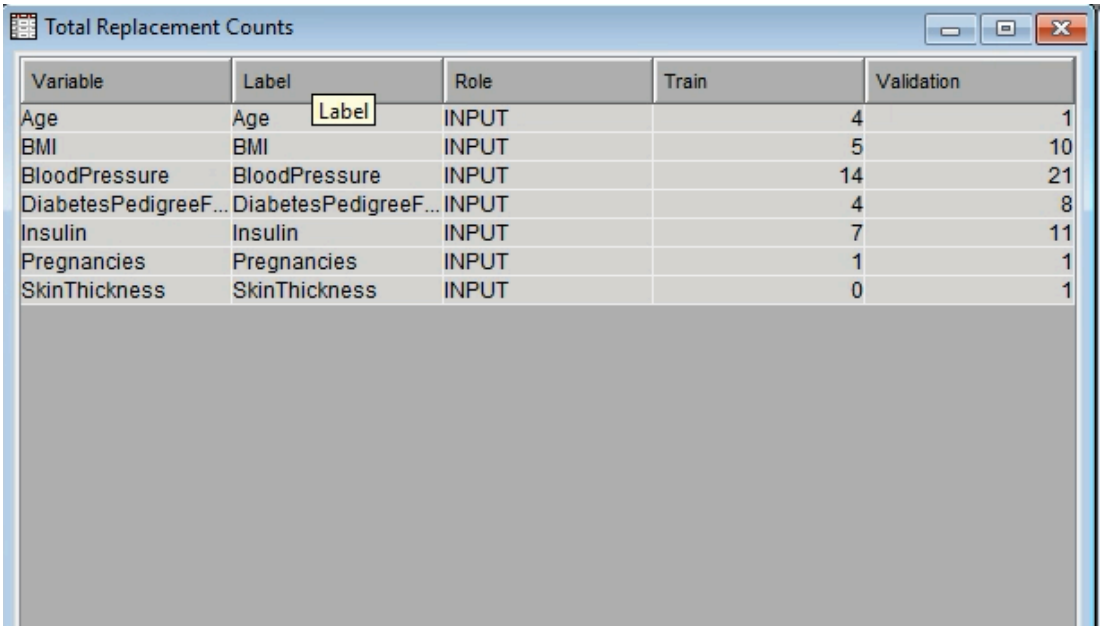
Unlike decision trees, regression models are very sensitive to different missing or extreme values, and skewness. Outcomes would become distorted if no actions were taken. Before proceeding with constructing regression models, it is required to spot the outliers and skewness of each variable to ensure our later analysis is unbiased and accurate. Since we do not have any missing values in this dataset, we can now only focus on tackling outliers and reducing the skewness of the corresponding variables.

Replacement— Cap & Floor

To deal with outliers, we have connected **Data Partition** node to **Replacement – Cap & Floor** node.



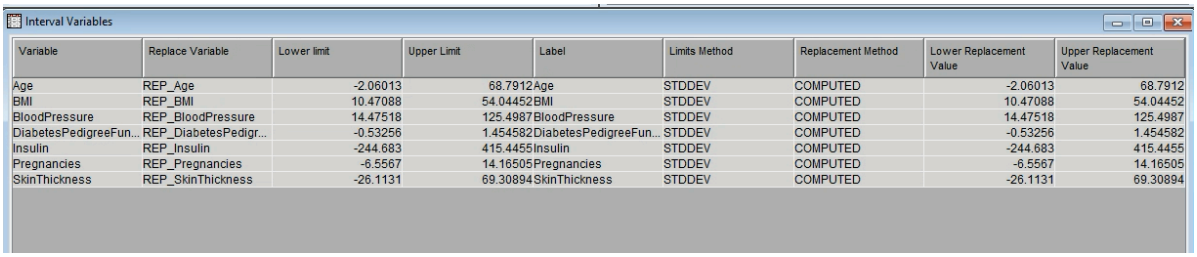
The below screenshot explains the replaced variables and their counts.



Total Replacement Counts

Variable	Label	Role	Train	Validation
Age	Age	INPUT	4	1
BMI	BMI	INPUT	5	10
BloodPressure	BloodPressure	INPUT	14	21
DiabetesPedigreeF...	DiabetesPedigreeF...	INPUT	4	8
Insulin	Insulin	INPUT	7	11
Pregnancies	Pregnancies	INPUT	1	1
SkinThickness	SkinThickness	INPUT	0	1

Here is the screenshot of the names of the replaced variables for references:

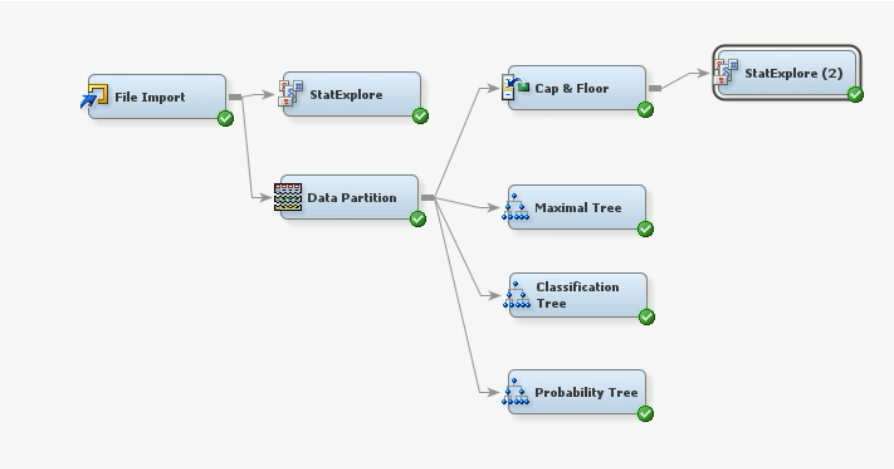


Interval Variables

Variable	Replace Variable	Lower limit	Upper Limit	Label	Limits Method	Replacement Method	Lower Replacement Value	Upper Replacement Value
Age	REP_Age	-2.06013	68.7912	Age	STDDEV	COMPUTED	-2.06013	68.7912
BMI	REP_BMI	10.47088	54.04452	BMI	STDDEV	COMPUTED	10.47088	54.04452
BloodPressure	REP_BloodPressure	14.47518	125.4987	BloodPressure	STDDEV	COMPUTED	14.47518	125.4987
DiabetesPedigreeFun...	REP_DiabetesPedigr...	-0.53256	1.454582	DiabetesPedigreeFun...	STDDEV	COMPUTED	-0.53256	1.454582
Insulin	REP_Insulin	-244.683	415.4455	Insulin	STDDEV	COMPUTED	-244.683	415.4455
Pregnancies	REP_Pregnancies	-6.5567	14.16505	Pregnancies	STDDEV	COMPUTED	-6.5567	14.16505
SkinThickness	REP_SkinThickness	-26.1131	69.30894	SkinThickness	STDDEV	COMPUTED	-26.1131	69.30894

Transform

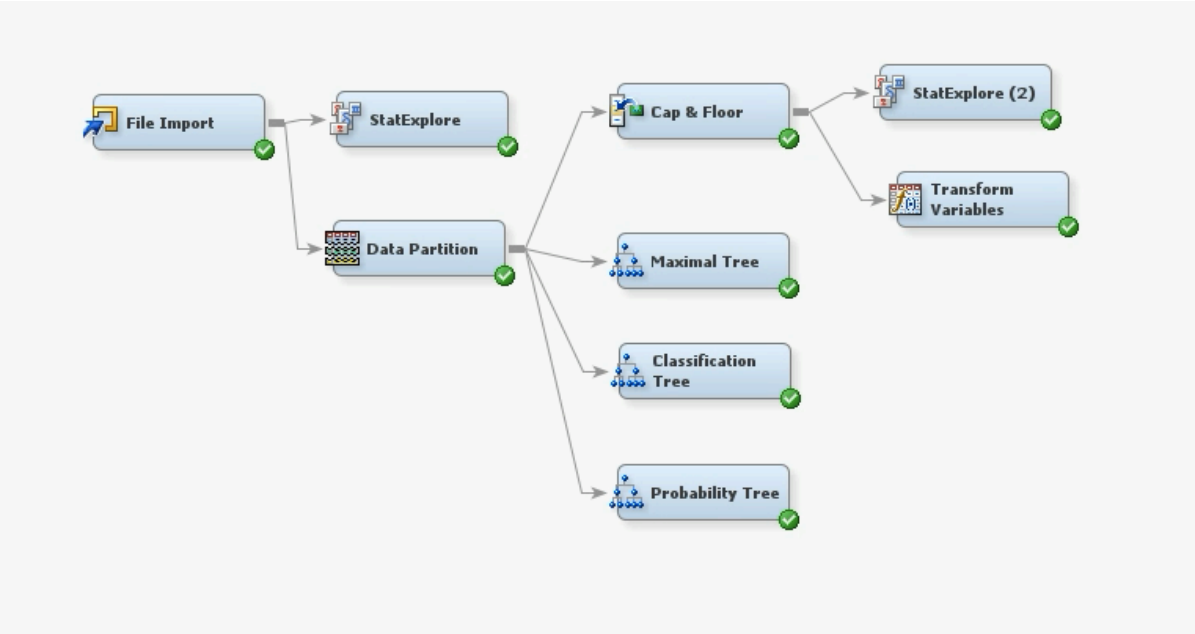
Before dealing with skewness, we connected the **Cap & Floor** node to another **StatExplore** node to understand the skewness of our variables.



Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role
TRAIN	Outcome	0	REP_Insulin	50	0	250	0	415.4455	69.72756	85.79975	1.512276	2.619402	INPUT
TRAIN	Outcome	1	REP_Insulin	96	0	133	0	415.4455	107.9566	117.8631	0.868111	-0.04343	INPUT
TRAIN	Outcome	0	REP_Pregnancies	2	0	250	0	13	3.32	3.13959	1.18464	0.796089	INPUT
TRAIN	Outcome	1	REP_Pregnancies	4	0	133	0	14.16505	4.69297	3.765333	0.466736	-0.66164	INPUT
TRAIN	Outcome	0	REP_DiabetesPedigreeFunction	0.315	0	250	0.085	1.454582	0.409193	0.270468	1.60803	2.904137	INPUT
TRAIN	Outcome	1	REP_DiabetesPedigreeFunction	0.412	0	133	0.088	1.454582	0.540603	0.344542	1.005324	0.080123	INPUT
TRAIN	Outcome	0	REP_Age	27	0	250	21	68.7912	31.72549	11.9475	1.455754	1.374488	INPUT
TRAIN	Outcome	1	REP_Age	34	0	133	21	68.7912	36.32174	10.49368	0.652813	-0.00719	INPUT
TRAIN	Outcome	0	REP_SkinThickness	23	0	250	0	60	20.656	15.40229	-0.00566	-1.03925	INPUT
TRAIN	Outcome	1	REP_SkinThickness	27	0	133	0	63	23.36842	16.72164	-0.25111	-1.14732	INPUT
TRAIN	Outcome	0	REP_BMI	30.4	0	250	10.47088	47.9	30.90777	6.535446	0.073328	-0.08642	INPUT
TRAIN	Outcome	1	REP_BMI	34.2	0	133	10.47088	54.04452	34.89293	6.41576	0.208463	1.452834	INPUT
TRAIN	Outcome	0	REP_BloodPressure	70	0	250	14.47518	122	69.24721	15.90716	-0.97725	3.617232	INPUT
TRAIN	Outcome	1	REP_BloodPressure	74	0	133	14.47518	114	72.90114	17.64327	-1.39937	3.701792	INPUT

Looking at the results, the skewness of Insulin, Pregnancies, DiabetesPedigreeFunction, Age, and BloodPressure is outside of the normal range (–1 and 1). In order to reduce their skewness, we have performed modification by adding the **Transform Variables** node afterwards as shown per below SAS Miner screenshot.



Below screenshot shows the method and the variables selected for Log transformation.

Variables - Trans

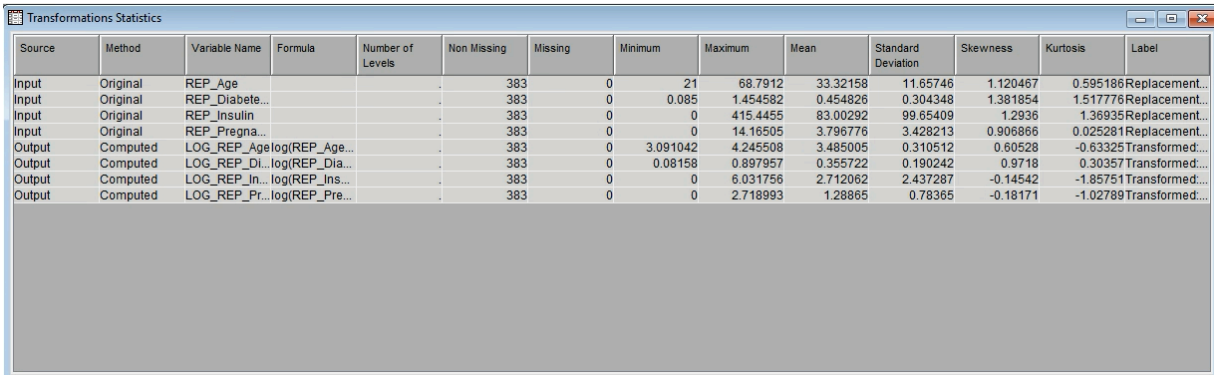
(none) not Equal to

Columns: Label Mining Basic

Name	Method	Number of Bins	Role	Level
Pregnancies	**Default**	**4**	Rejected	Interval
SkinThickness	**Default**	**4**	Rejected	Interval
Outcome	**Default**	**4**	Target	Binary
REP_BloodPressure	**Default**	**4**	Input	Interval
REP_SkinThickness	**Default**	**4**	Input	Interval
REP_BMI	**Default**	**4**	Input	Interval
BMI	**Default**	**4**	Rejected	Interval
BloodPressure	**Default**	**4**	Rejected	Interval
Age	**Default**	**4**	Rejected	Interval
Insulin	**Default**	**4**	Rejected	Interval
DiabetesPedigreeFunction	**Default**	**4**	Rejected	Interval
REP_Age	**Log**	**4**	Input	Interval
REP_DiabetesPedigreeFunction	**Log**	**4**	Input	Interval
REP_Insulin	**Log**	**4**	Input	Interval
REP_Pregnancies	**Log**	**4**	Input	Interval

We have selected 4 variables with high skewness namely Insulin, Pregnancies, DiabetesPedigreeFunction and Age for a log transformation. It is noticeable that when we took log on BloodPressure, the skewness was not improved but become worse (from –1.39 to over –3). To address the issue, we also tried using standardized method, but it did not work well as it creates the concern of having the uncertain range size. It becomes more complicated to interpret for our clinical clients using the standardized method as each interval is uneven. To keep it interpretable, we have decided to keep BloodPressure as Default to prevent distorted and inaccurate outcomes.

Below screenshot shows the updated skewness.


Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	REP_Age		.	383	0	21	68.7912	33.32158	11.65746	1.120467	0.595186	Replacement...
Input	Original	REP_Diabete...		.	383	0	0.085	1.454582	0.454826	0.304348	1.381854	1.517776	Replacement...
Input	Original	REP_Insulin		.	383	0	0	415.4455	83.00292	99.65409	1.2936	1.36935	Replacement...
Input	Original	REP_Pregna...		.	383	0	0	14.16505	3.796776	3.428213	0.906866	0.025281	Replacement...
Output	Computed	LOG_REP_Age	log(REP_Age...	.	383	0	3.091042	4.245508	3.485005	0.310512	0.60528	-0.63325	Transformed:...
Output	Computed	LOG_REP_Di...	log(REP_Dia...	.	383	0	0.08158	0.897957	0.355722	0.190242	0.9718	0.30357	Transformed:...
Output	Computed	LOG_REP_In...	log(REP_Ins...	.	383	0	0	6.031756	2.712062	2.437287	-0.14542	-1.85751	Transformed:...
Output	Computed	LOG_REP_Pr...	log(REP_Pre...	.	383	0	0	2.718993	1.28865	0.78365	-0.18171	-1.02789	Transformed:...

Now the skewness of all four logged variables are back to normal range of –1 and 1.

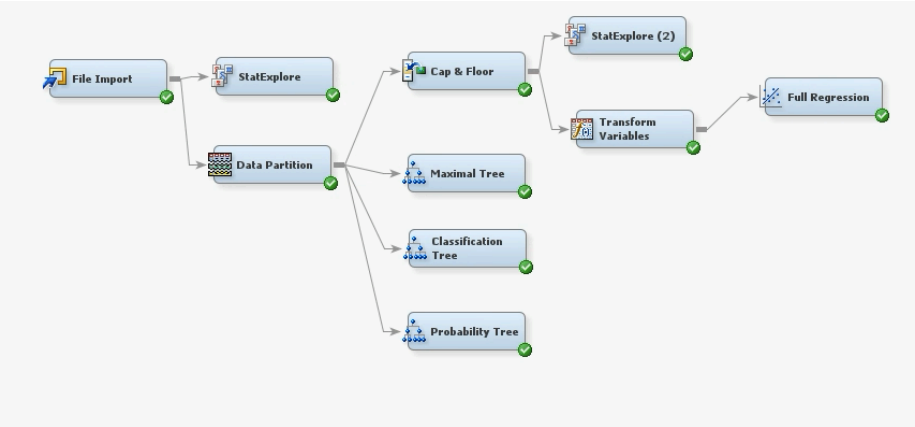
Replace Dummies (collapse)

Not required as there are no nominal and categorical data types in the analysis dataset.

Full Regression

After performing the above-mentioned data preparation, we can now create the following four regression models to figure out which one has the lowest average square error and the highest efficiency in explaining our dataset.

First, "**Full Regression**" node is added after "Transform Variables" in the SaS Miner:



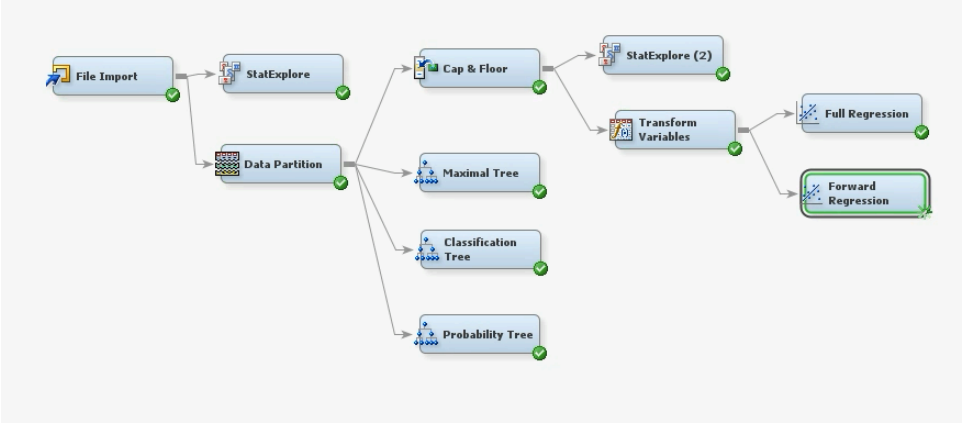
Full Regression- ASE (0.182999)

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_AIC_	Akaike's Information Criterion	441.2292	.	.
Outcome		_ASE_	Average Squared Error	0.188842	0.182999	.
Outcome		_AVERR_	Average Error Function	0.55513	0.537879	.
Outcome		_DFE_	Degrees of Freedom for Error	375	.	.
Outcome		_DFM_	Model Degrees of Freedom	8	.	.
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_ERR_	Error Function	425.2292	414.1667	.
Outcome		_FPE_	Final Prediction Error	0.1969	.	.
Outcome		_MAX_	Maximum Absolute Error	0.944128	0.942666	.
Outcome		_MSE_	Mean Square Error	0.192871	0.182999	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_NW_	Number of Estimate Weights	8	.	.
Outcome		_RASE_	Root Average Sum of Squares	0.43456	0.427783	.
Outcome		_RFPE_	Root Final Prediction Error	0.443734	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.439171	0.427783	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	472.8135	.	.
Outcome		_SSE_	Sum of Squared Errors	144.6533	140.9089	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_MISC_	Misclassification Rate	0.289817	0.293506	.

Sequential

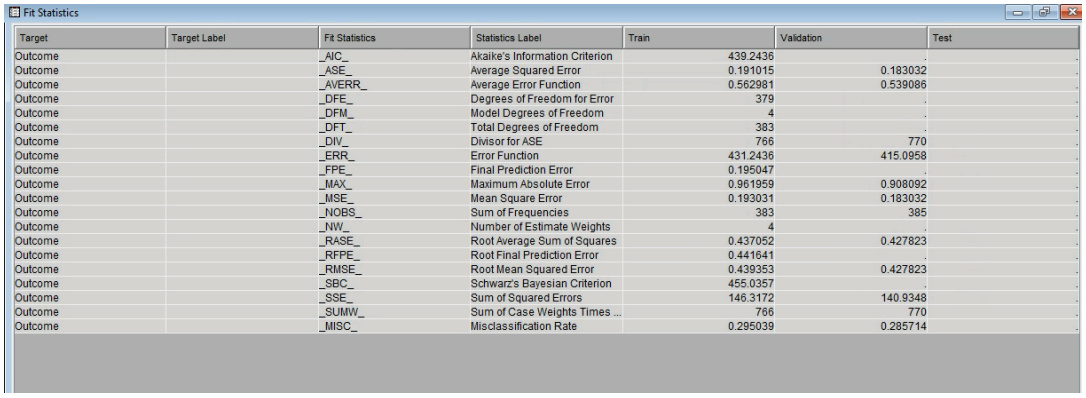
Forward Regression



Adding "**Forward Regression**" node in SaS Miner.

Forward Regression- ASE (0.183032)

The following screenshot shows an overview of the fit statistics for forward regression and the Average Squared Error is 0.183032.

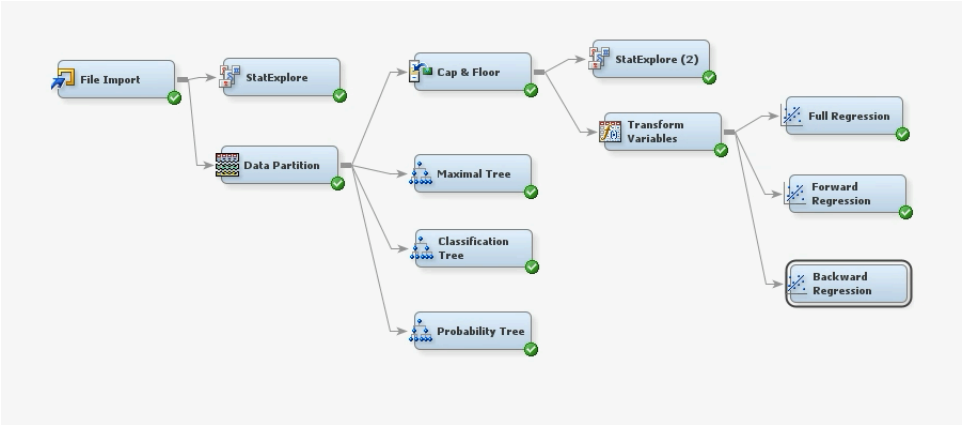


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_AIC_	Akaike's Information Criterion	439.2436	.	.
Outcome		_ASE_	Average Squared Error	0.191015	0.183032	.
Outcome		_AVERR_	Average Error Function	0.562981	0.539086	.
Outcome		_DFE_	Degrees of Freedom for Error	379	.	.
Outcome		_DFM_	Model Degrees of Freedom	4	.	.
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_ERR_	Error Function	431.2436	415.0958	.
Outcome		_FPE_	Final Prediction Error	0.195047	.	.
Outcome		_MAX_	Maximum Absolute Error	0.961959	0.908092	.
Outcome		_MSE_	Mean Square Error	0.193031	0.183032	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_NW_	Number of Estimate Weights	4	.	.
Outcome		_RASE_	Root Average Sum of Squares	0.437052	0.427823	.
Outcome		_RFPE_	Root Final Prediction Error	0.441641	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.439353	0.427823	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	455.0357	.	.
Outcome		_SSE_	Sum of Squared Errors	146.3172	140.9348	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_MISC_	Misclassification Rate	0.295039	0.285714	.

Backward Regression

Adding "**Backward Regression**" node in SaS Miner:



Backward Regression – ASE (0.182483)

The screenshot shows an overview of the fit statistics for backward regression and the Average Squared Error is 0.182483.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_AIC_	Akaike's Information Criterion	439.0902	.	.
Outcome		_ASE_	Average Squared Error	0.190253	0.182483	.
Outcome		_AVERR_	Average Error Function	0.56017	0.535932	.
Outcome		_DFE_	Degrees of Freedom for Error	378	.	.
Outcome		_DFM_	Model Degrees of Freedom	5	.	.
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_ERR_	Error Function	429.0902	412.6678	.
Outcome		_FPE_	Final Prediction Error	0.195286	.	.
Outcome		_MAX_	Maximum Absolute Error	0.953123	0.912566	.
Outcome		_MSE_	Mean Square Error	0.19277	0.182483	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_NW_	Number of Estimate Weights	5	.	.
Outcome		_RASE_	Root Average Sum of Squares	0.43618	0.42718	.
Outcome		_RFPE_	Root Final Prediction Error	0.441912	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.439056	0.42718	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	458.8303	.	.
Outcome		_SSE_	Sum of Squared Errors	145.734	140.5118	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_MISC_	Misclassification Rate	0.295039	0.290909	.

Backward Regression Process

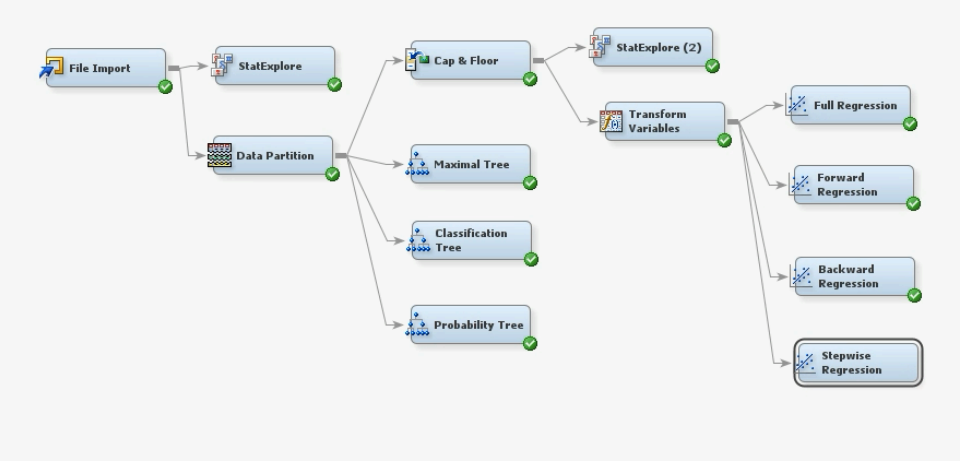
Step 1: Effect REP_BloodPressure removed
Step 2: Effect LOG_REP_Insulin removed
Step 3: Effect REP_SkinThickness removed
Step 4: Effect LOG_REP_Prgenancies removed

Remaining Variables

LOG_REP_Age
LOG_REP_DiabetesPedigreeFunction
REP_BMI

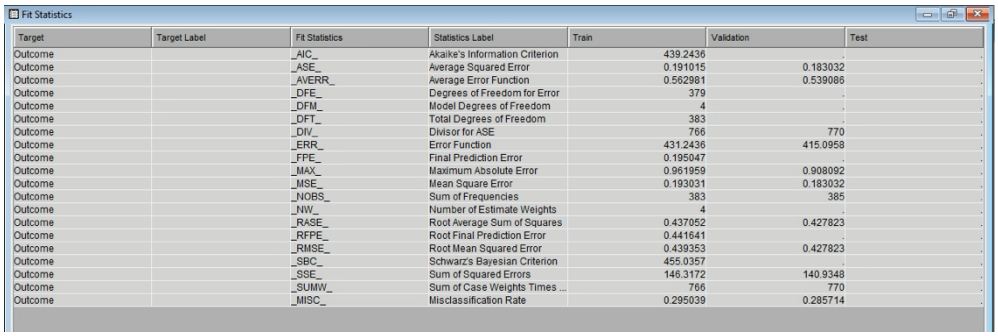
Stepwise Regression

Adding "**Stepwise Regression**" node in SaS Miner:



Stepwise Regression – ASE (0.183032)

The screenshot shows an overview of the fit statistics for stepwise regression and the Average Squared Error is 0.183032.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_AIC_	Akaike's Information Criterion	439.2436	.	.
Outcome		_ASE_	Average Squared Error	0.191015	0.183032	.
Outcome		_AVERR_	Average Error Function	0.562981	0.539086	.
Outcome		_DFE_	Degrees of Freedom for Error	379	.	.
Outcome		_DFM_	Model Degrees of Freedom	4	.	.
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_ERR_	Error Function	431.2436	415.0958	.
Outcome		_FPE_	Final Prediction Error	0.195047	.	.
Outcome		_MAX_	Maximum Absolute Error	0.961959	0.908092	.
Outcome		_MSE_	Mean Square Error	0.193031	0.183032	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_NW_	Number of Estimate Weights	4	.	.
Outcome		_RASE_	Root Average Sum of Squares	0.437052	0.427823	.
Outcome		_RFPE_	Root Final Prediction Error	0.441641	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.439353	0.427823	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	455.0357	.	.
Outcome		_SSE_	Sum of Squared Errors	146.3172	140.9348	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_MISC_	Misclassification Rate	0.295039	0.285714	.

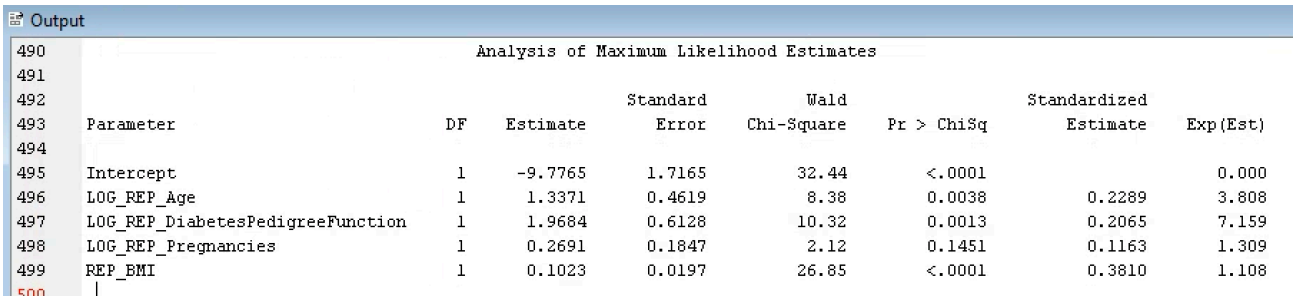
Summary

Types of Regression Models	**Average Square Error**
Full Regression	0.182999
Forward Regression	0.183032
Backward Regression	0.182483 (**Lowest)
Stepwise Regression	0.183032

Interpretation of Regression

The Backward Regression model would be our optimal choice in this step. In order to further understand the correlation between each variable and our target, we would interpret the Pr > ChiSq and odd ratio estimates of our backward regression.

Pr > ChiSq (Backward)



Output

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-9.7765	1.7165	32.44	<.0001		0.000
LOG_REP_Age	1	1.3371	0.4619	8.38	0.0038	0.2289	3.808
LOG_REP_DiabetesPedigreeFunction	1	1.9684	0.6128	10.32	0.0013	0.2065	7.159
LOG_REP_Pregnancies	1	0.2691	0.1847	2.12	0.1451	0.1163	1.309
REP_BMI	1	0.1023	0.0197	26.85	<.0001	0.3810	1.108

The output result shows that the Pr > ChiSq of BMI is less than 0.0001 or 0.01% chance of being not statistically significant. In other words, BMI is 0.9999 or 99.99% of the time is statistically significant to our outcome in this dataset, which is the chance of getting diabetes.

Odd Ratio Estimate (Backward)

Odds Ratio Estimates

Effect	Point Estimate
LOG_REP_Age	3.808
LOG_REP_DiabetesPedigreeFunction	7.159
LOG_REP_Pregnancies	1.309
REP_BMI	1.108

Odd ratio is also a good indicator for us to understand the data pattern thoroughly. The interpretation are as follows:

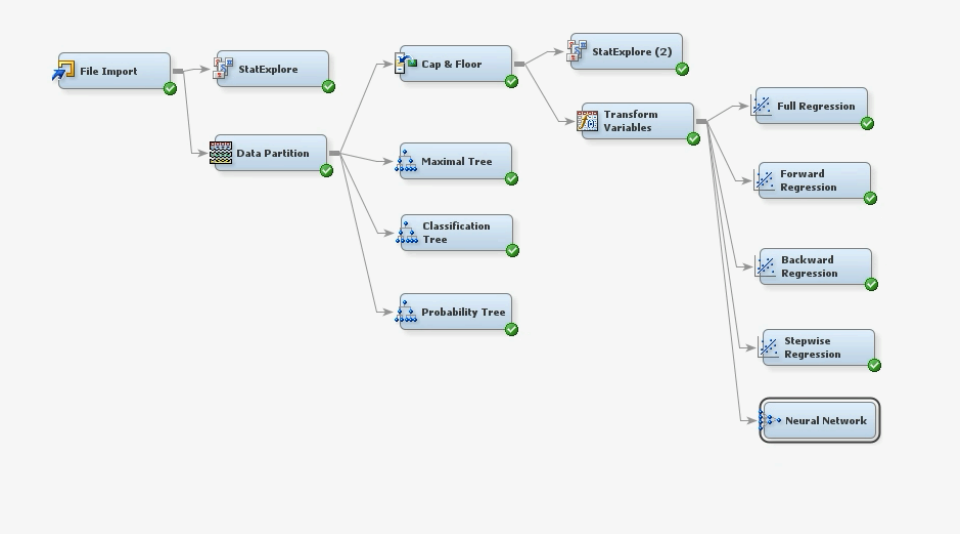
- For every age multiply by 2.74, the chance of getting diabetes is 3.808 times higher than not having it.
- For every DiabetesPedigreeFunction (diabetes record based on family history) multiply by 2.74, the chance of getting diabetes is 7.159 times higher than not having it.
- For every pregnancy multiply by 2.74, the chance of getting diabetes is 30.9% higher than not having it.
- For every 1 unit increase in BMI index, the chance of getting diabetes is 10.8% higher than not having it.

Independent Variables	**Correlation with Diabetes**
Age	Positively Correlated
Diabetes Pedigree Function	Positively Correlated
Pregnancies	Positively Correlated
BMI	Positively Correlated

At this stage, we have had a backward regression as our optimal model with the lowest ASE at 0.182483. However, we should keep creating neural networks to see if there are any potential models with lower ASE in place.

Neural Network

Adding "**Neural Network**" node in SaS Miner:



Full Neural Network (ASE= 0.196105)

The screenshot shows an overview of the fit statistics for full neural network and the Average Squared Error is 0.196105.

Fit Statistics

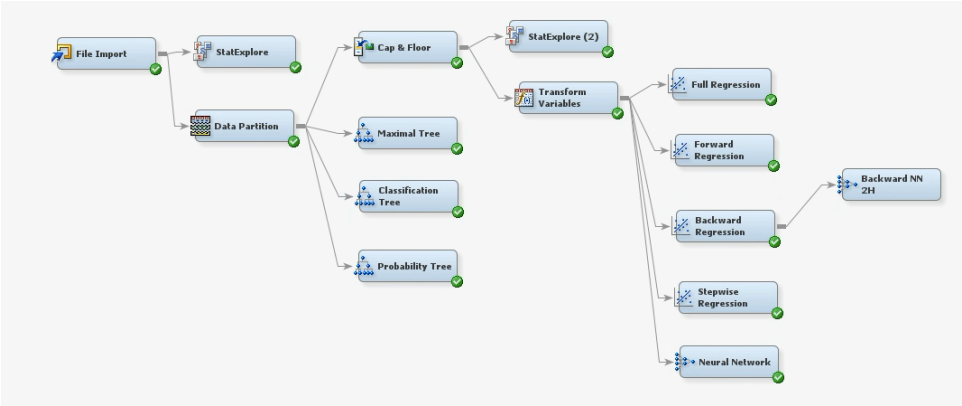
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DFE_	Degrees of Freedom for Error	355	.	.
Outcome		_DFM_	Model Degrees of Freedom	28	.	.
Outcome		_NW_	Number of Estimated Weights	28	.	.
Outcome		_AIC_	Akaike's Information Criterion	442.5618	.	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	553.1068	.	.
Outcome		_ASE_	Average Squared Error	0.168066	0.196105	.
Outcome		_MAX_	Maximum Absolute Error	0.965184	0.982015	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_RASE_	Root Average Squared Error	0.409958	0.442837	.
Outcome		_SSE_	Sum of Squared Errors	128.7383	151.0005	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_FPE_	Final Prediction Error	0.194577	.	.
Outcome		_MSE_	Mean Squared Error	0.181322	0.196105	.
Outcome		_RFPE_	Root Final Prediction Error	0.441109	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.425819	0.442837	.
Outcome		_AVERR_	Average Error Function	0.50465	0.580167	.
Outcome		_ERR_	Error Function	386.5618	446.7285	.
Outcome		_MISC_	Misclassification Rate	0.245431	0.296104	.
Outcome		_WRONG_	Number of Wrong Classificati...	94	114	.

Reduced Variable Set NN (Backward)

NN 2H

Since backward regression has the lowest ASE, we added the neural network node after backward regression.

First, we try adding neural network with 2 hidden units



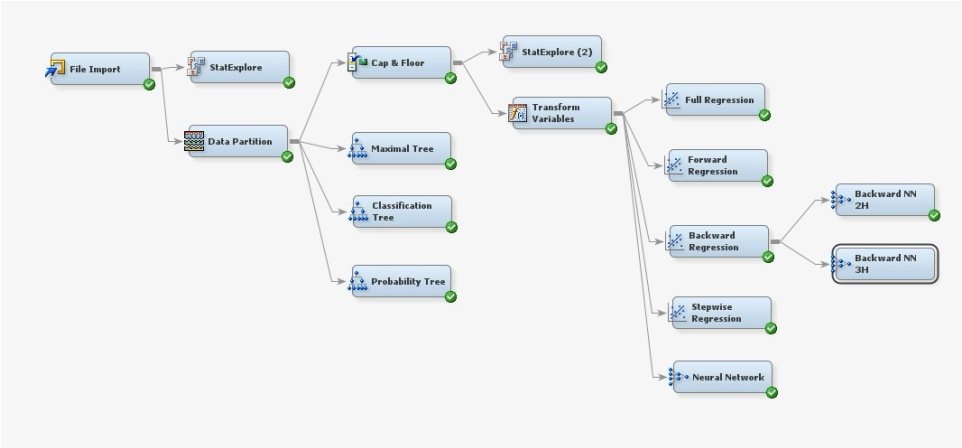
NN 2 Hidden Units ASE (0.181272)

Fit Statistics

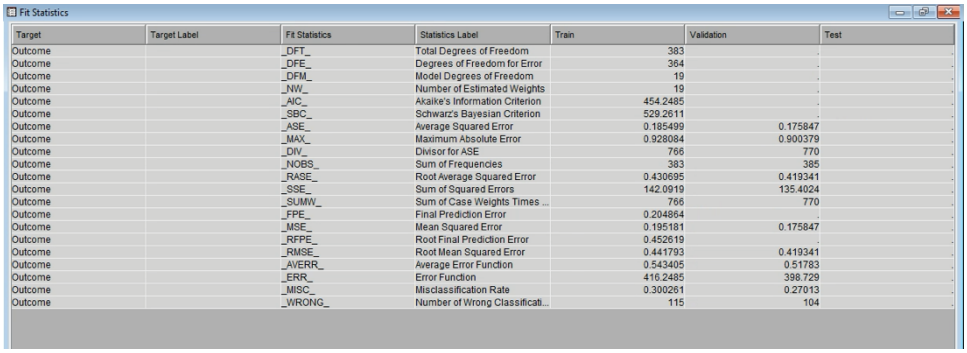
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DFE_	Degrees of Freedom for Error	370	.	.
Outcome		_DFM_	Model Degrees of Freedom	13	.	.
Outcome		_NW_	Number of Estimated Weights	13	.	.
Outcome		_AIC_	Akaike's Information Criterion	444.5636	.	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	495.888	.	.
Outcome		_ASE_	Average Squared Error	0.183295	0.181272	.
Outcome		_MAX_	Maximum Absolute Error	0.962951	0.88356	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_RASE_	Root Average Squared Error	0.42813	0.425761	.
Outcome		_SSE_	Sum of Squared Errors	140.4039	139.5795	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_FPE_	Final Prediction Error	0.196175	.	.
Outcome		_MSE_	Mean Squared Error	0.189735	0.181272	.
Outcome		_RFPE_	Root Final Prediction Error	0.442917	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.435586	0.425761	.
Outcome		_AVERR_	Average Error Function	0.546428	0.532682	.
Outcome		_ERR_	Error Function	418.5636	410.1652	.
Outcome		_MISC_	Misclassification Rate	0.276762	0.285714	.
Outcome		_WRONG_	Number of Wrong Classificati...	106	110	.

NN 3H

3 Hidden Units



NN 3 Hidden Units ASE (0.175847)

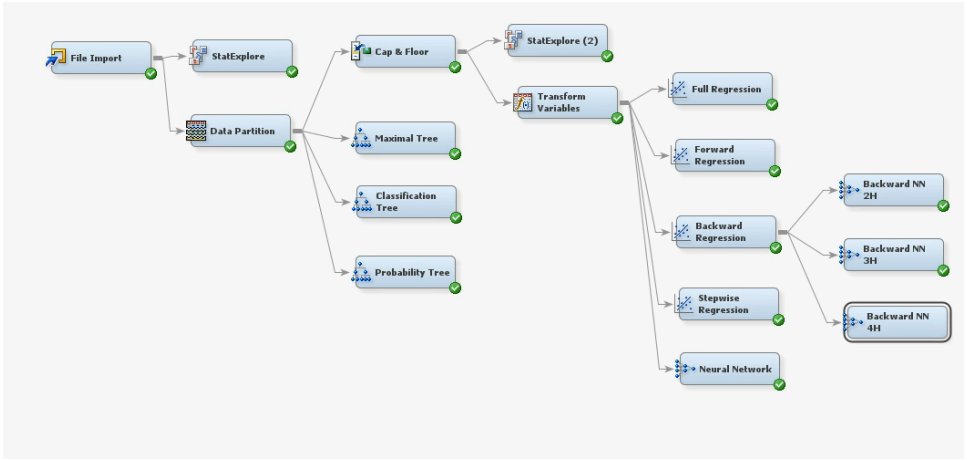


Fit Statistics

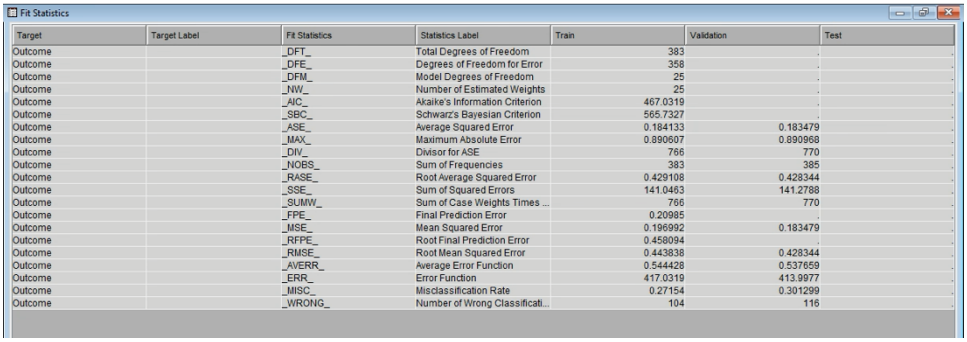
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DFE_	Degrees of Freedom for Error	364	.	.
Outcome		_DFM_	Model Degrees of Freedom	19	.	.
Outcome		_NW_	Number of Estimated Weights	19	.	.
Outcome		_AIC_	Akaike's Information Criterion	454.2485	.	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	529.2611	.	.
Outcome		_ASE_	Average Squared Error	0.185499	0.175847	.
Outcome		_MAX_	Maximum Absolute Error	0.928084	0.900379	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_RASE_	Root Average Squared Error	0.430695	0.419341	.
Outcome		_SSE_	Sum of Squared Errors	142.0919	135.4024	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_FPE_	Final Prediction Error	0.204864	.	.
Outcome		_MSE_	Mean Squared Error	0.195181	0.175847	.
Outcome		_RFPE_	Root Final Prediction Error	0.452619	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.441793	0.419341	.
Outcome		_AVERR_	Average Error Function	0.543405	0.51783	.
Outcome		_ERR_	Error Function	416.2485	398.729	.
Outcome		_MISC_	Misclassification Rate	0.300261	0.27013	.
Outcome		_WRONG_	Number of Wrong Classificati...	115	104	.

NN 4H

4 Hidden Units



NN 4 Hidden Units ASE (0.183479)

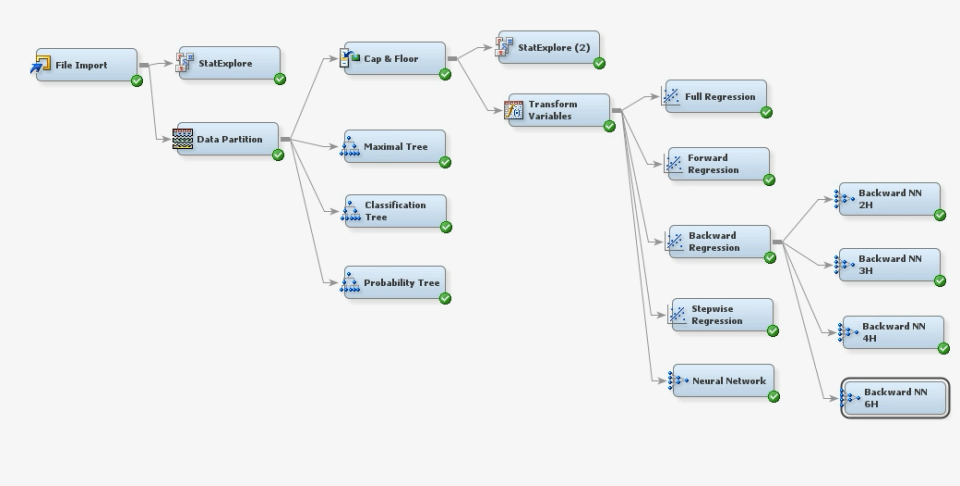


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_DFT_	Total Degrees of Freedom	383	.	.
Outcome		_DFE_	Degrees of Freedom for Error	358	.	.
Outcome		_DFM_	Model Degrees of Freedom	25	.	.
Outcome		_NW_	Number of Estimated Weights	25	.	.
Outcome		_AIC_	Akaike's Information Criterion	467.0319	.	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	565.7327	.	.
Outcome		_ASE_	Average Squared Error	0.184133	0.183479	.
Outcome		_MAX_	Maximum Absolute Error	0.890607	0.890968	.
Outcome		_DIV_	Divisor for ASE	766	770	.
Outcome		_NOBS_	Sum of Frequencies	383	385	.
Outcome		_RASE_	Root Average Squared Error	0.429108	0.428344	.
Outcome		_SSE_	Sum of Squared Errors	141.0463	141.2788	.
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	.
Outcome		_FPE_	Final Prediction Error	0.20985	.	.
Outcome		_MSE_	Mean Squared Error	0.196992	0.183479	.
Outcome		_RFPE_	Root Final Prediction Error	0.458094	.	.
Outcome		_RMSE_	Root Mean Squared Error	0.443838	0.428344	.
Outcome		_AVERR_	Average Error Function	0.544428	0.537659	.
Outcome		_ERR_	Error Function	417.0319	413.9977	.
Outcome		_MISC_	Misclassification Rate	0.27154	0.301299	.
Outcome		_WRONG_	Number of Wrong Classificati...	104	116	.

NN 6H

6 Hidden Units



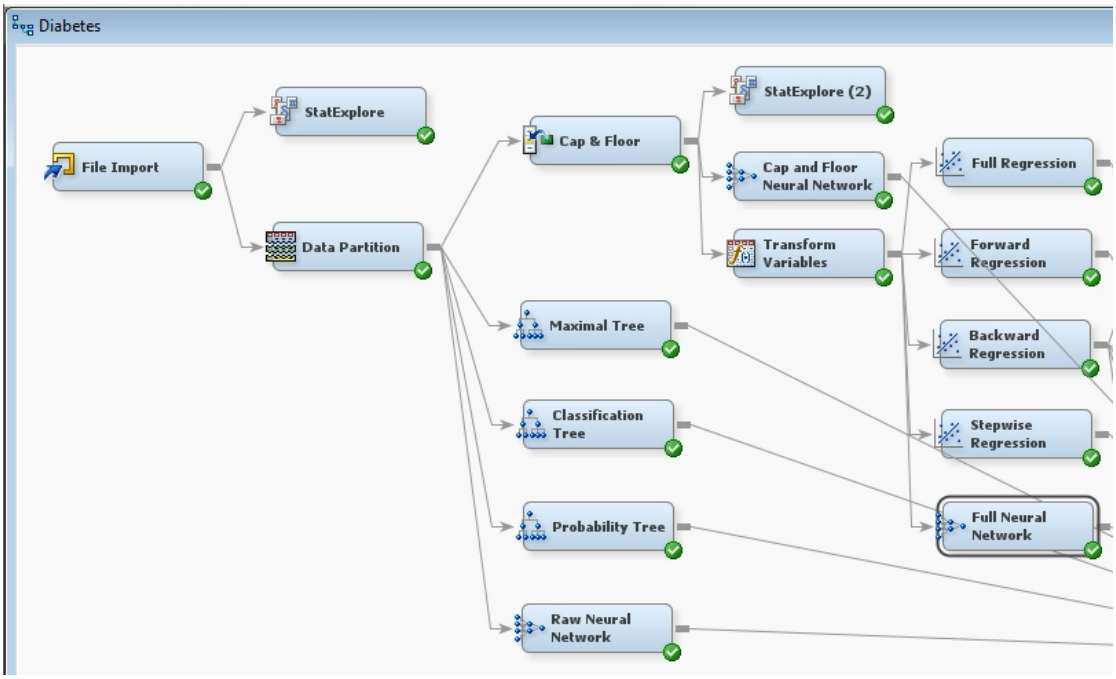
6 Hidden Units ASE (0. 185659)

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Outcome		_DFT_	Total Degrees of Freedom	383		
Outcome		_DFE_	Degrees of Freedom for Error	346		
Outcome		_DFM_	Model Degrees of Freedom	37		
Outcome		_NW_	Number of Estimated Weights	37		
Outcome		_AIC_	Akaike's Information Criterion	485.024		
Outcome		_SBC_	Schwarz's Bayesian Criterion	631.1013		
Outcome		_ASE_	Average Squared Error	0.179624	0.185659	
Outcome		_MAX_	Maximum Absolute Error	0.930195	0.928899	
Outcome		_DIV_	Divisor for ASE	766	770	
Outcome		_NOBS_	Sum of Frequencies	383	385	
Outcome		_RASE_	Root Average Squared Error	0.423821	0.430881	
Outcome		_SSE_	Sum of Squared Errors	137.5922	142.9573	
Outcome		_SUMW_	Sum of Case Weights Times ...	766	770	
Outcome		_FPE_	Final Prediction Error	0.218041		
Outcome		_MSE_	Mean Squared Error	0.198833	0.185659	
Outcome		_RFPE_	Root Final Prediction Error	0.466949		
Outcome		_RMSE_	Root Mean Squared Error	0.445907	0.430881	
Outcome		_AVERR_	Average Error Function	0.536585	0.548937	
Outcome		_ERR_	Error Function	411.024	422.6818	
Outcome		_MISC_	Misclassification Rate	0.281984	0.296104	
Outcome		_WRONG_	Number of Wrong Classificati...	108	114	

Raw Neural Network and Cap & Floor Neural Network

To acquire a more accurate model, we have constructed another two Neural Network after Data Partition (Raw Neural Network) and Cap & Floor (Cap & Floor Neural Network)



Raw Neural Network (ASE=0.202369)

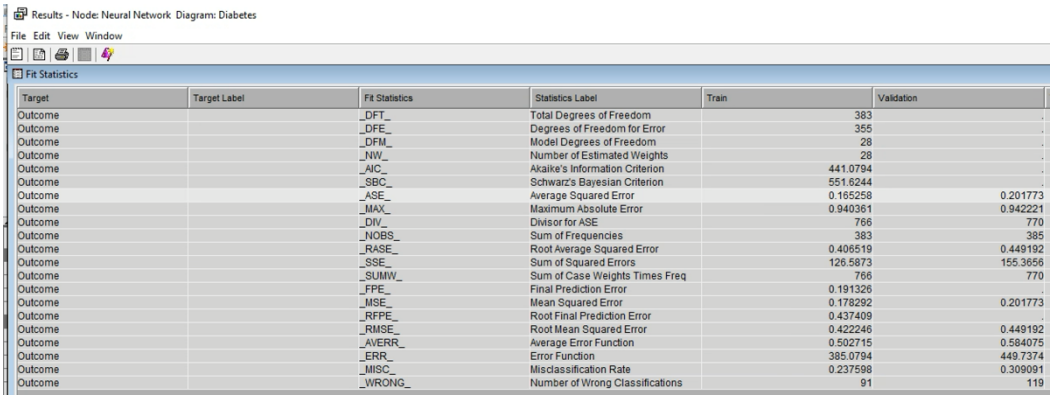
Results - Node: Raw Neural Network Diagram: Diabetes

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Outcome		_DFT_	Total Degrees of Freedom	383	.
Outcome		_DFE_	Degrees of Freedom for E...	355	.
Outcome		_DFM_	Model Degrees of Freedom	28	.
Outcome		_NW_	Number of Estimated Wei...	28	.
Outcome		_AIC_	Akaike's Information Crite...	433.1854	.
Outcome		_SBC_	Schwarz's Bayesian Criter...	543.7304	.
Outcome		_ASE_	Average Squared Error	0.161129	0.202369
Outcome		_MAX_	Maximum Absolute Error	0.897846	0.973543
Outcome		_DIV_	Divisor for ASE	766	770
Outcome		_NOBS_	Sum of Frequencies	383	385
Outcome		_RASE_	Root Average Squared Err...	0.401409	0.449854
Outcome		_SSE_	Sum of Squared Errors	123.4248	155.824
Outcome		_SUMW_	Sum of Case Weights Ti...	766	770
Outcome		_FPE_	Final Prediction Error	0.186546	.
Outcome		_MSE_	Mean Squared Error	0.173838	0.202369
Outcome		_RFPE_	Root Final Prediction Error	0.43191	.
Outcome		_RMSE_	Root Mean Squared Error	0.416938	0.449854
Outcome		_AVERR_	Average Error Function	0.492409	0.598767
Outcome		_ERR_	Error Function	377.1854	461.0504
Outcome		_MISC_	Misclassification Rate	0.229765	0.296104
Outcome		_WRONG_	Number of Wrong Classifi...	88	114

Cap & Floor Neural Network (ASE=0.201773)


Results - Node: Neural Network Diagram: Diabetes

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Outcome		_DFT_	Total Degrees of Freedom	383	.
Outcome		_DFE_	Degrees of Freedom for Error	355	.
Outcome		_DFM_	Model Degrees of Freedom	28	.
Outcome		_NW_	Number of Estimated Weights	28	.
Outcome		_AIC_	Akaike's Information Criterion	441.0794	.
Outcome		_SBC_	Schwarz's Bayesian Criterion	551.6244	.
Outcome		_ASE_	Average Squared Error	0.165258	0.201773
Outcome		_MAX_	Maximum Absolute Error	0.940361	0.942221
Outcome		_DIV_	Divisor for ASE	766	770
Outcome		_NOBS_	Sum of Frequencies	383	385
Outcome		_RASE_	Root Average Squared Error	0.406519	0.449192
Outcome		_SSE_	Sum of Squared Errors	126.5873	155.3656
Outcome		_SUMW_	Sum of Case Weights Times Freq	766	770
Outcome		_FPE_	Final Prediction Error	0.191326	.
Outcome		_MSE_	Mean Squared Error	0.178292	0.201773
Outcome		_RFPE_	Root Final Prediction Error	0.437409	.
Outcome		_RMSE_	Root Mean Squared Error	0.422246	0.449192
Outcome		_AVERR_	Average Error Function	0.502715	0.584075
Outcome		_ERR_	Error Function	385.0794	449.7374
Outcome		_MISC_	Misclassification Rate	0.237598	0.309091
Outcome		_WRONG_	Number of Wrong Classifications	91	119

Summary

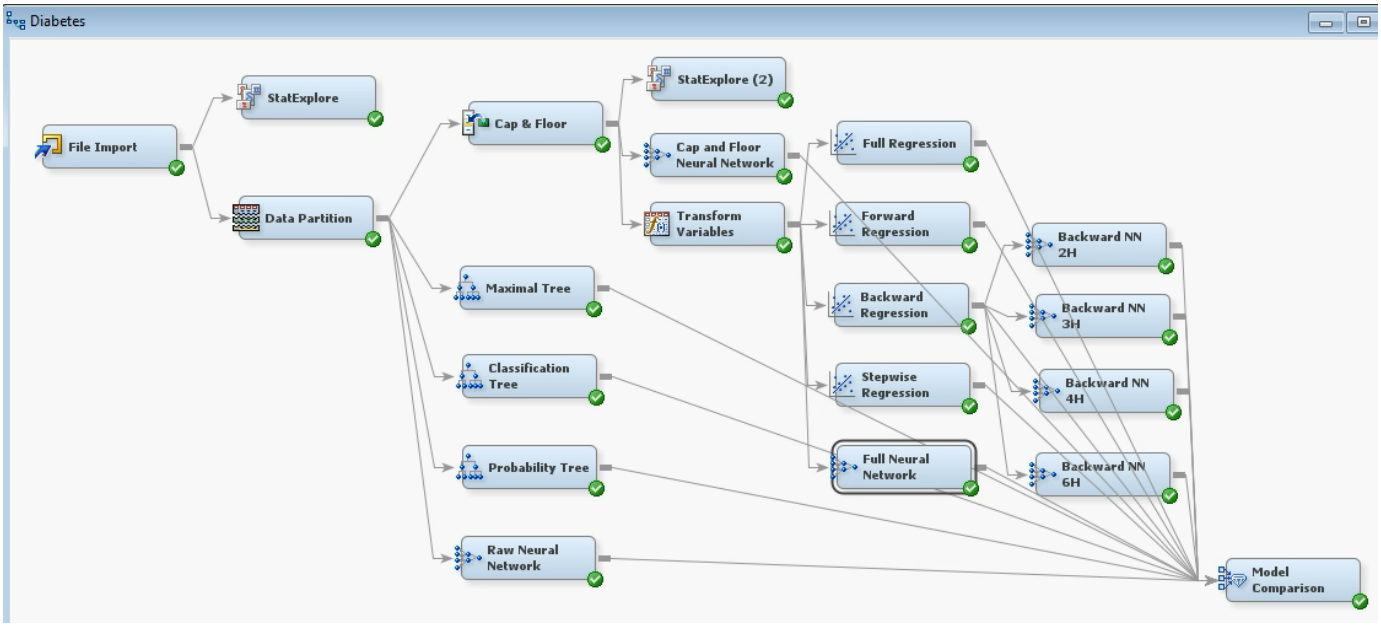
Neural Network	**Average Square Error**
Full Neural Network	0.196105
Backward NN 2 Hidden Units	0.181272
Backward NN 3 Hidden Units	0.175847 (lowest)
Backward NN 4 Hidden Units	0.183479
Backward NN 6 Hidden Units	0.185659
Cap & Floor Neural Network	0.201773
Raw Neural Network (after Data Partition)	0.202369

Now we have figured out that the backward regression neural network with three hidden units has the lowest ASE at 0.175847 among all seven neural network models. This would be our final optimal model for interpreting this diabetes dataset with the lowest error between training and validation data. Our next step would be connecting all models to the model comparison node and generating a ROC curve and index to confirm our optimal choice is accurate and precise.

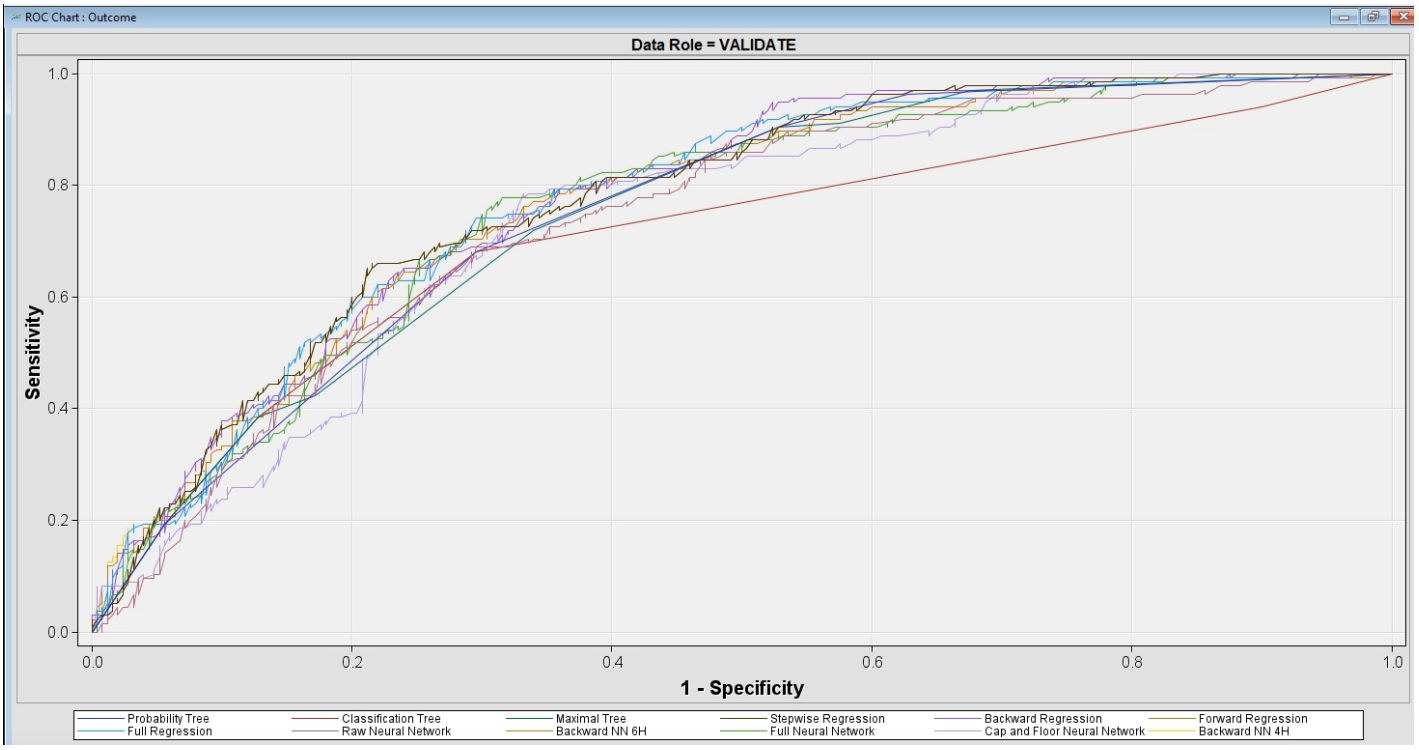
Assessment

Model Comparison

Adding "**Model Comparison**" node and connect it to all the models that were discussed above:



ROC Index Chart



Backward NN 3 Hidden Units is the optimal model (lowest ASE & highest ROC index)

Results - Node: Model Comparison Diagram: Diabetes

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Roc Index ▼	Selection Criterion: Valid: Average Squared Error
Y	Neural2	Neural2	Backward NN 3H	Outcome	0.789	0.175847
	Neural	Neural	Backward NN 2H	Outcome	0.779	0.181272
	Neural3	Neural3	Backward NN 4H	Outcome	0.778	0.183479
	Reg3	Reg3	Backward Regression	Outcome	0.777	0.182483
	Reg2	Reg2	Forward Regression	Outcome	0.776	0.183032
	Reg4	Reg4	Stepwise Regression	Outcome	0.776	0.183032
	Reg	Reg	Full Regression	Outcome	0.775	0.182999
	Neural6	Neural6	Backward NN 6H	Outcome	0.769	0.185659
	Neural5	Neural5	Full Neural Network	Outcome	0.758	0.196105
	Tree3	Tree3	Probability Tree	Outcome	0.753	0.190623
	Tree	Tree	Maximal Tree	Outcome	0.749	0.200281
	Neural7	Neural7	Raw Neural Network	Outcome	0.738	0.202369
	Neural4	Neural4	Cap and Floor Neural Network	Outcome	0.737	0.201773
	Tree2	Tree2	Classification Tree	Outcome	0.703	0.202916

It is hard to tell which ROC curve is the best visually at first glance. However, a higher ROC index means a bigger area between the percentage of the record and the target. Our optimal model has the highest ROC index and lowest average squared error simultaneously, which are 0.789 and 0.175847 respectively. The model comparison node confirms that the backward neural network with three hidden units would be our optimal model in this research by having the largest area under the ROC curve and the lowest error between training and validation.

Conclusion

The aim of our project is to provide our clinical clients a reliable predictive model to help screen whether their 21-year-old or above female patients should receive more attention and clinical advice to monitor their health conditions with relates to diabetes.

After all the data modeling processes, a backward neural network with three hidden units stands out from the crowd with the highest ROC index and lowest average squared error. Using the dataset, the outcome of this model is the prediction of whether one has diabetes. Such a Neural Network model could be of good use if people are concerned about their health conditions and want to understand the likelihood of being diagnosed with diabetes. While variable reduction takes place in the backward selection process, the remaining four independent variables (BMI, Pregnancies, Age, DiabetesPedigreeFunction) become essential indicators to determine and predict the likelihood of having diabetes.

In other words, whether one has a higher probability of being diagnosed with diabetes depends on age, diabetes pedigree function, number of pregnancies, and BMI. Since all four of these factors have a positive correlation with diabetes, the lower the values represent the lower risk of having diabetes. In addition, the Backward Neural Network with three hidden units' model is our optimal model since it has the lowest Average Squared Error, indicating the lowest error between training and validation data.

Although this neural network helps predict the outcome of whether the patients have diabetes or not, it is not the best methodology to represent the relationship between each variable and the outcome. Supposedly if our clients intend to interpret the correlation between each variable and the chance of having diabetes thoroughly to their patients, it is suggested that the backward regression model is a good fit.

If time and resources are allowed, it would be best if we could expand our sample size to male and younger population to predict the disease likelihood. With the number of pregnancies being an indicator for diabetes for female population, studies show that Erectile Dysfunction (ED) is a vital and unique indicator for men to determine whether they have diabetes (CDC, 2022).

In conclusion, out of the data mining models explored above, the Backward Neural Network with three hidden units' model projects a relatively more reliable prediction.

References

Akturk, M. (2020, August 5). *Diabetes dataset*. Kaggle. Retrieved December 15, 2022, from https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Centers for Disease Control and Prevention. (2022, March 15). *"Diabetes and Men."* Centers for Disease Control and Prevention. Retrieved December 15, 2022, from https://www.cdc.gov/diabetes/library/features/diabetes-and-men.html

Centers for Disease Control and Prevention. (2022, July 7). *What is diabetes?* Centers for Disease Control and Prevention. Retrieved December 15, 2022, from https://www.cdc.gov/diabetes/basics/diabetes.html