

Cardiovascular Disease Prediction Mode

Candidate Number: XXXXXXXXXX

## Table of Contents

<b>Individual Part of Assignment .....</b>	<b>5</b>
1 Introduction .....	5
2 Business Objectives .....	5
2.1 Targeted Marketing .....	5
2.2 Public Awareness .....	5
2.3 Entrepreneurship .....	5
3 Construction of Predictive Features .....	5
4 Baseline Method .....	9
5 Hyperparameter Tunning .....	10
6 Decision Tree .....	10
6.1 First Decision Tree .....	11
6.2 Second Decision Tree .....	16
6.3 Third Decision Tree .....	18
6.4 Fourth Decision Tree .....	19
6.5 Fifth Decision Tree .....	20
6.6 Sixth Decision Tree .....	22
6.7 Cross Validation for All Decision trees .....	23
7 Neural Network .....	26
8 Regression .....	29
8.1 First Regression Analysis .....	29
8.2 Second Regression Analysis .....	33
8.3 Third Regression Model .....	36
8.4 Fourth Regression Model .....	36
8.5 Fifth Regression Model .....	37
9 Decision Tree with Selected Variables .....	37
10 Original Dataset Neural Network .....	40
11 Original Dataset Regression .....	41
12 Model Comparison .....	43
13 Possible Future Improvements .....	44
14 To deploy the model in real-world application .....	44

## Big Data for Decision Making

15	References .....	45
	Appendix I.....	46
Figure 1	Blood Pressure Values Categories.....	7
Figure 2	BMI Calculation Formula.....	7
Figure 3	BMI.....	8
Figure 4	Top 10 observations .....	9
Figure 5	Updated Dataset.....	9
Figure 6	Baseline .....	10
Figure 7	Maximal Decision Tree Variable Selection.....	11
Figure 8	Interactive Maximal Decision Tree .....	12
Figure 9	Splitting Rules for Maximal Tree .....	12
Figure 10	Maximal Tree .....	13
Figure 11	Maximal Tree First Split.....	13
Figure 12	Maximal Tree Variable Importance .....	14
Figure 13	Maximal Tree Performance .....	14
Figure 14	Maximal Tree Subtree Assessment .....	15
Figure 15	Decision Tree node properties.....	16
Figure 16	Decision Tree with Depth set to 10. ....	17
Figure 17	Results of Second Decision Tree .....	17
Figure 18	Third Decision Tree Results with subtree assessment set of Avg .....	18
Figure 19	Decision Tree with Entropy Split Rule .....	19
Figure 20	Decision Tree (Entropy) Results.....	20
Figure 21	Fifth Decision Tree Properties.....	21
Figure 22	Fifth Decision Tree Results.....	22
Figure 23	Sixth Decision Tree Using Entropy as Split Rule .....	23
Figure 24	Second Decision Tree with Cross Validation.....	24
Figure 25	Third Decision Tree with Cross Validation .....	24
Figure 26	Fourth Decision Tree with Cross Validation.....	25
Figure 27	Fifth Decision Tree with Cross Validation.....	25
Figure 28	Sixth Decision Tree with Cross Validation .....	26
Figure 29	Neural Network Node Property .....	26
Figure 30	Neural Network .....	27
Figure 31	Neural Network Optimisation.....	28
Figure 32	Neural Network Result .....	28
Figure 33	Regression Node Properties .....	29
Figure 34	Regression results .....	30
Figure 35	Regression with Baseline.....	30
Figure 36	Regression Analysis .....	31

## Big Data for Decision Making

Figure 37 Regression Model Significance.....	32
Figure 38 Variables Skewness .....	32
Figure 39 BMI Skewness Removed.....	33
Figure 40 BMI Transformed. ....	33
Figure 41 Regression with no skewness Analysis.....	34
Figure 42 Fit Statistics of Second Regression .....	34
Figure 43 Second Regression Model Output .....	35
Figure 44 Third Regression Fit Statistics .....	36
Figure 45 Third Regression Model Fit Statistics .....	36
Figure 46 Fifth Regression Model Fit Statistics.....	37
Figure 47 Variable Selection with Selection Node.....	37
Figure 48 Decision Tree with Selected Variables .....	38
Figure 49 Original Dataset Variables .....	39
Figure 50 Original Variable Decision Tree .....	39
Figure 51 Decision Tree without Cross Validation.....	40
Figure 52 Decision Tree with Original Dataset and Depth 10.....	40
Figure 53 Original Dataset Neural Network .....	41
Figure 54 Regression Analysis with Original Dataset .....	42
Figure 55 Original Dataset Variable Skewness.....	42
Figure 56 Original Dataset without age and height Skewness .....	42
Figure 57 Original Dataset with Skewness removed.....	43
Figure 58 Model Comparison Node Results.....	43
Table 1 Features for the model.....	6

## Individual Part of Assignment

### 1 Introduction

Cardiovascular disease (CVD) in general refers to conditions affecting the heart or blood vessels. It is associated with building of fatty deposits in the arteries and the increased risk of blood clotting. CVD is one of the main causes of deaths and disabilities in UK however this can be prevented by leading a healthy lifestyle. Different types of CVD are coronary heart disease, strokes and TIAs, peripheral arterial disease and aortic disease (NHS, 2018).

### 2 Business Objectives

Business objective of this model is to build a predictive model that can analyse the activities of public and can predict if a particular person is likely to have a positive cardiovascular disease (CVD). The use of this model is not limited to the organisational levels for promotions but to the non-profit organisations to create an awareness in people and to guide them to a healthy lifestyle. Some possible usages of model are described in section 2.1, 2.2 and 2.3.

#### 2.1 Targeted Marketing

Organisations who are in health industry can use this model to get a glimpse to the daily activities of people and can use the predictions for the marketing purposes to create an awareness for their products which can lead to increase in sales.

#### 2.2 Public Awareness

Non-profit organisations and government health sector can use this model to create awareness and by getting data from the public and with the help of model can get predictions for the people who are likely to develop CVD, can awareness them and guide them to healthy lifestyle. On the other side government health sector can use this model while allocating the budgets for different departments and can better prepare for the patients can have positive CVD.

#### 2.3 Entrepreneurship

This model can also develop the entrepreneurship opportunities, model can be used in the development of a mobile application where people can enter the data of daily basis and check if they are in the risk to develop CVD and can seek medical help on time or can use the results of application for the change in daily activities.

### 3 Construction of Predictive Features

The dataset that is used for the development of this model consist of 70,000 entries and analyse the 10 features regarding the target variable to generate predictions. Features used in the model and their attributes are shown in table 1.

## Big Data for Decision Making

Variable/Features	Type	Description
Age	Continuous	days
Height	Continuous	CM
Gender	Binary	0 = Female 1 = Male
Ap_hi	Continuous	Systolic Blood Pressure
Ap_lo	Continuous	Diastolic Blood Pressure
Cholesterol	Scale	1 = Low 2 = Medium 3 = High
Gluc	Scale	Contain the values of Glucose levels. 1 = Low 2 = Medium 3 = High
Smoke	Binary	0 = No Smoking 1 = Smoking
Alco	Binary	Information if people have intake or not. 0 = No 1 = Yes
active	Binary	Information about the physical activities 0 = Inactive 1 = Active

*Table 1 Features for the model.*

These features/variables are used with the target variable cardio which takes binary values 1 for positive CVD and 0 for negative CVD.

Original dataset that obtained from the Kaggle contained the continuous values for ap\_hi (Systolic blood pressure) and ap\_lo (Diastolic blood pressure) however for the purpose of the modelling we changed the values to categorical values in the group part of assignment where 1 is for low or normal blood pressure, 2 is for medium blood pressure and 3 is for the high blood pressure. Reference to change the values are given in Figure 1:

# Blood Pressure Categories



BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
<b>NORMAL</b>	<b>LESS THAN 120</b>	<b>and</b>	<b>LESS THAN 80</b>
<b>ELEVATED</b>	<b>120 – 129</b>	<b>and</b>	<b>LESS THAN 80</b>
<b>HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1</b>	<b>130 – 139</b>	<b>or</b>	<b>80 – 89</b>
<b>HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2</b>	<b>140 OR HIGHER</b>	<b>or</b>	<b>90 OR HIGHER</b>
<b>HYPERTENSIVE CRISIS (consult your doctor immediately)</b>	<b>HIGHER THAN 180</b>	<b>and/or</b>	<b>HIGHER THAN 120</b>

Figure 1 Blood Pressure Values Categories

During the group part of the assessment, we created another feature BMI which took the values of weight and height, BMI calculation formula is shown in the figure 2:

$$\text{BMI} = \text{Weight}(\text{kg}) / [\text{Height}(\text{m})]^2$$

Figure 2 BMI Calculation Formula

The process to create a new column BMI with SAS Enterprise Miner is shown in figure 3:

## Big Data for Decision Making

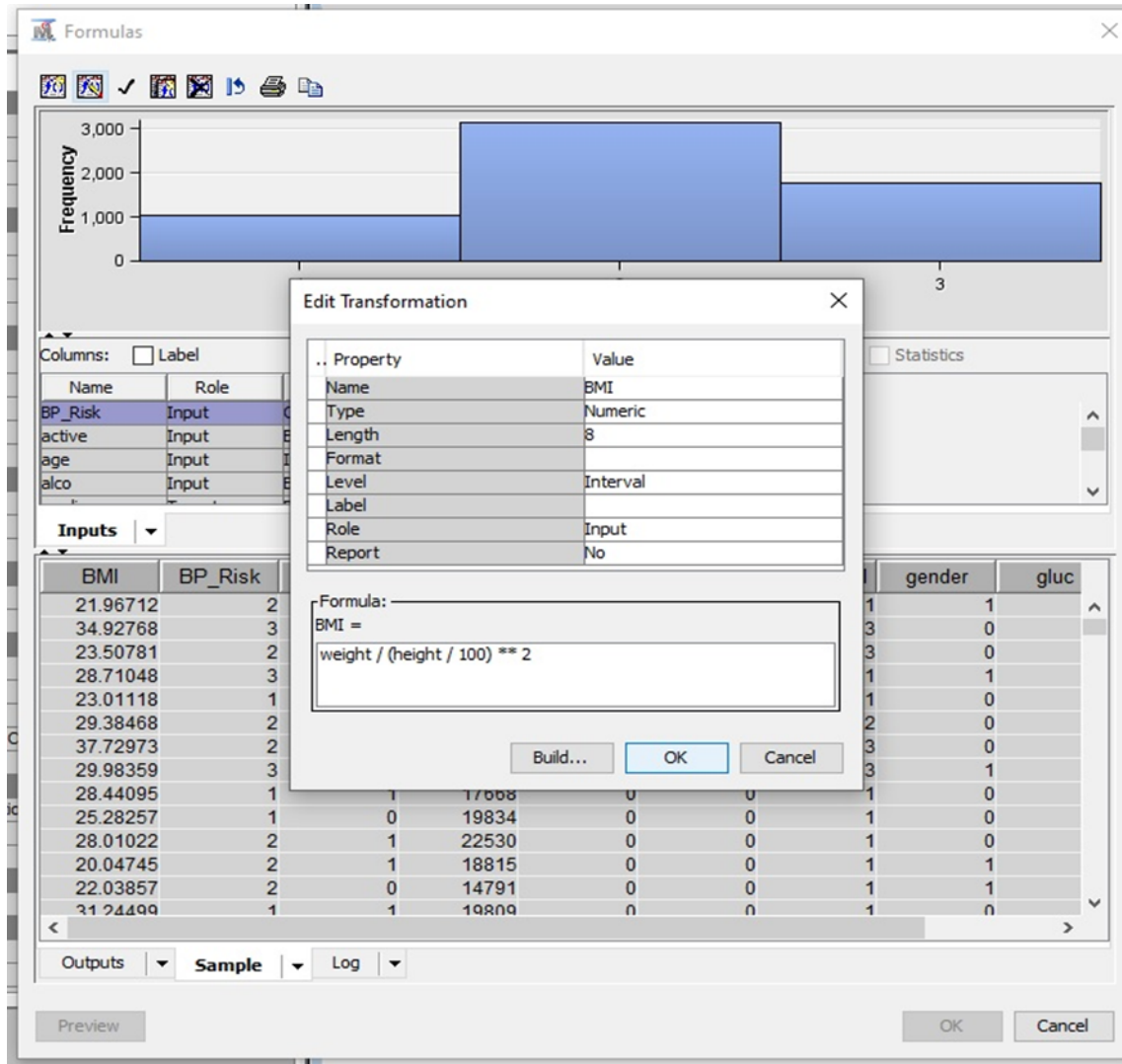


Figure 3 BMI

While working on the group assessment part we created another column using the SQL Code for the BP\_Risk which takes the average values of ap\_hi and ap\_lo. Code to create the feature is as follow:

```
PROC SQL; /* Creating Total BP risk Column */
```

```
    ALTER TABLE IMPORT
```

```
    ADD BP_Risk INT;
```

```
QUIT;
```

```
PROC SQL; /* Adding data to BP Risk column */
```

```
    UPDATE IMPORT
```



## Big Data for Decision Making

```
SET BP_Risk = ROUND((ap_hi + ap_lo) / 2);
```

```
QUIT;
```

Top 10 observations of the modified dataset are shown in Figure 4:

► Table of Contents

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BP_Risk
0	18393	1	188	82	1	2	1	1	0	0	1	0	2
1	20228	0	156	85	3	3	3	1	0	0	1	1	3
2	18857	0	165	64	2	1	3	1	0	0	0	1	2
3	17623	1	169	82	3	3	1	1	0	0	1	1	3
4	17474	0	156	56	1	1	1	1	0	0	0	0	1
8	21014	0	151	67	1	2	2	2	0	0	0	0	2
9	22113	0	157	93	2	2	3	1	0	0	1	0	2
12	22584	1	178	95	2	3	3	3	0	0	1	1	3
13	17668	0	158	71	1	1	1	1	0	0	1	0	1
14	19834	0	164	68	1	1	1	1	0	0	0	0	1

Figure 4 Top 10 observations

Updated dataset with all the modified features is shown in Figure 5:

Columns: ☐ Label ☐ Mining

Name	Partition Role	Role	Level
BMI	Default	Input	Interval
BP_Risk	Default	Input	Ordinal
active	Default	Input	Binary
age	Default	Input	Interval
alco	Default	Input	Binary
ap_hi	Default	Input	Interval
ap_lo	Default	Input	Interval
cardio	Default	Target	Binary
cholesterol	Default	Input	Ordinal
gender	Default	Input	Binary
gluc	Default	Input	Ordinal
height	Default	Input	Interval
smoke	Default	Input	Binary
weight	Default	Input	Interval

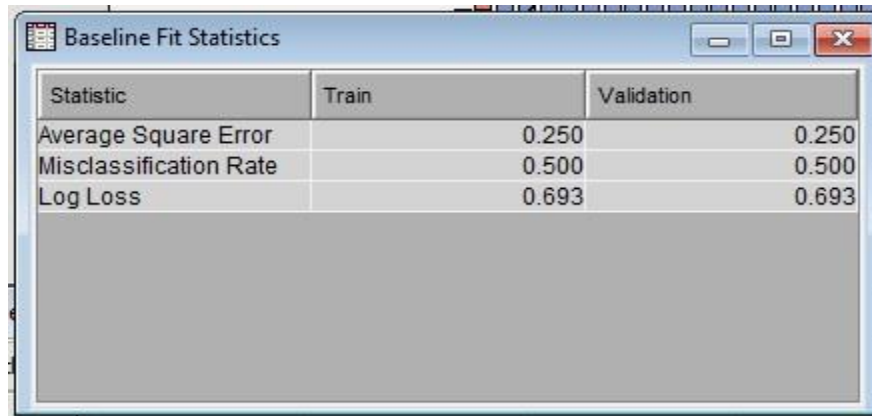
Figure 5 Updated Dataset.

For the building, a better model which is good at prediction new modified features and original features are used interchangeably.

## 4 Baseline Method

Baseline methods are used to evaluate the performance of the model. Baseline is the value that is used to compare with the value predicted by model to check if the model has learnt something and how much model has learnt. During this assignment different

baseline methods are used according to the models and defined in each section of the model. Baseline used for the binary nature variable is shown in Figure 6:



The image shows a software window titled "Baseline Fit Statistics". It contains a table with three columns: "Statistic", "Train", and "Validation". The table lists three statistics: "Average Square Error" with values 0.250 for both Train and Validation, "Misclassification Rate" with values 0.500 for both, and "Log Loss" with values 0.693 for both. The window has standard Windows-style controls (minimize, maximize, close) in the top right corner.

Statistic	Train	Validation
Average Square Error	0.250	0.250
Misclassification Rate	0.500	0.500
Log Loss	0.693	0.693

Figure 6 Baseline

## 5 Hyperparameter Tunning

In machine learning hyperparameters are used to govern the modelling algorithms and do not have default setting. Different parameters are used for different models. Decision Tree algorithm hyperparameter can be depth of the tree, for high performance forest the number of trees, number of hidden layers can neurons in neural network and different regulatory measures to prevent the model overfitting. These parameters not only govern the working of model but also work to enhance the quality of each model's predictions (Koch, et al., 2017).

There are no missing values in the dataset however there are some variables in data that required to remove the skewness.

First model that is used for the prediction of the CVD cases is Decision Tree.

## 6 Decision Tree

Decision tree represents data in hierarchical segmentation. The first segment or original segment is called root node of tree, by applying the simple rules original segment can be divided into different segments as a result each segment is divided into sub-segments and each sub-segment is further divided into each sub-segment and segment that is divided into sub-segment is called parent segment or parent node and resultant segments or nodes are called child nodes. The last node in the decision tree which is not further segmented into sub-segments is call leaf node or when multiple nodes it is called leaves. Decision tree node working comprised of different steps which include, node definition, rule, assignment of each record of dataset to the leaf.

## Big Data for Decision Making

### 6.1 First Decision Tree

First decision tree that is developed in the assignment had used the variables that were chosen in the group part of assignment screenshot of chosen variables is shown in Figure 7:

Name	Use	Report	Role	Level
BMI	Default	No	Input	Interval
BP_Risk	Default	No	Input	Ordinal
_dataobs_	Default	No	ID	Interval
active	Default	No	Input	Binary
age	Default	No	Input	Interval
alco	Default	No	Input	Binary
ap_hi	No	No	Input	Interval
ap_lo	No	No	Input	Interval
cardio	Yes	No	Target	Binary
cholesterol	Default	No	Input	Ordinal
gender	Default	No	Input	Binary
gluc	Default	No	Input	Ordinal
height	No	No	Input	Interval
smoke	Default	No	Input	Binary
weight	No	No	Input	Interval

Figure 7 Maximal Decision Tree Variable Selection

First Decision Tree developed interactively by training the root node to the maximum depth. Due to the binary nature of the target variable ProbChisq is used as splitting rule to split the tree nodes. Screenshot the of the process in shown in Figure 8:

# Big Data for Decision Making

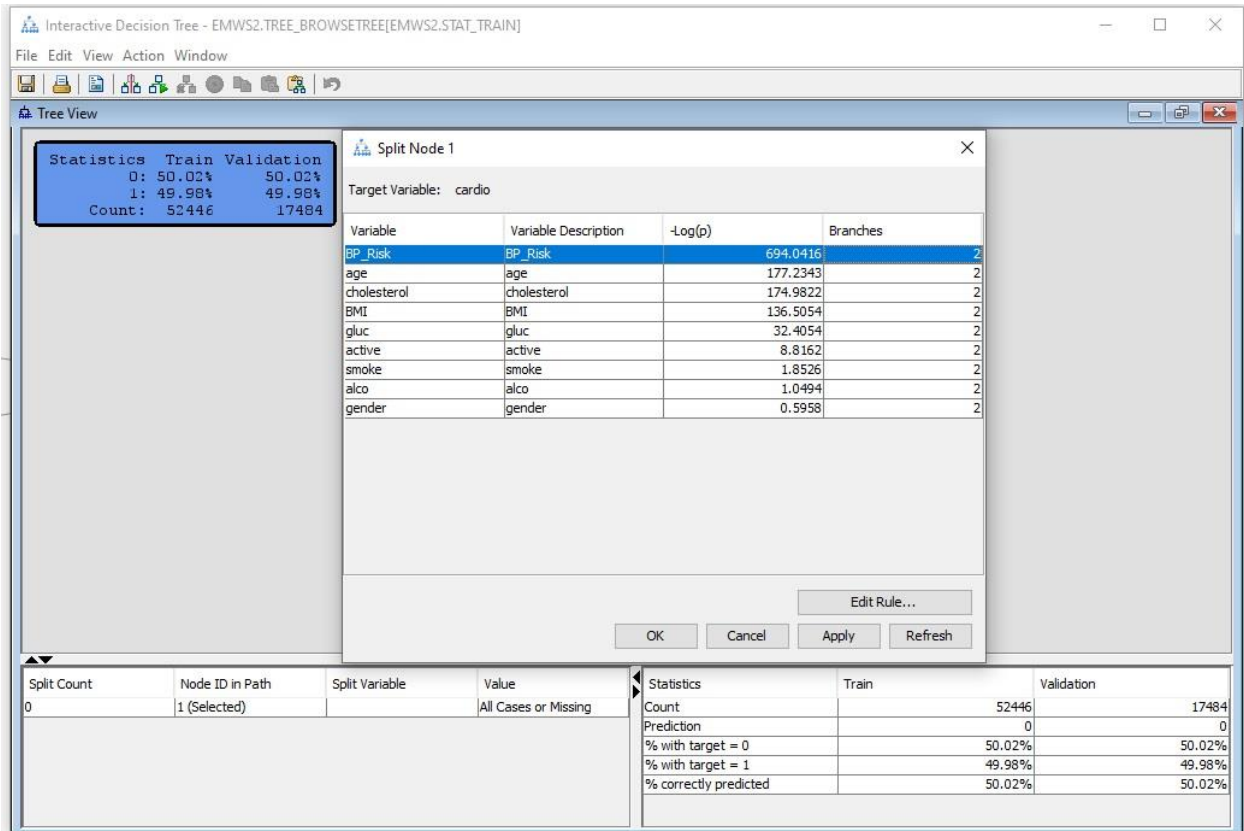


Figure 8 Interactive Maximal Decision Tree

The splitting rules for each node are shown in Figure 9:

Competing Rules Details For Node 1		
Branch	Split Variable	Value
1	BP_Risk	<3 or Missing
2	BP_Risk	>= 3
1	age	<19379.5000
2	age	>= 19379.5000 or Missing
1	cholesterol	<2 or Missing
2	cholesterol	>=2
1	BMI	<27.7093 or Missing
2	BMI	>=27.7093
1	gluc	<2 or Missing
2	gluc	>=2
1	active	0
2	active	1 or Missing
1	smoke	0 or Missing
2	smoke	1
1	alco	0 or Missing
2	alco	1
1	gender	0 or Missing
2	gender	1

Figure 9 Splitting Rules for Maximal Tree

Screenshot of the maximal tree is shown in Figure 10:

# Big Data for Decision Making

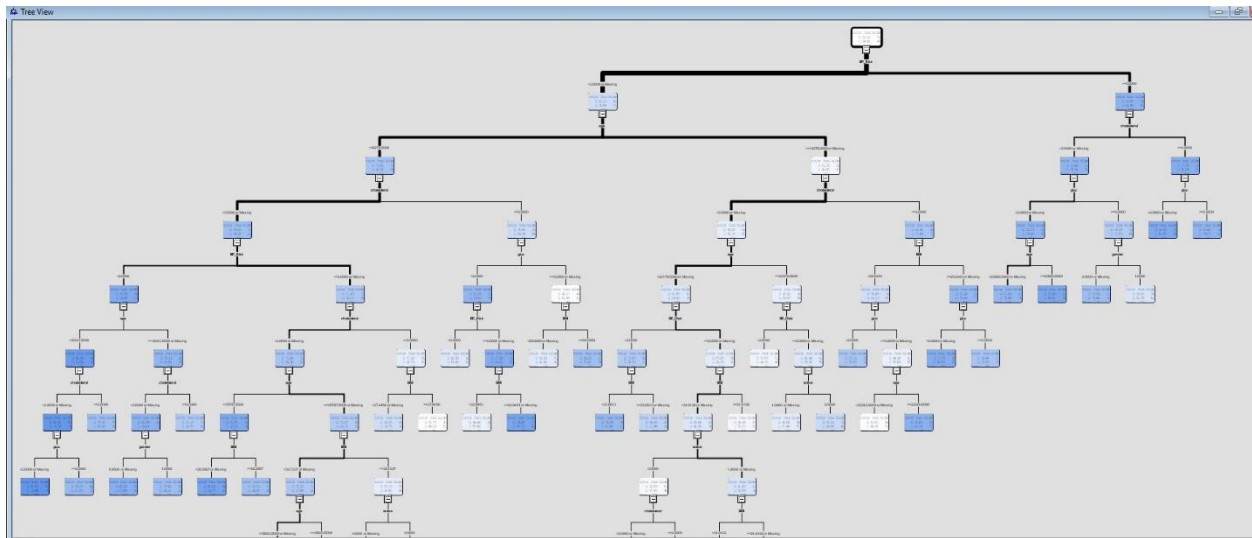


Figure 10 Maximal Tree

In the maximal tree root node is divided according to the most important variable BP\_Risk and divided in two branches, first highest flow of data is divided between the values of less than 3 which contains the 36685 values from train data and 17791 from the validate data. The split branch has the dominating values of people who did not had CVD with 63.11% for train data and 62.86% for validate data. Further this branch is divided based on age having the values of less than 1978 days and this node also dominate the values with the negative CVD, 73.65% for train data and 73.36% for validate data. The details of the first two branches are shown in figure 11.

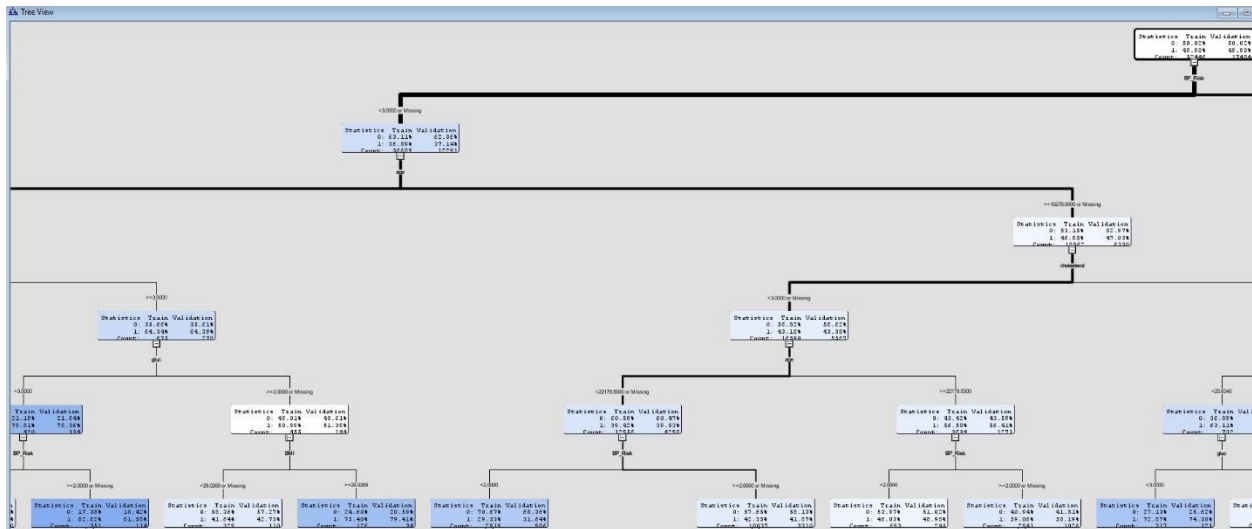
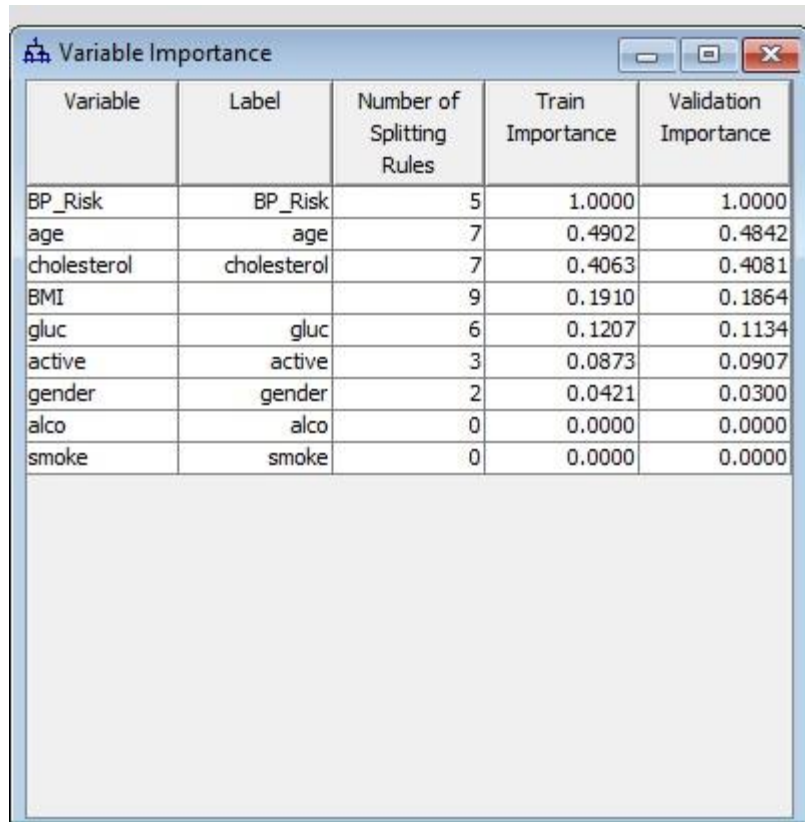


Figure 11 Maximal Tree First Split

Variable importance for the selection of each node is shown in Figure 12:

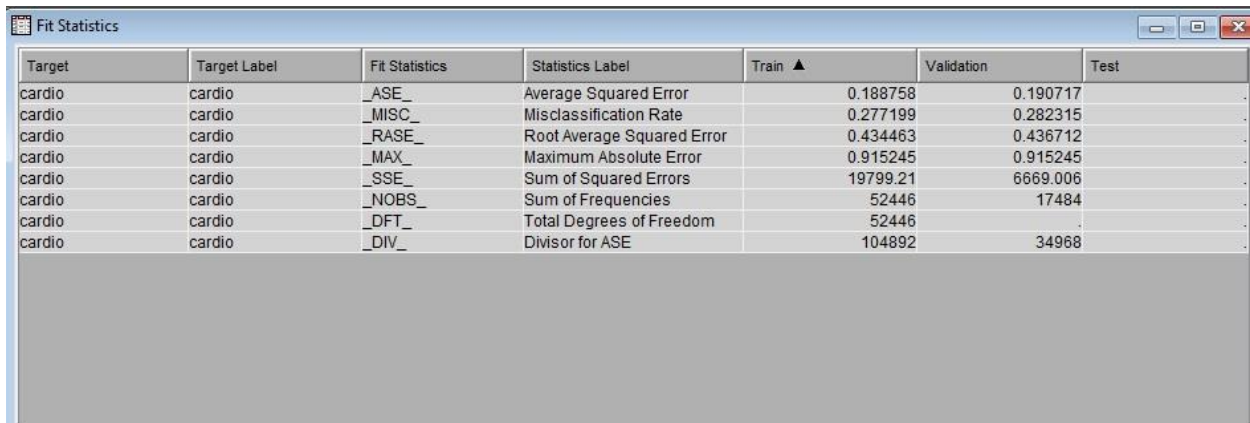
## Big Data for Decision Making



Variable	Label	Number of Splitting Rules	Train Importance	Validation Importance
BP_Risk	BP_Risk	5	1.0000	1.0000
age	age	7	0.4902	0.4842
cholesterol	cholesterol	7	0.4063	0.4081
BMI		9	0.1910	0.1864
gluc	gluc	6	0.1207	0.1134
active	active	3	0.0873	0.0907
gender	gender	2	0.0421	0.0300
alco	alco	0	0.0000	0.0000
smoke	smoke	0	0.0000	0.0000

Figure 12 Maximal Tree Variable Importance

Maximal tree results are shown in Figure 13:



Target	Target Label	Fit Statistics	Statistics Label	Train ▲	Validation	Test
cardio	cardio	_ASE_	Average Squared Error	0.188758	0.190717	.
cardio	cardio	_MISC_	Misclassification Rate	0.277199	0.282315	.
cardio	cardio	_RASE_	Root Average Squared Error	0.434463	0.436712	.
cardio	cardio	_MAX_	Maximum Absolute Error	0.915245	0.915245	.
cardio	cardio	_SSE_	Sum of Squared Errors	19799.21	6669.006	.
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	.
cardio	cardio	_DFT_	Total Degrees of Freedom	52446	.	.
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	.

Figure 13 Maximal Tree Performance

For the interactive tree with maximal length has Misclassification Rate of 27% for the train data and 28% for the validation data. Misclassification is used as a measure due to binary nature of the target variable. The subtree assessment is shown in Figure 14:

## Big Data for Decision Making

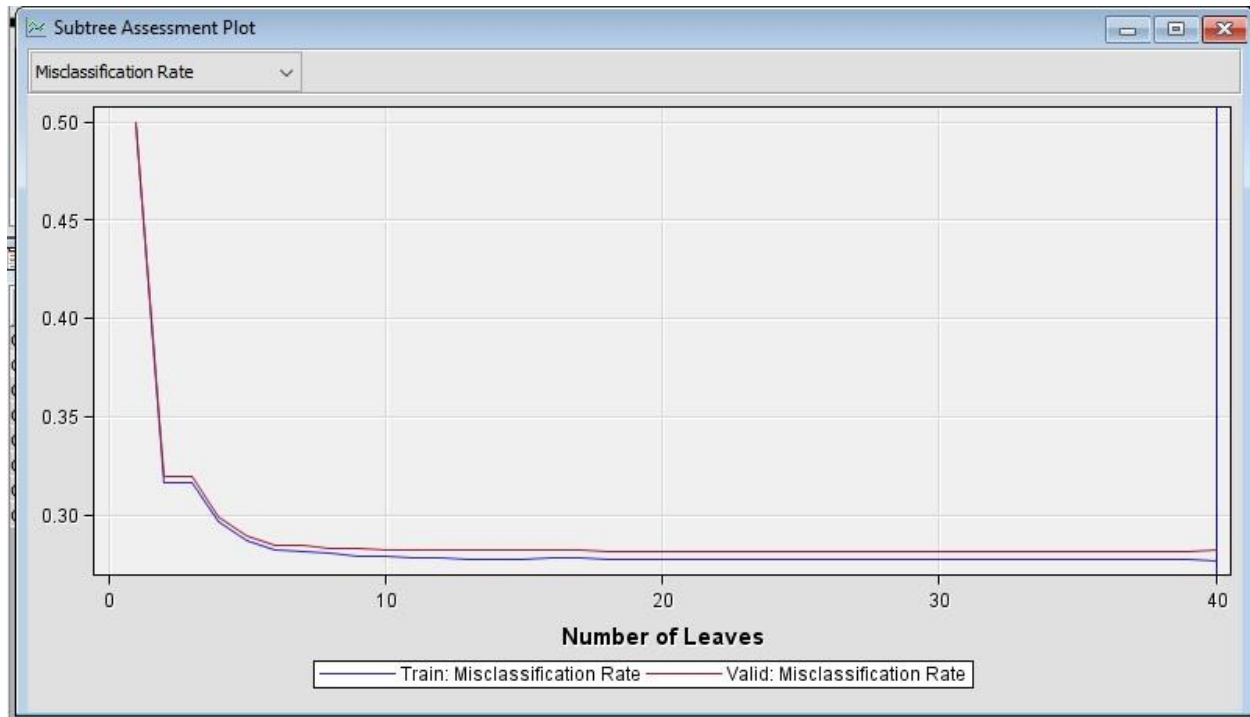


Figure 14 Maximal Tree Subtree Assessment

The current results of the misclassification which is:

$$\text{Accuracy (Train data)} = 1 - \text{misclassification}$$

$$\text{Accuracy (Train Data)} = 1 - 0.27$$

$$= 0.73 \text{ or } 73\%$$

$$\text{Accuracy (Validation Data)} = 1 - \text{misclassification}$$

$$= 1 - 0.28$$

$$= 0.72 \text{ or } 72\%$$

Maximal tree has obtained the accuracy of 73% for the train data and 72% for the validation data with the generation of 40 leaves.

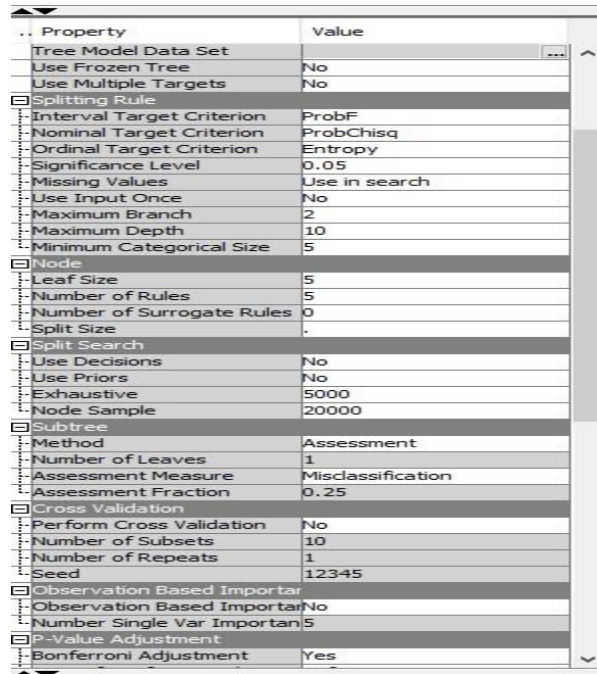
This maximal tree can be manually pruned to the optimal size however for this assessment SAS Enterprise Miner automation is used to obtain the optimal tree with hit and trail method.

As the nature of the target variable is binary. The baseline for the decision tree can be 50% misclassification if the model can perform better than 50% misclassification it can be assumed that the model has learnt something. Further the hyperparameters of the trees will be set with the depth of tree.



## 6.2 Second Decision Tree

Second decision tree that is developed uses the same variables as maximal tree and hyperparameter which is depth of the tree is set to 10, subtree tree assessment criteria is set to misclassification which is probability type, for splitting rule it is chi squared and significance for splitting is set to 0.05 p-value. Screenshot of SAS properties is shown in Figure 15:



Property	Value
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
<b>Splitting Rule</b>	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.05
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
<b>Node</b>	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
<b>Split Search</b>	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
<b>Subtree</b>	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
<b>Cross Validation</b>	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
<b>Observation Based Importance</b>	
Observation Based Importance	No
Number Single Var Importance	5
<b>P-Value Adjustment</b>	
Bonferroni Adjustment	Yes

Figure 15 Decision Tree node properties

Screenshot of the second decision tree is shown in Figure 16:



# Big Data for Decision Making

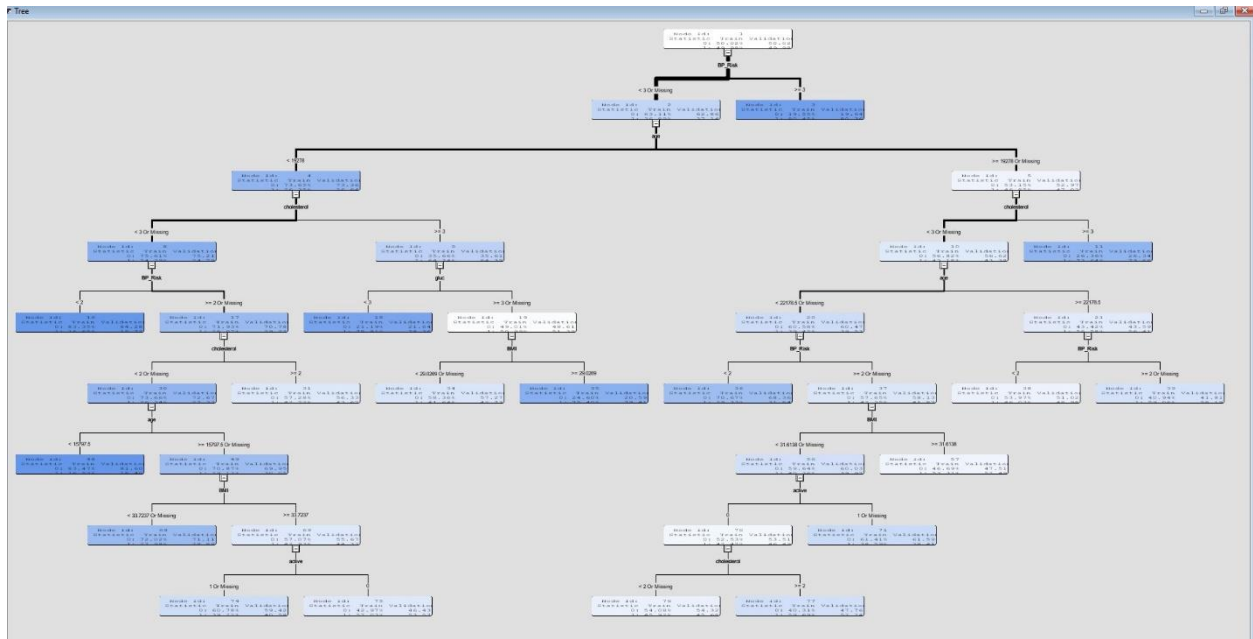


Figure 16 Decision Tree with Depth set to 10.

This tree has also split the variables according to the importance and first split is the done based on BP\_Risk variable. Fit statistics, subtree assessment and leaf statistics of the tree are shown in Figure 17:

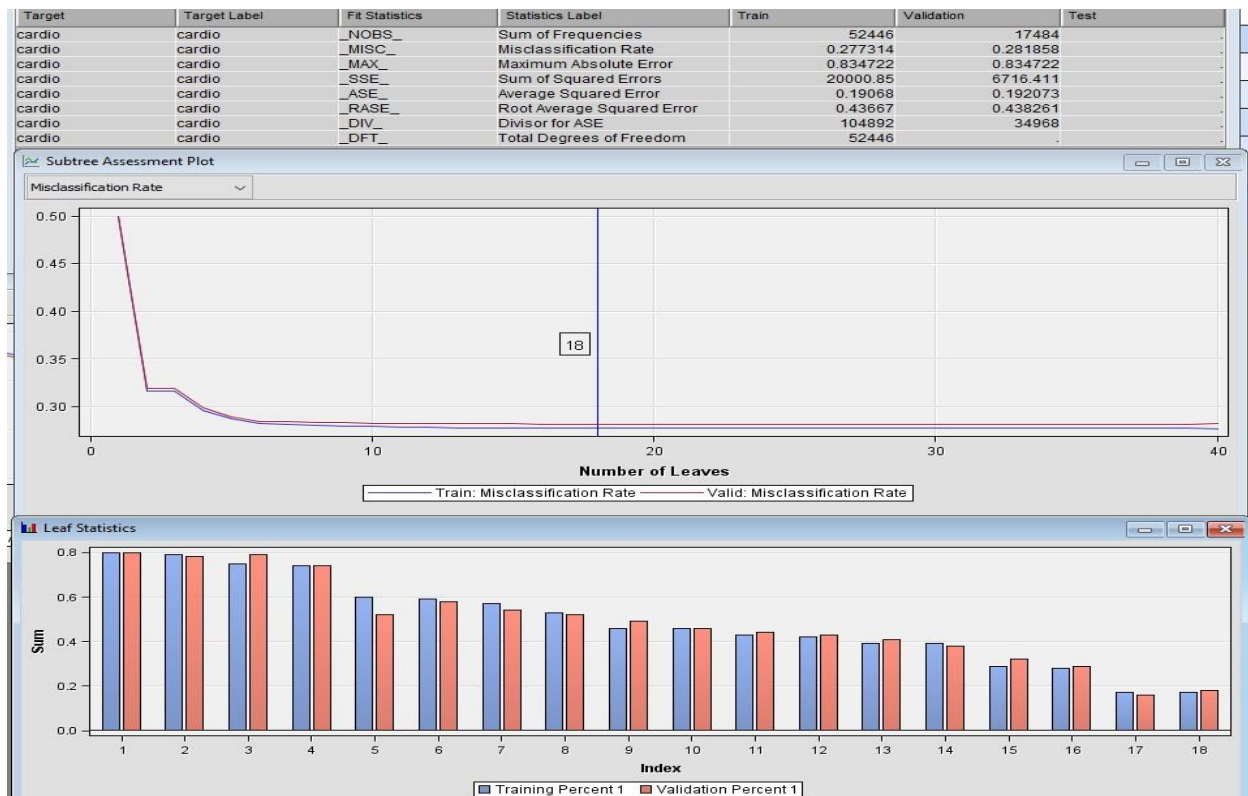


Figure 17 Results of Second Decision Tree

## Big Data for Decision Making

From the figure 16 fit statistics, the performance of three has decreased which mean current tree is not performance as compared the maximal tree. Leaf statistics shows that the algorithm started with the maximum of leaves and achieved this performance with the 18 leaves.

### 6.3 Third Decision Tree

Third decision tree that is included in the modelling performance use the same parameter as second decision tree however for the subtree assessment criteria average squared error is used as a selection criterion to compare the performance of tree and results are shown in Figure 18:

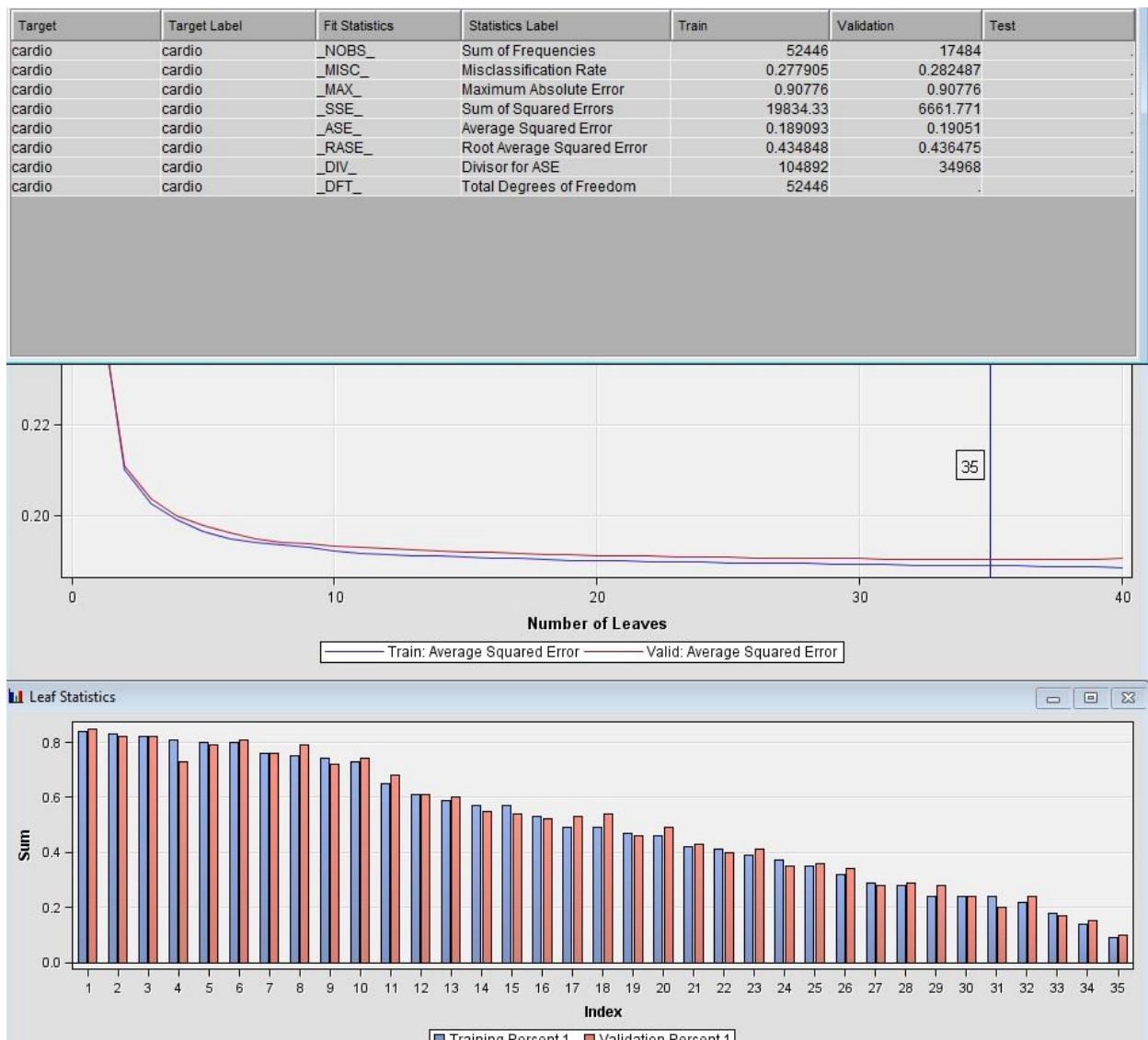


Figure 18 Third Decision Tree Results with subtree assessment set of Avg

With the subtree selection criteria set to averaged squared error the performance of the model has decreased and model used 35 leaves to reach to the results.

### 6.4 Fourth Decision Tree

Fourth decision tree is developed with the same hyperparameter of depth 10 as in decision tree 2 and 3. However the splitting rule is Entropy is used which splits the node based on purity of the node as compared to probchisq where it split the node based on p-values. Complete tree is shown in Figure 19:

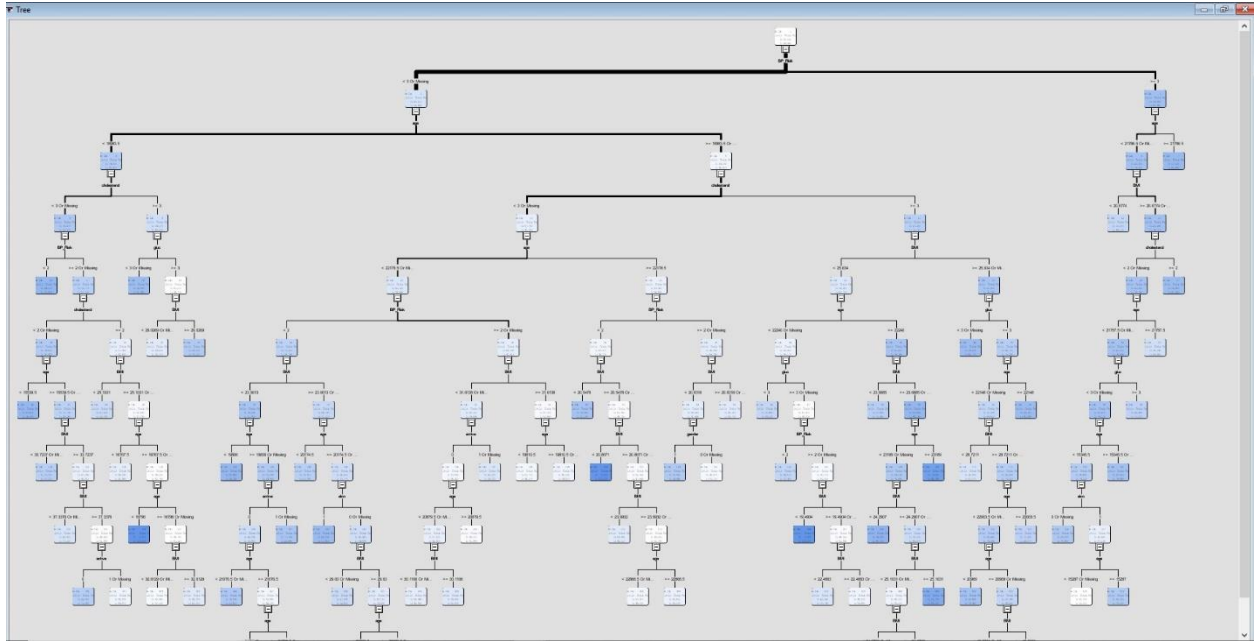


Figure 19 Decision Tree with Entropy Split Rule

And the results of the tree are shown in figure 20:

## Big Data for Decision Making

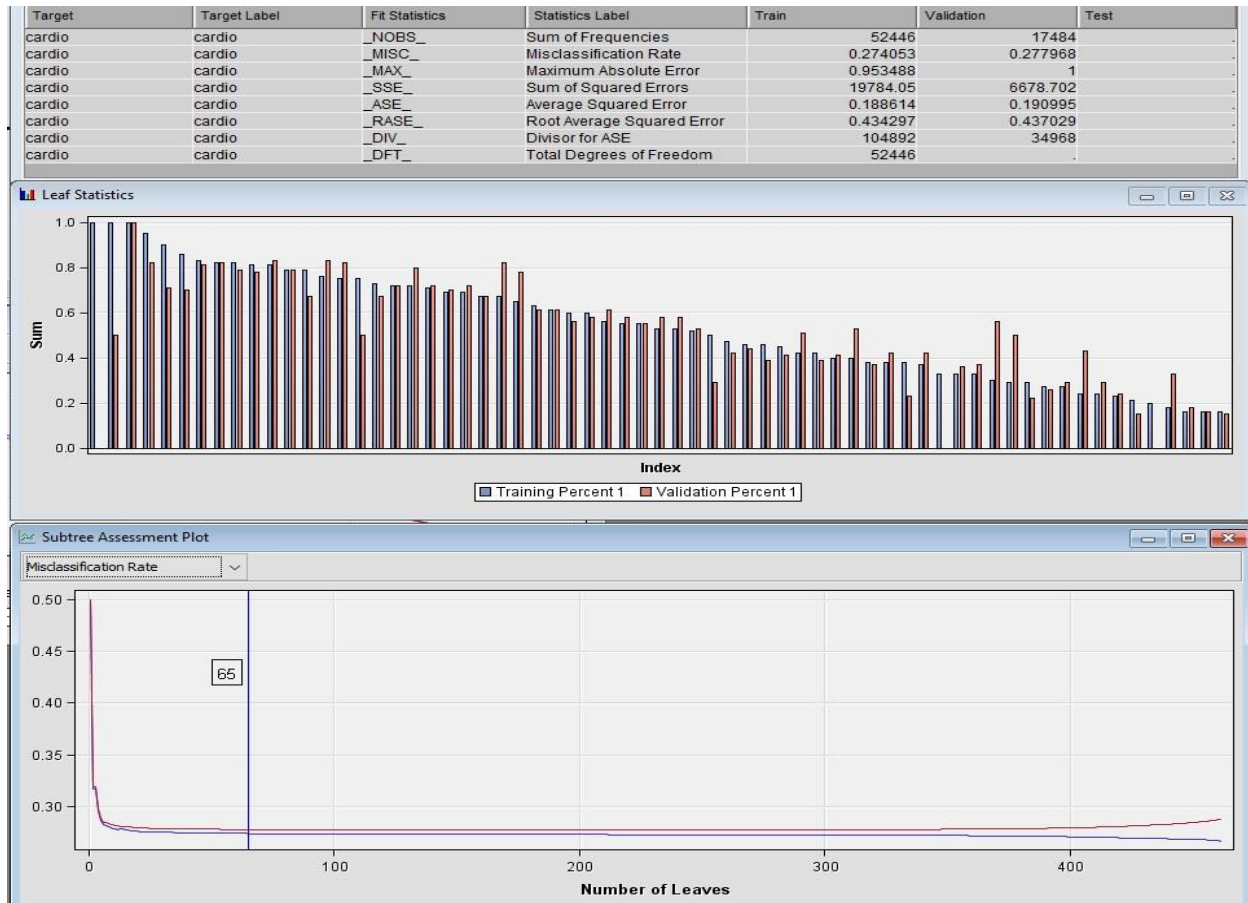


Figure 20 Decision Tree (Entropy) Results

Results shows that it has improve the performance of the tree has slightly improve to 27% misclassification rate in the validation rule and tree has used 65 leaves to produce the results.

### 6.5 Fifth Decision Tree

Fifth decision tree is developed with the depth of the tree set to default value of 6 and splitting rule set to probchisq and subtree assessment set to misclassification.

Properties of the tree are shown in Figure 21:

## Big Data for Decision Making

Property	Value
<b>General</b>	
Node ID	Tree5
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
<b>Splitting Rule</b>	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.05
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
<b>Node</b>	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
<b>Split Search</b>	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
<b>Subtree</b>	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
<b>Cross Validation</b>	

Figure 21 Fifth Decision Tree Properties

And the results of the tree are shown in Figure 22:

## Big Data for Decision Making

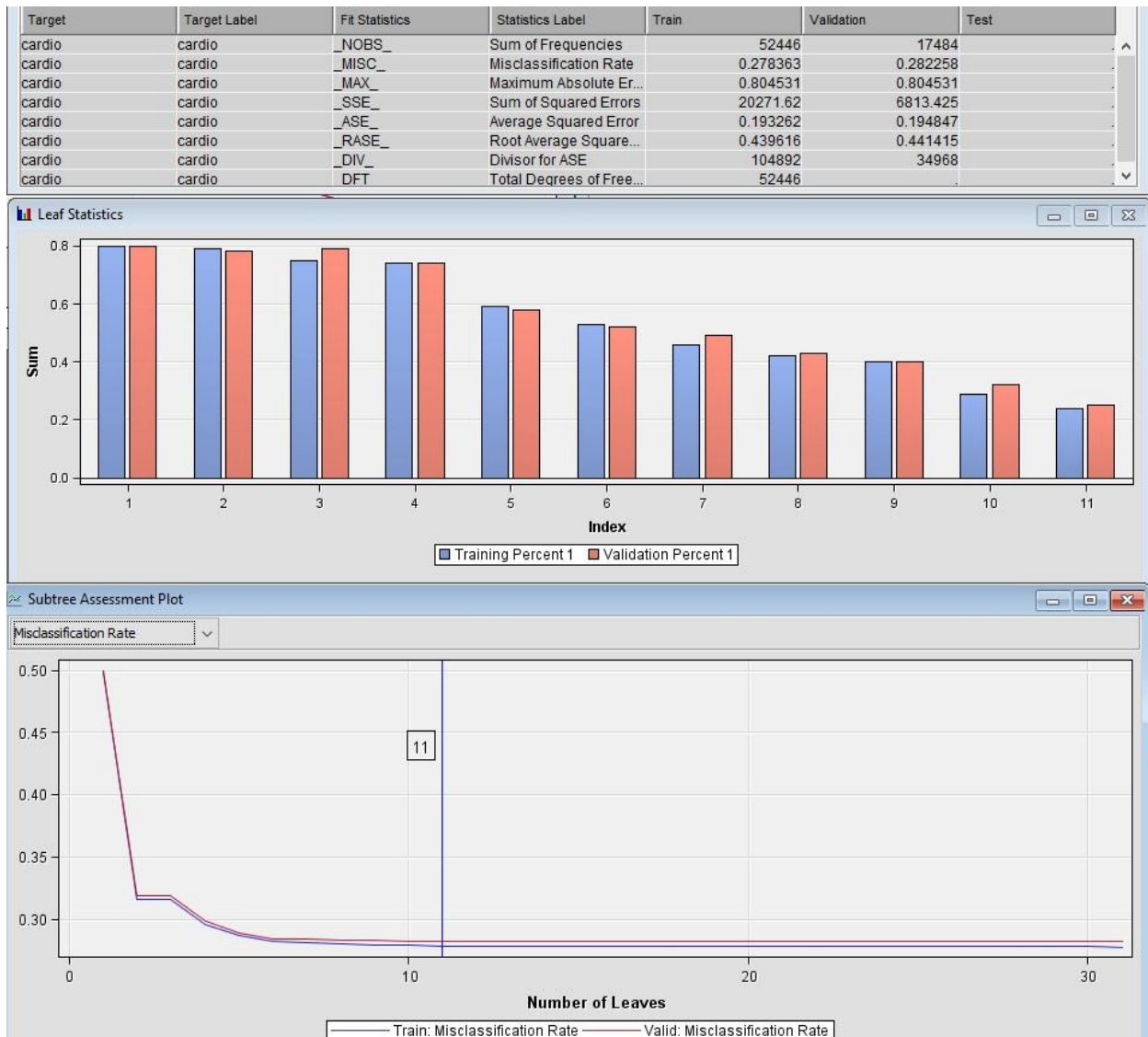


Figure 22 Fifth Decision Tree Results

Results shows that the performance of the model has decreases with misclassification of the validation data went to 28%.

### 6.6 Sixth Decision Tree

Sixth decision tree is developed by using the same depth of 6 as in fifth decision tree however splitting rule property is set to Entropy. Results are shown in Figure 23:

## Big Data for Decision Making

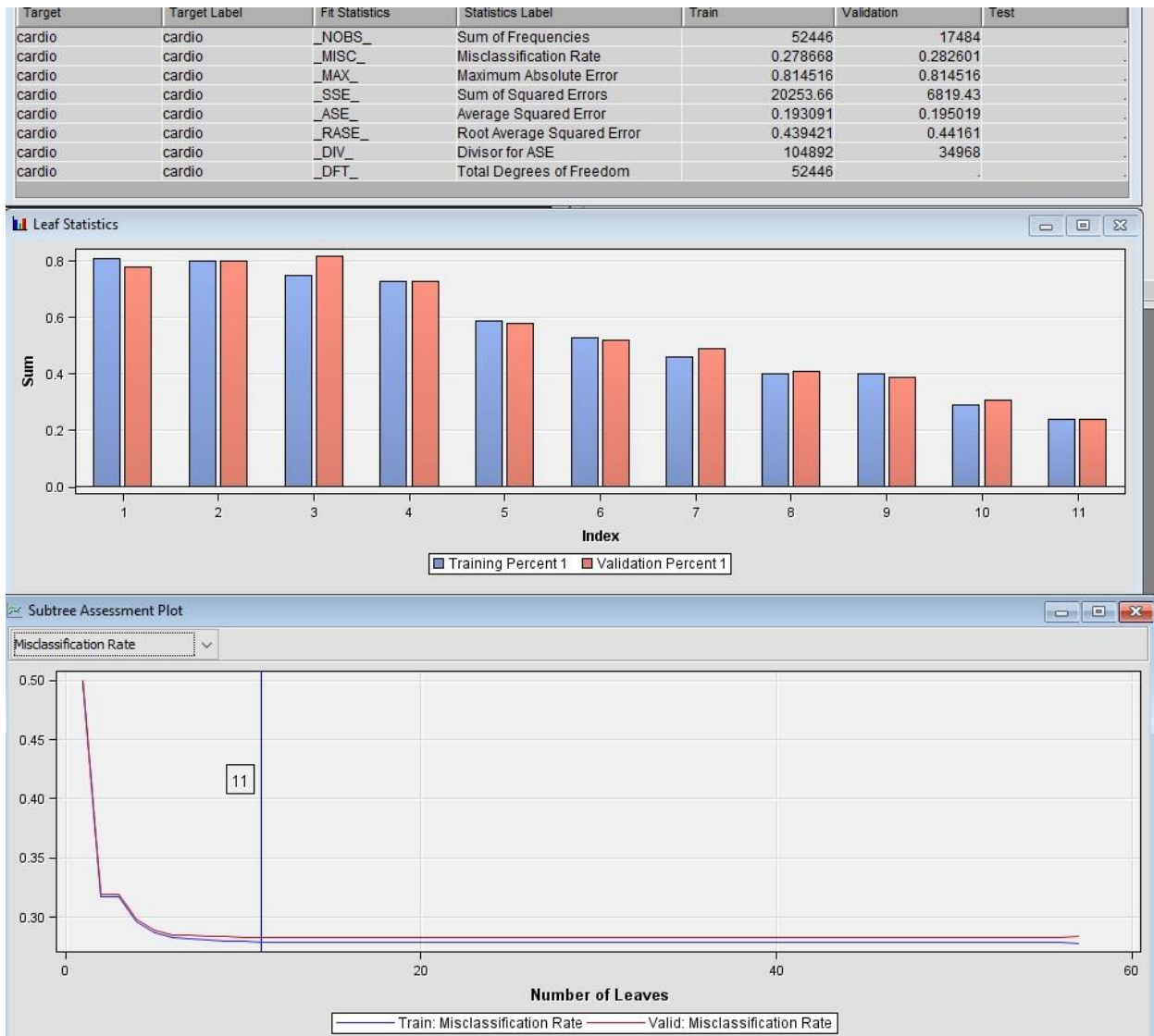


Figure 23 Sixth Decision Tree Using Entropy as Split Rule

Results shows that the performance of the model has not improved it is still at 28% for the validation data and model as achieved these results with 11 leaves.

All the results of the Decision Tree 1, 2, 3, 4, and 5 are without the cross validation.

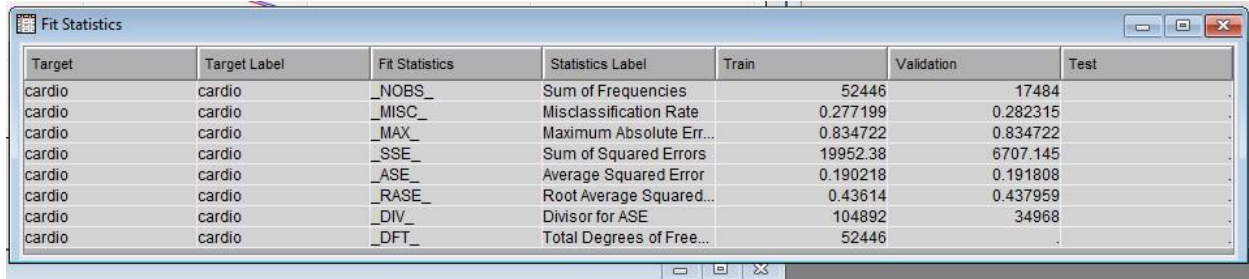
### 6.7 Cross Validation for All Decision trees

Cross validation can also be used to control and improve the performance of the model. All the previous trees are assessed with the cross-validation property set to Yes.

Results of the second decision tree with the cross validation set to Yes are shown in figure 24:



## Big Data for Decision Making



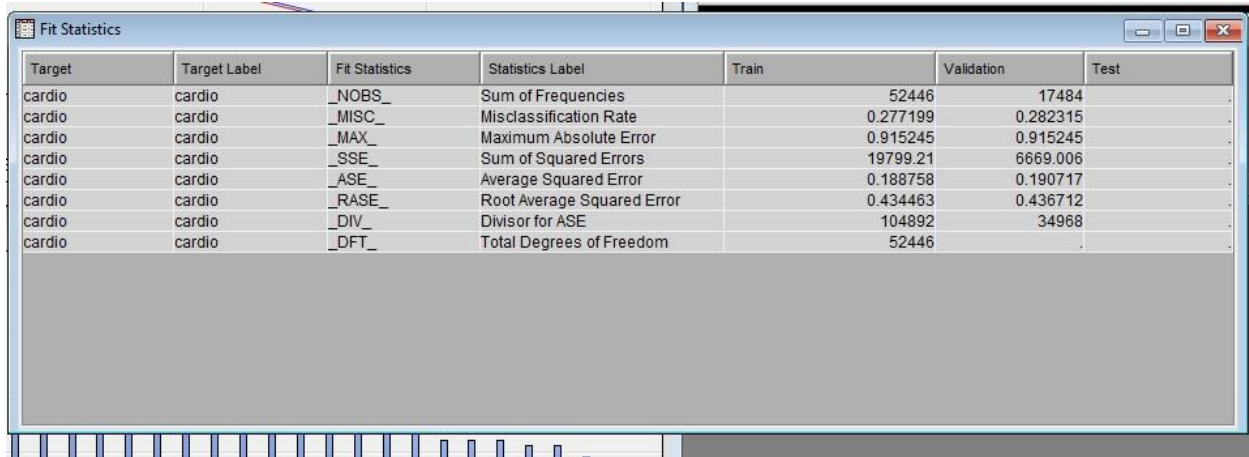
The image shows a 'Fit Statistics' window from a statistical software package. It displays a table of fit statistics for a decision tree model. The table has columns for Target, Target Label, Fit Statistics, Statistics Label, Train, Validation, and Test. The data is as follows:

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.277199	0.282315	
cardio	cardio	_MAX_	Maximum Absolute Error	0.834722	0.834722	
cardio	cardio	_SSE_	Sum of Squared Errors	19952.38	6707.145	
cardio	cardio	_ASE_	Average Squared Error	0.190218	0.191808	
cardio	cardio	_RASE_	Root Average Squared Error	0.43614	0.437959	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Freedom	52446		

Figure 24 Second Decision Tree with Cross Validation

Misclassification rate for the second decision tree with cross validation has increased to 0.2823 for validation data and 0.2772 for train data from the tree that achieved the results without cross validation which had the results of 0.281 for validation data and 0.2773 for train data. As a result, second decision tree with cross validation is not included in the final assessment of models.

Results of the third decision tree with cross validation to Yes are shown in Figure 25:



The image shows a 'Fit Statistics' window from a statistical software package. It displays a table of fit statistics for a decision tree model. The table has columns for Target, Target Label, Fit Statistics, Statistics Label, Train, Validation, and Test. The data is as follows:

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.277199	0.282315	
cardio	cardio	_MAX_	Maximum Absolute Error	0.915245	0.915245	
cardio	cardio	_SSE_	Sum of Squared Errors	19799.21	6669.006	
cardio	cardio	_ASE_	Average Squared Error	0.188758	0.190717	
cardio	cardio	_RASE_	Root Average Squared Error	0.434463	0.436712	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Freedom	52446		

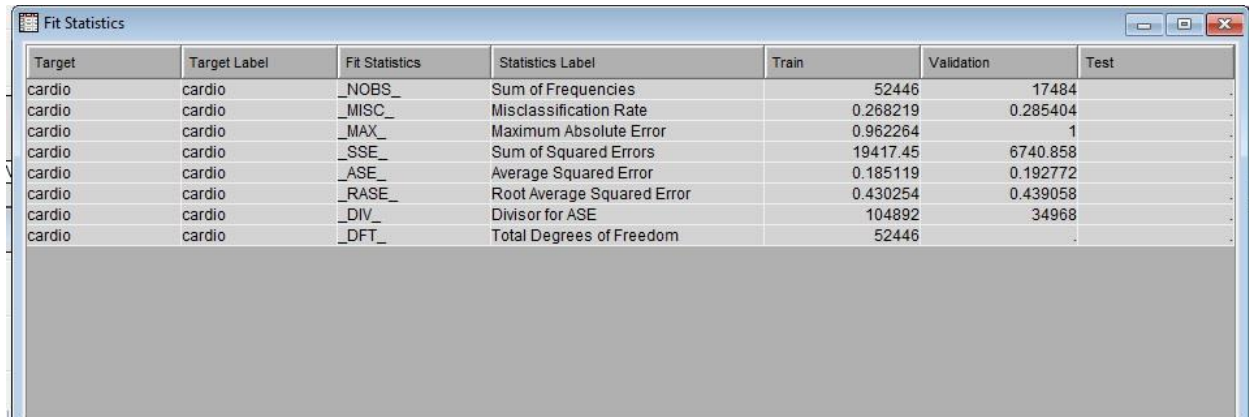
Figure 25 Third Decision Tree with Cross Validation

Performance of third decision tree with the cross validation has slightly improved from the without cross validation values of misclassification 0.2825 for validation data and 0.2778 for train data to the new values of 0.2772 for train data and 0.2823 of validation data. As a result, the third decision tree with cross validation will be used to the final comparison of the models.

Results of the fourth decision tree with cross validation are shown in Figure 26:



## Big Data for Decision Making

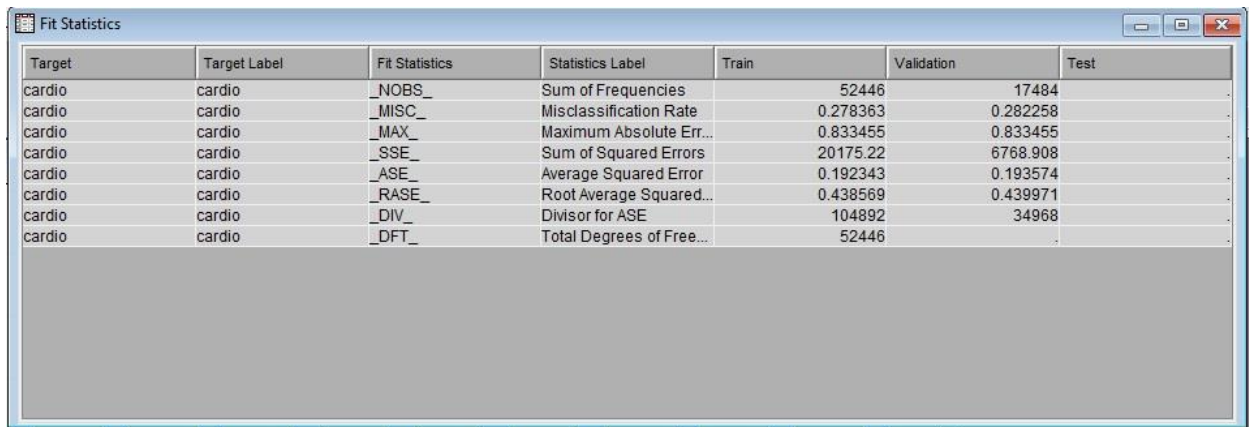


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.268219	0.285404	
cardio	cardio	_MAX_	Maximum Absolute Error	0.962264	1	
cardio	cardio	_SSE_	Sum of Squared Errors	19417.45	6740.858	
cardio	cardio	_ASE_	Average Squared Error	0.185119	0.192772	
cardio	cardio	_RASE_	Root Average Squared Error	0.430254	0.439058	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Freedom	52446		

Figure 26 Fourth Decision Tree with Cross Validation

Performance of the 4<sup>th</sup> decision tree has been decreased for the validation data to 0.2854 and improved for train data to 0.2682 from the values of tree without cross validation of 0.2778 for validation data and 0.2740 for the train data. As a result, decision tree without cross validation is kept for the final evaluation.

Results of the fifth decision tree with cross validation are shown in Figure 27:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.278363	0.282258	
cardio	cardio	_MAX_	Maximum Absolute Err...	0.833455	0.833455	
cardio	cardio	_SSE_	Sum of Squared Errors	20175.22	6768.908	
cardio	cardio	_ASE_	Average Squared Error	0.192343	0.193574	
cardio	cardio	_RASE_	Root Average Squared...	0.438569	0.439971	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Free...	52446		

Figure 27 Fifth Decision Tree with Cross Validation

Performance of the fifth decision tree with cross validation is exactly same as in without cross validation.

Results of sixth decision tree with cross validation are shown in Figure 28:

## Big Data for Decision Making

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.278305	0.282716	
cardio	cardio	_MAX_	Maximum Absolute Err...	0.814516	0.814516	
cardio	cardio	_SSE_	Sum of Squared Errors	20246.49	6820.342	
cardio	cardio	_ASE_	Average Squared Error	0.193022	0.195045	
cardio	cardio	_RASE_	Root Average Squared...	0.439343	0.441639	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Free...	52446		

Figure 28 Sixth Decision Tree with Cross Validation

Performance of sixth decision tree with cross validation has exactly same as without cross validation as a result, tree without cross validation will be used for the final assessment of models.

## 7 Neural Network

Neural network model uses the mathematical functions to map the inputs into outputs. In this assessment target variable is categorical and the results will be the probabilities of each even happening. Neural network is consisting of layers where the first layer is called input layer and last layer is called target layer.

Due to the running time of the neural network node only one neural network is use for the modelling. Properties of the neural network are shown in Figure 29:

.. Property	Value
<b>General</b>	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Misclassification
Suppress Output	No
<b>Score</b>	
Hidden Units	No
Residuals	Yes
Standardization	No

Figure 29 Neural Network Node Property

Variable was used same as decision tree which are shown in Figure 7. Model selection property is selected to the misclassification. Network properties are shown in Figure 30:

## Big Data for Decision Making

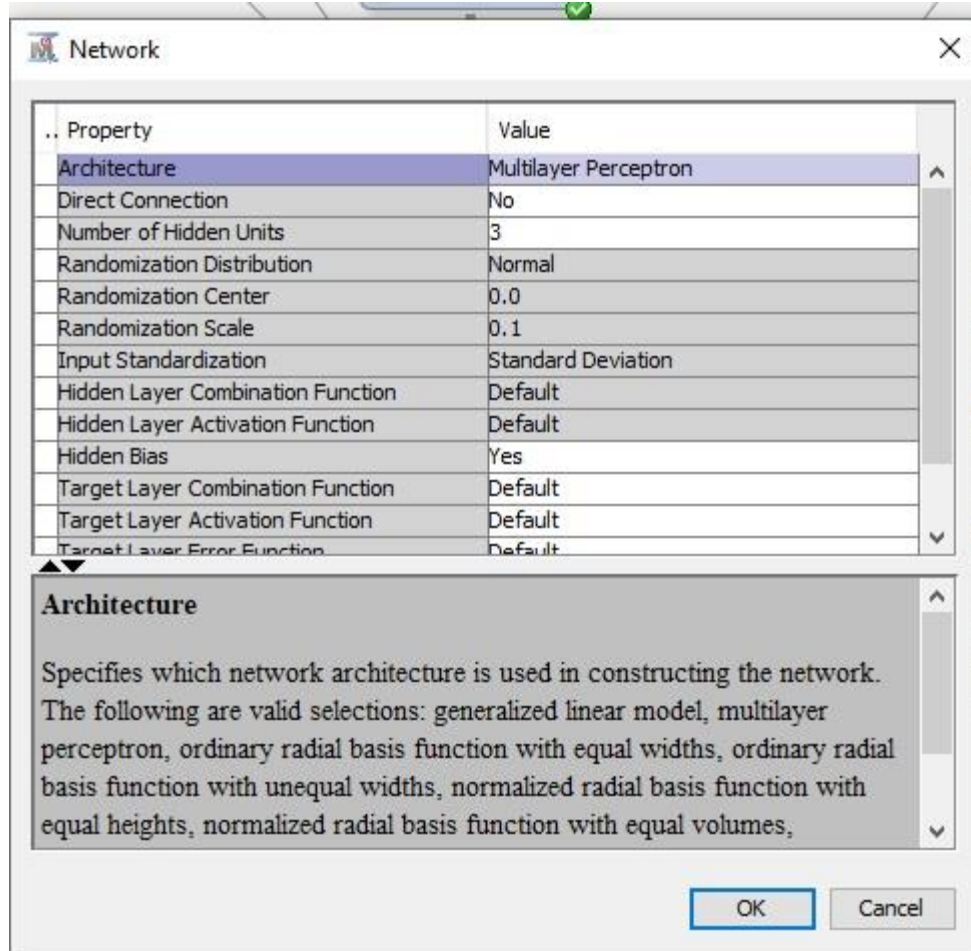


Figure 30 Neural Network

Multilayer perception (MLP) model is used as an architecture of the neural network. MLP consist of three layers Start layer, Hidden layer, and target layer. MLP is used as this can work on nonlinear functions. Hidden layer property is set to 3 as hyperparameter. Optimisation properties are shown in Figure 31:

# Big Data for Decision Making

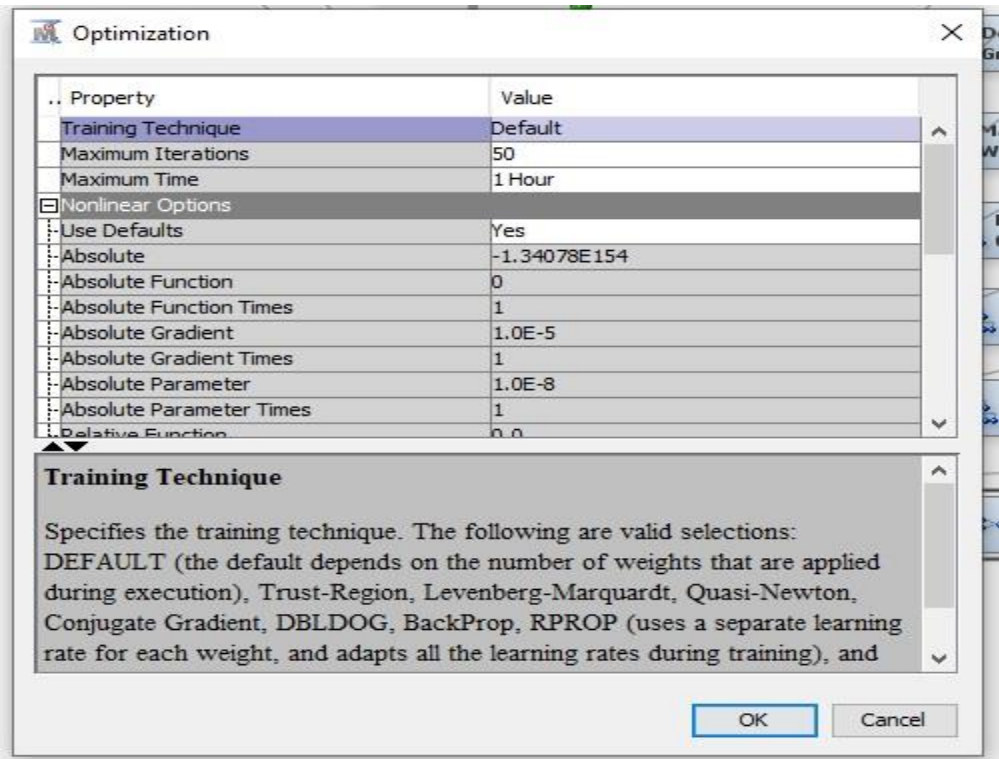


Figure 31 Neural Network Optimisation

Number of iterations for network are set to 50 and maximum time is set to 1 Hour. Results of neural network are shown in Figure 32:

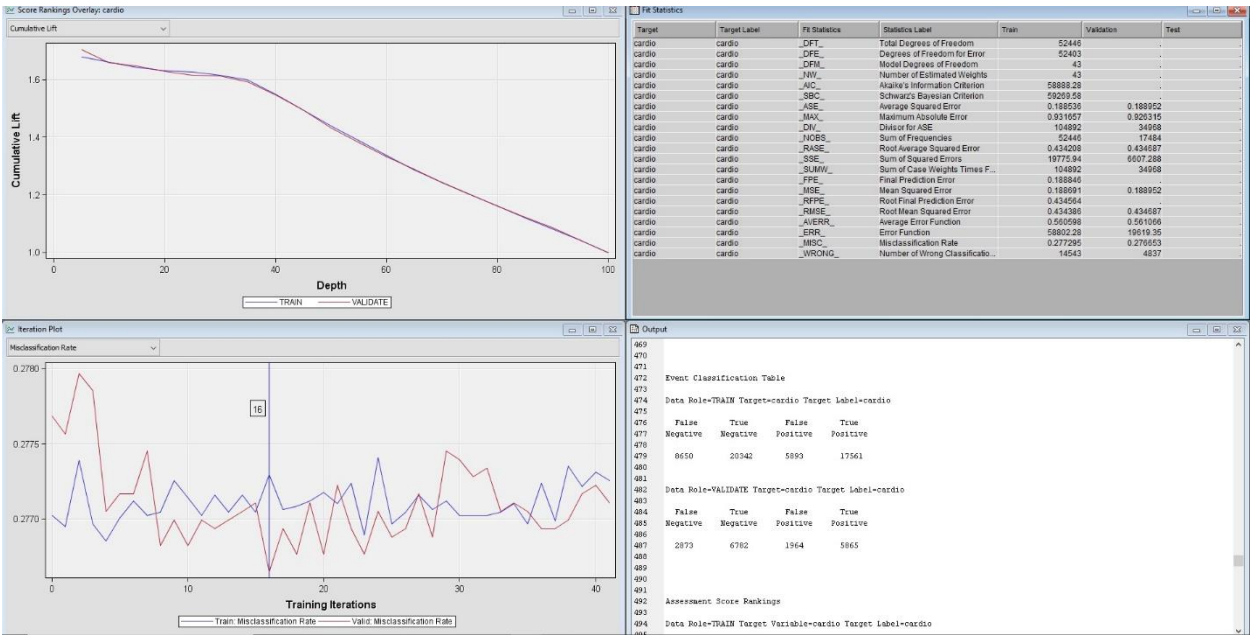


Figure 32 Neural Network Result

Results of the neural network shows the misclassifications for validation data of 27.67% and 27.73% for train data. Which makes the neural network as the best working model in comparison to the decision trees. Neural networks achieve this results in 16 iterations.

## 8 Regression

Regression analysis is used to prediction and in this assessment logistic regression analysis is used for prediction due to binary nature of the target variable.

### 8.1 First Regression Analysis

First regression analysis is used to predict the CVD values by using the same variables as shown in figure 7. First regression analysis is done without removing the skewness of the variables and one of the model selection criteria is used. Property of node are shown in Figure 33:

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

Figure 33 Regression Node Properties



## Big Data for Decision Making

Result of the regression are shown in figure 34:

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information Criterion	59758.52	.	.
cardio	cardio	_ASE_	Average Squared Error	0.191509	0.192166	.
cardio	cardio	_AVERR_	Average Error Function	0.569467	0.570866	.
cardio	cardio	_DFE_	Degrees of Freedom for Error	52433	.	.
cardio	cardio	_DFM_	Model Degrees of Freedom	13	.	.
cardio	cardio	_DFT_	Total Degrees of Freedom	52446	.	.
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	.
cardio	cardio	_ERR_	Error Function	59732.52	19962.04	.
cardio	cardio	_FPE_	Final Prediction Error	0.191604	.	.
cardio	cardio	_MAX_	Maximum Absolute Error	0.968679	0.974857	.
cardio	cardio	_MSE_	Mean Square Error	0.191556	0.192166	.
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	.
cardio	cardio	_NW_	Number of Estimate Weights	13	.	.
cardio	cardio	_RASE_	Root Average Sum of Squares	0.437617	0.438368	.
cardio	cardio	_RFPE_	Root Final Prediction Error	0.437726	.	.
cardio	cardio	_RMSE_	Root Mean Squared Error	0.437672	0.438368	.
cardio	cardio	_SBC_	Schwarz's Bayesian Criterion	59873.8	.	.
cardio	cardio	_SSE_	Sum of Squared Errors	20087.76	6719.674	.
cardio	cardio	_SUMW_	Sum of Case Weights Times Freq	104892	34968	.
cardio	cardio	_MISC_	Misclassification Rate	0.280174	0.281801	.

Figure 34 Regression results

Results shows that the averaged squared error is 0.1915 for train data and 0.1922 for the validation data. That shows that trained model has the errors of 19% for both train and validation data. Misclassification rate for the validation data is 0.2818 and train data is 0.2802. cumulative lift of the model as compared to baseline is shown in figure 35:

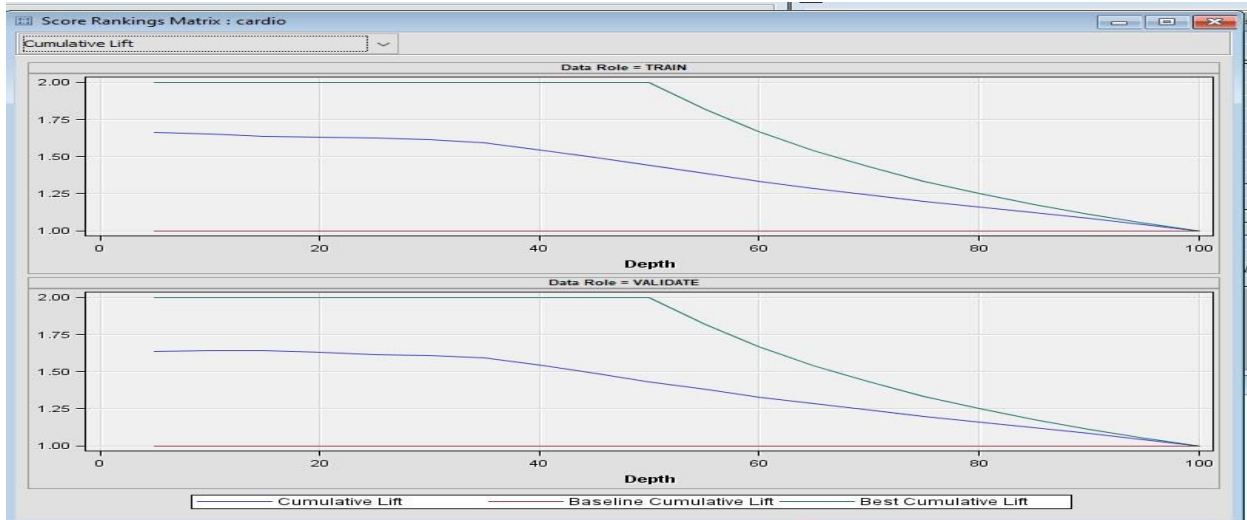


Figure 35 Regression with Baseline

This shows that the model is performing better as compared to the base line. Analysis of the model in shown in figure 36:

## Big Data for Decision Making

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
BMI	1	306.9041	<.0001
BP_Risk	2	5702.6907	<.0001
active	1	67.7630	<.0001
age	1	1339.4247	<.0001
alco	1	13.1070	0.0003
cholesterol	2	820.7025	<.0001
gender	1	18.3720	<.0001
gluc	2	52.8510	<.0001
smoke	1	13.0575	0.0003

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-3.7721	0.1037	1322.31	<.0001		0.023
BMI	1	0.0354	0.00202	306.90	<.0001	0.1026	1.036
BP_Risk 1	1	-0.9226	0.0184	2514.32	<.0001		0.397
BP_Risk 2	1	-0.3439	0.0137	629.19	<.0001		0.709
active 0	1	0.1024	0.0124	67.76	<.0001		1.108
age	1	0.000154	4.205E-6	1339.42	<.0001	0.2091	1.000
alco 0	1	0.0867	0.0239	13.11	0.0003		1.091
cholesterol 1	1	-0.4944	0.0181	745.01	<.0001		0.610
cholesterol 2	1	-0.1307	0.0231	32.11	<.0001		0.878
gender 0	1	-0.0479	0.0112	18.37	<.0001		0.953
gluc 1	1	0.0948	0.0212	20.09	<.0001		1.099
gluc 2	1	0.1307	0.0300	19.05	<.0001		1.140
smoke 0	1	0.0716	0.0198	13.06	0.0003		1.074

Odds Ratio Estimates		
Effect		Point Estimate
BMI		1.036
BP_Risk	1 vs 3	0.112
BP_Risk	2 vs 3	0.200
active	0 vs 1	1.227
age		1.000
alco	0 vs 1	1.189
cholesterol	1 vs 3	0.326
cholesterol	2 vs 3	0.470
gender	0 vs 1	0.909
gluc	1 vs 3	1.378
gluc	2 vs 3	1.428
smoke	0 vs 1	1.154

Figure 36 Regression Analysis

## Big Data for Decision Making

First table in the figure shows the correlation of each variable as compared to the target variable and identified age as the most important variable. Odds ratios having the values of greater than 1 show that the increment in these variables can change the model output and it does have effect on the predicted output. As the odd ratio is calculated on the basis of retain over not retained and value greater then 1 mean variable is making impact to the predictions. Significance of the model is shown in figure 37:

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood Ratio		DF	Pr > ChiSq
Intercept Only	Intercept & Covariates	Chi-Square			
72705.583	59732.519	12973.0639		12	<.0001

Figure 37 Regression Model Significance

Chi-square value is 12973.06 which shows the high significance of the model.

As seen in the statexplore node from the group part of assignment that there are some variable values that skewed, and skewness need to be removed to help the model make the better predictions. Summary of the interval variables in shown in the Figure 38:

Interval Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
BMI	INPUT	27.49662	5.266293	52446	0	12.25447	26.39798	64.51613	1.228916	2.69538
age	INPUT	19471.17	2464.718	52446	0	10798	19705	23713	-0.30957	-0.82059
ap_hi	INPUT	1.689071	0.877953	52446	0	1	1	3	0.647434	-1.39106
ap_lo	INPUT	2.043492	0.700214	52446	0	1	2	3	-0.05989	-0.96079
height	INPUT	164.3963	7.95342	52446	0	108	165	200	-0.01041	0.881783
weight	INPUT	74.20467	14.35912	52446	0	35	72	183	0.979029	2.239962

Data Role=VALIDATE

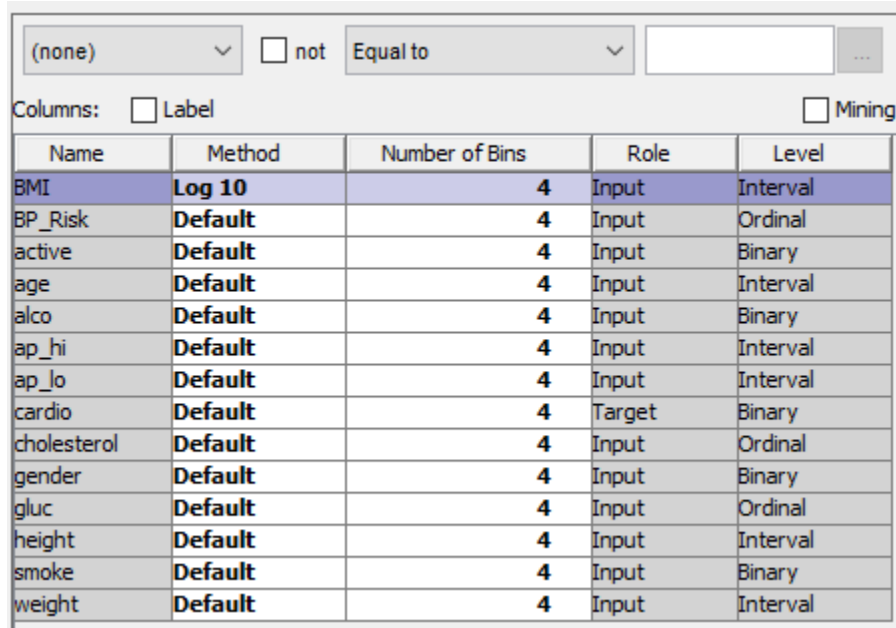
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
BMI	INPUT	27.47264	5.218787	17484	0	14.57726	26.34649	63.9754	1.254165	2.778871
age	INPUT	19461.22	2474.611	17484	0	10859	19698	23701	-0.29961	-0.83209
ap_hi	INPUT	1.684054	0.875696	17484	0	1	1	3	0.658906	-1.37332
ap_lo	INPUT	2.037005	0.702626	17484	0	1	2	3	-0.05147	-0.97442
height	INPUT	164.456	7.93577	17484	0	100	165	207	-0.04839	1.025024
weight	INPUT	74.18074	14.15221	17484	0	36	72	200	0.986511	2.411803

Figure 38 Variables Skewness



## Big Data for Decision Making

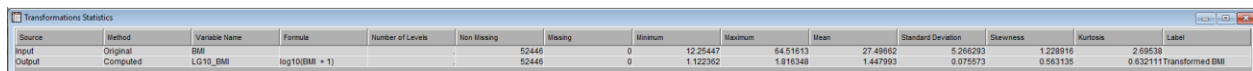
As a rule of thumb if skewness of variables is less than -1 and greater than 1 then the distribution of the variable is highly skewed and if the skewness is between -1 and -0.5 or 0.5 to 1 it is moderately skewed. As shown in the figure 38 BMI variable has skewed values and need to be change the distribution using the log transformation. To remove the skewness of BMI the variable transformation node is used, and settings of the node are shown in Figure 39:



Name	Method	Number of Bins	Role	Level
BMI	Log 10	4	Input	Interval
BP_Risk	Default	4	Input	Ordinal
active	Default	4	Input	Binary
age	Default	4	Input	Interval
alco	Default	4	Input	Binary
ap_hi	Default	4	Input	Interval
ap_lo	Default	4	Input	Interval
cardio	Default	4	Target	Binary
cholesterol	Default	4	Input	Ordinal
gender	Default	4	Input	Binary
gluc	Default	4	Input	Ordinal
height	Default	4	Input	Interval
smoke	Default	4	Input	Binary
weight	Default	4	Input	Interval

Figure 39 BMI Skewness Removed

Result of the transformation node are shown in Figure 40:



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	BMI			52446	0	12.25447	64.51613	27.49562	5.266293	1.228916	2.69538	
Output	Computed	LOG10_BMI	log10(BMI + 1)		52446	0	1.122362	1.810348	1.447993	0.075573	0.563135	0.632111	Transformed BMI

Figure 40 BMI Transformed.

Now the skewness of BMI has been decreased from 1.2289 to 0.5632 which is an acceptable range.

### 8.2 Second Regression Analysis

Variable used for second regression analysis are shown in Figure 41:

## Big Data for Decision Making

Variables - Reg2

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mini

Name	Use	Report	Role	Level
BP_Risk	No	No	Input	Ordinal
LG10_BMI	Default	No	Input	Interval
active	Default	No	Input	Binary
age	Default	No	Input	Interval
alco	Default	No	Input	Binary
ap_hi	No	No	Input	Interval
ap_lo	No	No	Input	Interval
cardio	Yes	No	Target	Binary
cholesterol	Default	No	Input	Ordinal
gender	Default	No	Input	Binary
gluc	Default	No	Input	Ordinal
height	No	No	Input	Interval
smoke	Default	No	Input	Binary
weight	No	No	Input	Interval

Figure 41 Regression with no skewness Analysis

This regression is run without selecting any model selection criteria and fit statistics of the analysis are shown in Figure 42:

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information C...	59742.85		
cardio	cardio	_ASE_	Average Squared Error	0.191461	0.192052	
cardio	cardio	_AVERR_	Average Error Function	0.569318	0.57049	
cardio	cardio	_DFE_	Degrees of Freedom f...	52433		
cardio	cardio	_DFM_	Model Degrees of Fre...	13		
cardio	cardio	_DFT_	Total Degrees of Free...	52446		
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_ERR_	Error Function	59716.85	19948.9	
cardio	cardio	_FPE_	Final Prediction Error	0.191556		
cardio	cardio	_MAX_	Maximum Absolute Err...	0.968234	0.971036	
cardio	cardio	_MSE_	Mean Square Error	0.191508	0.192052	
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_NW_	Number of Estimate ...	13		
cardio	cardio	_RASE_	Root Average Sum of ...	0.437563	0.438237	
cardio	cardio	_RFPE_	Root Final Prediction ...	0.437671		
cardio	cardio	_RMSE_	Root Mean Squared E...	0.437617	0.438237	
cardio	cardio	_SBC_	Schwarz's Bayesian C...	59858.13		
cardio	cardio	_SSE_	Sum of Squared Errors	20082.72	6715.666	
cardio	cardio	_SUMW_	Sum of Case Weights ...	104892	34968	
cardio	cardio	_MISC_	Misclassification Rate	0.280098	0.281114	

Figure 42 Fit Statistics of Second Regression

## Big Data for Decision Making

As a result, performance of the model has improved slightly as the misclassification rate has dropped a little. Odds ratio, Chi-square value of model is shown in Figure 43:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
BP_Risk	2	5668.4986	<.0001
LG10_BMI	1	324.0818	<.0001
active	1	67.7671	<.0001
age	1	1332.9348	<.0001
alco	1	13.2772	0.0003
cholesterol	2	819.0359	<.0001
gender	1	17.8651	<.0001
gluc	2	52.6917	<.0001
smoke	1	12.7797	0.0004

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-6.4342	0.2207	850.11	<.0001		0.002
BP_Risk 1	1	-0.9194	0.0184	2491.54	<.0001		0.399
BP_Risk 2	1	-0.3443	0.0137	630.16	<.0001		0.709
LG10_BMI	1	2.5135	0.1396	324.08	<.0001	0.1047	12.348
active	0	0.1024	0.0124	67.77	<.0001		1.108
age	1	0.000154	4.206E-6	1332.93	<.0001	0.2087	1.000
alco	0	0.0872	0.0239	13.28	0.0003		1.091
cholesterol 1	1	-0.4938	0.0181	743.23	<.0001		0.610
cholesterol 2	1	-0.1308	0.0231	32.20	<.0001		0.877
gender	0	-0.0472	0.0112	17.87	<.0001		0.954
gluc	1	0.0950	0.0212	20.19	<.0001		1.100
gluc	2	0.1300	0.0299	18.84	<.0001		1.139
smoke	0	0.0708	0.0198	12.78	0.0004		1.073

Odds Ratio Estimates		
Effect		Point Estimate
BP_Risk	1 vs 3	0.113
BP_Risk	2 vs 3	0.200
LG10_BMI		12.348
active	0 vs 1	1.227
age		1.000
alco	0 vs 1	1.191
cholesterol	1 vs 3	0.327
cholesterol	2 vs 3	0.470
gender	0 vs 1	0.910
gluc	1 vs 3	1.377
gluc	2 vs 3	1.426
smoke	0 vs 1	1.152

Figure 43 Second Regression Model Output

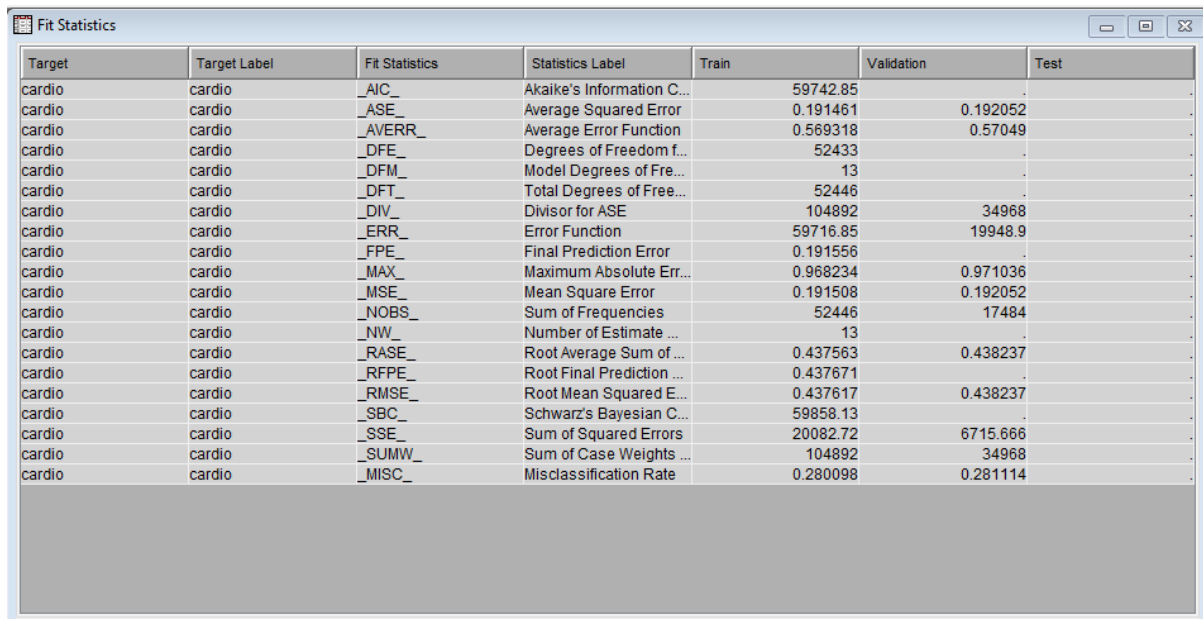
By removing the skewness of BMI its significance has been increased in the model with chi-square value of 324.08 and BP\_Risk is still as the variable with the highest significance. Odds ratio of the modified BMI has been increased as well.

For the next three regression model variables are used same as second regression model however for third regression model selection criterion is backward selection, fourth regression model criterion is forward selection, and fifth model criterion is stepwise selection.

## Big Data for Decision Making

### 8.3 Third Regression Model

Fit statistics of the model are shown in the Figure 44:



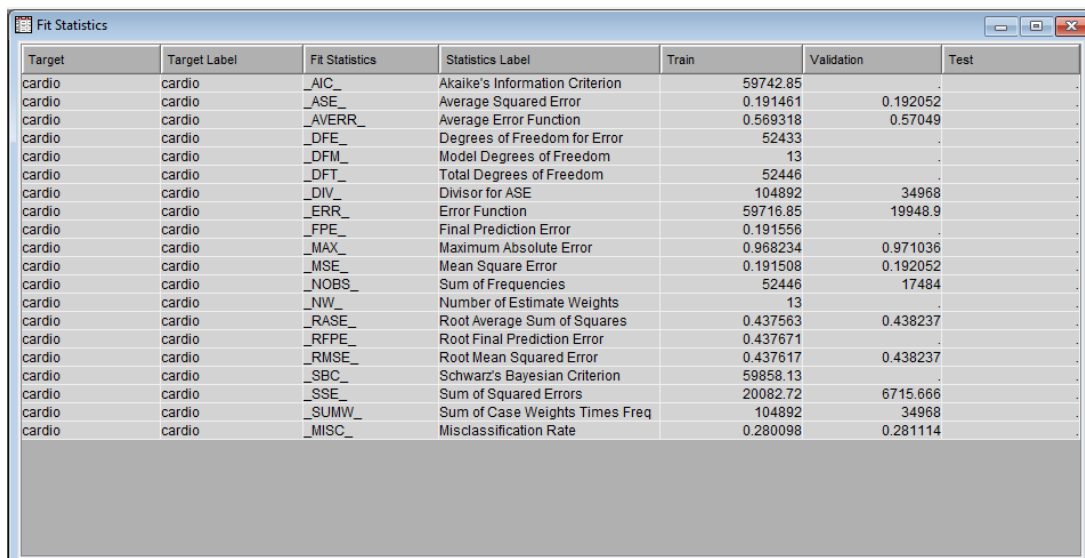
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information C...	59742.85	.	.
cardio	cardio	_ASE_	Average Squared Error	0.191461	0.192052	.
cardio	cardio	_AVERR_	Average Error Function	0.569318	0.57049	.
cardio	cardio	_DFE_	Degrees of Freedom f...	52433	.	.
cardio	cardio	_DFM_	Model Degrees of Fre...	13	.	.
cardio	cardio	_DFT_	Total Degrees of Free...	52446	.	.
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	.
cardio	cardio	_ERR_	Error Function	59716.85	19948.9	.
cardio	cardio	_FPE_	Final Prediction Error	0.191556	.	.
cardio	cardio	_MAX_	Maximum Absolute Err...	0.968234	0.971036	.
cardio	cardio	_MSE_	Mean Square Error	0.191508	0.192052	.
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	.
cardio	cardio	_NW_	Number of Estimate ...	13	.	.
cardio	cardio	_RASE_	Root Average Sum of ...	0.437563	0.438237	.
cardio	cardio	_RFPE_	Root Final Prediction ...	0.437671	.	.
cardio	cardio	_RMSE_	Root Mean Squared E...	0.437617	0.438237	.
cardio	cardio	_SBC_	Schwarz's Bayesian C...	59858.13	.	.
cardio	cardio	_SSE_	Sum of Squared Errors	20082.72	6715.666	.
cardio	cardio	_SUMW_	Sum of Case Weights ...	104892	34968	.
cardio	cardio	_MISC_	Misclassification Rate	0.280098	0.281114	.

Figure 44 Third Regression Fit Statistics

Performance of the model with backward selection criterion has not improve it has same performance as the second regression model by having the same misclassification rate.

### 8.4 Fourth Regression Model

Same variables are used for this model as in second and third model, only difference is the model selection criterion which is forward for third regression model, and results are shown in Figure 45:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information Criterion	59742.85	.	.
cardio	cardio	_ASE_	Average Squared Error	0.191461	0.192052	.
cardio	cardio	_AVERR_	Average Error Function	0.569318	0.57049	.
cardio	cardio	_DFE_	Degrees of Freedom for Error	52433	.	.
cardio	cardio	_DFM_	Model Degrees of Freedom	13	.	.
cardio	cardio	_DFT_	Total Degrees of Freedom	52446	.	.
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	.
cardio	cardio	_ERR_	Error Function	59716.85	19948.9	.
cardio	cardio	_FPE_	Final Prediction Error	0.191556	.	.
cardio	cardio	_MAX_	Maximum Absolute Error	0.968234	0.971036	.
cardio	cardio	_MSE_	Mean Square Error	0.191508	0.192052	.
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	.
cardio	cardio	_NW_	Number of Estimate Weights	13	.	.
cardio	cardio	_RASE_	Root Average Sum of Squares	0.437563	0.438237	.
cardio	cardio	_RFPE_	Root Final Prediction Error	0.437671	.	.
cardio	cardio	_RMSE_	Root Mean Squared Error	0.437617	0.438237	.
cardio	cardio	_SBC_	Schwarz's Bayesian Criterion	59858.13	.	.
cardio	cardio	_SSE_	Sum of Squared Errors	20082.72	6715.666	.
cardio	cardio	_SUMW_	Sum of Case Weights Times Freq	104892	34968	.
cardio	cardio	_MISC_	Misclassification Rate	0.280098	0.281114	.

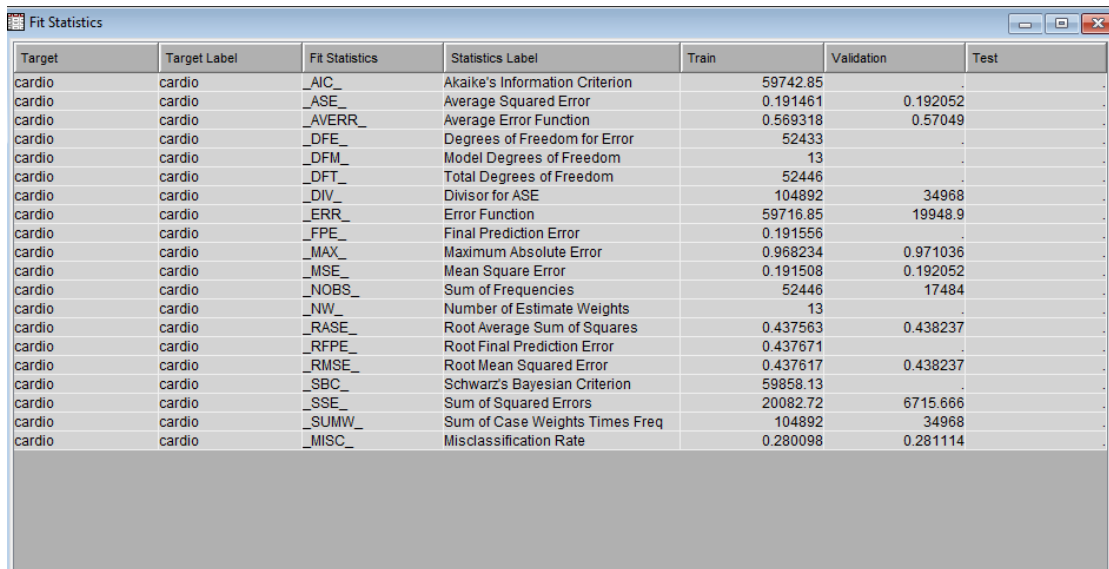
Figure 45 Third Regression Model Fit Statistics

## Big Data for Decision Making

The results of the model are same as third and second regression model which have the same values of misclassification and mean squared error for both models.

### 8.5 Fifth Regression Model

The variables used in this model are same as previous regression model the difference is the model selection criterion which use the stepwise selection. Results of the model are shown in Figure 46:



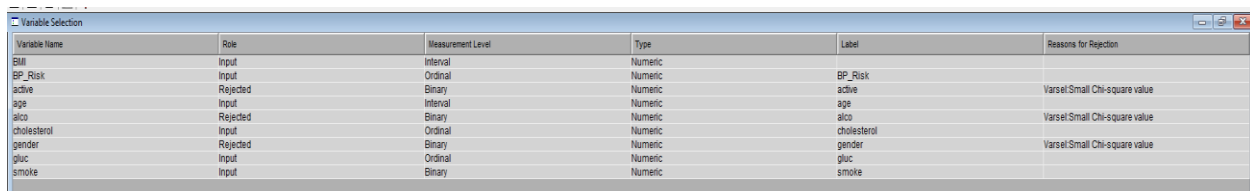
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information Criterion	59742.85	.	.
cardio	cardio	_ASE_	Average Squared Error	0.191461	0.192052	.
cardio	cardio	_AVERR_	Average Error Function	0.569318	0.57049	.
cardio	cardio	_DFE_	Degrees of Freedom for Error	52433	.	.
cardio	cardio	_DFM_	Model Degrees of Freedom	13	.	.
cardio	cardio	_DFT_	Total Degrees of Freedom	52446	.	.
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	.
cardio	cardio	_ERR_	Error Function	59716.85	19948.9	.
cardio	cardio	_FPE_	Final Prediction Error	0.191556	.	.
cardio	cardio	_MAX_	Maximum Absolute Error	0.968234	0.971036	.
cardio	cardio	_MSE_	Mean Square Error	0.191508	0.192052	.
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	.
cardio	cardio	_NW_	Number of Estimate Weights	13	.	.
cardio	cardio	_RASE_	Root Average Sum of Squares	0.437563	0.438237	.
cardio	cardio	_RFPE_	Root Final Prediction Error	0.437671	.	.
cardio	cardio	_RMSE_	Root Mean Squared Error	0.437617	0.438237	.
cardio	cardio	_SBC_	Schwarz's Bayesian Criterion	59858.13	.	.
cardio	cardio	_SSE_	Sum of Squared Errors	20082.72	6715.666	.
cardio	cardio	_SUMW_	Sum of Case Weights Times Freq	104892	34968	.
cardio	cardio	_MISC_	Misclassification Rate	0.280098	0.281114	.

Figure 46 Fifth Regression Model Fit Statistics

The result of the regression model is same as previous model there is no difference in the model performance.

## 9 Decision Tree with Selected Variables

For this purpose, variable selection node has been used where the variables with the highest chi-square values are used and then these variables are used for decision tree if it can improve the model performance. Variable selection node results are shown in Figure 47:



Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
BMI	Input	Interval	Numeric		
BP_Risk	Input	Ordinal	Numeric	BP_Risk	
active	Rejected	Binary	Numeric	active	Varsel Small Chi-square value
age	Input	Interval	Numeric		
alco	Rejected	Binary	Numeric	alco	Varsel Small Chi-square value
cholesterol	Input	Ordinal	Numeric	cholesterol	
gender	Rejected	Binary	Numeric	gender	Varsel Small Chi-square value
gluc	Input	Ordinal	Numeric	gluc	
smoke	Input	Binary	Numeric	smoke	

Figure 47 Variable Selection with Selection Node

Variable selection node has dropped the age, alco and gender due to low chi-square values.

## Big Data for Decision Making

For this decision depth of tree as hyperparameter is set to 10 and Entropy rule splitting criteria is used with the cross assessment to Yes. Results are shown in Figure 48:

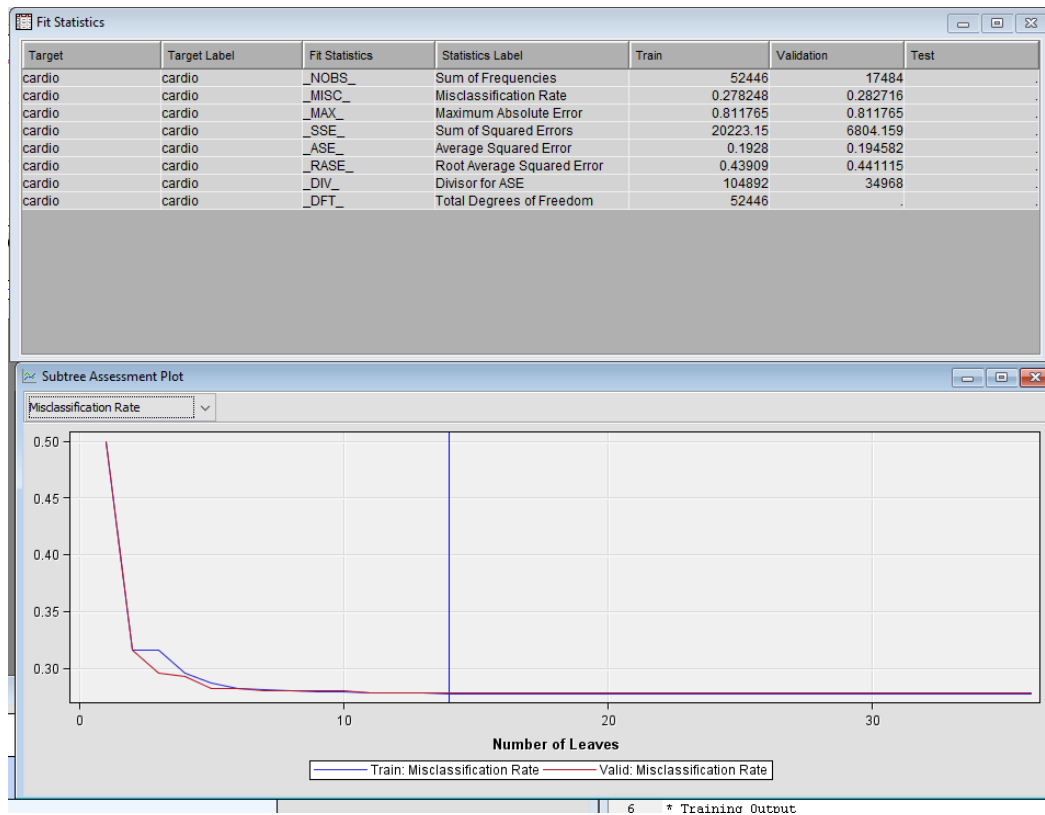


Figure 48 Decision Tree with Selected Variables

Performance of the tree has not been improved it is same as previously develop models.

A different approach has been used to train the model and make better predictions. The dataset is used in original form with separate weight and heigh values in place of BMI and separate ap\_hi (Upper blood pressure) values and ap\_lo (lower blood pressure) values in place of using BP\_Risk which was the average of both ap\_hi and ap\_lo values.

Variables that will be used for the following models are shown in Figure 49:

## Big Data for Decision Making

2. Variables - trees

(none) ☐ not ☐ Equal to

Columns: ☐ Label

Name	Use	Report	Role	Level
ap_hi	No	No	Input	Interval
ap_lo	No	No	Input	Interval
active	Default	No	Input	Binary
alco	Default	No	Input	Binary
ap_hi	Default	No	Input	Interval
ap_lo	Default	No	Input	Interval
cardio	Yes	No	Target	Binary
cholesterol	Default	No	Input	Interval
gender	Default	No	Input	Binary
gluc	Default	No	Input	Interval
height	Default	No	Input	Interval
smoke	Default	No	Input	Binary
weight	Default	No	Input	Interval

Figure 49 Original Dataset Variables

In this modelling process I have dropped the variables that were created in the feature engineering process. Results of the first decision tree with these variables of depth 20 are cross validation Yes and assessment criteria misclassification is used and results are shown in Figure 50:

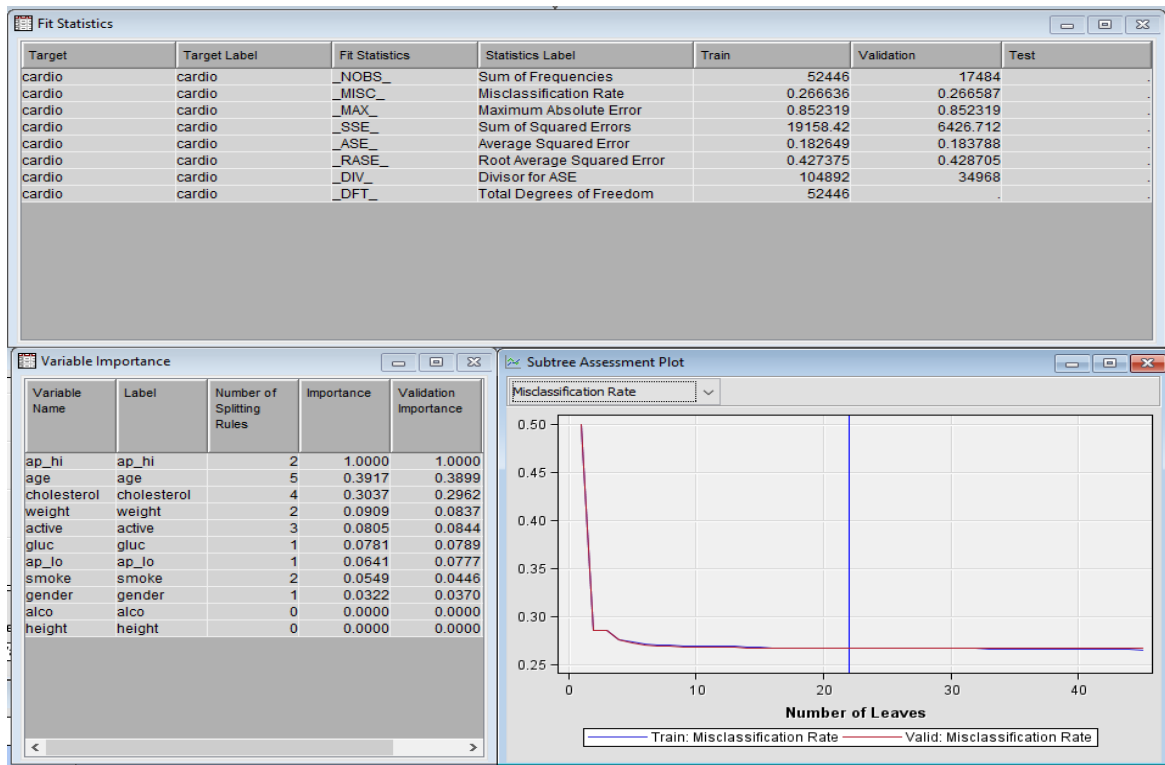


Figure 50 Original Variable Decision Tree

Performance of tree with original dataset variables has been improve by 2% for both training and validation data, and Tree achieved the misclassification of 26% by the 22 leaves. Variable importance table shows the variable that are significant in the model prediction with ap\_hi as the most important variable, alco and height as the most insignificant variables. Results of the same decision tree with No cross validation are shown in Figure 51:

## Big Data for Decision Making

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_MISC_	Misclassification Rate	0.266636	0.266587	
cardio	cardio	_MAX_	Maximum Absolute Err...	0.852319	0.852319	
cardio	cardio	_SSE_	Sum of Squared Errors	19158.42	6426.712	
cardio	cardio	_ASE_	Average Squared Error	0.182649	0.183788	
cardio	cardio	_RASE_	Root Average Squared...	0.427375	0.428705	
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_DFT_	Total Degrees of Free...	52446		

Figure 51 Decision Tree without Cross Validation

Now the third tree with original data set with the depth of 10 is built and results are shown in Figure 52:

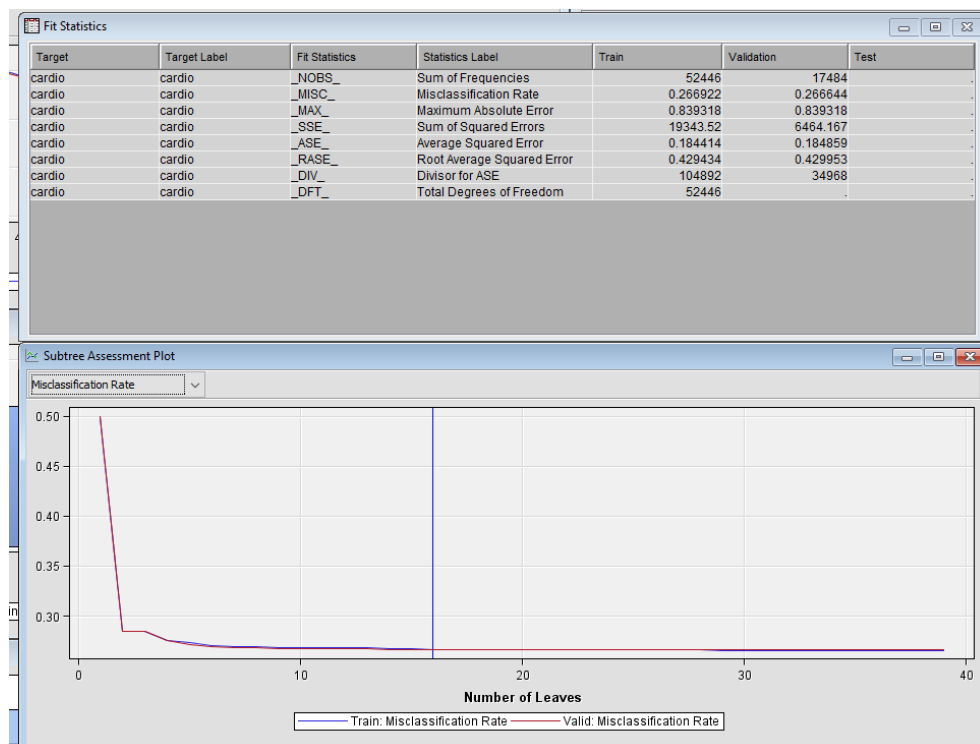


Figure 52 Decision Tree with Original Dataset and Depth 10

Now the performance of the in very small decimal points and tree and obtained this result with the max of 18 leaves.

## 10 Original Dataset Neural Network

Same settings of neural network have been used as the previous neural network, with MLP model, 3 hidden layers and max iteration to 50. Results are shown in Figure 53:



# Big Data for Decision Making

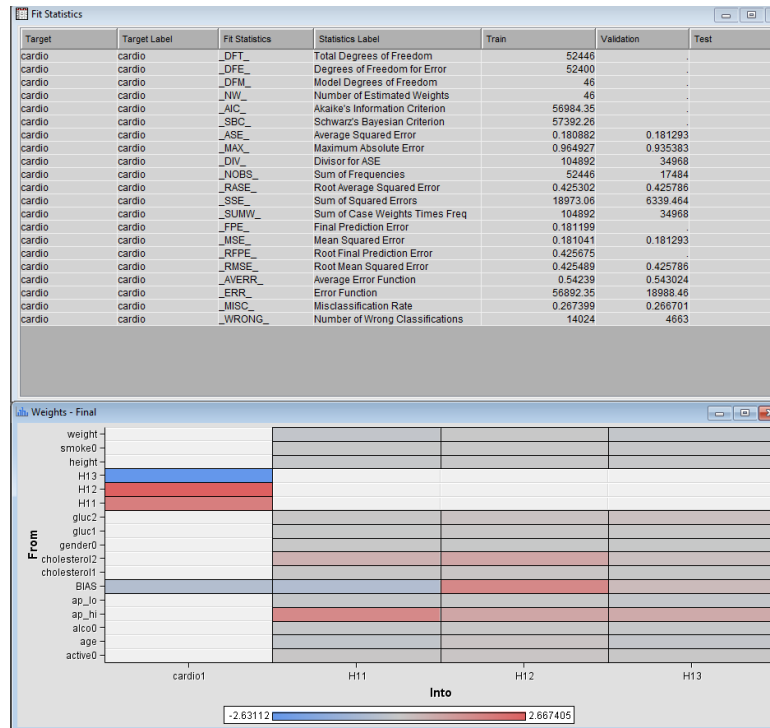


Figure 53 Original Dataset Neural Network

Form the result of neural network it shows that model has done misclassification to 26% for both training and validation data and average squared error is 18% for both as well. From the weight table it shows the number of hidden layers in the network and weight that neural network has assigned them as given instruction in the model property 3 hidden layers were selected and those hidden layers are shown with name H11, H12 and H13.

## 11 Original Dataset Regression

Same variables as shown I figure 49 are used for the modelling. As observed from previous regression analysis that model selection process does not impact much on this dataset predictions. So, no model is selected for the analysis and results are shown in Figure54:

## Big Data for Decision Making

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information Criterion	57994.98		
cardio	cardio	_ASE_	Average Squared Error	0.18426	0.184761	
cardio	cardio	_AVER_	Average Error Function	0.552635	0.553728	
cardio	cardio	_DFE_	Degrees of Freedom for Error	52432		
cardio	cardio	_DFM_	Model Degrees of Freedom	14		
cardio	cardio	_DFT_	Total Degrees of Freedom	52446		
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_ERR_	Error Function	57966.98	19362.77	
cardio	cardio	_FPE_	Final Prediction Error	0.184358		
cardio	cardio	_MAX_	Maximum Absolute Error	0.970989	0.963725	
cardio	cardio	_MSE_	Mean Square Error	0.184309	0.184761	
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_NW_	Number of Estimate Weights	14		
cardio	cardio	_RASE_	Root Average Sum of Squares	0.429255	0.429838	
cardio	cardio	_RFPE_	Root Final Prediction Error	0.42937		
cardio	cardio	_RMSE_	Root Mean Squared Error	0.429312	0.429838	
cardio	cardio	_SBC_	Schwarz's Bayesian Criterion	58119.12		
cardio	cardio	_SSE_	Sum of Squared Errors	19327.39	6460.717	
cardio	cardio	_SUMW_	Sum of Case Weights Times Freq	104892	34968	
cardio	cardio	_MISC_	Misclassification Rate	0.27045	0.272306	

Figure 54 Regression Analysis with Original Dataset

This regression model while comparing to previous regression models with featured engineered dataset works better and have better misclassification result which is less than 1%.

Regression model with original dataset and skewness removed. Statexplore results are shown in Figure 55:

Data Role=TRAIN										
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	19471.17	2464.718	52446	0	10798	19705	23713	-0.30957	-0.82059
ap_hi	INPUT	1.689071	0.877953	52446	0	1	1	3	0.647434	-1.39106
ap_lo	INPUT	2.043492	0.700214	52446	0	1	2	3	-0.05989	-0.96079
height	INPUT	164.3963	7.95342	52446	0	108	165	200	-0.01041	0.881783
weight	INPUT	74.20467	14.35912	52446	0	35	72	183	0.979029	2.239962

Figure 55 Original Dataset Variable Skewness

As explained in rule of thumb and by looking at the skewness value of age of and height need to properly distribute. As for the values of ap\_hi and ap\_lo are in categorical order with 1, 2 and 3 values. Values are transformed using the transformed node and results are shown in Figure 56:

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	age			52446	0	10798	23713	19471.17	2464.718	-0.30957	-0.82059	age
Input	Original	height			52446	0	108	200	164.3963	7.95342	-0.01041	0.881783	height
Output	Computed	LG10_age	log10(age + 1)		52446	0	4.033384	4.375005	4.285777	0.056885	-0.52524	-0.60953	Transformed: age
Output	Computed	LG10_height	log10(height + 1)		52446	0	2.037426	2.303196	2.218021	0.020975	-0.24321	1.59595	Transformed: height

Figure 56 Original Dataset without age and height Skewness

## Big Data for Decision Making

Regression analysis with transformed variables is run and results are shown in Figure 57:

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
cardio	cardio	_AIC_	Akaike's Information Criterion	58007.23		
cardio	cardio	_AVERR_	Average Error Function	0.552752	0.553903	
cardio	cardio	_ASE_	Average Squared Error	0.18437	0.184898	
cardio	cardio	_DFE_	Degrees of Freedom for Error	52432		
cardio	cardio	_DIV_	Divisor for ASE	104892	34968	
cardio	cardio	_ERR_	Error Function	57979.23	19368.9	
cardio	cardio	_FPE_	Final Prediction Error	0.184469		
cardio	cardio	_MAX_	Maximum Absolute Error	0.96976	0.963433	
cardio	cardio	_MSE_	Mean Square Error	0.18442	0.184898	
cardio	cardio	_MISC_	Misclassification Rate	0.271003	0.27305	
cardio	cardio	_DFM_	Model Degrees of Freedom	14		
cardio	cardio	_NW_	Number of Estimate Weights	14		
cardio	cardio	_RASE_	Root Average Sum of Squares	0.429384	0.429998	
cardio	cardio	_RFPE_	Root Final Prediction Error	0.429498		
cardio	cardio	_RMSE_	Root Mean Squared Error	0.429441	0.429998	
cardio	cardio	_SBC_	Schwarz's Bayesian Criterion	58131.38		
cardio	cardio	_SUMW_	Sum of Case Weights Times Freq	104892	34968	
cardio	cardio	_NOBS_	Sum of Frequencies	52446	17484	
cardio	cardio	_SSE_	Sum of Squared Errors	19338.99	6465.511	
cardio	cardio	_DFT_	Total Degrees of Freedom	52446		

Figure 57 Original Dataset with Skewness removed

Model has performance better as compared to the previous regression model and has the misclassification rate of 27.1% for train data and 27.3% for the validation data.

## 12 Model Comparison

To compare the performance of developed models and compare the performance Model Comparison node is used from Assess part of SEMMA modelling process and Results are shown in Figure 58:

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree8	Decision Tree with AP values Depth (20)	0.26659	0.18265	0.26664	0.18379
	Tree9	Decision Tree with AP values Depth (10)	0.26664	0.18441	0.26692	0.18486
	Neural2	Neural Network with APV values	0.26670	0.18088	0.26740	0.18129
	Reg6	Regression with APV values	0.27231	0.18426	0.27045	0.18476
	Reg7	Regression with skewness and AP values	0.27305	0.18437	0.27100	0.18490
	Neural	Neural Network	0.27665	0.18854	0.27729	0.18895
	Tree4	Decision Tree Group Variables Depth (10) Splitting Entropy	0.27797	0.18861	0.27405	0.19099
	Reg2	Regression with no Skewness	0.28111	0.19146	0.28010	0.19205
	Reg3	Regression with NS and Backward Model	0.28111	0.19146	0.28010	0.19205
	Reg4	Regression with NS and Forward selection	0.28111	0.19146	0.28010	0.19205
	Reg5	Regression with NS and stepwise Model selection	0.28111	0.19146	0.28010	0.19205
	Reg	Regression with none model selection	0.28180	0.19151	0.28017	0.19217
	Tree2	Decision Tree Group Variables Depth (10)	0.28186	0.19068	0.27731	0.19207
	Tree5	Decision Tree Group Variables Depth (6)	0.28226	0.19326	0.27836	0.19485
	Tree	Maximal Tree With Group Variables	0.28232	0.18876	0.27720	0.19072
	Tree3	Decision Tree Group Variables Depth (10) Assessment AVG With Cross Validation	0.28232	0.18876	0.27720	0.19072
	Tree7	Decision Tree with Selected Variables	0.28272	0.19280	0.27825	0.19458
	Tree6	Decision Tree Group Variables Depth (6) Entropy	0.28272	0.19302	0.27831	0.19505

Figure 58 Model Comparison Node Results

## Big Data for Decision Making

Model comparison node has selected Decision Tree which used the original dataset values and depth to 20 as the best performing model which shows the misclassification rate for the validation data of 0.26659 and for train data misclassification rate is 0.2664. SAS code generated for this decision tree can be used for the prediction purposes. This result shows that model will classify:

Classification Accuracy (validation Data) = 1 – misclassification

$$\begin{aligned}\text{Classification Accuracy (validation Data)} &= 1 - 0.26659 \\ &= 0.73341 \text{ or } 73.341\%\end{aligned}$$

This model can predict the result to the 73% correct. Second decision tree with original dataset is second best model having the efficiency of 73% only minor difference in the fourth decimal point. Neural network with that is developed with the original data set is the third best model, having the same classification accuracy. Poor performing model is the decision tree that was developed using the featured engineering dataset and had the depth of 10 and Entropy was used as a splitting rule with the efficiency of 71.1728%.

### 13 Possible Future Improvements

Possible future improvements for the model can be the vastest dataset having the more diverse values, as the target variable had almost 50% values for both positive CVD and negative CVD. Whereas the values of the feature that are shown graphically had the limited variation, alcohol intake for the people in the dataset had 95% values with the people who were not taking the alcohol and only 5% of the people in the dataset had alcohol intake. Feature engineering can be done in more efficient way for the improvement of the predictions with the help of field experts they can provide the better guidance in classification of high and low blood pressure values. Overall spread of the data with real time values can enhance the model working.

### 14 To deploy the model in real-world application

As the models efficiency is not very high still this model can be used for the awareness creation in public or can be used the prediction model code to develop an mobile application that can help people to track if they fall in the risk of getting CVD and with the significance of variables defined in the results evaluation of model application can highlight the most significant variable if the user is falling under getting the positive CVD and can send a notification to user to take precautionary step or consult an medical specialist.

## 15 References

Koch, P., Wujek, B., Golovidov, O. & Gardner, S., 2017. *Automated Hyperparameter Tuning for Effective Machine Learning*, USA: SAS Institute.

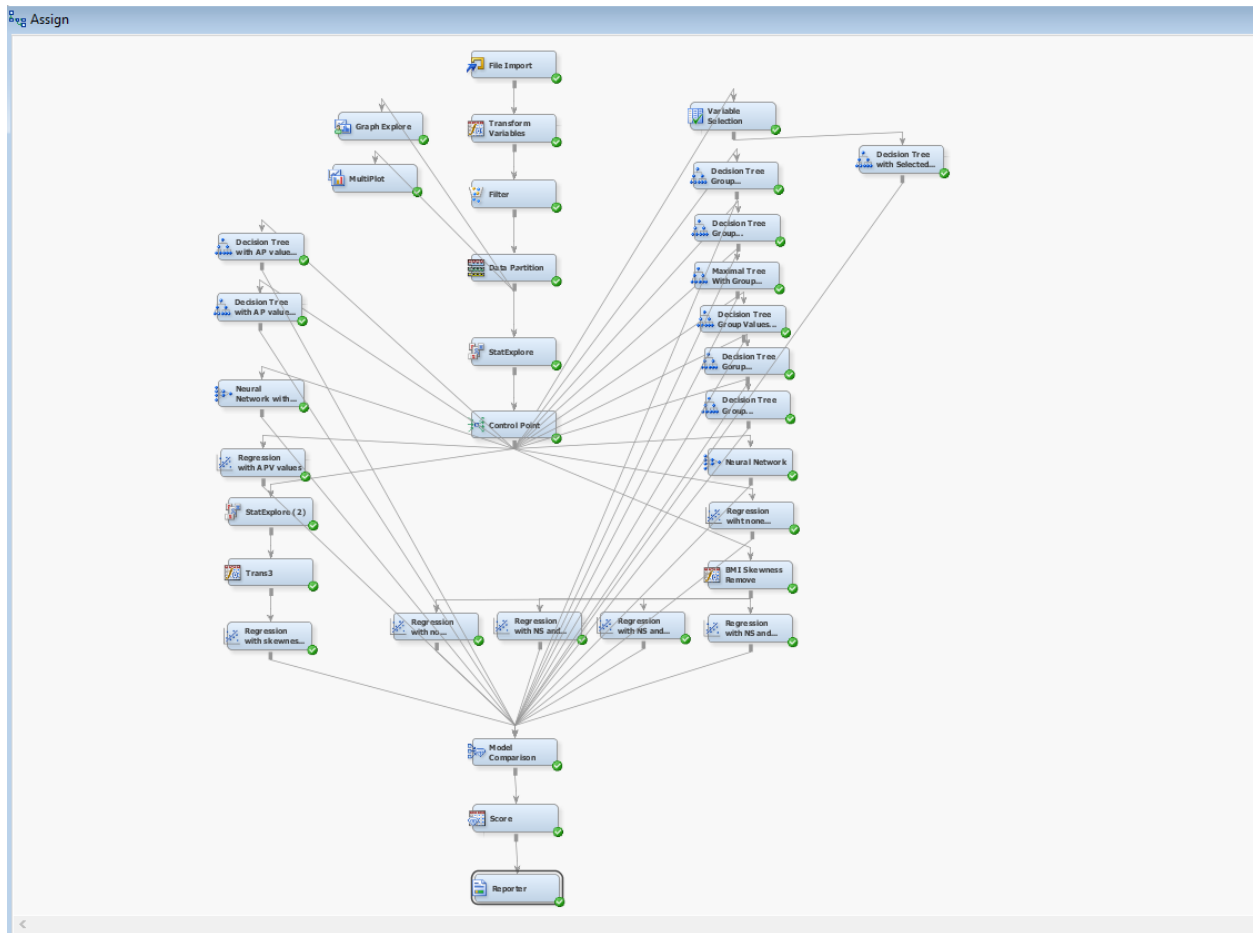
NHS, 2018. *Cardiovascular Disease*. [Online]

Available at: <https://www.nhs.uk/conditions/cardiovascular-disease/>

[Accessed 23 06 2021].

## Appendix I

Screenshot of the model is shown below:



XML Diagram and PDF report of the model node is attached in the file:

Report is generated using the score node of the SAS with properties set to summary and Reporter node is used to generate the report. Screenshots of the report are shown below

SAS Enterprise Miner Report

Model Summary  
Data

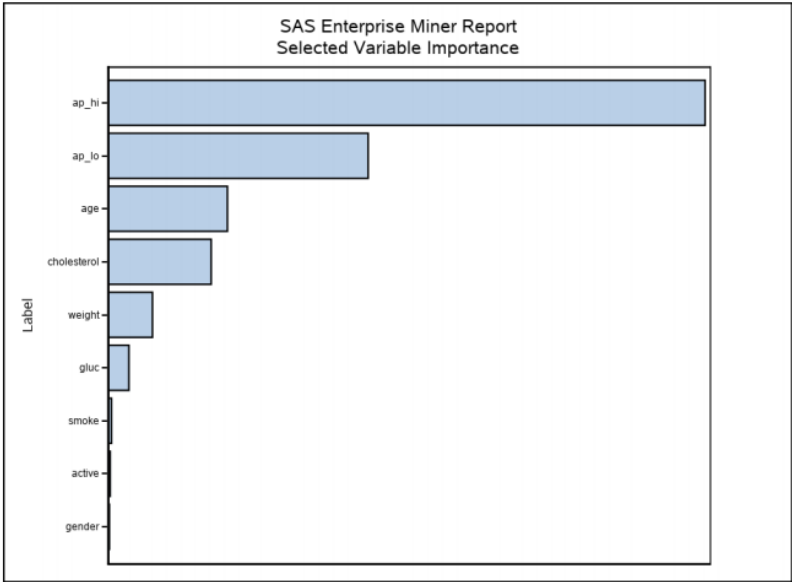
Property	Value
Input Data Source	EMW52.FIMPORT_DATA
Target Variable	cardio
Event Level	1
Observations	69952
Original Variables	14
Selected Variables	9

Target: cardio

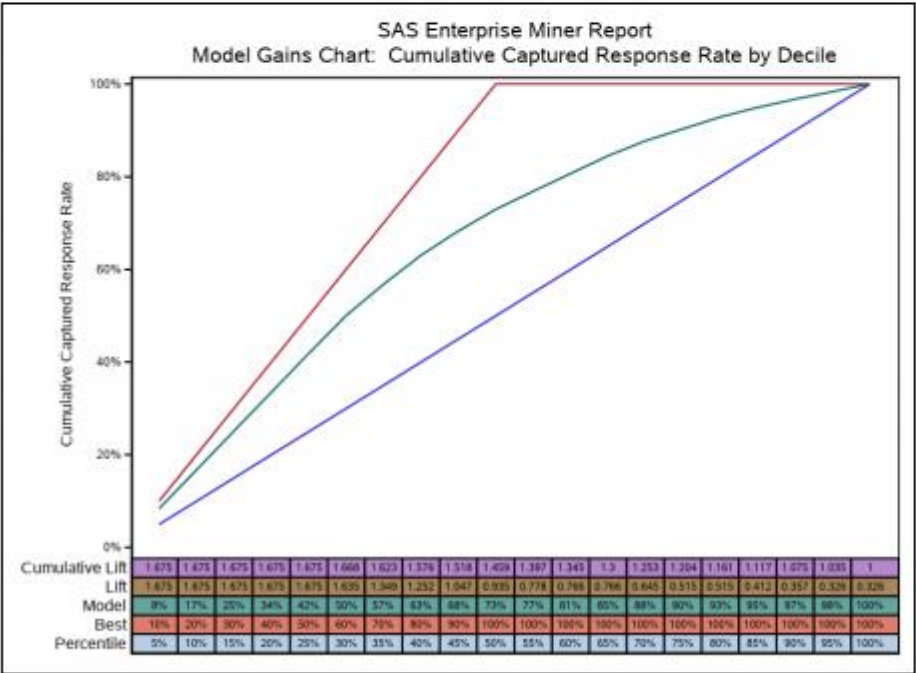
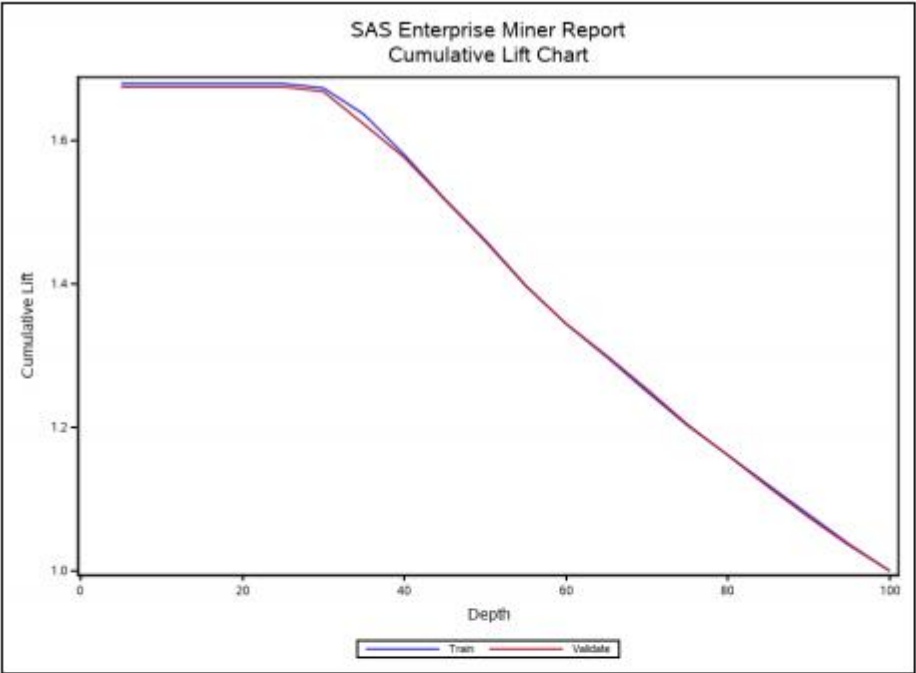
Value	Count	Data Percentage
1	26211	49.9771
0	26235	50.0229

Variable Summary

Role	Level	Original Count	Selected Input Count
ID	Internal	1	0
Input	Binary	4	3
Input	Internal	6	4
Input	Ordinal	3	2
Target	Binary	1	0

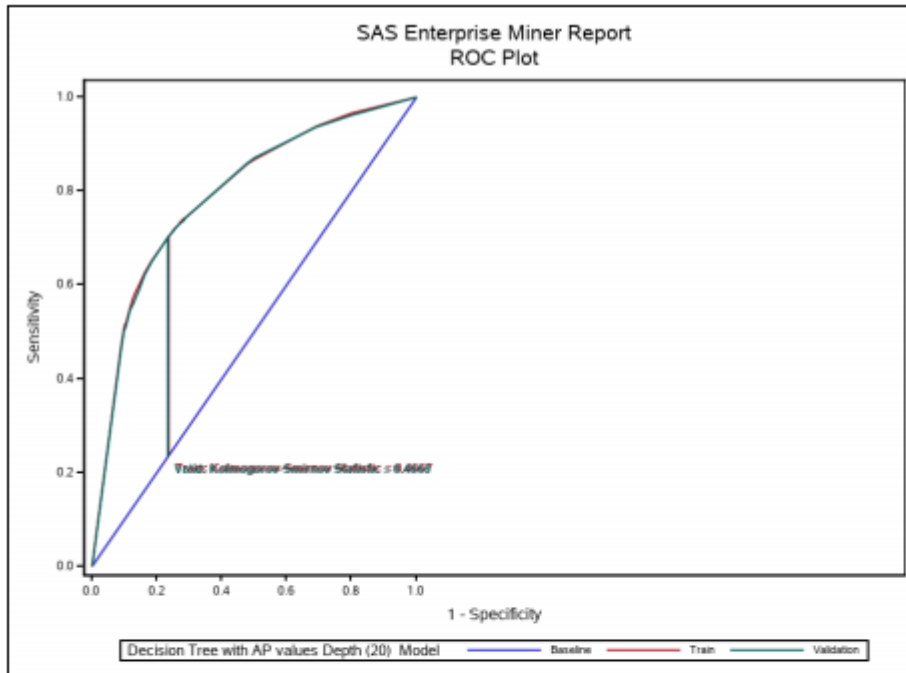






# Big Data for Decision Making

3



## Variable Attribute Importance and Crosstabulations

		Scorecard Points	Overall N	Overall %	cardio			
					0		1	
					N	%	N	%
active	0	32.00	-	-	4825.00	18.39	5484.00	20.92
	1	0.00	-	-	21410.00	81.61	20727.00	79.08
age	1: LOW - 16871	4.00	-	-	6708.00	25.57	3117.00	11.89
	2: 16871 - 18751.5	10.00	-	-	5070.00	19.33	3796.00	14.48
	3: 18751.5 - 19943	0.00	-	-	5167.00	19.70	4914.00	18.75
	4: 19943 - 22178	47.00	-	-	7057.00	26.90	9187.00	35.05
	5: 22178 - HIGH	158.00	-	-	2233.00	8.51	5197.00	19.83
ap_hi	1: LOW - 1.5	0.00	-	-	21083.00	80.36	9818.00	37.46
	2: 1.5 - 2.5	190.00	-	-	2807.00	10.70	4144.00	15.81
	3: 2.5 - HIGH	436.00	-	-	2345.00	8.94	12249.00	46.73
ap_lo	1: LOW - 1.5	2.00	-	-	7748.00	29.53	4018.00	15.33
	2: 1.5 - 2.5	0.00	-	-	15265.00	58.19	11368.00	43.37
	3: 2.5 - HIGH	7.00	-	-	3222.00	12.28	10825.00	41.30
cholesterol	1	2.00	-	-	21950.00	83.67	17302.00	66.01
	2	0.00	-	-	2865.00	10.92	4244.00	16.19
	3	169.00	-	-	1420.00	5.41	4665.00	17.80
gender	0	44.00	-	-	17248.00	65.74	16965.00	64.72
	1	0.00	-	-	8987.00	34.26	9246.00	35.28

(Continued)

# Big Data for Decision Making

4

		Scorecard Points	Overall N	Overall %	cardio			
					0		1	
					N	%	N	%
gluc	1	53.00	.	.	23146.00	88.23	21438.00	81.79
	2	55.00	.	.	1569.00	5.98	2252.00	8.50
	3	0.00	.	.	1520.00	5.79	2521.00	9.62
smoke	0	83.00	.	.	23814.00	90.77	24021.00	91.64
	1	0.00	.	.	2421.00	9.23	2190.00	8.36
weight	1: LOW - 70.5	0.00	.	.	14182.00	54.06	10286.00	39.24
	2: 70.5 - 80.5	3.00	.	.	6629.00	25.27	7078.00	27.00
	3: 80.5 - 92.5	7.00	.	.	3645.00	13.89	5214.00	19.89
	4: 92.5 - HIGH	17.00	.	.	1779.00	6.78	3633.00	13.86

## Classification Matrix Target=cardio

	Data Role			
	TRAIN		VALIDATE	
	Predicted		Predicted	
	0	1	0	1
Target				
0	76.44	23.56	76.68	23.32
1	29.77	70.23	30.00	70.00

## Model Fit Statistics

Statistic	Train	Validation
Sum of Frequencies	52446.0000	17484.0000
Misclassification Rate	0.2666	0.2666
Maximum Absolute Error	0.8523	0.8523
Sum of Square Errors	19158.4155	6426.7123
Average Squared Error	0.1826	0.1838
Root Average Square Error	0.4274	0.4287
Roc Index	0.7920	0.7900
Gini Coefficient	0.5840	0.5810
Kolmogorov-Smirnov Statistic	0.4670	0.4670
Kolmogorov-Smirnov Probability Cutoff	0.4570	.
Lift at 10%	1.6794	1.6750
Cumulative Lift at 10%	1.6794	1.6750
Captured Response at 10%	8.3961	8.3733
Cumulative % Captured Response at 10%	16.7953	16.7562

## Model Selection based on Valid: Misclassification Rate

Selected Model	Model Node	Model Description	Target Label	Train: Akaike's Information Criterion	Train: Lift	Valid: Lift
Y	Tree8	Decision Tree with AP values Depth (20)	cardio	.	1.67940	1.67505
	Tree9	Decision Tree with AP values Depth (10)	cardio	.	1.67940	1.67505
	Neural2	Neural Network with APV values	cardio	56984.35	1.71322	1.66667
	Reg6	Regression with APV values	cardio	57994.98	1.68269	1.69872

## Big Data for Decision Making

Selected Model	Model Node	Model Description	Target Label	Train: Akaike's Information Criterion	Train: Lift	Valid: Lift
	Reg7	Regression with skewness and AP values	cardio	58007.23	1.68651	1.70330
	Neural	Neural Network	cardio	58888.28	1.64072	1.61401
	Tree4	Decision Tree Group Variables Depth (10) Splitting Entropy	cardio	.	1.65244	1.62961
	Reg2	Regression with no Skewness	cardio	59742.85	1.64911	1.65751
	Reg3	Regression with NS and Backward Model	cardio	59742.85	1.64911	1.65751
	Reg4	Regression with NS and Forward selection	cardio	59742.85	1.64911	1.65751
	Reg5	Regression with NS and stepwise Model selection	cardio	59742.85	1.64911	1.65751
	Reg	Regression with none model selection	cardio	59758.52	1.64606	1.64835
	Tree2	Decision Tree Group Values Depth (10)	cardio	.	1.60980	1.60790
	Tree5	Decision Tree Group Variables Depth (6)	cardio	.	1.60980	1.60790
	Tree	Maximal Tree With Group Variables	cardio	.	1.67668	1.69359
	Tree3	Decision Tree Group Variables Depth (10) Assessment AVG With Cross Validation	cardio	.	1.67668	1.69359
	Tree7	Decision Tree with Selected Variables	cardio	.	1.60980	1.60790
	Tree6	Decision Tree Group Variables Depth (6) Entropy	cardio	.	1.60980	1.60790

### Project Information

Property	Value
Name	Big
Diagram	Assign
Path	/home/u58552588/Big
Date Created	26Jun2021:00:16:44

End of Report