Business Machine Learning I

Group Project Report

Group Members:

1. Sai Asish Inturi (Team Lead)
2. Sai Sudheer Neelam
3. Mani Krishna Bollam
4. Rajasekhar Padakandla

Heart Disease Prediction

Background Study

Heart illness is a catch-all phrase for a variety of ailments that have an impact on the heart. Congenital heart defects, arrhythmia, coronary artery disease, and other ailments are included. In the United States, heart disease is the top cause of death for both men and women. Smoking, diabetes, high cholesterol, high blood pressure, and a sedentary lifestyle are all risk factors for heart disease (O., 2020). Depending on the kind and severity of the condition, treatment options for heart disease may include lifestyle modifications, drugs, surgery, and lifestyle modifications.

In the medical industry, using machine learning to predict cardiac disease is growing in popularity. Based on a patient's age, gender, and medical history, machine learning algorithms can be used to forecast the likelihood that they will develop heart disease. Machine learning algorithms can precisely forecast the possibility of a patient acquiring heart disease by examining data from their medical records. Doctors can utilize this technology to identify patients who are at risk of developing cardiac disease so that preventive steps can be taken, as well as to better inform decisions regarding a patient's care and treatment.

Problem Statement

The goal of this project is to create a predictive model that can accurately classify whether a patient has a heart disease or not based on their medical history, symptoms, and other factors. The model should be able to identify the factors that are most important in the diagnosis of heart disease and accurately predict the risk of a patient having the condition. The model should be able to provide accurate results for both existing and new patients.

Objectives

1. Visual analysis will be performed using Tableau to achieve the aim
2. SAS E Miner will be used for implementation of machine learning to achieve the aim
3. Different models will be evaluated with error metrics in SAS and final model will be analyzed

Data Source

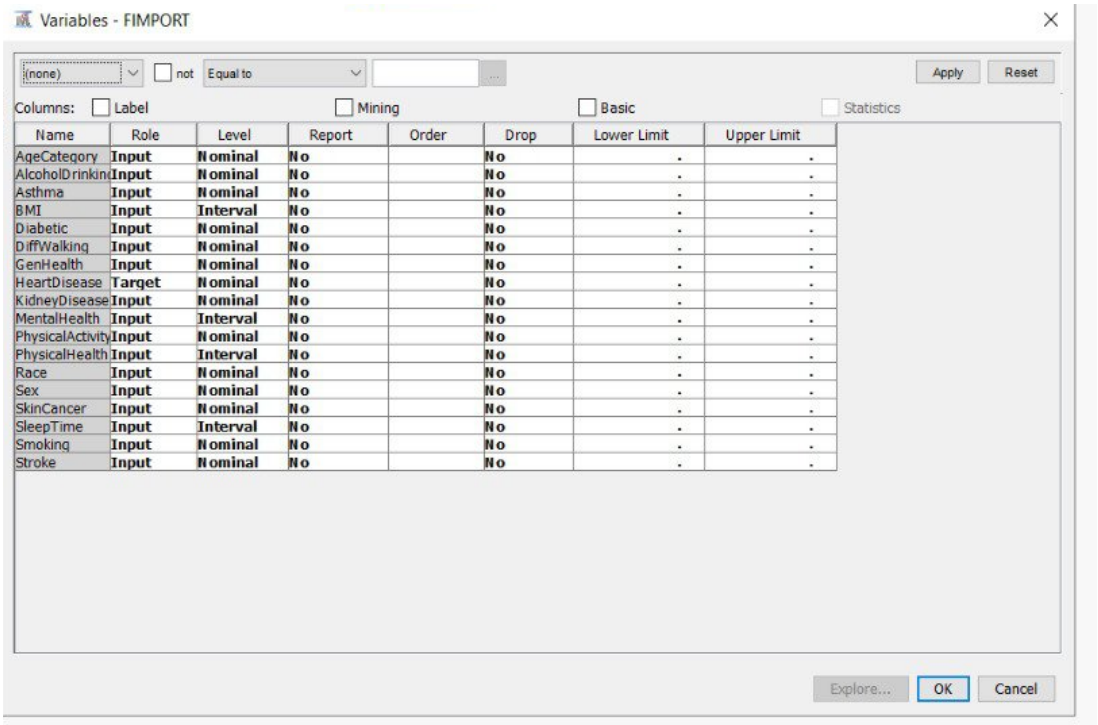
https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Originally, the dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

The data set is recorded from telephonic survey conducted by CDC. The data set contain 18 features where the target variable is the heart disease. Out of 18 variables, 9 variables are Boolean, 5 variables are categorical and four variables are numeric in nature. The class of the data set are heavily imbalanced where there are very less samples of heart disease patients.

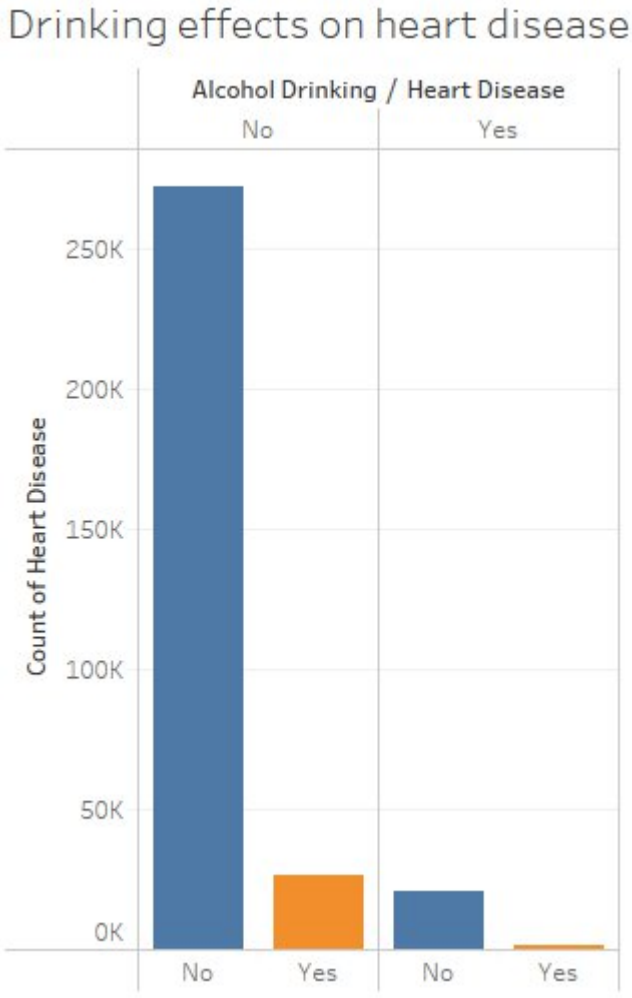
Input and Target Variables

The target variable in this study was the "Heart Disease". The remaining variables were considered to be independent variables in the study. The following table is a detailed view of it:

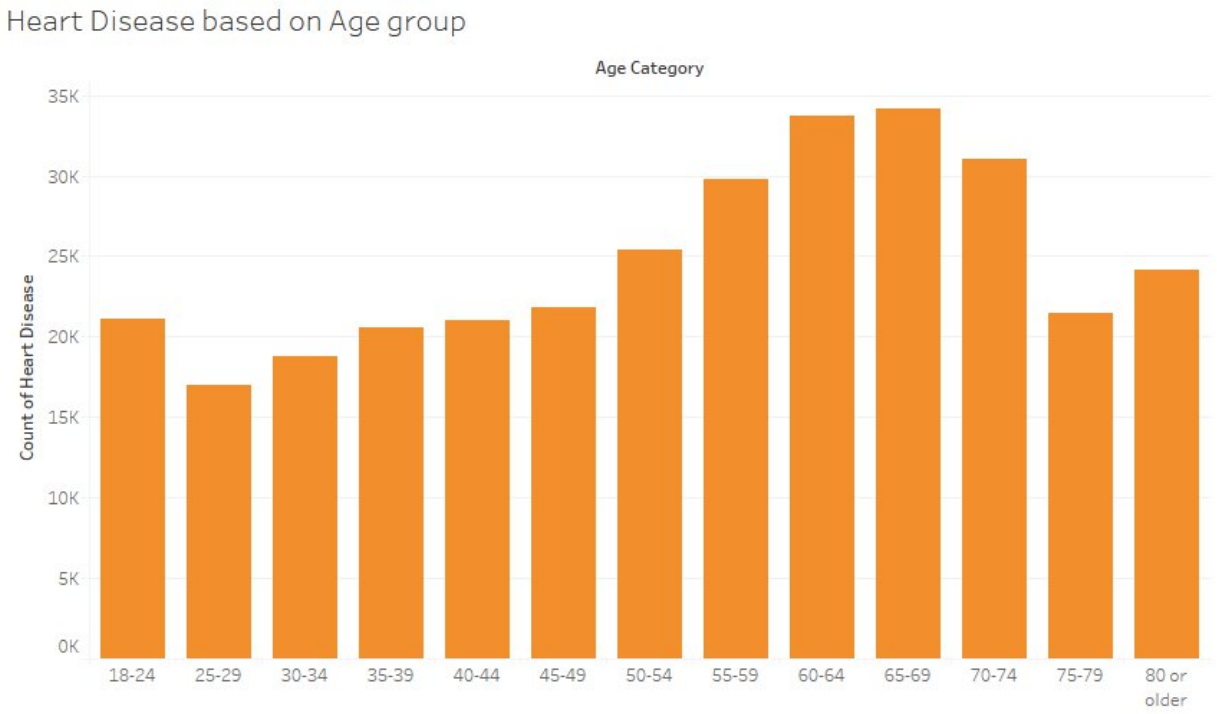


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AgeCategory	Input	Nominal	No		No	.	.
AlcoholDrinkin	Input	Nominal	No		No	.	.
Asthma	Input	Nominal	No		No	.	.
BMI	Input	Interval	No		No	.	.
Diabetic	Input	Nominal	No		No	.	.
DiffWalking	Input	Nominal	No		No	.	.
GenHealth	Input	Nominal	No		No	.	.
HeartDisease	Target	Nominal	No		No	.	.
KidneyDisease	Input	Nominal	No		No	.	.
MentalHealth	Input	Interval	No		No	.	.
PhysicalActivity	Input	Nominal	No		No	.	.
PhysicalHealth	Input	Interval	No		No	.	.
Race	Input	Nominal	No		No	.	.
Sex	Input	Nominal	No		No	.	.
SkinCancer	Input	Nominal	No		No	.	.
SleepTime	Input	Interval	No		No	.	.
Smoking	Input	Nominal	No		No	.	.
Stroke	Input	Nominal	No		No	.	.

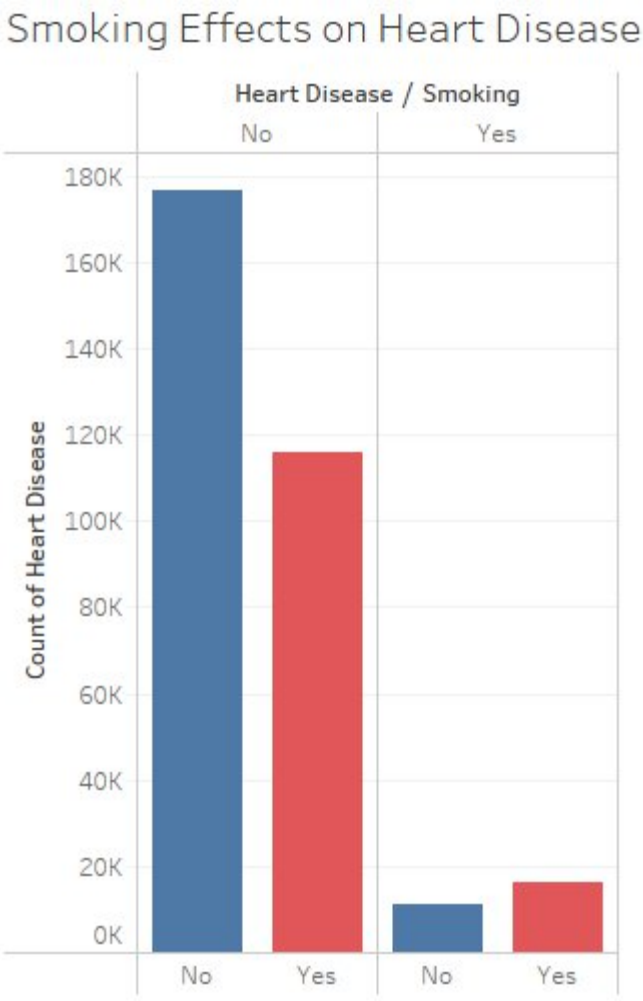
Findings using Tableau



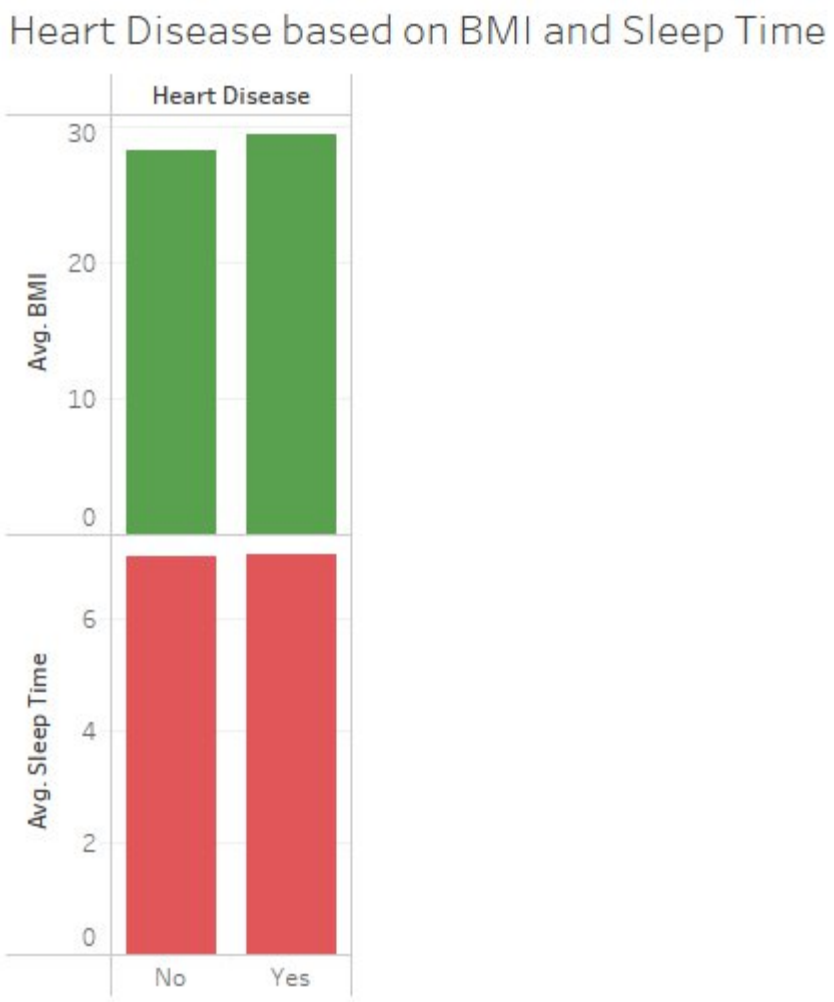
The graph shows there are a greater number of samples who do not drink alcohol and they do not have heart disease. Out of the samples who drink alcohol, it is found there are more samples who are normal than the samples who have heart disease. This indicates alcohol drinking do not have any relation with heart disease.



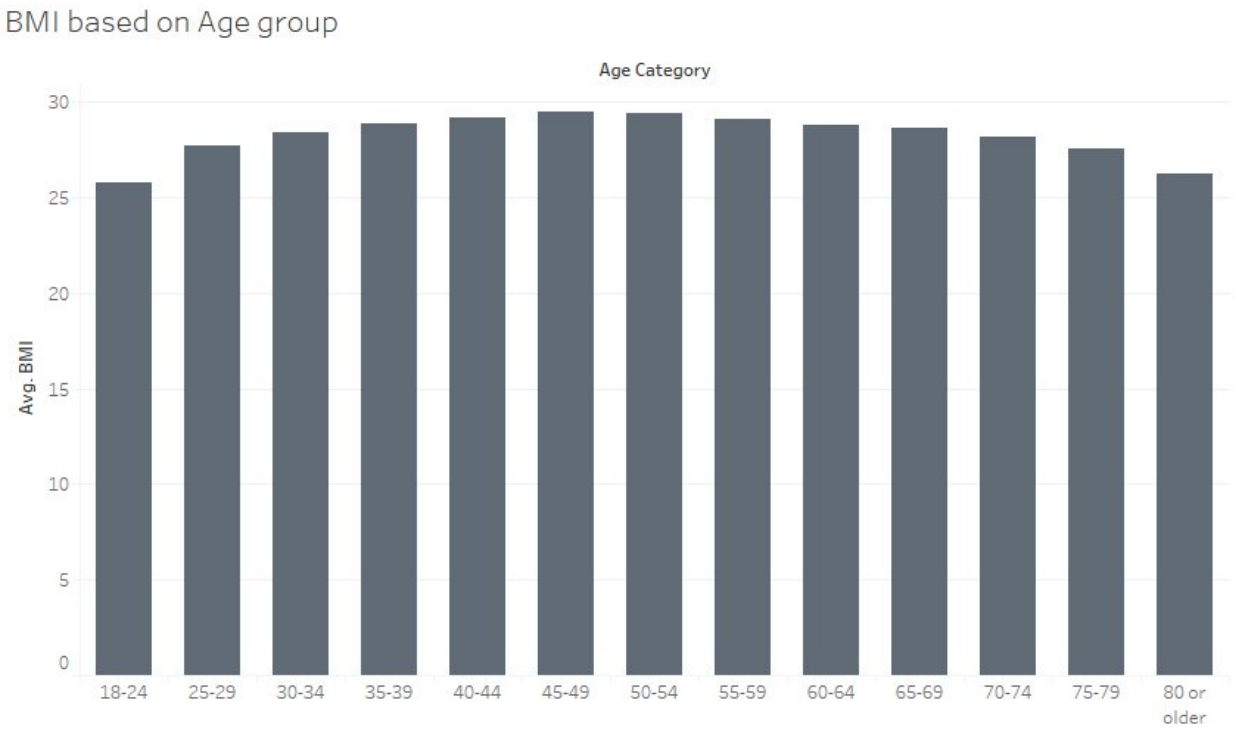
Heart disease can be related to age group so this visualization is used to look into the heart disease patients. Based on age group, it is seen that the patients who are ranging from 60 to 69 years of age are having more heart disease. This indicates older patients are riskier of having heart disease.



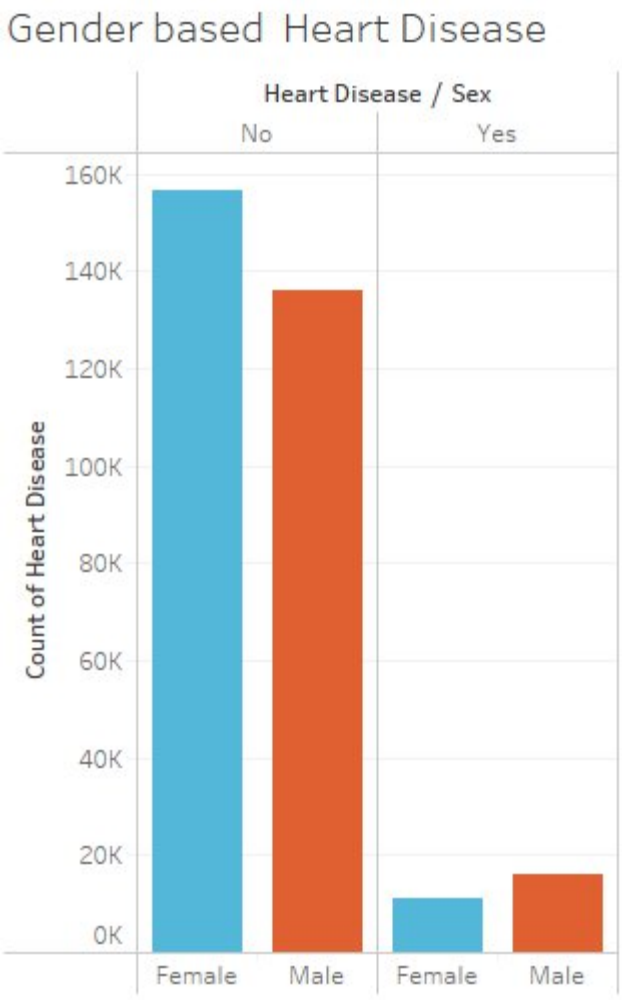
Patients who do not have heart disease and do not smoke are more than the patients who smoke. Also, in case of patients having heart disease that comparatively more patients who smoke than the non-smoking patients. This shows that smoking can be related to heart disease.



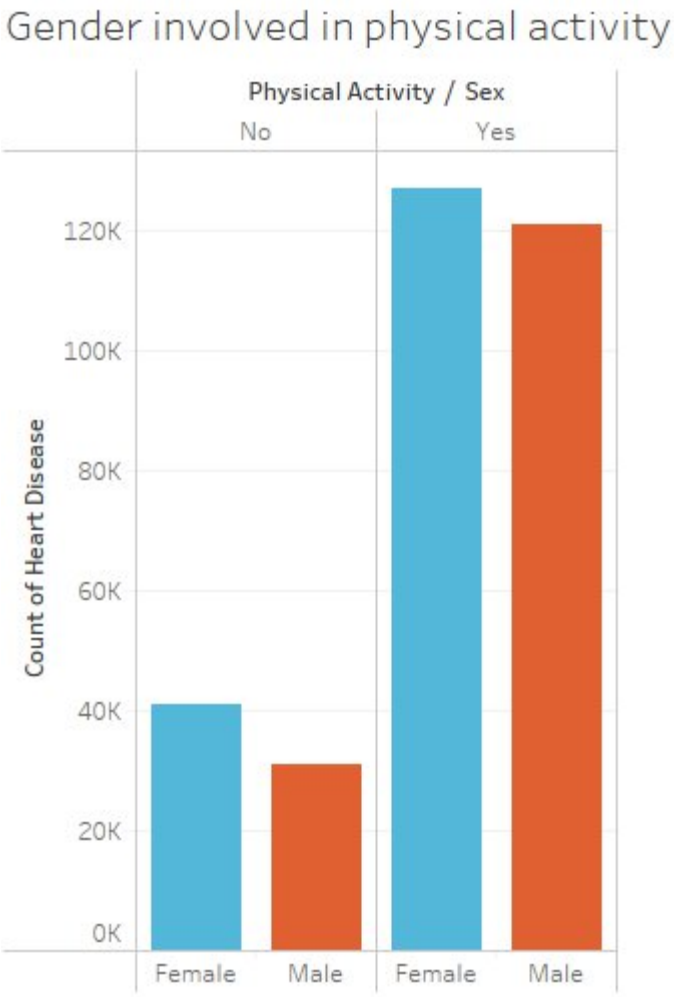
BMI and average sleep time can have relation with the heart disease. The plot shows the average BMI of patients having heart disease are higher than the normal patients. Also, the average sleep time is comparatively higher than the sleep time of normal patients.



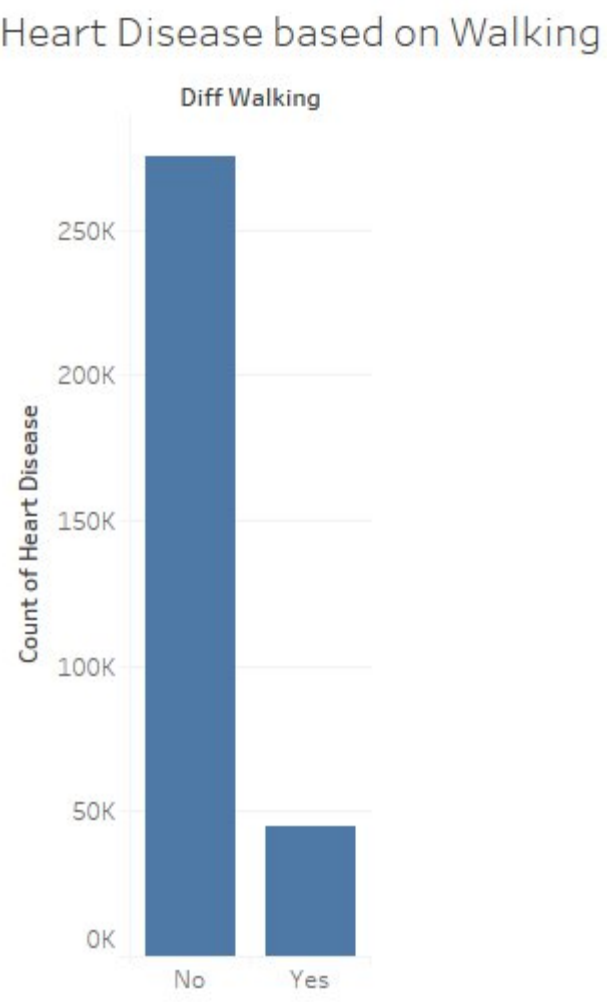
Since age is found to have relation with the heart disease, BMI is analyzed see that average BMI is present more in patients ranging from 45 to 65 years of age. This indicates that both BMI and age can have good correlation with heart disease and they can be used to decide whether a patient is going to have heart disease or not.



In normal patients, it is seen that there are more female patients then male who do not have heart disease. But in heart disease patient it is seen then male patients are comparatively higher than female patients. So, this data can only be helpful to predict heart disease of male patients than female.



Physical activities can also improve the health in heart disease. The patients who are not involved in physical activities are comparatively female more than male. Also, patients who are having physical activities are also female comparatively to male. This indicates that female patients are more involved in physical activities than male patients.



Difficulty in working can also contribute to heart disease where there are a greater number of patients who do not have difficulties in walking in this data. So, this data can be suitable to predict only heart disease for those patients who do not have any difficulty in walking.

Data Exploration and Preprocessing

For exploration of missing values, Excel is used. For scaling or other techniques, SAS will be used.

1. Missing Data

The data do not contain null values in any column.

2. Data Inconsistency

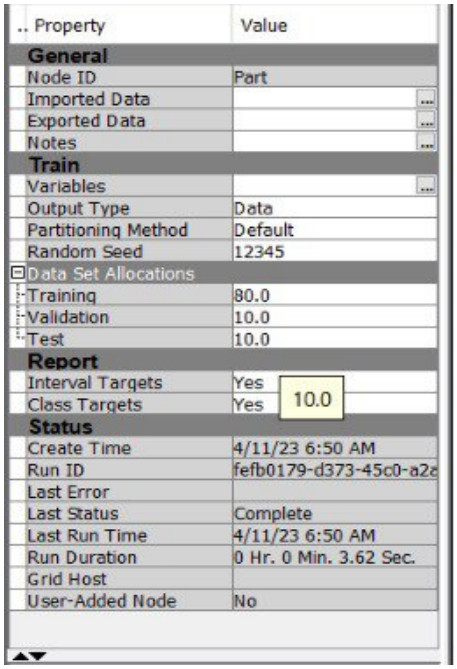
The data contain feature which have relation to the target. Also, there could be more columns like cholesterol, BP, etc. that can predict heart disease more efficiently. Also, the data is imbalanced which can give biased estimates.

3. Data Reduction

The data contain 18 features which are adequate. Also, the data is suitable to present US population. So, no reduction is necessary.

There were no missing values in the dataset therefore only preprocessing done on the dataset was splitting it into Training, Testing, and Validation datasets. The training dataset was 80% of the data, the validation dataset was 10%, and the testing dataset is 10% of the data.

The following table gives a glimpse of the data partition:

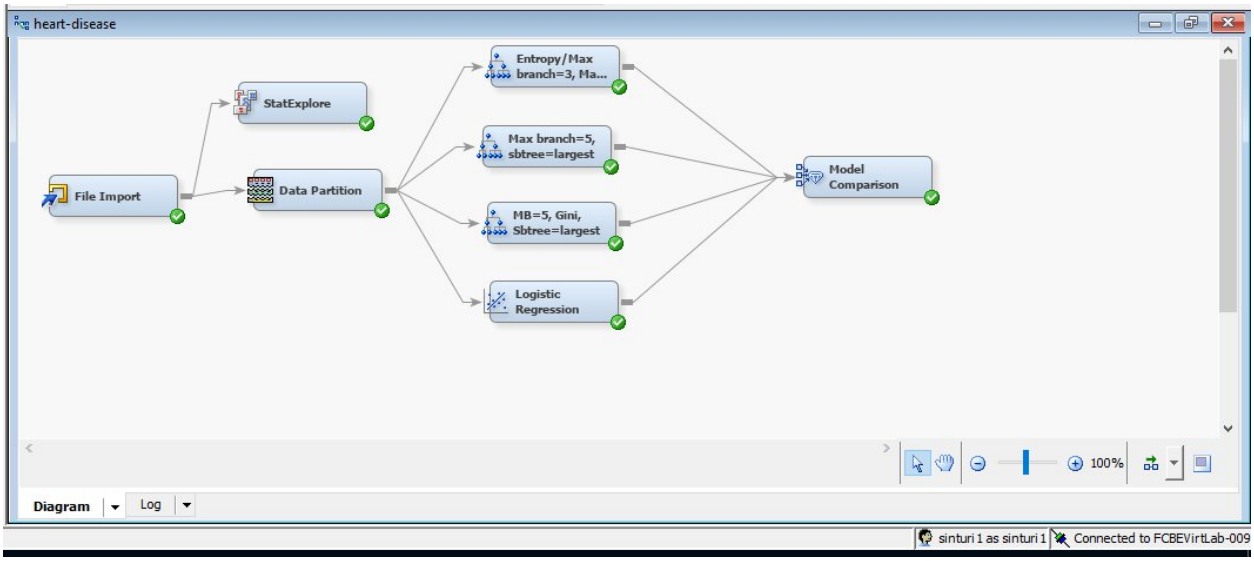


Predictive Modelling

Heart disease comes under classifications. So, for predictive modelling techniques, classification algorithms like Logistic regression or decision tree can be used for implementation of machine learning where all the data will be fed into SAS E Miner and trained the model using different classifiers.

Overview of Models built

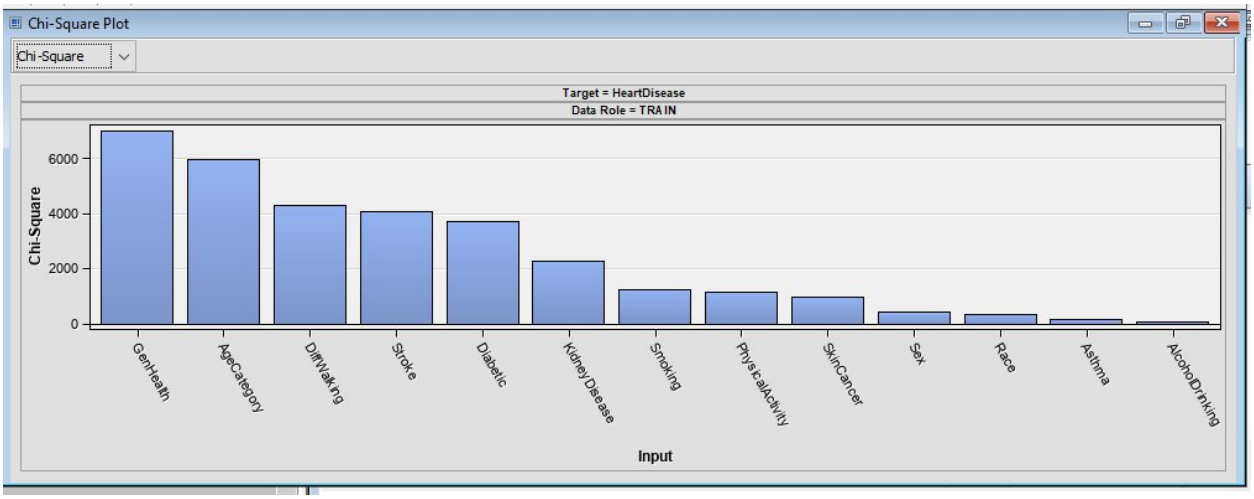
Using SAS, the following models are built as see in the plot below.



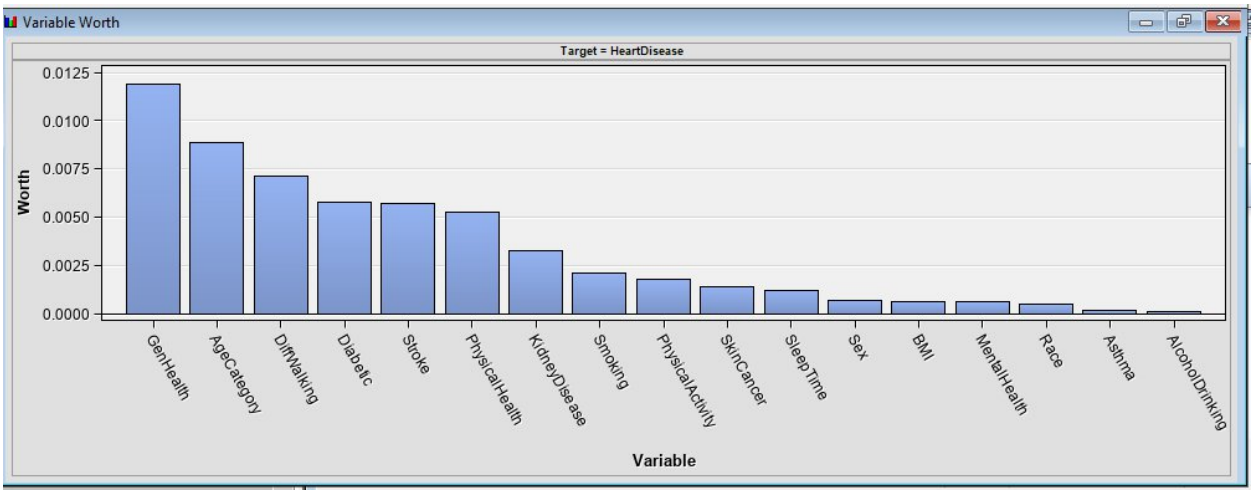
For predictive modelling purposes, SAS EM is used where exploration of the data is carried out and data is partitioned into training and validation. Also, different decision trees are implemented in training data and predicted in test data. The results of all the trees are the evaluated based on the error they gave and the best model is chosen based on the least error given by the model.

Variable Exploration

It is used to look into the importance of each input variables.



From Chi square Analysis, the variables like Genhealth, Age category, Diffwalking, stroke, diabetic are some of the variables that decide the rate of heart disease in patients. Chi square is suitable to find the significance is categorical variables and these are the categorical variables from which the significance to the output variable is found.

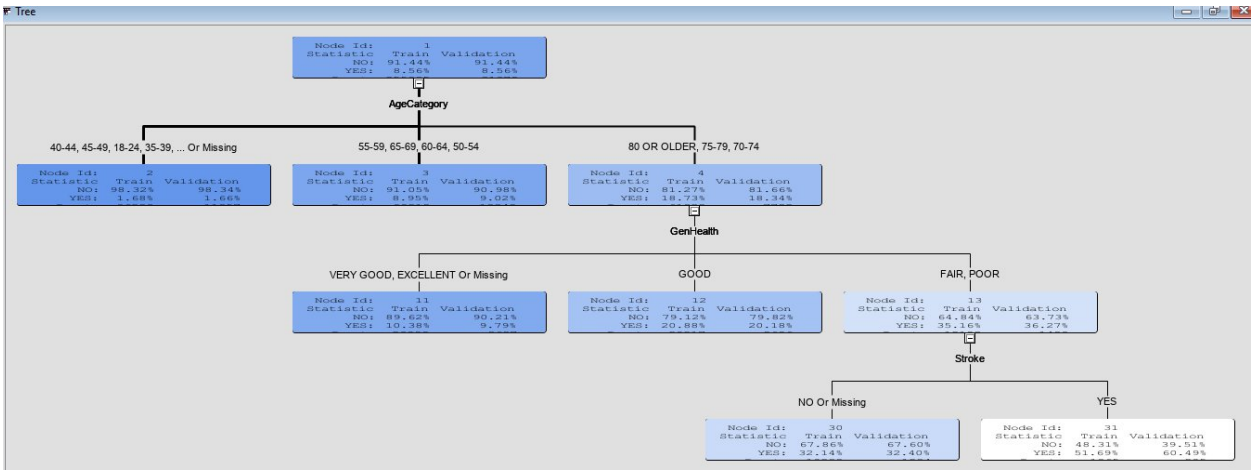


GenHealth, Age Category and DiffWalking are also the worthiest variables found from the data. In future prediction of heart disease, these variables can be used.

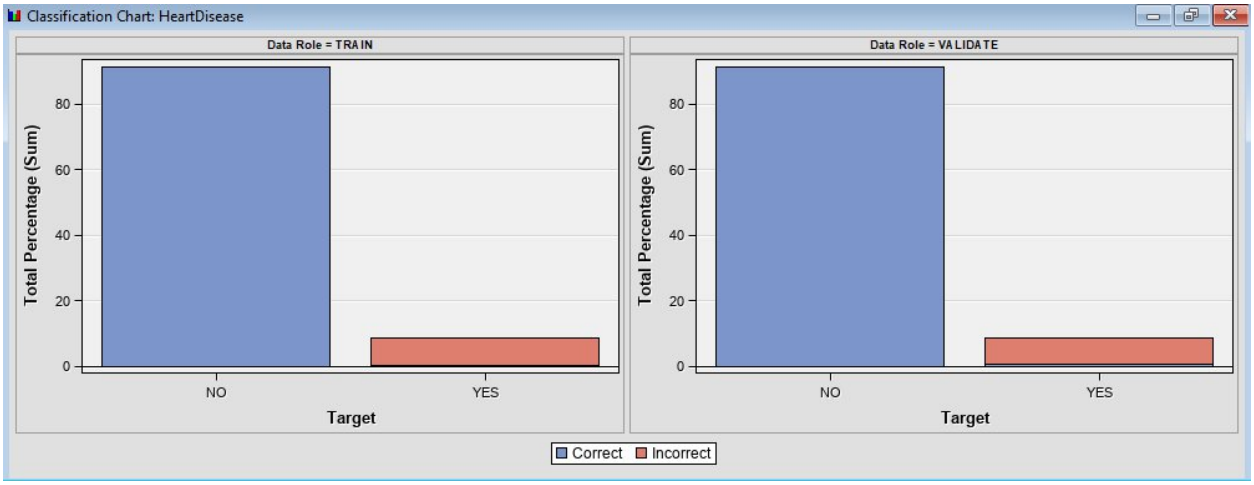
Model Training Results

1) Decision Tree 1

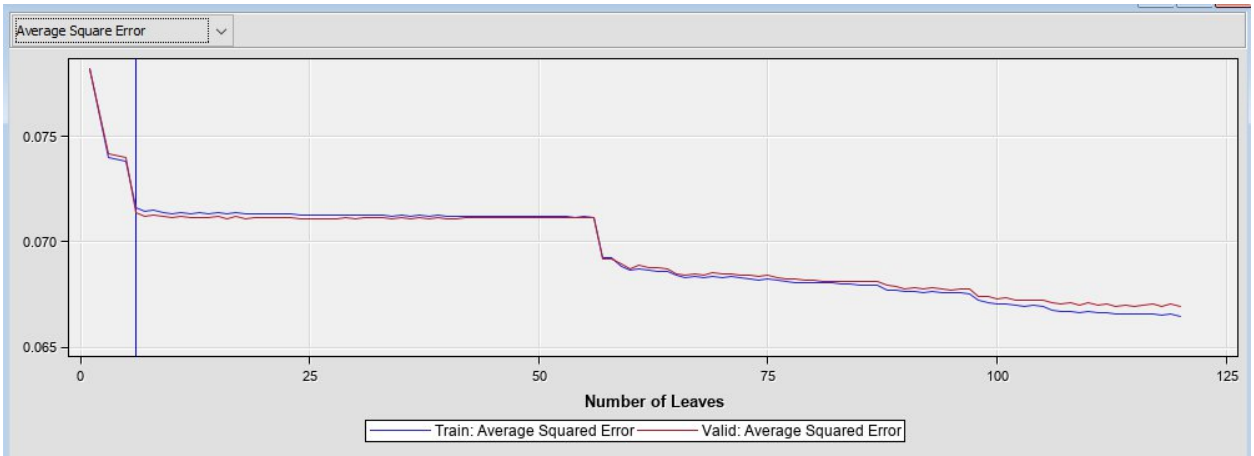
The first decision tree is created with maximum branch of 3 and maximum depth of 5 and nominal target criteria set to entropy and trained in the data.



From the tree, we can see that Age category is used for the first split. So, Age category is the most important variable as per the given tree.



The data looks imbalanced where there are only few patients who are having heart diseases in the data. The mis classification rate is highly observed in patients having heart disease. This is because the decision tree cannot be well trained on samples due to minimum presence of observations in train data.



The average square error is observed in different leaves where the tree gives the minimum error in 6 leaves but after that the error also decreased after 50 leaves. So, this tree with 6 leaves gives the optimum performance and the valid average squared error is lower than the train average squared error given by this tree.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
HeartDisea...		_NOBS_	Sum of Fre...	255835	31979	31981
HeartDisea...		_MISC_	Misclassific...	0.085344	0.084243	0.085176
HeartDisea...		_MAX_	Maximum A...	0.983176	0.983176	0.983176
HeartDisea...		_SSE_	Sum of Squ...	36666.19	4567.794	4579.631
HeartDisea...		_ASE_	Average Sq...	0.07166	0.071419	0.071599
HeartDisea...		_RASE_	Root Avera...	0.267694	0.267243	0.26758
HeartDisea...		_DIV_	Divisor for ...	511670	63958	63962
HeartDisea...		_DFT_	Total Degre...	255835	.	.

The test error is slightly lower than the train error that shows that the model is giving good performance with lower error in test data.

Event Classification Table

Data Role=TRAIN Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
20933	233037	901	964

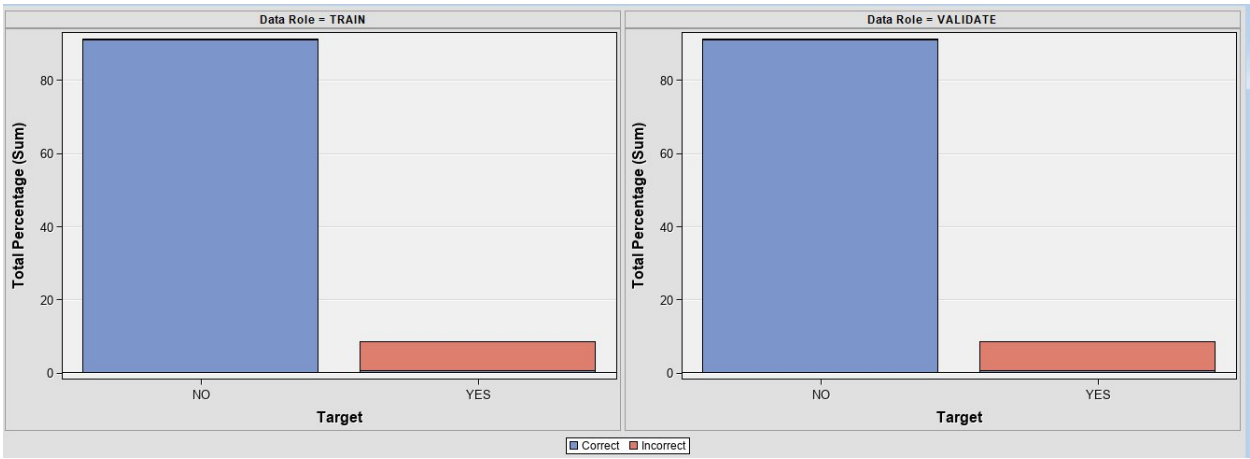
Data Role=VALIDATE Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
2613	29161	81	124

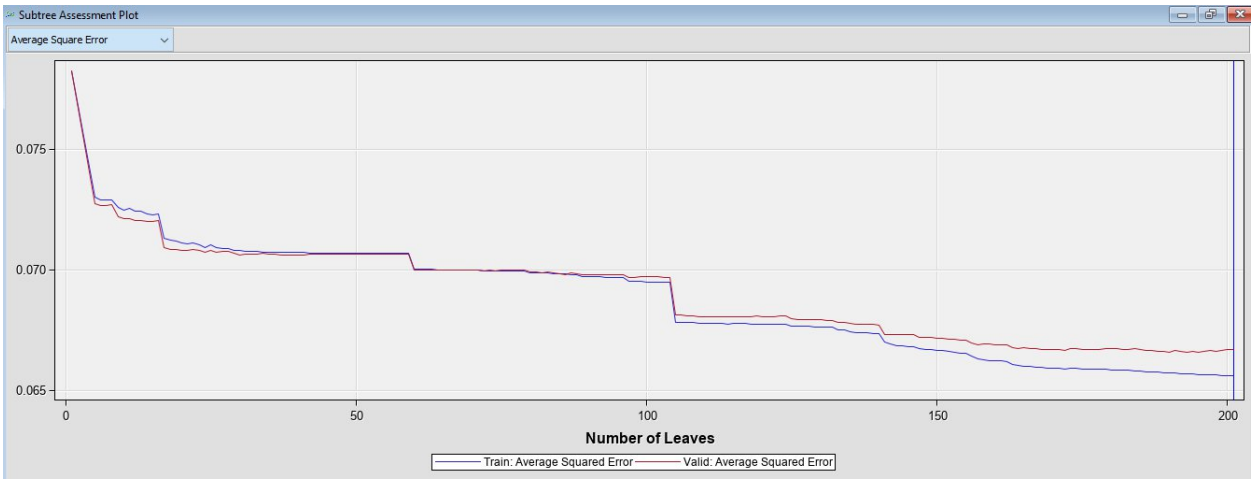
There are few heart disease samples and almost half of the samples are misclassified in train data and more than half of the samples are miss classified in validation data.

2) Decision Tree 2

The second decision tree is created with maximum branch of 5 and subtree is assessed based on the largest method.



The classification chart shows similar performance to the previous decision tree where the tree wrongly classified the samples with heart disease. The tree is suffering in learning the patterns in imbalanced samples.



In 201 leaves, the tree gives the optimal performance but it seems that the validation average squared error is increasing compared to the train error. This tree is giving a good indication of overfitting at high number of leaves.

Statistics Label	Train	Validation	Test
Sum of Frequencies	255835	31979	31981
Misclassification Rate	0.083964	0.084462	0.084581
Maximum Absolute Error	0.995846	1	0.995846
Sum of Squared Errors	33559.19	4266.308	4247.931
Average Squared Error	0.065588	0.066705	0.066413
Root Average Squared Error	0.256101	0.258273	0.257708
Divisor for ASE	511670	63958	63962
Total Degrees of Freedom	255835	.	.

There error in train and test data is slightly lower to the previous tree which shows that increasing the branches improve the performance in prediction of heart disease.

Event Classification Table

Data Role=TRAIN Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
20465	232922	1016	1432

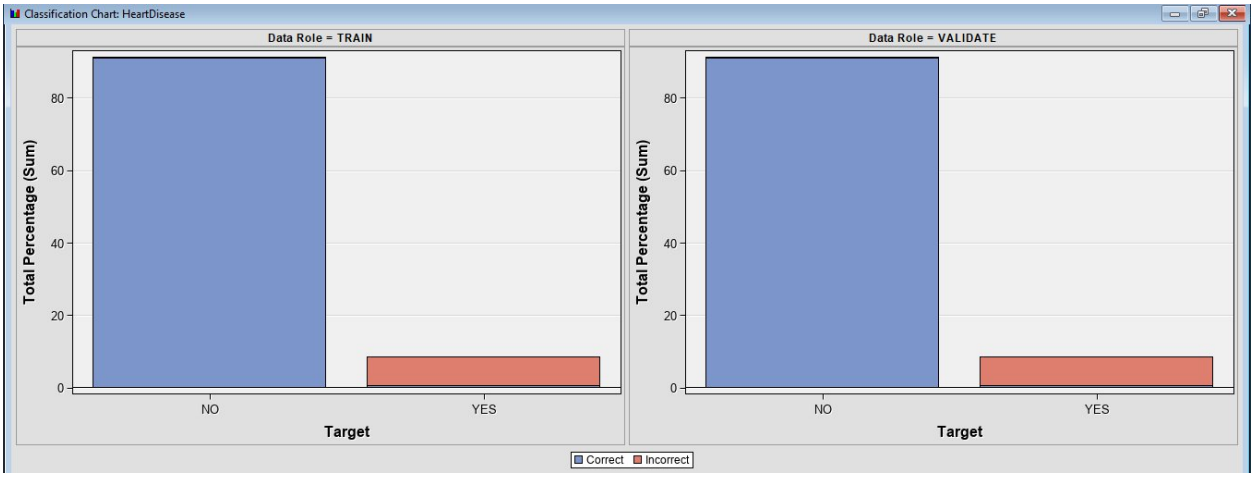
Data Role=VALIDATE Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
2564	29105	137	173

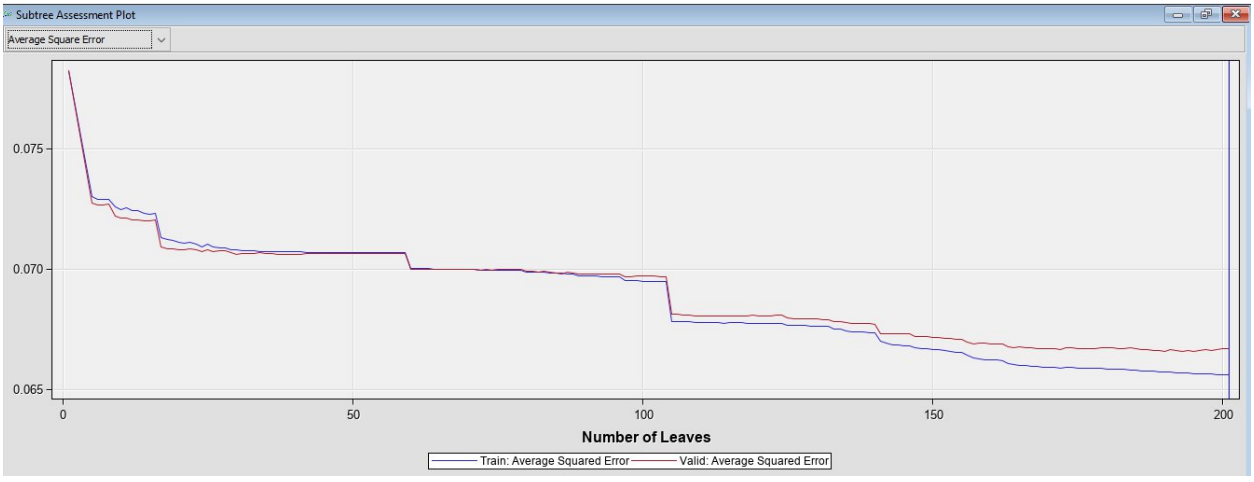
The classification table also shows a good true positive rate indicating a good improvement in prediction of patients having heart disease.

3) Decision Tree 3

The third tree is created with maximum branch of 5 and entropy measure is said to Gini and also the subtree is assessed using the largest method.



Using Gini as the nominal target criteria the model does not show any improvement as seen from the plot. However, the exact classification rate can be seen from the classification table to see the exact number of observations that are misclassified.



The same number of leaves is considered to be optimal where the average square error in valid data is likely higher than train data. After 100 leaves the model seems to show a good indication of overfitting where the validation error is increasing compared to the train error.

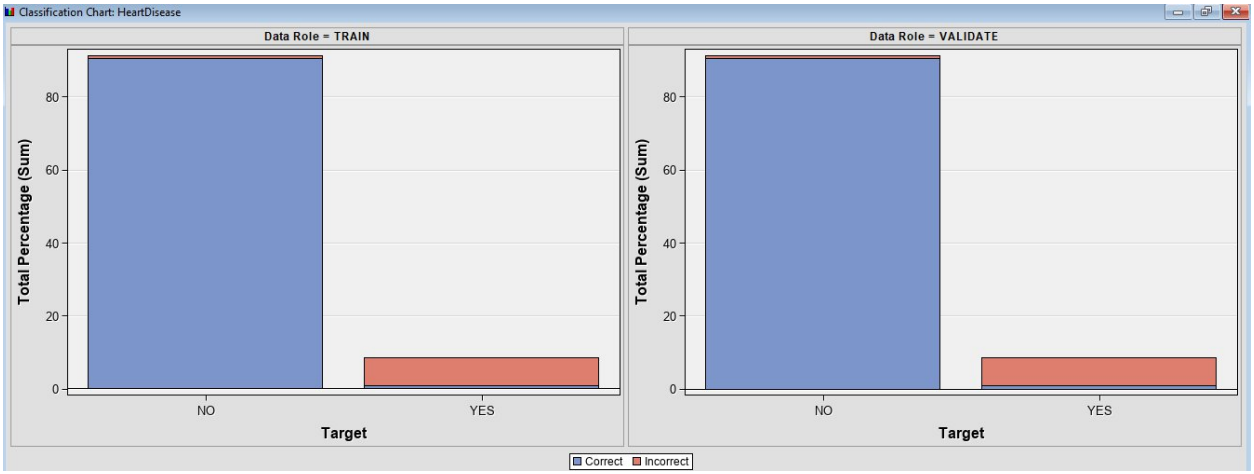
Statistics Label	Train	Validation	Test
Sum of Frequencies	255835	31979	31981
Misclassification Rate	0.083964	0.084462	0.084581
Maximum Absolute Error	0.995846	1	0.995846
Sum of Squared Errors	33559.19	4266.308	4247.931
Average Squared Error	0.065588	0.066705	0.066413
Root Average Squared Error	0.256101	0.258273	0.257708
Divisor for ASE	511670	63958	63962
Total Degrees of Freedom	255835	.	.

The fit statistics also show similar performance using Gini impurity measures where the test error is a bit higher than the train error.

```
Event Classification Table

Data Role=TRAIN Target=HeartDisease Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

  20465      232922       1016        1432


Data Role=VALIDATE Target=HeartDisease Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

  2564        29105        137         173
```

The classification table shows no improvement had been observed using Gini impurities measures compared to the previous decision tree with same parameters. This indicates using different nominal target criterion does not improve the performance.

4) Logistic Regression



Using logistic regression, the misclassification rate in patients having heart disease is slightly improved which can be seen from the plot. The red region in minimum observations had been reduced compared to the previous models.

Statistics Label	Train	Validation	Test
Akaike's Information Criterion	116096.4	.	.
Average Squared Error	0.065627	0.065815	0.06599
Average Error Function	0.226748	0.227331	0.22825
Degrees of Freedom for Error	255797	.	.
Model Degrees of Freedom	38	.	.
Total Degrees of Freedom	255835	.	.
Divisor for ASE	511670	63958	63962
Error Function	116020.4	14539.67	14599.35
Final Prediction Error	0.065647	.	.
Maximum Absolute Error	0.998457	0.998594	0.998424
Mean Square Error	0.065637	0.065815	0.06599
Sum of Frequencies	255835	31979	31981
Number of Estimate Weights	38	.	.
Root Average Sum of Squares	0.256178	0.256544	0.256885
Root Final Prediction Error	0.256216	.	.
Root Mean Squared Error	0.256197	0.256544	0.256885
Schwarz's Bayesian Criterion	116493.6	.	.
Sum of Squared Errors	33579.41	4209.379	4220.86
Sum of Case Weights Times Freq	511670	63958	63962
Misclassification Rate	0.084101	0.084243	0.083987

From the fit statistics, the mean squared error can be seen where the mean squared error in test data is slightly higher than the train data. After testing the model in real time observations, Logistic regression can give over fitting performance.

Event Classification Table

Data Role=TRAIN Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
19546	231968	1970	2351

Data Role=VALIDATE Target=HeartDisease Target Label=' '

False Negative	True Negative	False Positive	True Positive
2445	28993	249	292

The classification table shows must higher true positive samples which indicates that the highest number of patients having heart disease is correctly classified by the Logistic regression compared to the previous decision trees.

Model Evaluation Criteria and Final Model

The statistics comparison had been used to evaluate the final model where the misclassification rate in valid data is chosen. The lowest mis classification rate in valid data is given by the Logistic regression. Hence, this makes Logistic regression as our final model.

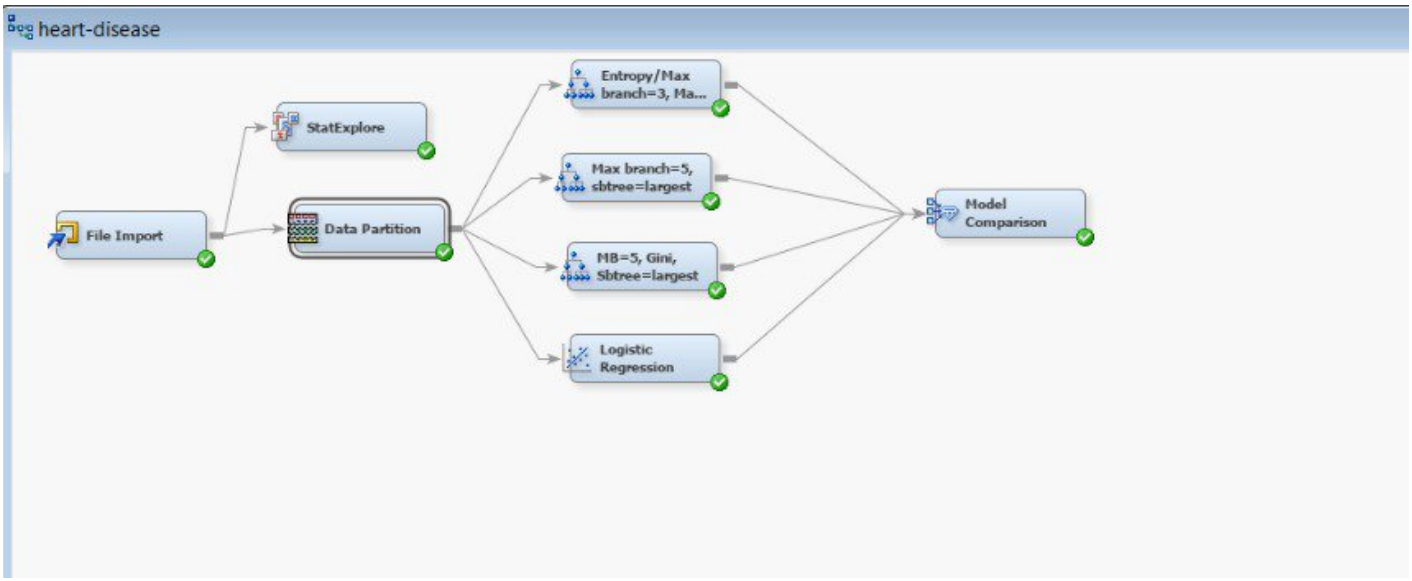
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Reg	Reg	Logistic Regression	HeartDisea...		0.084243
	Tree	Tree	Entropy/Max branch=3, Max depth=5	HeartDisea...		0.084243
	Tree2	Tree2	Max branch=5, sbtree=largest	HeartDisea...		0.084462
	Tree3	Tree3	MB=5, Gini, Sbtree=largest	HeartDisea...		0.084462

The misclassified samples can be seen from the classification table of all the models where we can see that Logistic regression gave the highest true positive samples that indicates correctly classified samples of patients having heart disease. We have focused on the prediction performance in those patients having heart disease as we wanted to inspect the performance of models in imbalanced data.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Entropy/Max branch=3, Max depth=5	TRAIN	HeartDisease		20933	233037	901	964
Tree	Entropy/Max branch=3, Max depth=5	VALIDATE	HeartDisease		2613	29161	81	124
Tree2	Max branch=5, sbtree=largest	TRAIN	HeartDisease		20465	232922	1016	1432
Tree2	Max branch=5, sbtree=largest	VALIDATE	HeartDisease		2564	29105	137	173
Tree3	MB=5, Gini, Sbtree=largest	TRAIN	HeartDisease		20465	232922	1016	1432
Tree3	MB=5, Gini, Sbtree=largest	VALIDATE	HeartDisease		2564	29105	137	173
Reg	Logistic Regression	TRAIN	HeartDisease		19546	231968	1970	2351
Reg	Logistic Regression	VALIDATE	HeartDisease		2445	28993	249	292

Final Model Diagram



Interpretation of the findings

From the visualization of variables and predictive performance, we noticed the following

- The age category of the patients had been divided into too many groups which can bring bias to the prediction
- The data contain very few samples of patients having disease that affected the overall performance of models
- The model mostly contain categorical variables from which ordinality issues can be observed

- Decision trees with maximum branches helps to bring improvement in prediction of heart disease

Recommendations

The following approaches can be recommended for future work

- The data should be added with more samples of patients having heart disease making the data balance the data
- Instead of Label encoding the categorical variable, one hot encoding can be used to prevent ordinary issues
- Neural network techniques and other classification technique such as ensemble techniques, bagging techniques or boosting techniques can be applied to determine the predictive performance
- Different other variables relevant to the heart disease such as cholesterol, blood pressure and other indicators can be studied for better analysis.

Summary and Lessons Learned

1. Age of a patient can decide occurrence of heart disease
2. BMI of a patient can also decide the presence of heart disease
3. Age and BMI also have good relation where BMI is found higher in older age patients
4. There are more female patients in the data who are involved in physical activities
5. Smoking habits have also good relation with the heart disease
6. Decision tree with increasing branches increases the chance of overfitting
7. It is difficult to bring good accuracy in imbalanced samples.

Future Work

1. There are more relevant features like cholesterol, blood pressure that can be added in future
2. The data of heart disease patients can be added in order to balance the sample
3. The features with more correlation with heart disease can be updated for better prediction
4. The parameters of the model can be tuned in case of improving the prediction

References

O., B. (2020). Risk Feature Aware Accurate Heart Disease Prediction System Using Fuzzy Extreme Learning Machine. *Journal of Advanced Research in Dynamical and Control Systems*, 12(SP3), pp.36–44. doi:https://doi.org/10.5373/jardcs/v12sp3/20201236.