**Unveiling Heart Attack using Data Mining Techniques from key Health Factors**

**Abstract and Highlights**

The main motto of this project is to predict the heart disease based on different key factors by considering the health factors of that person. To predict this, we have used different data Mining techniques such as Data Visualization, Exploratory Analysis and Predictive Analysis. We have discussed briefly on these techniques in the below.

1. Data Visualization helps to visualize the scenario in such a way that we can analyze different parameters in different ways by using different charts.

2. In exploratory Analysis, we have explored how often a person gets heart attack based on different factors.

3. In predictive Analysis, we have analyzed to predict the heart disease by considering different parameters.

As we know that there are different health parameters that influence the body disease, here we have considered few parameters to predict the heart disease. From the results that we have obtained from this analysis, we have noticed that the certain factors have high impact in occurring heart diseases. And once we able to figure out this, we can make a structure to prevent this.

**Problem Description:**

Heart diseases are a major contributor to global mortality rates. They not only affect an individual's well-being but also burden healthcare expenses. Over the past ten years, India has experienced significant economic losses due to cardiovascular diseases. Therefore, it is crucial to have reliable and precise predictions for heart-related ailments. 1. The heart plays a crucial role in the human body, pumping blood to all parts of the body. If it malfunctions, the brain and other organs will cease to function, and the person could die within moments.

As per the recent survey, the people who are dying with heart attack are increasing significantly. And WHO (World Health Organization) is still finding the cause for this issue. So, predicting heart attack by considering various health parameters helpful to save a life. Though we cannot give 100% assurance that these parameters have great impact but if it helps to unfold the research that will be a great achievement. (Heart disease prediction using data mining, 2021)

**Brief Description of Data Set and Data Source:**

This dataset was prepared by CDC (Centers for Disease Control and Prevention) which is part of Behavioral Risk Factors Surveillance Systems (BRFSS). The data was collected by conducting the telephonic surveys about the status of health of US Citizens. And this data set was from the year 2022.

With an aim to get a good data set with original data in it a lot of research has been done and the data set was obtained from Kaggle Website. It was a pre-cleaned data set, but there may be chances of having the duplicate values, null values which need to be taken care of. If these steps are ignored there may be chances of having errors in the output.

The Data set was having the 18 variables and large number of records. Now let us look at the variables present in the data set and their description.

|  |  |
| --- | --- |
| **PREDICTOR VARIABLE** | **DESCRIPTION** |
| HeartDisease | Had HeartAttack in past |
| BMI | Body Mass Index |
| Smoking | Smoking Status |
| Alcohol Drinking | Alcohol Drinking Status |
| Stroke | Had Stroke in past |
| PhysicalHealth | Number of days of Physical Illness in past 30 days like injuries etc. |
| MentalHealth | Number of days of Mental Illness in past 30 days like depression, Anxiety. |
| DifficultWalking | Has Difficulty in Walking |
| Sex | Gender of the Person |
| AgeCategory | Category of Age of Person which a person belongs to |
| Race | Ethnicity Category |
| Diabetic | Had Diabetics in past |
| Physical Activity | Does Physical Exercise |
| General Health | Condition of Overall Health |
| SleepTime | Average number of hours of sleep in 24 hours |
| Asthma | Had Asthma in past |
| KidneyDisease | Had KidneyDisease in past |
| SkinCancer | Had SkinCancer in past |

**Table 1: Predictor Variables and their description**

The dataset was obtained from the following website <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

**Data Cleaning, Preparation and Modification:**

Data Cleaning is one of the most important and basic steps in data mining. Apart from cleaning, preparation and modification should be done to predict the outcome with the highest accuracy.

The data set which we had taken has about 319,796 records. It was very large data set so; random sampling should be done on the data set to obtain a sample data set using excel. We have mainly used Excel to clean the data set and also for picking up the sample data set.

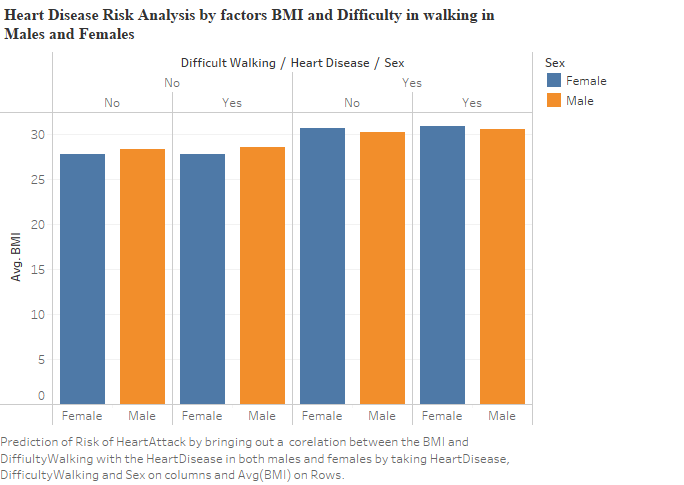
And other point to be noted is that the data set was having the variables which are class imbalance which means the variables which has binary values like YES or NO or may be 0 0r 1 were having more records with more importance towards only one value. Mainly, in our data set the Target Variable HeartDisease was highly imbalanced. The variable had nearly about 292,422 records which indicates NO and 27,373 records which indicates YES. So almost 90% percent of the records are NO for HeartDisease and only 10% are having YES for HeartDisease. So, if directly random sampling was done there may be chances of selecting all the records which has NO HeartDisease, which may not give accurate output for our research questions. So, balancing of the sample data set was done by choosing both the classes YES and NO in approximately equal proportions. This was done by using filter node and Random function in Excel. The original data set had also nearly about 18,700 duplicate values which may give in appropriate results, so they are removed using excel by remove duplicate’s function and the data set also contained some null values which were deleted using some functions in excel.

Data Preparation is one such step that would consume almost 80% of the time required for the project. The various steps involved in the Data Preparation are extracting the required and appropriate data, verifying the integrity of the data, and creating the new variables if necessary.

And another crucial step is to prepare the score data. The scoring data that we have considered in this project is the 20% sampling from the original Data Set. The main use of scoring the data is to get the appropriate results.

Finally, the balanced sample data set required for our project was obtained after performing all the above steps and the data set was containing 29,945 records. With this number of balanced records, it will be very easy to analyze the results and they will be more reliable.

**Data Visualization to explore the data sets (Graphs):**

****

**Figure 1: Graph 1(Analysing Heart Disease risk by BMI, DifficultWalking in Male & Female)**

In the first sheet we have used stacked bars to get a clear understanding between the variables we are going to analyse by the visualization. For this we have taken difficulty in walking, HeartDisease and sex in columns and Average BMI on Rows. We have put the sex in the colors to identify males and females distinctly. Females are represented in blue color and Male are displayed in orange colors as shown in figure 1.

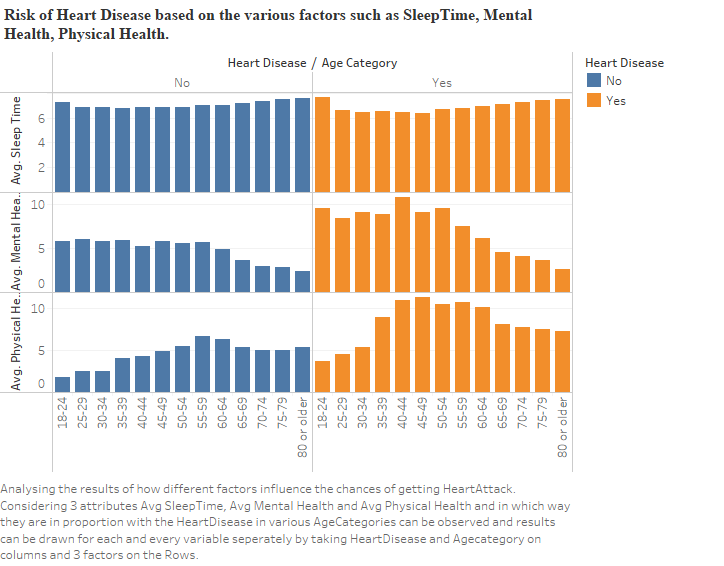
From the figure 1 it can be clearly observed that the Males who have no heartdisease and no difficulty walking are having Avg BMI of 28.330 and the ones with no heartdisease and has Difficult walking has an Avg BMI of 30.299. Which clearly portrays that the Males who are having the Difficult in walking has Higher BMI irrespective of the heartdisease.

For Females who has no heartdisease and no DifficultWalking has a Avg BMI of 27.820 and for ones with no heart disease and having DifficultWalking has an Avg BMI of 30.693 which means that irrespective of heartdisease the females who are having DifficultWalking has a Higher BMI.

For Males irrespective of whether they have DifficultWalking or not but one with HeartDisease has a higher BMI. For example, from the graph the Males who have DifficultWalking and no HeartDisease has BMI of 30.299 and for Males who has DifficultWalking and also HeartDisease has BMI of 30.639 which indicates that ones with the HeartDisease have higher BMI.

In case of females also the same observation can be noticed similar to males, the female’s with DifficultWalking and no heart disease has a BMI of 30.693 and one with DifficultWalking and HeartDisease has BMI of 30.947. These results suggest that female’s having the HeartDisease has higher BMI.

From all the observations taken from both females and males the findings clearly suggest that there is a relation existing between DifficultWalking, HeartDisease and BMI. So, we can clearly understand that BMI and DifficultWalking are the important factors in determining the risk of heart attack. People who are having difficulty in walking and higher BMI may be obese which means they are having high risk of having a heart attack.  **(World Health Orginazation Cardiovascular diseases, 2024)**



**Figure 2: Predicting HeartDisease Risk by factors SleepTime, Physical health, Mental Health in different Age groups.**

As shown in figure 2 for sheet 2 we have used horizontal bars for better understanding the relation existing between the various factors such as Avg SleepTime, Avg PhysicalHealth, Avg MentalHealth with HeartDisease in different Age Groups. Avg SleepTime, Avg PhysicalHealth, Avg MentalHealth are taken on Rows. In columns we have considered the HeartDisease and Age Category. And also, HeartDisease is taken in colors. “NO” HeartDisease is indicated in blue color and “YES” HeartDisease is indicated in orange color.

Let us first look at the bars comparing Avg SleepTime and HeartDisease in different Age Categories. By observing the results, we can say that except in the case of Age Category 18-24, people who are having the HeartDisease has less Avg SleepTime than the people who are not having the HeartDisease. For instance, let us consider an Age category (40-44), in this category people who are having the HeartDisease are having Avg SleepTime of 6.421 where as people who are not having the HeartDisease are having Avg SleepTime of 6.881. So, we can say that risk of heart attack is high in people who are having less SleepTIme.

Now let us consider another comparison between Avg MentalHealth and HeartDisease in different Age Categories. Avg MentalHealth is nothing but number of times a person was mentally ill in past 30 days like Depression, Anxiety, etc. By analysing the chart, it clearly portrays that people having the HeartDisease are those who are mentally ill for more days than the people who are not having the heart disease. Let us take an example of people who are in Age Category of (18-24). In this category people who are not having the HeartDisease are having the Avg MentalHealth 5.779, those who are having the HeartDisease are having it as 9.522. There is a large difference in the results, so people who are mentally ill for more days are having high risk of getting heart attack.

Now let us explore the comparison between Avg PhysicalHealth and HeartDisease in different Age Categories. Here Avg PhysicalHealth is nothing but number of days a person was ill physically like injuries in past 30 days. By careful observation we can easily say that people who are having the HeartDisease are physically ill for a greater number of days than the people who are not having the HeartDisease in all Age Categories. For example, consider the Age Category of (80 or Older), in this category people who are having the HeartDisease are having the Avg PhysicalHealth as 7.310 and those who are not having the HeartDisease are having the Avg Physical Health as 5.399. So, the people who are physically ill for a greater number of days are more prone to have a HeartAttack.

By comparing HeartDisease with three main factors we got to know that HeartDisease is inversely proportional to Avg SleepTime and directly proportional to Avg Physical Health and Avg Mental Health. So, a person who is mentally or physically ill and not having a proper sleep time are having high chances of getting attacked by HeartAttack.

**Exploratory Analysis**

Exploratory analysis, which includes understanding basic statistics like mean, median, mode, standard deviation, minimum, maximum, etc. for numerical variables and frequency distributions for categorical variables, is a vital part of unsupervised data mining and analysis. It helps us quickly gain insights by helping us understand the main characteristics of the dataset and explore patterns, trends, outliers, and relationships between variables.

By the process of doing exploratory analysis, it is possible to identify patterns and trends in the data, as well as correlations between two variables, handle missing data, and identify outliers in the dataset. Additionally, we can investigate correlations between variables and provide insights into how such relationships represent themselves in correlation learning.

When working on an exploratory analysis, the primary focus is on clustering techniques using the unsupervised learning method. The primary goal of cluster analysis is to create homogenous subgroups of the dataset, which comprise observations and records that are more like each other than the entire dataset, which compares all different observations at one place.

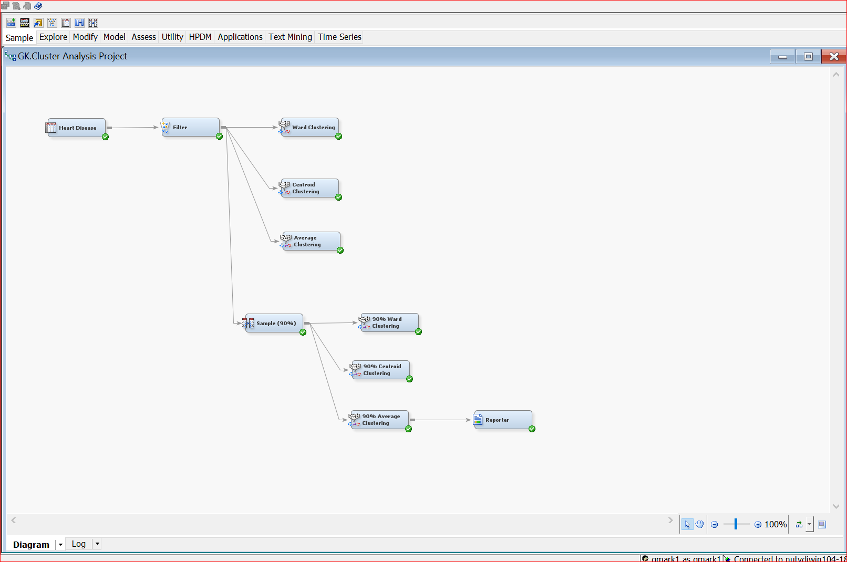


Figure -3

The clustering analysis process can be seen in the following image, where two different clustering approaches are used: hierarchical clustering and non-hierarchical clustering (K mean).

**Hierarchical clustering:**

By grouping objects together according to their similarities, we may continue to arrange them into a family tree-like structure through the process of hierarchical clustering.

**Hierarchical clustering (K mean):**

Sorting individual data points into individual categories without further grouping them into subgroups shows what non-hierarchical clustering is like.

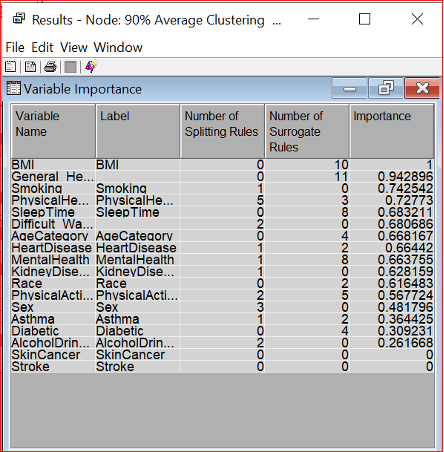
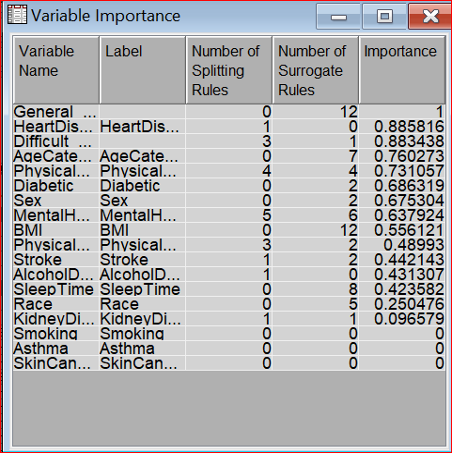
On the other hand, with 90% centroid and 90% average with same segment size of 4 segments with highest being segment is 3 with over 1162 frequency and 90% ward clustering as 5 segments and highest size segment is 2 with over 1177 frequency, there is an unsupervised Hierarchical clustering with centroid and average with same segment size of 4 segments with highest being segment is 1 with over 11605 frequency and ward clustering as 7 segments and highest size segment is 5 with over 8192 frequency.

The research aims to make a prediction model that can tell how likely a person is to have heart disease based on their health. The research will use data mining to find out which factors are related to having or not having heart disease.

We made sure the data was clean and easy to work with when getting the Indicators of Heart Disease dataset ready for analysis. This means we had to clean the data carefully to make sure it was accurate and didn't have any errors or missing information. Furthermore, stratification methods were used to solve the issue of unequal numbers of different classes in many datasets, especially those related to medical conditions like heart disease. To make sure the data is fair, we divided it into smaller groups based on certain characteristics. Each group has a similar number of each type of data. This made it possible to look at the dataset's information more carefully and accurately.

The study involved the use of clustering methods such as centroid, average, and ward without the guidance of a supervisor to organize the data. The best method was found to be Ward Clustering. It found four groups in a 90% of the data and three groups in all of the data.

The BMI rate and how often people went to the doctor were least in Cluster 2. The BMI rate was highest in Cluster 1. Most people in Cluster 2 did not have diabetes. More people in Clusters 1 and 3 were smokers. People who exercised regularly were more common in all groups. The research also focused on the importance of various factors and determined that BMI, smoking, and exercise were the most crucial elements in forecasting the risk of heart disease.



Centroid and average (Figure-4) 90% centroid and 90% average (Figure-5)

The various factors considered General Health and BMI emerged as the most important variable in Centroid clustering and average clustering, according to our analysis, which we used after analyzing the clustering and getting around to the view tab. We then looked at the cluster profiles, focusing on average clustering and variable importance.

This result emphasizes how important it is for BMI and general health to shape the clustering dynamics in the dataset. We identified that the most significant variable is BMI by giving the 90% average clustering and 90% centroid clustering priority.

We additionally found that the patterns and structures present in the data provide crucial insights. We also go deep into the characteristics that contribute to the formation of clusters, and from this we can conclude that BMI is important in defining heart disease.

**Average Clustering**

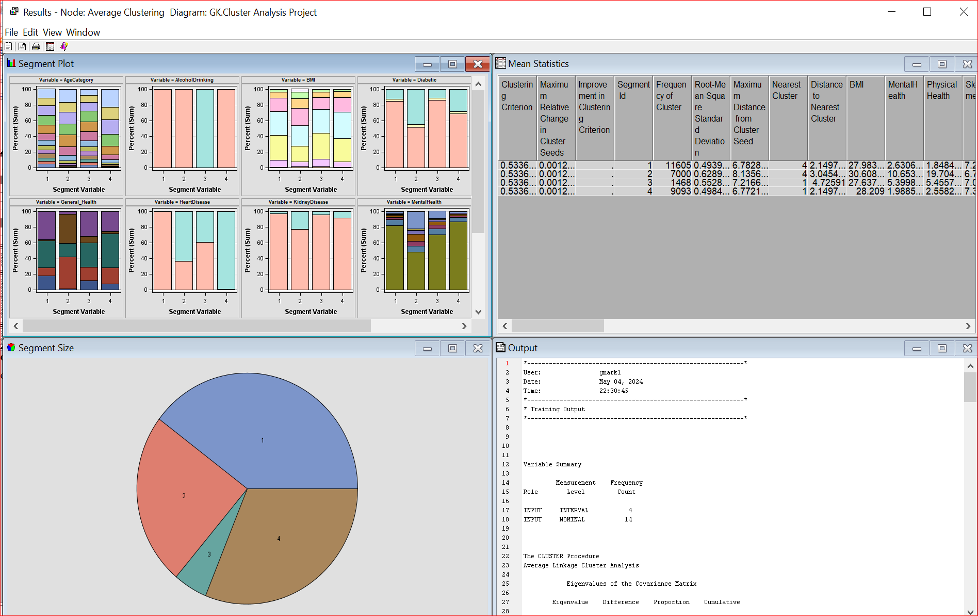


Figure -6

**90% Average Clustering**

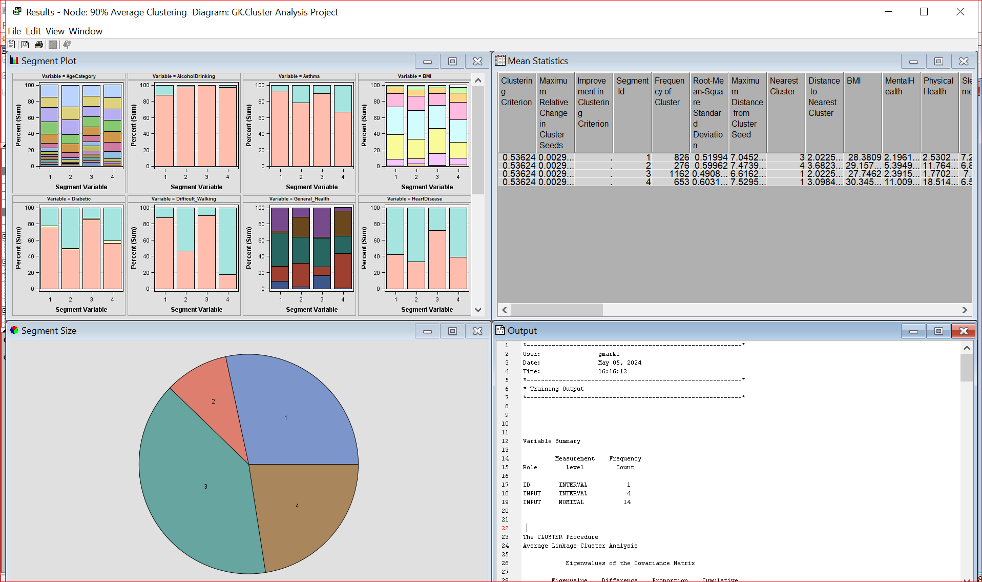


Figure -7

Ultimately, the primary dataset and the 90% subset dataset display comparable cluster values, with a minor variance in one cluster. This shows that the number of clusters is stable and nearly the same in both datasets. This makes us even more sure that the clusters made by Ward Clustering are stable. This means that the information from the groups of data can help us see how likely it is for someone to have heart disease. One important factor that stands out in both sets of data is the person's BMI.

**Predictive Analysis**

**Classification/Decision Tree:**

Classification/ Decision tree is a type of supervised learning where the algorithm looks like a tree structure. In this algorithm, we have root node, Decision Node and leaf nodes. Let’s discuss briefly on the nodes

* **Root Node:** It is the starting node where the entire data set is divided into different categories from this node.
* **Decision Node:** Decision node decides whether to make the final decision or to extend the tree based on the different conditions. Decision nodes will always have extended branches to connect to other nodes.
* **Leaf Node:** Leaf node is also known as Terminal node. It gives the output of the tree which is final. The leaf node does not have any branches to extend.

In our project, we have used four different classification/decision nodes and named as ClassDec Tree B2D6, ClassDec Tree B2D4, ClassDec Tree B2D2, ClassDec Tree B3D6 where B represents Branches and D represents Depth. We have changed the variables in the properties such as branches and depth as per the name suggess.

For ClassDec Tree B2D6, we have taken 2 branches and 6 depths. For ClassDec Tree B2D4, we have considered 2 branches and 4 depths. For ClassDec Tree B2D2, we have taken 2 branches and 2 depth nodes and finally for ClassDec Tree B3D6, we have considered 3 branches and 6 depths.

The best model among all these 4-decision trees is considered ClassDec Tree B3D6 where the miscellaneous rate of Train is “0.305428”, Validation is “0.313622” and Test is “0.309683”. The overfitting for this classification tree is very less as there is minimum difference between Average Square Error of Train and Validation and between Validation and Test. The Average square error of Train is “0.20388”, Validation is “0.207085” and Test is “0.20581”. The important variables in ClassDec Tree B3D6 are Age Category, General Health, Stroke, Sex, Difficulty in Walking, Diabetic and Sleep Time.

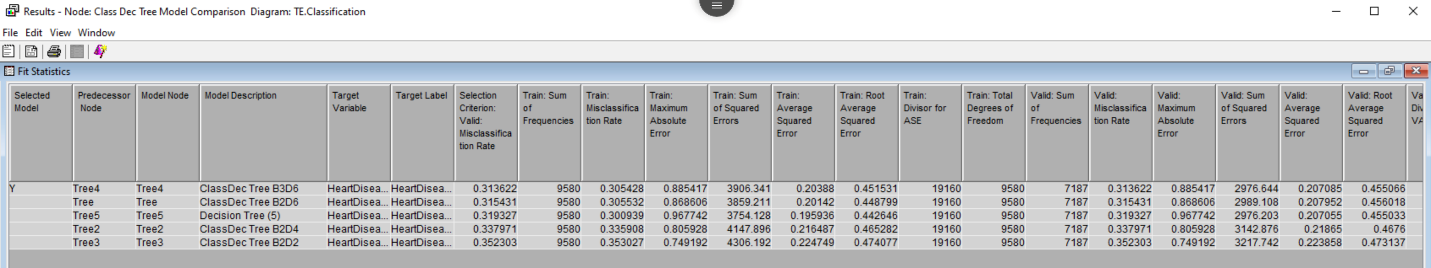


Fig 8 : Fit Statistics report of the Classification Tree

The above diagram shows the best classification in the entire 5 decision trees. It decides ClassDec Tree B3D6 as the best model.

Let’s calculate the accuracy and misclassification using confusion matrix. Let’s navigate to Output window and collect the values of True Positive, True Negative, False Positive and False Negative under “Event Misclassification Table”.

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix for Decision Tree** | | |
|  | **1** | **0** |
| **1** | 2983 | 1264 |
| **0** | 3671 | 1662 |

* Accuracy = TN+TP/ TN+TP+FP+FN

= 2983+3671/1/1264+1662+2983+3671

= 6654/9580

= 0.69 = 69%

* Miscellaneous = 100 – 69 = 31%

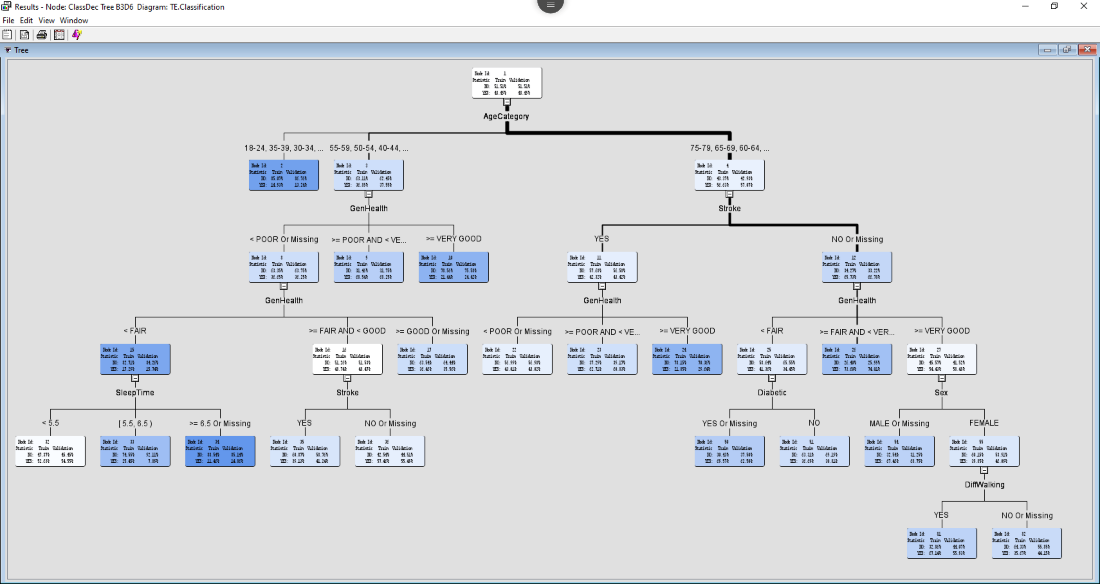


Fig 9: Decision Tree of ClassDec Tree B3D6

The above figure represents the tree diagram of ClasDec Tree B3D6. Here the root node is AgeCategory . The remaining nodes in this tree are General Health, Sleep Time, Stroke, Diabetic, Sex and Difficulty in walking. Based on the different factors the outcome will come. For example, a person with 25 years age, female, general health is Good, No diabetic, No difficulty in walking and sleep time is 6 hours a day. Let’s see how the above scenario will predict the heart attack. In this scenario, the leaf node is 2 and the heart attack prediction of yes is “14.92%” and No is “85.07%”. Here in this scenario, it doesn’t consider any other parameters, it is only predicted based on the age. That means the people with less than or equal to 40 years.

**Logistic Regression:**

The second technique that we have used in this project is Logistic Regression. In this project, we have used 4 logistic regression models which are Exhaustive Regression, Forward Regression, Backward Regression and Stepwise Regression by changing the properties of each node.

Exhaustive Regression is considered as the best model as per the comprehensive analysis we have performed. The miscellaneous rate of Train is “0.291336”, Validation is “0.300682” and Test is “0.28951”. Here the overfitting is very less as the difference between the miscellaneous rate of Train and validation and validation and test is very minimum.

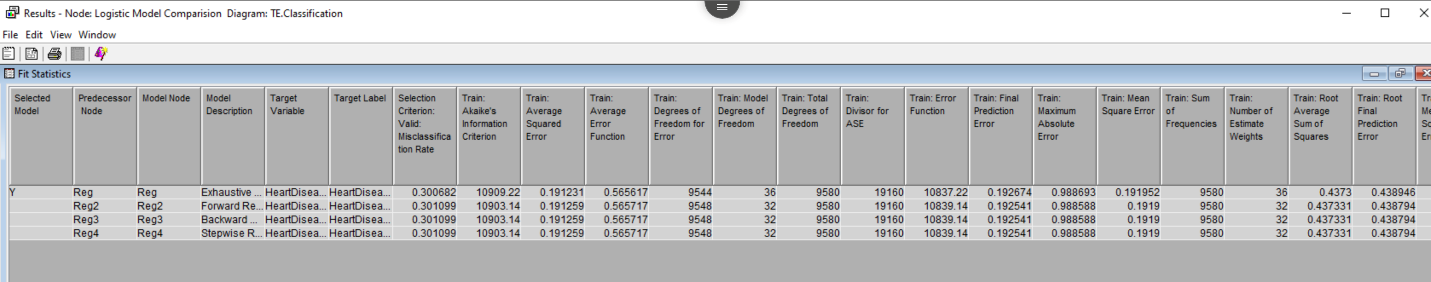


Fig 10 : Fit Statistics report of Logistic Regression

By performing confusion matrix analysis, we have found the accuracy and miscellaneous percentage.

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix for Exhaustive Regression** | | |
|  | **1** | **0** |
| **1** | 3250 | 1396 |
| **0** | 3539 | 1395 |

* Accuracy = TN+TP/ TN+TP+FP+FN

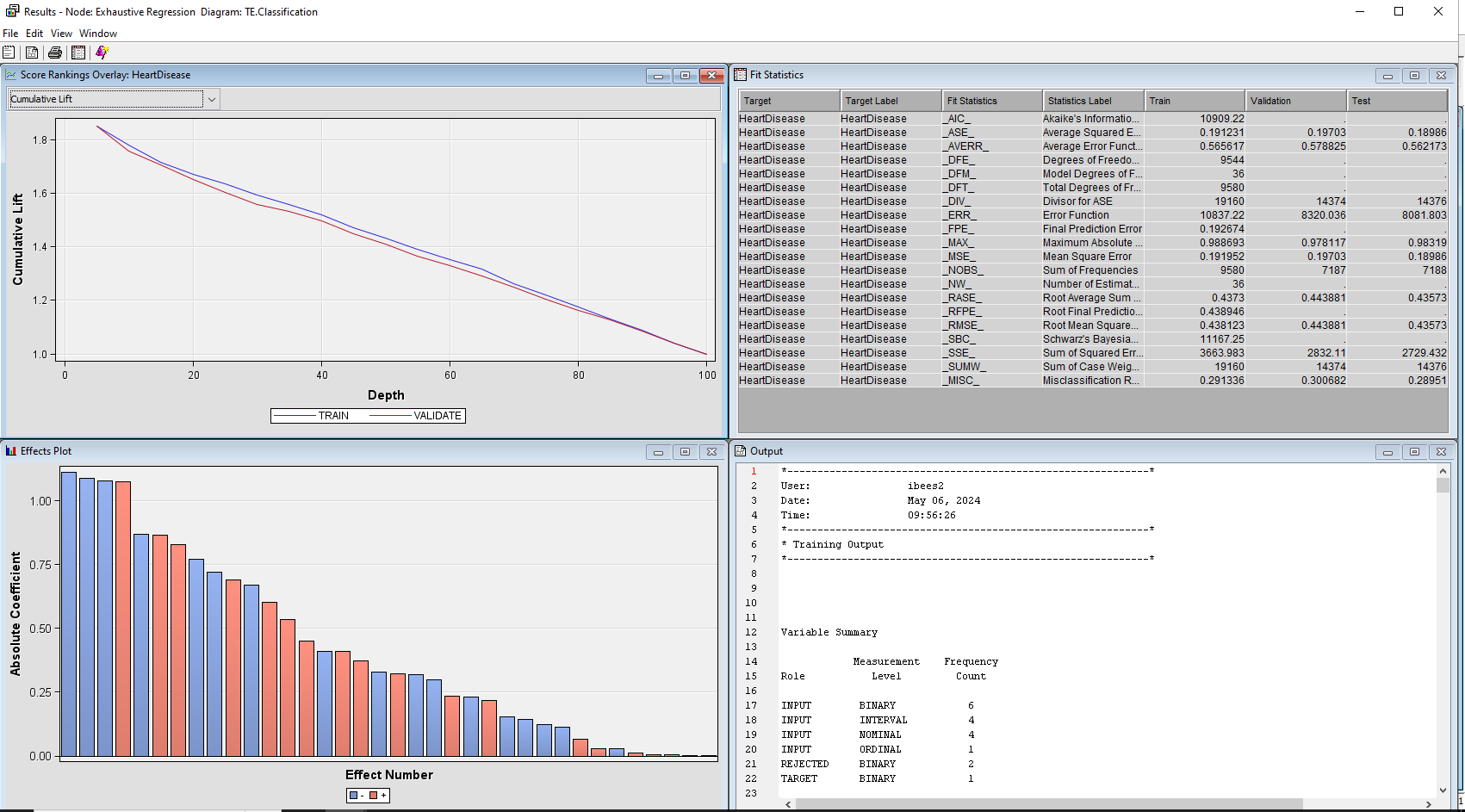
= 3539+3250/3539+3250+1396+1395

= 6789/9580

= 0.71= 71%

* Miscellaneous = 100 – 71= 29%

The Exhaustive Regression model is considered as the best model and the important variables are Alcohol Drinking, BMI, Physical Activity, Physical Health, Sleep Line.



**Fig 11: The output result of Exhaustive Regression**

The above figure represents the results of Exhaustive Regression. Here we can see that there is minimum Overfitting. The train and validate data curves have a very slight difference which represents that there is very less Overfitting in the model.

**Neural Network (Best Model):**

The other analysis that we have performed in our project is Neural Network. Here we have trained six different models such as Neural Network with 3 Hidden Units, Neural Network with 5 Hidden Units, Neural Network with 3 Hidden Units DataBase, Neural Network with 5 Hidden Units Database, Neural Network with 3 Hidden Units Back Prop, Neural Network with 5 Hidden Units Bak Prop by changing the variables in the property.

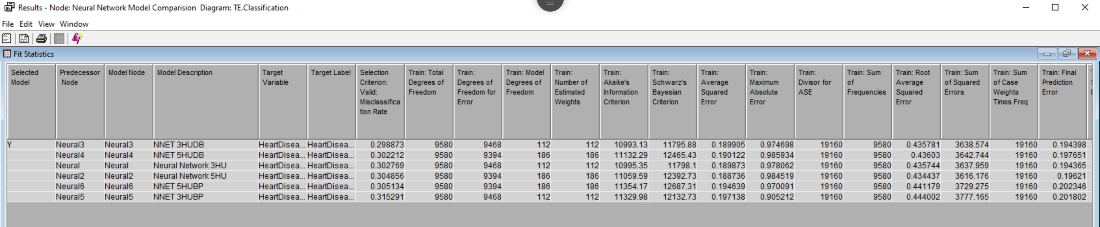


Fig 12: The Fit Statistics result window of Neural Network

From all the six models, Neural Network with 3 hidden units Data Base is considered as the best model. This model is not only considered as the best model in the neural network but from the entire 3 supervised techniques.

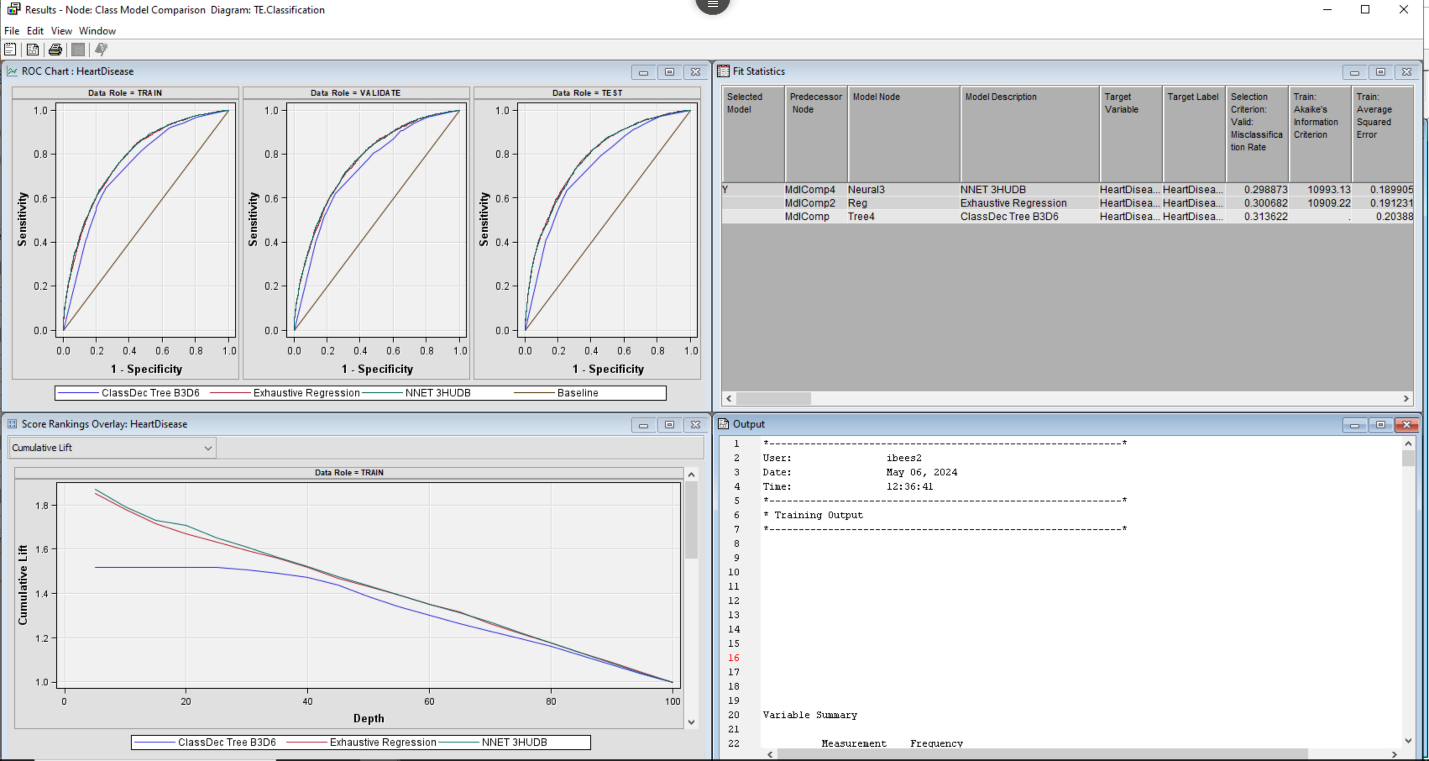


Fig 13 : The entire result window of Neural Network 3 HU with DB

From the confusion matrix below, we have calculated the accuracy and miscellaneous.

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix for Neural Network 3 Hidden Units** | | |
|  | **1** | **0** |
| **1** | 3227 | 1346 |
| **0** | 3589 | 1418 |

* Accuracy = TN+TP/ TN+TP+FP+FNFN

= 3589+3227/3589+3227+1346+1418

= 6816/9580

= 0.71= 71%

* Miscellaneous = 100 – 71= 29%

Though Neural Network and Logistic Regression has same accuracy percentage which is 71% but the model has chosen Neural Network as the best model by considering Overfitting and miscellaneous rates factors.­­­

**Conclusion:**

The aim of the project was to construct a model capable of anticipating an individual's susceptibility to heart disease by analyzing different aspects of their health. We used different methods to analyze data, like grouping similar things together, making decisions based on the data, and using a computer program to find patterns.

The analysis looked at how different things like weight, health, smoking, exercise, and heart disease risk are related using clustering methods like Ward, Average, and Centroid Clustering. The researchers found that Ward Clustering was the best method to group data into three different groups based on different characteristics.

Out of all the ways to predict things, the Neural Network model with 3 hidden units was the best. It was about 71% accurate at predicting the risk of heart disease. The Exhaustive Regression model is the best at Logistic Regression and performed similarly with an accuracy of about 71%.

According to the study findings, BMI emerges as the key factor in determining the likelihood of developing heart disease. Other factors that also matter include overall health, smoking, physical fitness, sleep, difficulty walking, age, and whether someone already has heart disease.

The results show that things you can change, like staying at a healthy weight, being active, and not smoking, are important for lowering the chance of heart disease. Also, when looking at an individual's overall risk, we need to think about things like their age and gender, which they can't change.

Finally, the project used data mining to find important health signs linked to the risk of heart disease. The new predictive models can help doctors find people at high risk and give them the right treatment to lower their risk. Additionally, what we learn from looking into things can help with making people healthier and reducing heart disease. This information can be used for making plans to help people live healthier lives.

References:

* Http://iopscience.iop.org/article/10.1088/1742-6596/1043/1/012003 | request PDF. (n.d.). <https://www.researchgate.net/publication/326028799_httpiopscienceioporgarticle1010881742-659610431012003> (Heart disease prediction using data mining, 2021)
* World Health Organization. (n.d.-a). Cardiovascular diseases. World Health Organization.<https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1>