Department of Computer Science and Engineering

Institute of Engineering

JK Lakshmipat University, Jaipur, India

Anamika Agrawal

anamikaagrawal@jklu.edu.in

Abstract

In this paper, we will present an application of machine learning in real life. We will introduce to a dataset on which we will work to understand the data and how it will help to predict whether the person will have a cardio disease or not. Then we will move towards data analysis and will get answers for a reason behind the person having a cardio disease. Moreover, we will predict the accuracy of the model by applying machine learning algorithms.

PREDICTING THE RISK OF HEART DISEASE

A case study on predicting whether person will have heart disease or not

1. Case Synopsis
2. Learning Objective of the case
3. Introduction to data
4. Study Questions

4.1 Selecting features for prediction

4.2 Developing a logistic regression model that can be used for predicting the probability of having cardio disease with only-

a. Cholesterol level

b. Smoke

c. Gender

d. Alcohol

e. Active

f. Glucose

4.3 Developing a stepwise logistic regression model including all the parameters.

4.4 Calculating the most accuracy of the model.

4.5 Developing a CHAID tree and defining business rules.

4.6 Calculating the accuracy of the CHAID tree.

4.7 Comparing the accuracy tell us the most efficient model.

4.8 Giving the final recommendations.

1. References

Appendix A: Description of Utilized technique: CHAID

Appendix B: CRISP Model Steps:

B.1. Business understanding

B.2. Data understanding

B.3. Data preparation

B.4. Modelling

B.5. Evaluation

Appendix C: Python Implementation

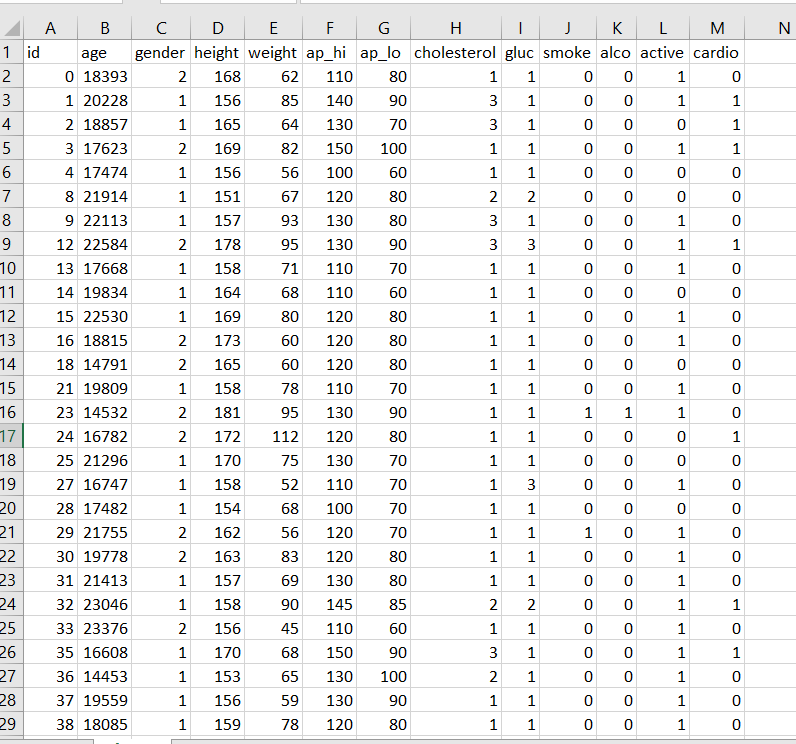
1. **Case Synopsis-**

At the flip of the century, cardiovascular diseases (CVDs) became the leading cause for mortality in India. This epidemiological transition is essentially due to the rise within the prevalence of CVDs and CVD risk factors in India. In 2016, the calculable prevalence of CVDs in India was estimated to be 54.5 million. One in 4 deaths in India are now due to CVDs with ischemic heart condition. These diseases tend to have an effect on patients inside the foremost economical years of their lives and finish in ruinous social and economic consequences. The most common causes of having a heart diseases are- smoking, high blood pressure, diabetes, excessive use of alcohol or coffee, cholesterol level and etc. In this case study we have collected various factor such as gender, glucose level, cholesterol level, etc. and then predict whether the person will have cardiac disease or not.

In this case classification techniques such as logistic regression as well as classification tree techniques such as chi-squared automated interaction detection(CHAID) and classification and regression tree(CART) are Prediction of classification probabilities, likelihood ratio tests, Wald’s test, classification table an classification plots are also used.

1. **Learning objectives of the case-**
2. Understand the application of decision tree(applied on classification problem or regression problem)
3. Understand the difference between CHAID technique and other decision trees.
4. Understand which statistical test should be used( Chi-Square test)and how similar or associated variables are.
5. Understand how this algorithm works and how it shows the output.
6. Understand the various statistical techniques(p-test , f-test or chi-squared test).
7. Demonstrate the application of Logistic regression in solving classification problem.
8. **Introduction to the Dataset-**

In this report we have used the cardio disease dataset from Kaggle. The dataset was available in csv format. The snapshot of the dataset is attached below-



1. **Suggested Questions:**

1.What are the most important features considered while predicting whether a person has a heart disease or not?

2. Develop a logistic regression model that can be used for predicting the probability of having cardio disease with only:

a. Cholesterol level

b. Smoke

c. Gender

d. Alcohol

e. Active

f. Glucose

3. Develop a stepwise logistic regression model by including all the appropriate parameters.

4. Comment on the model development; how do you measure the accuracy of the model?

5. Develop a chi-squared automatic interaction detection(CHAID) tree using the data provided. Discuss how this tree can be used?

6. Measure the accuracy of the CHAID decision tree.

7. Comparing the accuracy tell us the most efficient model.

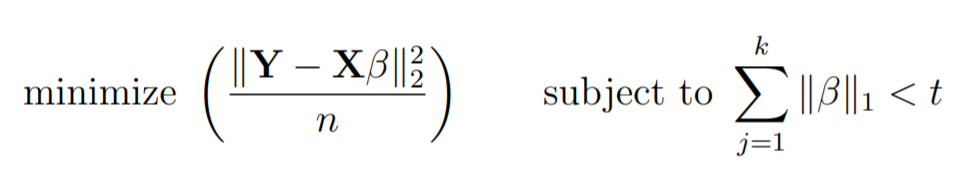
8. What are your final recommendation to keep heart healthy.

1. **Suggested Answers:**
2. What are the most important features considered while predicting whether a person has a heart disease or not?

The purpose of the question is to make realize that not all the parameters given in the dataset are used. Only few variables can be directly used for building the model. The variable used are-

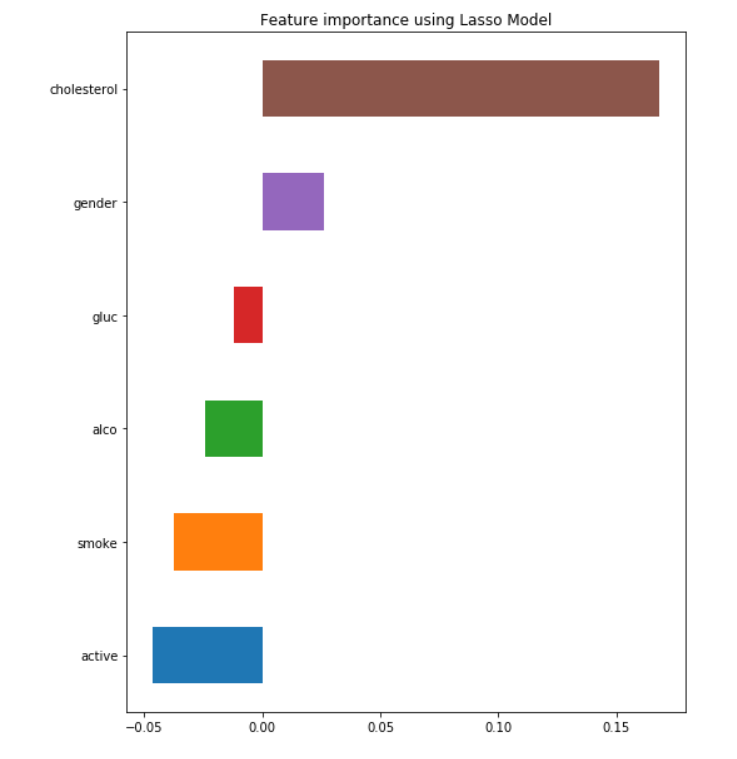
|  |  |
| --- | --- |
| VARIABLE NAME | MEANING |
| Gender | Tells about gender of the person |
| Cholesterol | Tells about cholesterol level of the person. |
| Gluc | Tells about glucose level of the person. |
| smoke | Tells whether person smokes or not |
| Alco | Tells whether person consumes alcohol or not. |
| Active | Tells whether person is active or not. |
| cardio | Decision (Predict whether person will have cardio problem or not) |

The method used for the feature selection is ’Lasso feature selection’. LASSO stands for ‘Least Absolute Shrinkage and Selection Operator’. The LASSO method puts a constraint on the sum of the absolute values of the model parameters, the sum has to be less than a fixed value (upper bound). The goal of the process is to minimize the prediction error. The lasso estimate is defined by the solution to the l1 optimization problem



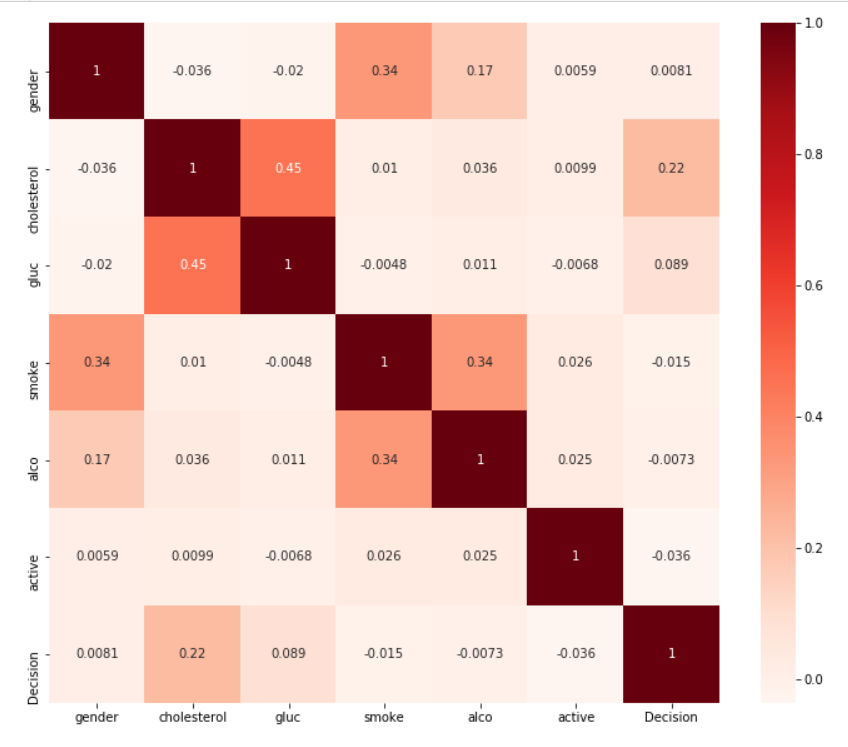
where t is the upper bound for the sum of the coefficients.

The output of the Lasso feature selection is shown below-



This graph is showing the importance of features used in the model.

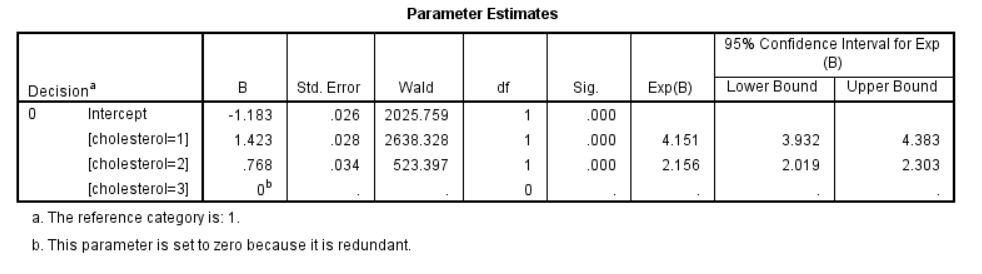
The heatmap below tells us the correlation of the variables-

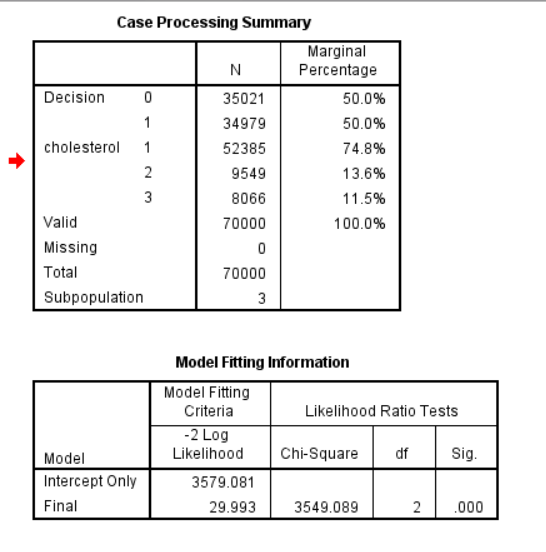


1. Develop a logistic regression model that can be used for predicting the probability of having cardio disease with only:

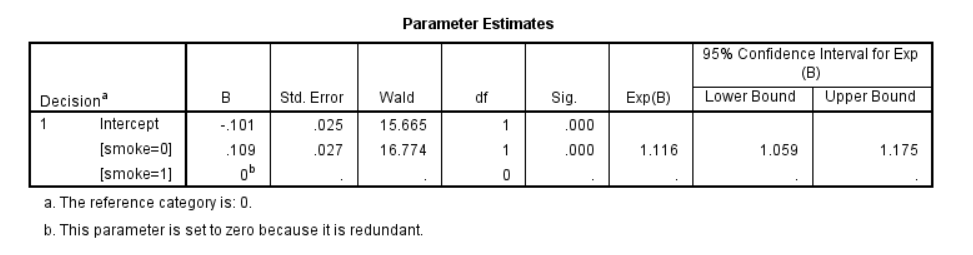
In the following tables the variables in the equations and the classification plot are shown-

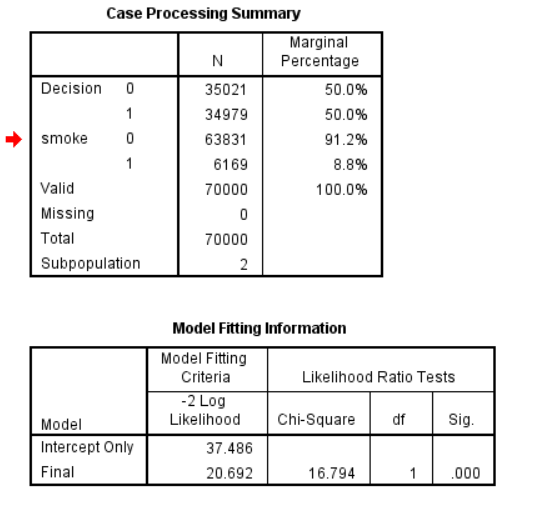
1. Cholesterol level-



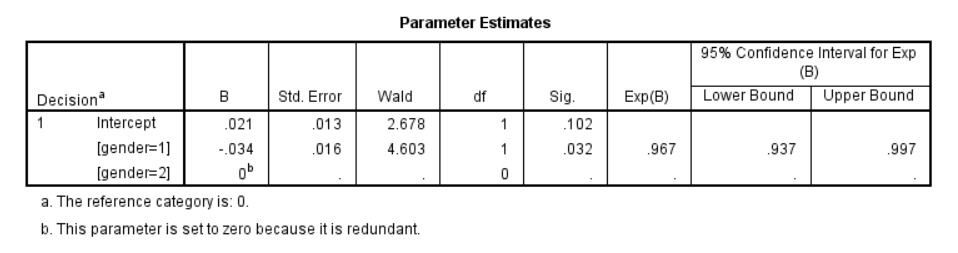


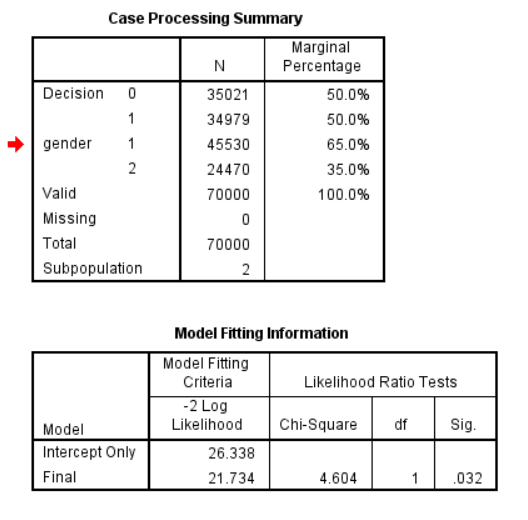
1. Smoke



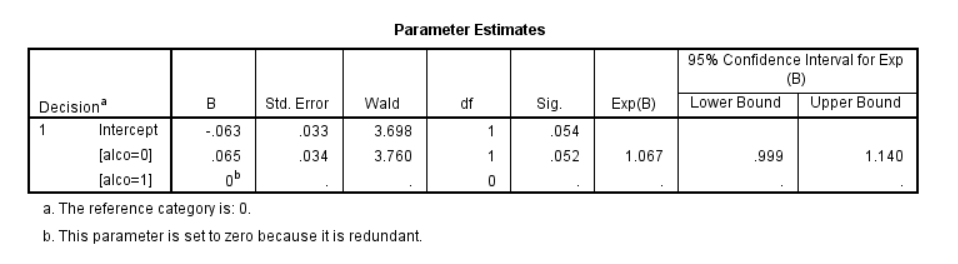


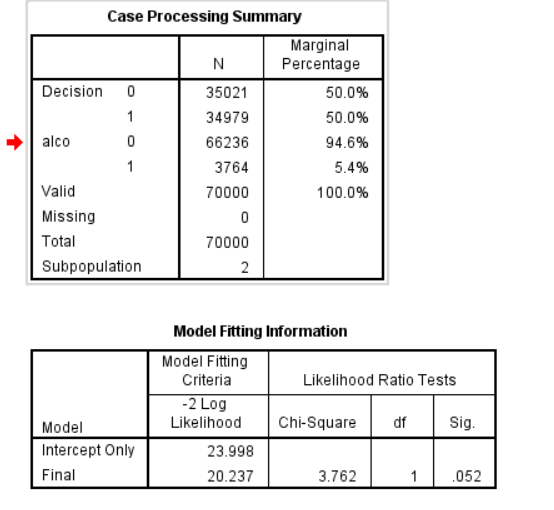
1. Gender



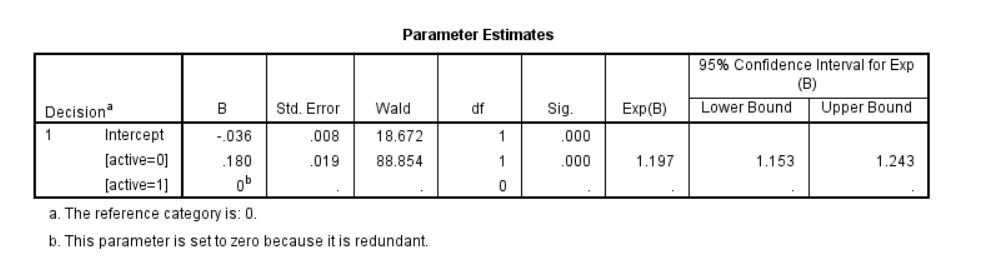


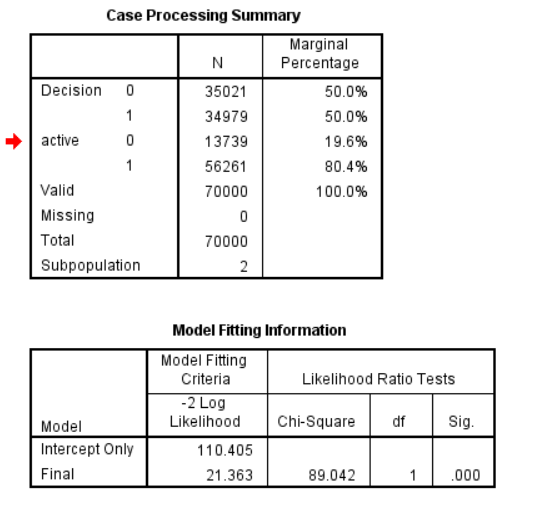
1. Alcohol-



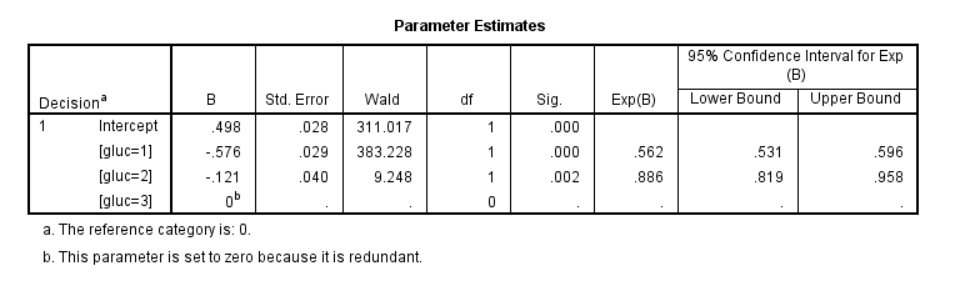


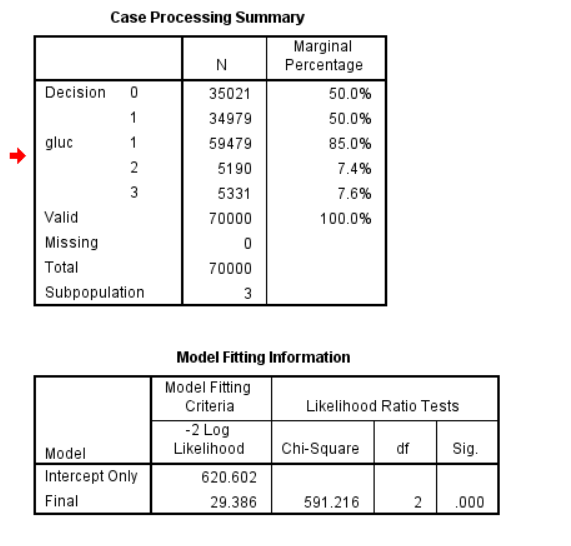
1. Active-





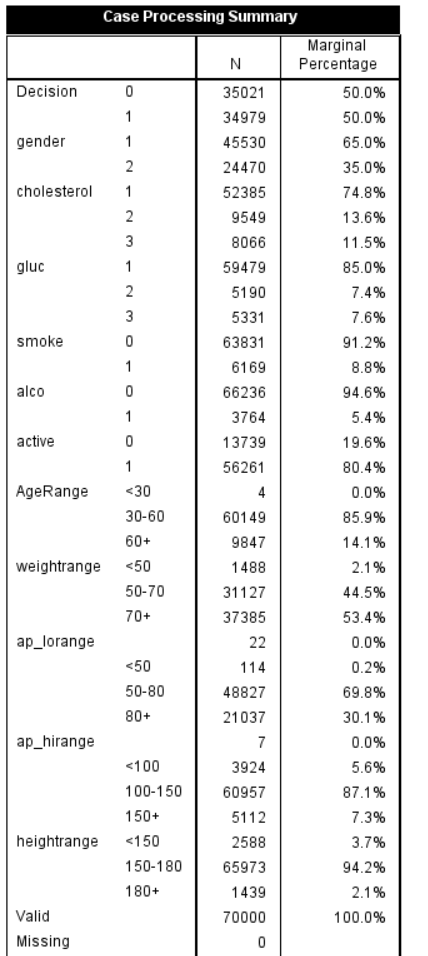
1. Glucose-

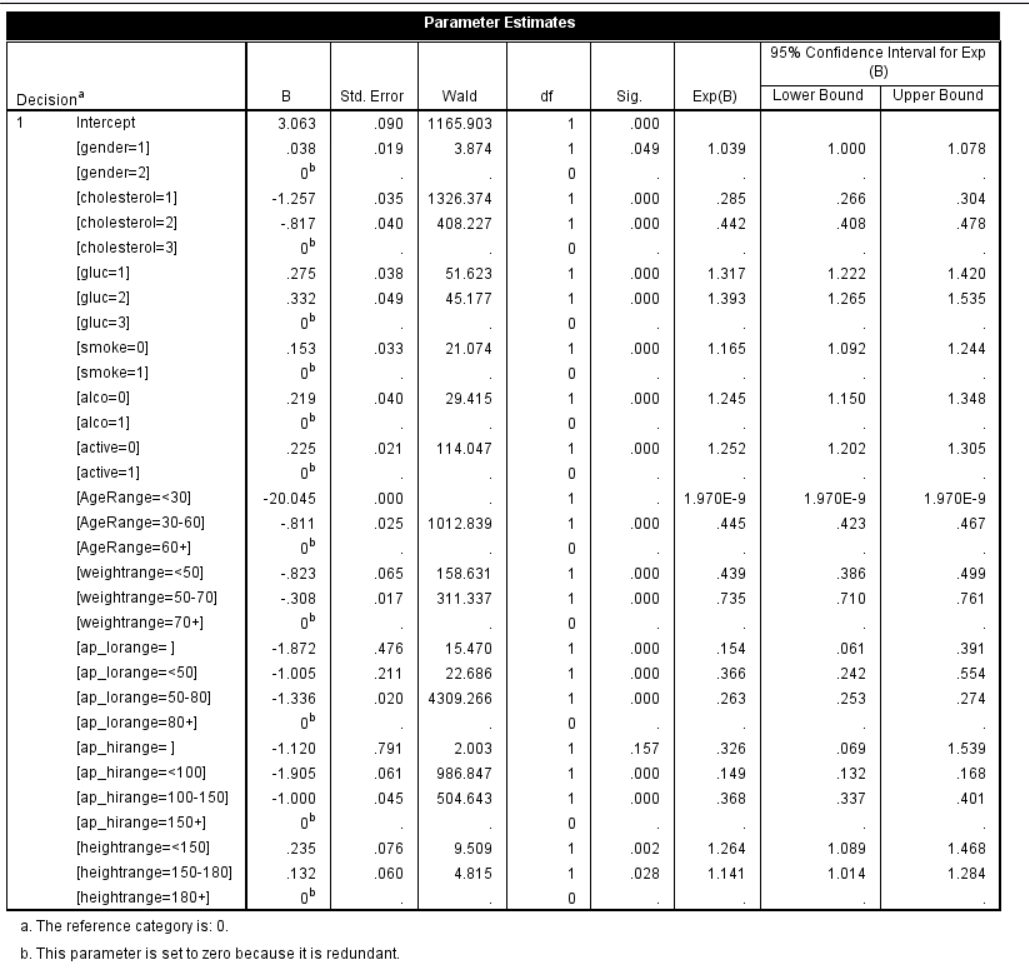




1. Develop a stepwise logistic regression model by including all the appropriate parameters.

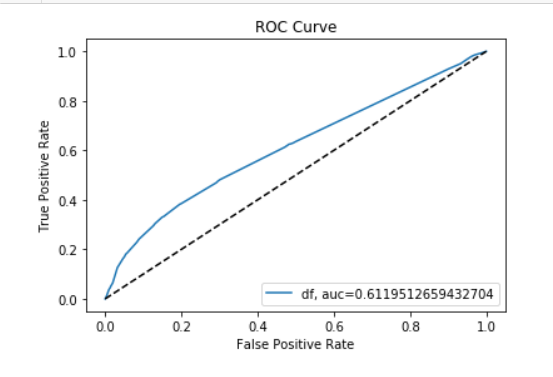
A stepwise regression output is shown below-





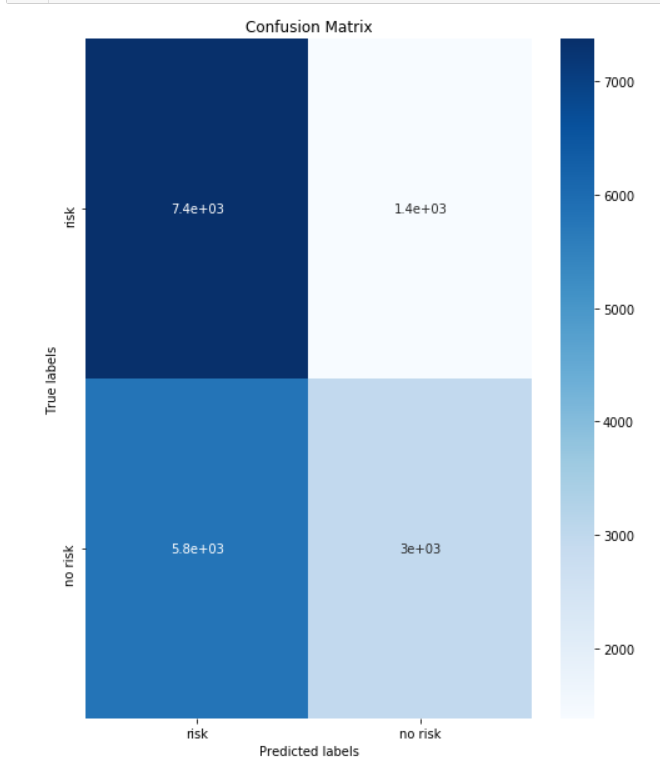
1. Comment on the model development; how do you measure the accuracy of the model?

Accuracy of the model is obtained measures such as sensitivity, specificity, and over all accuracy. These measures are dependent on the classification cut-off probability used. A way for measuring the accuracy of logistic regression is through the area under ROC curve.



Accuracy- 0.611951265

Another way for measuring the accuracy is- Confusion matrix



Accuracy: 0.5901142857142857

Precision: 0.6814404432132964

Recall: 0.33756432246998286

We have also calculated the MAE, MSE and RMSE.

Mean Absolute Error: 0.4098857142857143

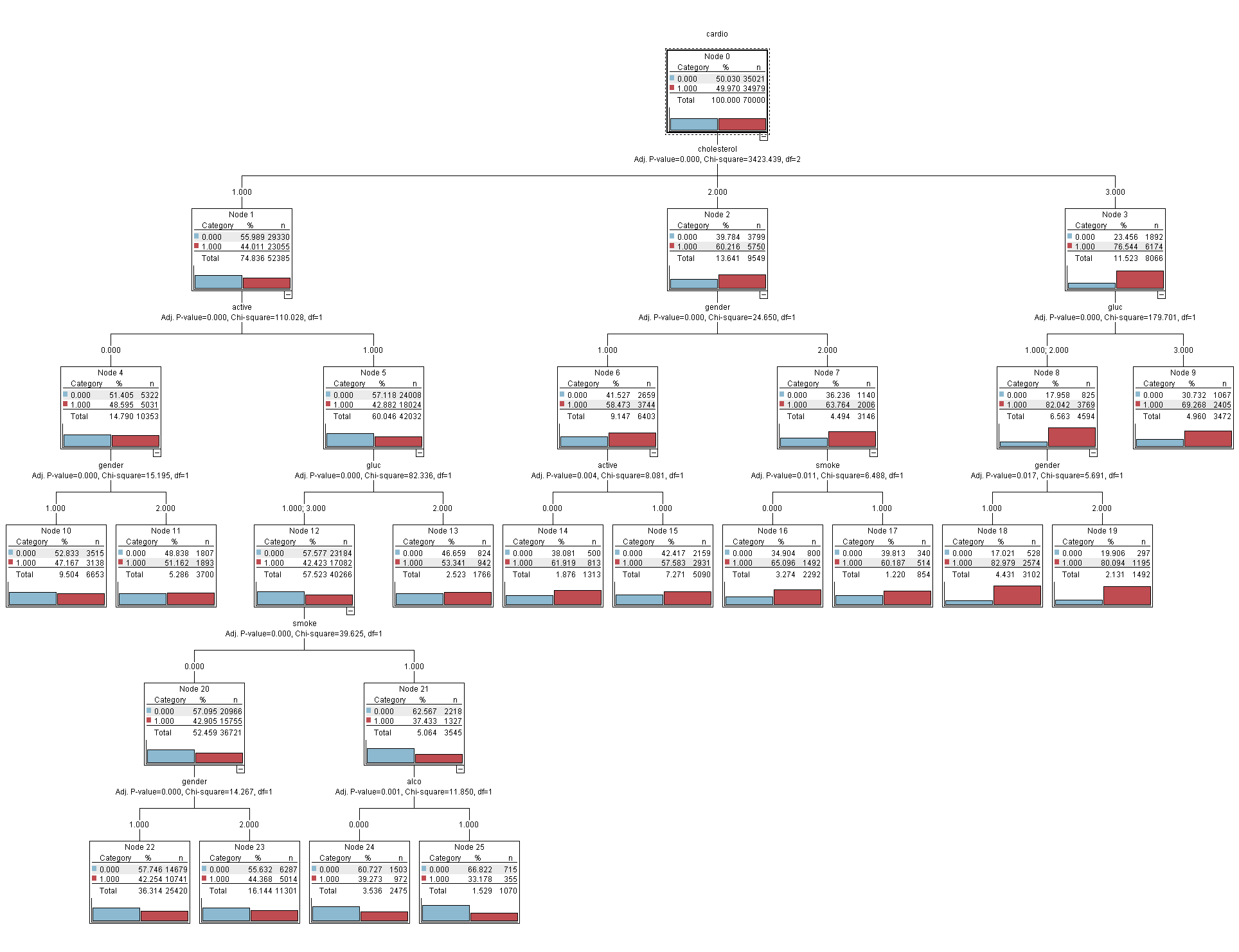
Mean Squared Error: 0.4098857142857143

Root Mean Squared Error: 0.6402231753738022

1. Develop a chi-squared automatic interaction detection(CHAID) tree using the data provided. Discuss how this tree can be used?

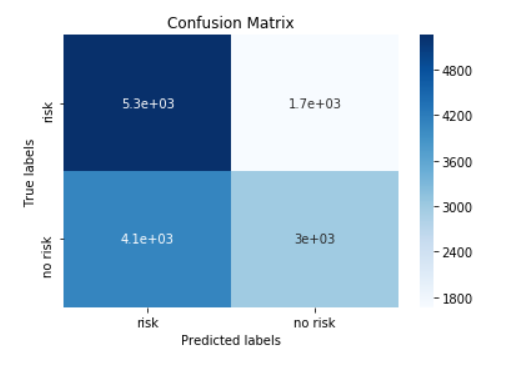
**Rules for leaf nodes of CHAID**

|  |  |
| --- | --- |
| **Cases** | **Rules** |
| 1 | If cholesterol =1, active=0,gender=women, then cardio= 47% |
| 2 | If cholesterol =1, active=0,gender=men, then cardio= 51% |
| 3 | If cholesterol =1, active=1, glucose = 2 , then cardio= 53% |
| 4 | If cholesterol =1, active=1, glucose = 1,3 , smoke=0 , gender = women, then cardio= 42% |
| 5 | If cholesterol =1, active=1, glucose = 1,3 , smoke=0 , gender = men, then cardio= 44% |
| 6 | If cholesterol =1, active=1, glucose = 1,3 , smoke=1 , alcohol = 0, then cardio= 39% |
| 7 | If cholesterol =1, active=1, glucose = 1,3 , smoke=1 , alcohol = 1, then cardio= 33% |
| 8 | If cholesterol =2, active=0,gender=women, then cardio= 61% |
| 9 | If cholesterol =2, active=1,gender=women, then cardio= 57% |
| 10 | If cholesterol =2,gender=women, smoke=0 then cardio= 65% |
| 11 | If cholesterol =2,gender=men, smoke=1 then cardio= 60% |
| 12 | If cholesterol =3,gender=women, glucose = 1,2 , then cardio= 82% |
| 13 | If cholesterol =3,gender=men, glucose = 1,2 , then cardio= 80% |
| 14 | If cholesterol =3, glucose = 3 , then cardio= 69% |



1. Measure the accuracy of the CHAID decision tree.

The accuracy of the decision tree is calculated by Confusion matrix-



Accuracy: 0.5892857142857143

Precision: 0.642075916791765

Recall: 0.42318021201413425

We have also calculated the MAE, MSE and RMSE.

Mean Absolute Error: 0.4107142857142857

Mean Squared Error: 0.4107142857142857

Root Mean Squared Error: 0.6408699444616557

1. Comparing the accuracy tell us the most efficient model.
2. Comparison of confusion matrix-

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | **PREDICTED** |
| CLASS 1 | TP | FN |
| CLASS 2 | FP | TN |

TP(True Positive)- Observation is positive, and is predicted to be positive.

FP(False Positive) - Observation is negative, but is predicted positive

TN(True Negative) - Observation is negative, and is predicted to be negative.

FN(False Negative) - Observation is positive, but is predicted negative.



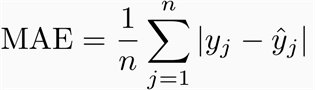
Accuracy of Model 1- 0.5901142857142857

Accuracy of Model 2 - 0.5892857142857143

Model 1(Logistic Regression) is more efficient that model 2(Decision Tree).

1. Comparison in terms of error-

Mean Absolute Error(MAE)- measures the average magnitude of the errors.



Error in Model 1- 0.4098857142857143

Error in Model 2 - 0.4107142857142857

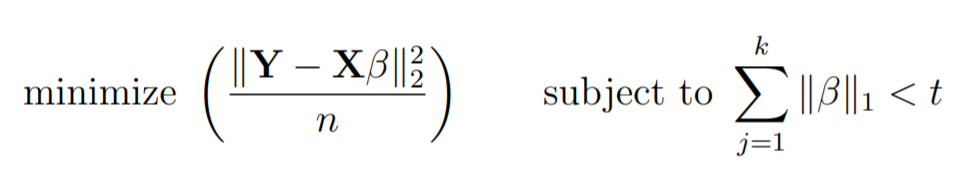
In this also, the error in model 1(Logistic regression) is less than model 2(Decision Tree)

**So, We can conclude that model 1 is more efficient than model 2.**

1. What are your final recommendation to keep heart healthy.

We have used logistic regression as well as CHAID decision tree to come up with our final recommendation. The following approaches are used in the model-

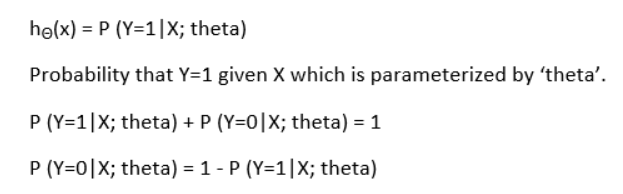
1. LASSO for feature selection- . LASSO stands for ‘Least Absolute Shrinkage and Selection Operator’. The LASSO method puts a constraint on the sum of the absolute values of the model parameters, the sum has to be less than a fixed value (upper bound). The goal of the process is to minimize the prediction error. The lasso estimate is defined by the solution to the l1 optimization problem

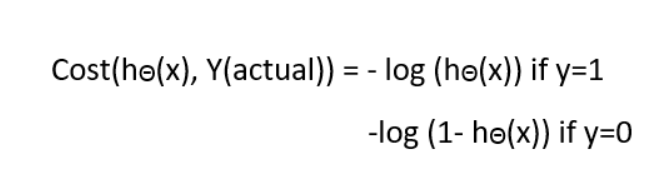


where t is the upper bound for the sum of the coefficients.

1. Logistic regression- Logistic regression may be a statistical model that in its basic form uses a logistic function to model a binary variable , although more complex extensions exist. In multivariate analysis , logistic regression (or logit regression) is estimating the parameters of a logistic model (a sort of binary regression).

The mathematical representation of the logistic regression is given below-





1. ROC Curve- ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability.

**Appendix A: Description of Utilized technique: CHAID**

The CHAID algorithms consist of three steps: merging, splitting and stopping. A tree is grown by repeatedly using these three steps on each node starting from the root node. It only accepts nominal or ordinal categorical predictors. When predictors are continuous, they are transformed into ordinal predictors before using the following algorithm.

**Merging**-

For each predictor variable X, merge non-significant categories. Each final category of X will result in one child node if X is used to split the node. The merging step also calculates the adjusted p-value that is to be used in the splitting step-

1) If X has 1 category only, stop and set the adjusted p-value to be 1.

2) If X has 2 categories, go to step 8.

3) Else, find the allowable pair of categories of X that is least significantly different the most similar pair is the pair whose test statistic gives the largest p-value with respect to the dependent variable Y.

4) For the pair having the largest p-value, check if its p-value is larger than a user specified alpha-level α merge (alpha merge). If it does, this pair is merged into a single compound category. Then a new set of categories of X is formed.

5) If the newly formed compound category consists of three or more original categories, then find the best binary split within the compound category which p-value is the smallest. Perform this binary split if its p-value is not larger than an alpha-level α split-merge.

6) Go to step 2.

7) The adjusted p-value is computed for the merged categories.

**Splitting**-

The “best” split for each predictor is found in the merging step. The splitting step selects which predictor to be used to best split the node. Selection is accomplished by comparing the adjusted p-value associated with each predictor. The adjusted p-value is obtained in the merging step.

1) Select the predictor that has the smallest adjusted p-value (i.e., most significant).

2) If this adjusted p-value is less than or equal to a user-specified alpha-level α split (alpha split), split the node using this predictor. Else, do not split and the node is considered as a terminal node.

**Stopping-**

The stopping step checks if the tree growing process should be stopped according to the following stopping rules.

1)If a node becomes pure; that is, all cases in a node have identical values of the dependent variable, the node will not be split.

2)If all cases in a node have identical values for each predictor, the node will not be split.

3)If the current tree depth reaches the user specified maximum tree depth limit value, the tree growing process will stop.

4)If the size of a node is less than the user-specified minimum node size value, the node will not be split.

5)If the split of a node results in a child node whose node size is less than the user specified minimum child node size value, child nodes that have too few cases (as compared with this minimum) will merge with the most similar child node as measured by the largest of the p-values. However, if the resulting number of child nodes is 1, the node will not be split.

**Appendix B: CRISP Model Steps:**

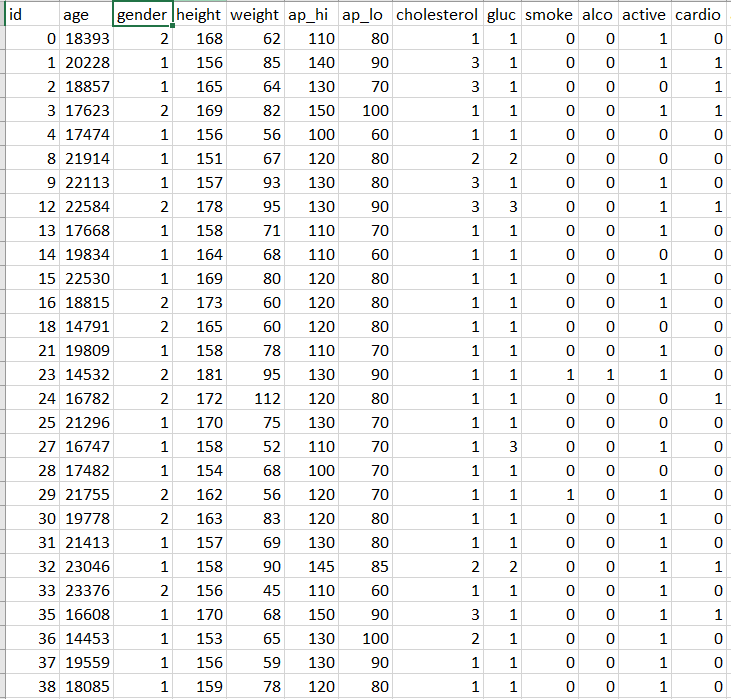
B.1. **Business understanding-**

The main objective of this step is to determine the goal of this case study and what can be achieved by the results. The CHAID decision tree algorithm is a classification algorithm and can be used to discover the relationship between the variables.

In this case study we have worked on the medical related data and can help a lot in this field. We have collected various factors that affect the risk of having a cardio disease. So, when a person arrives a hospital and tells us about the following features we can predict whether that person has a cardio disease or not.

B.2. **Data understanding-**

The medical cover type dataset falls under medical aspect. It has 13 attributes and 70,000 rows.



B.3. **Data preparation-**

In the dataset, the column age is age of the person given in days. It is the continuous value so we have converted it into the categorical values and changed age into years. The next column is of gender which tells us about the gender of the person. The next two columns are height and weight which are continuous values and need to be converted into categorical values. Ap\_hi and ap\_lo are the blood pressures of the persons which we have converted to the categorical value. The next columns are cholesterol which tells about the cholesterol level of the person, gluc is the glucose level, smoke and alcohol is categorical values which tells whether person smokes or consumes alcohol or not. The active is column is also a categorical value which tells about how active a person is. And at last the Cardio is the decision and the predicted value.

The steps in Data pre-processing includes- Lasso feature selection and converting the values of the columns.

B.4. **Modelling-**

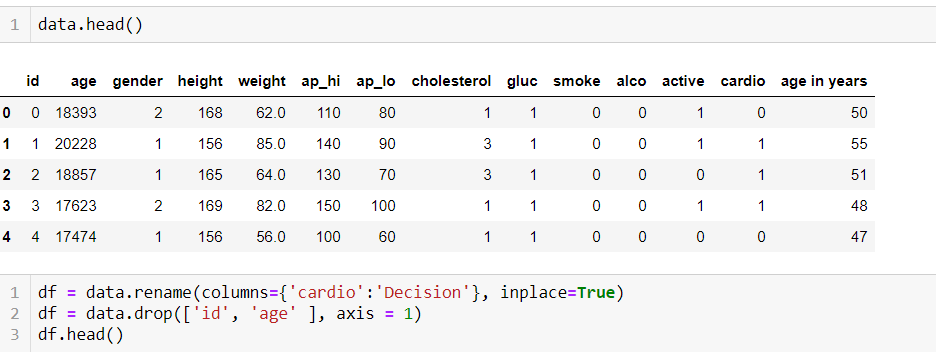
The fourth phase of CRISP model is modelling. We have to specify the techniques that would suit our assumptions. This is necessary to be a well-informed about all the possibilities. In the case study we have used Logistic regression as well as CHAID decision tree to build our model. After modelling we have checked accuracy of both the models and compared them.

B.5. **Evaluation-**

This is the last step of CRISP model which deals with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient.

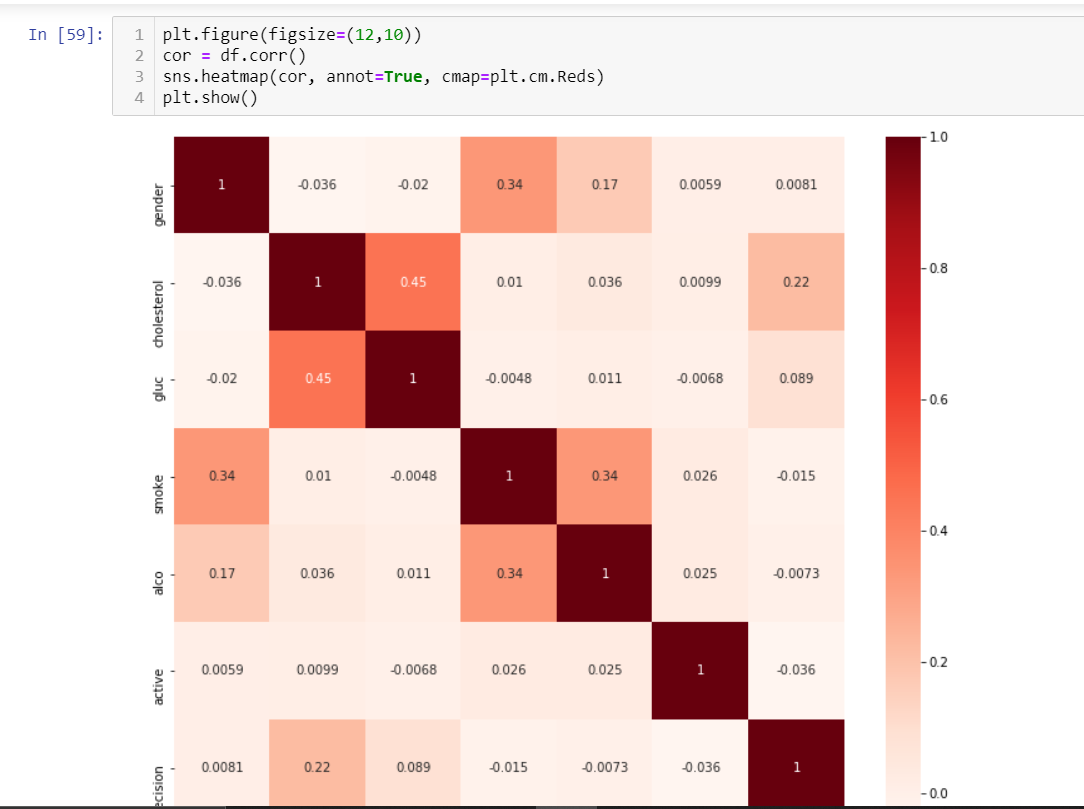
**Appendix C: Python Implementation-**

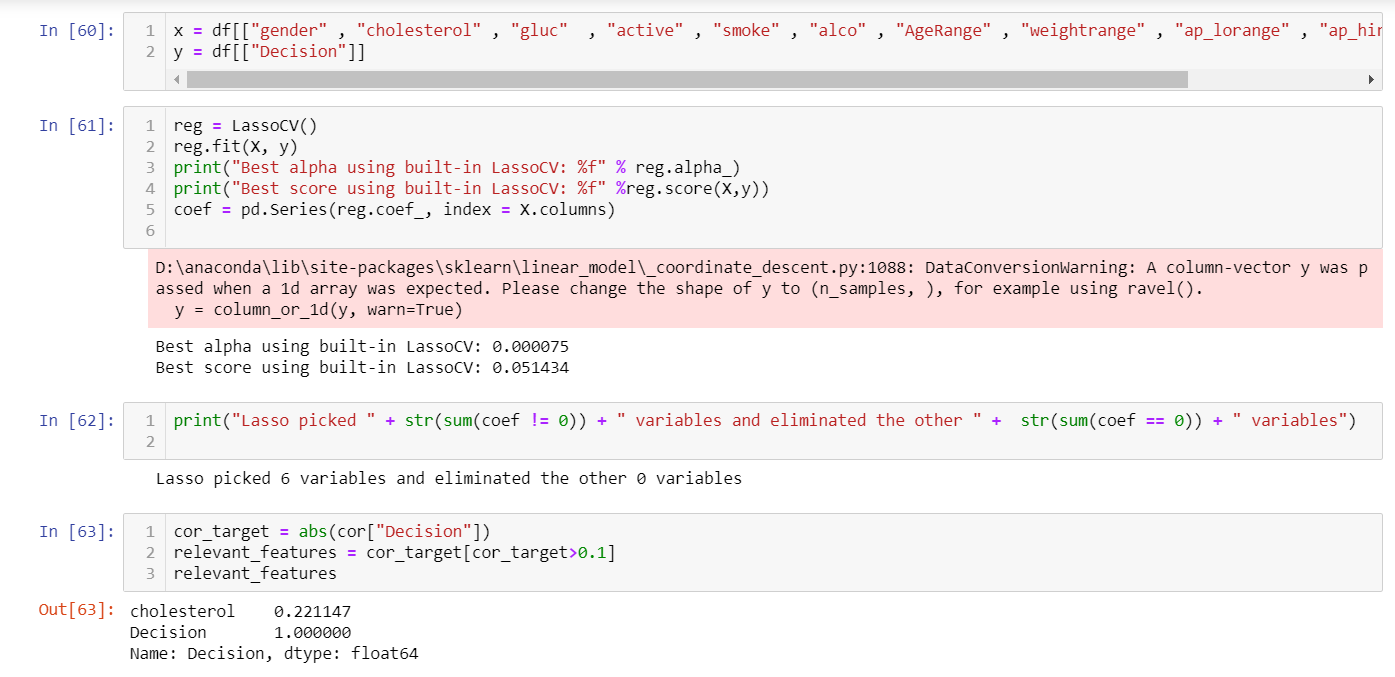
1. Data pre-processing

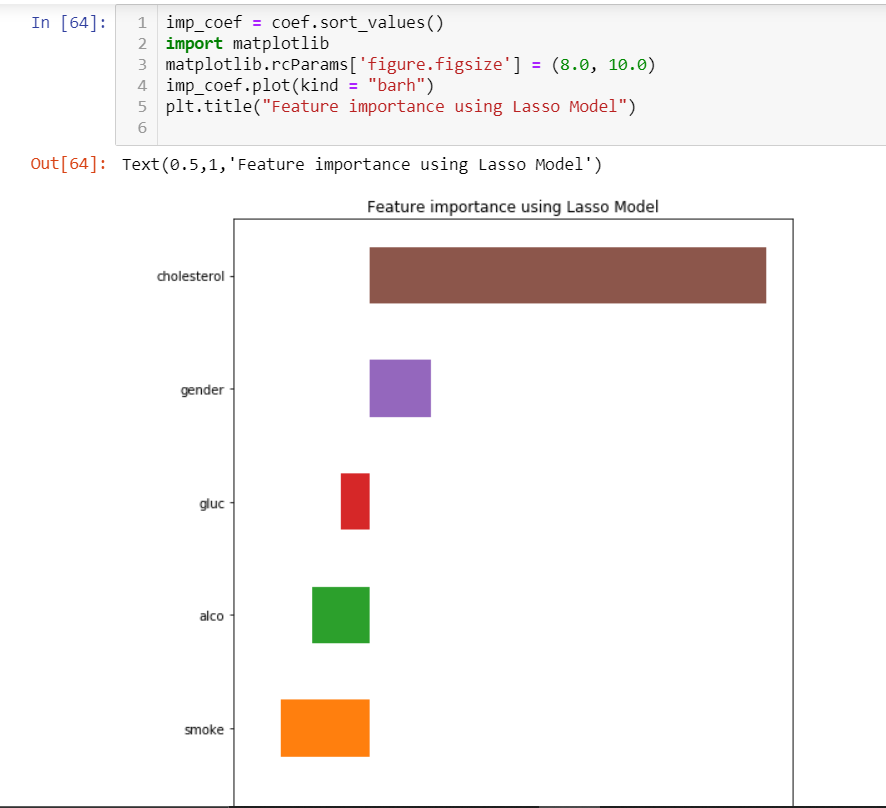




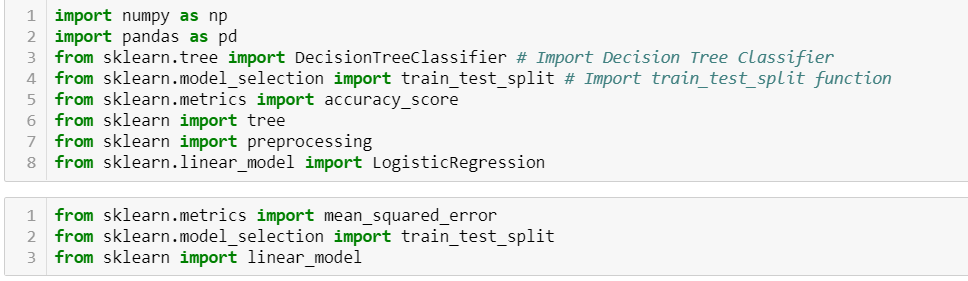
1. Feature selection-

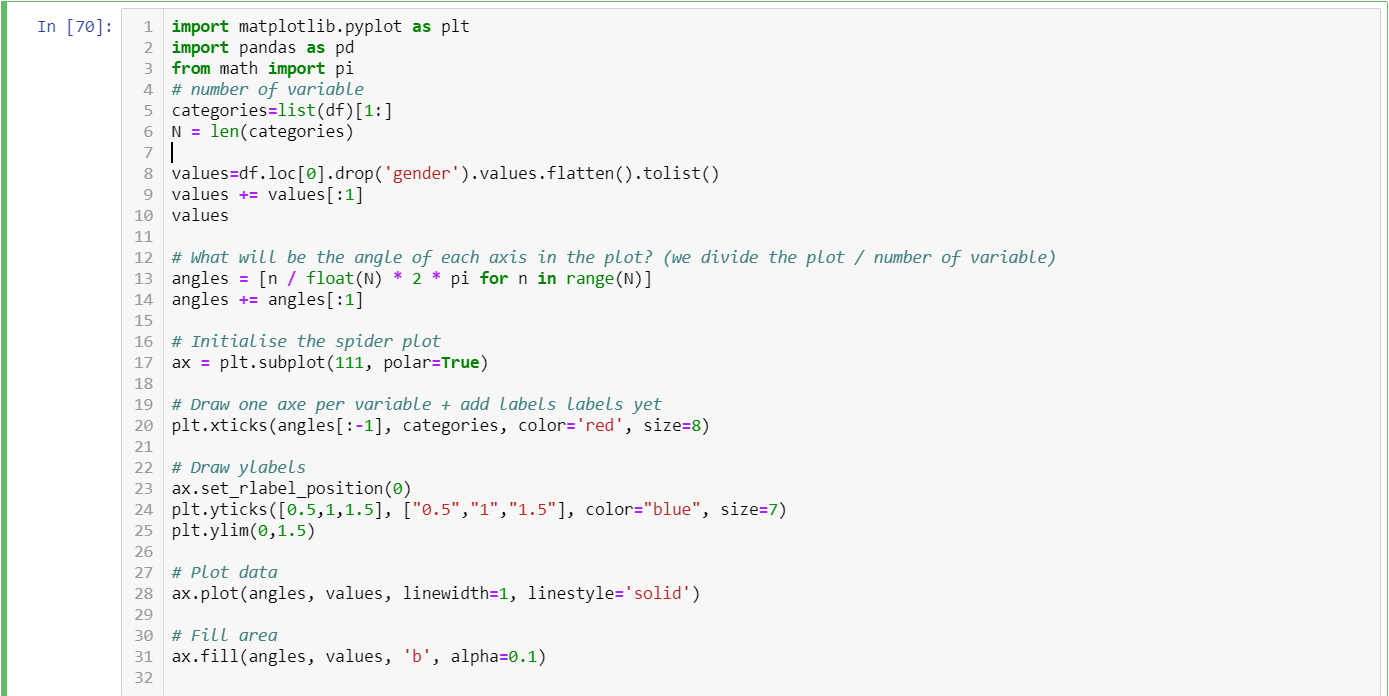


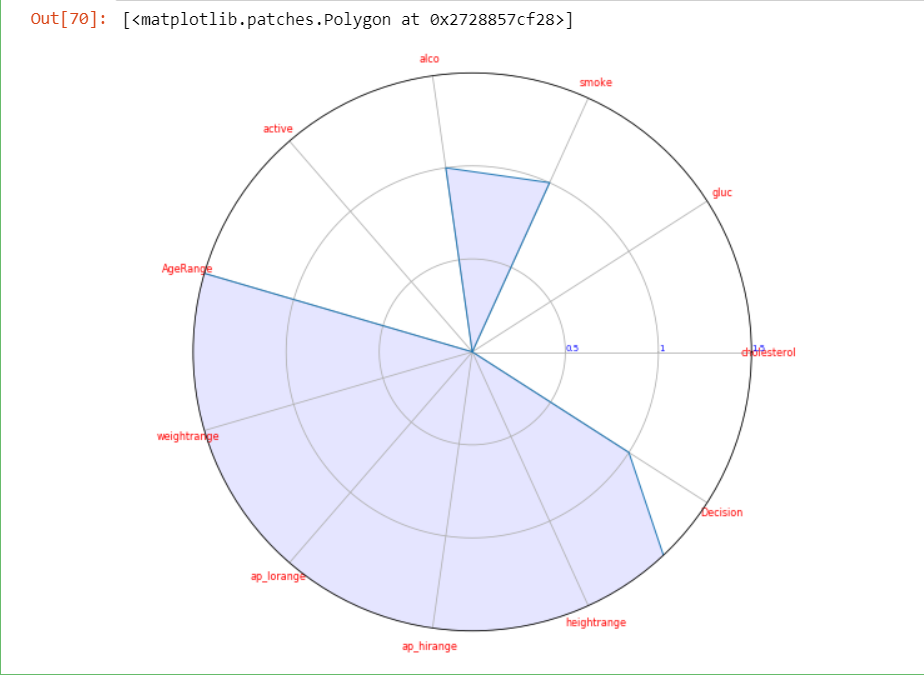




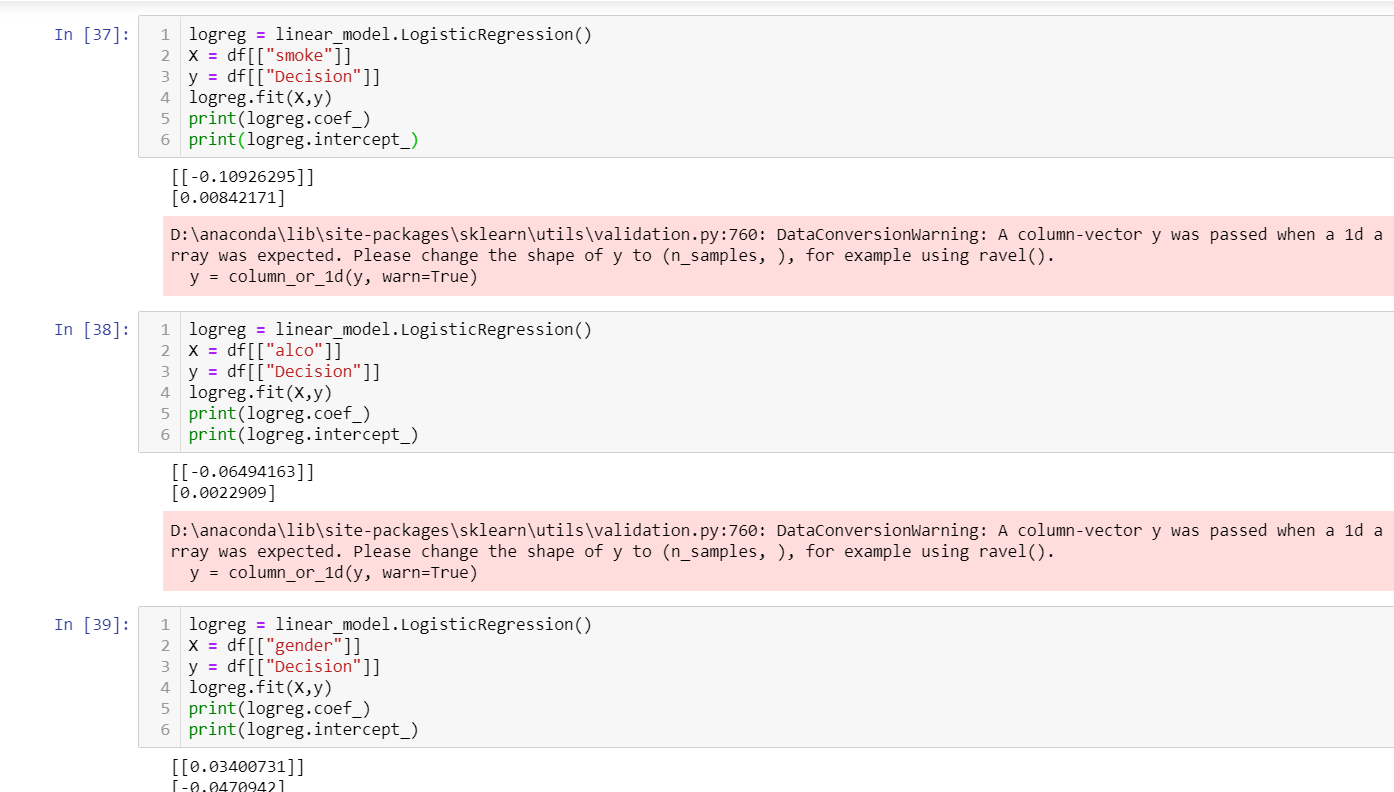
1. Data visualization-

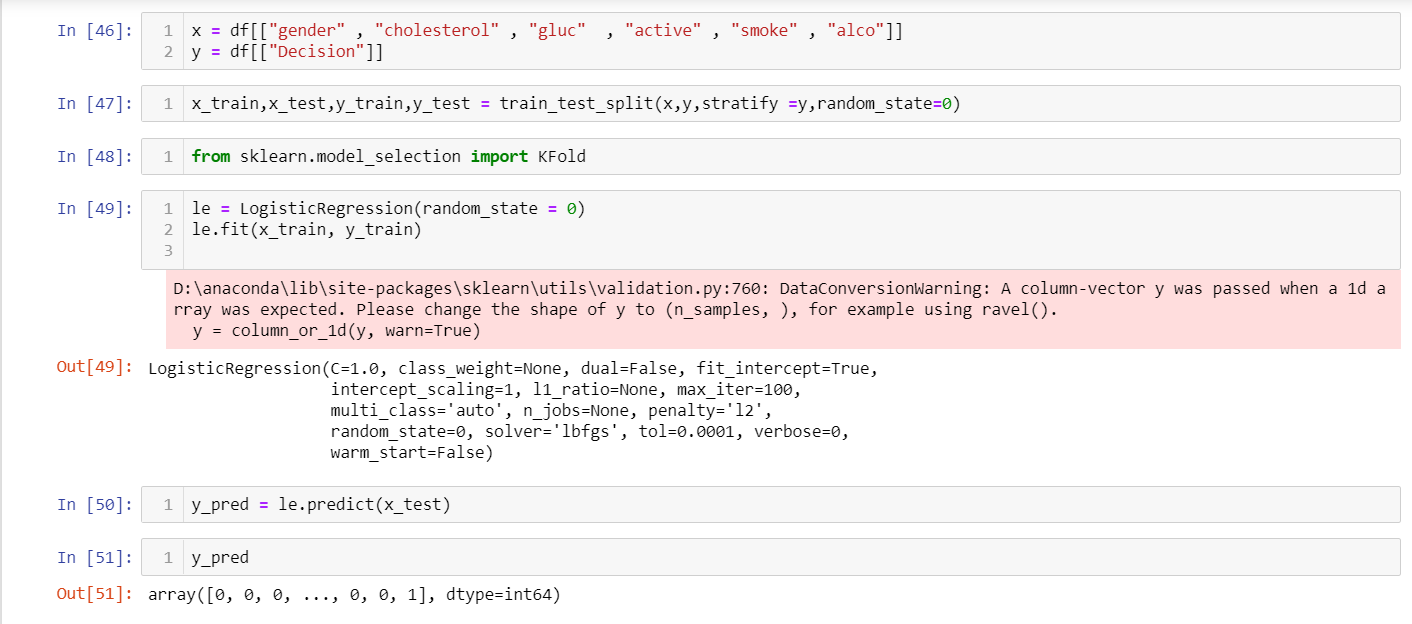




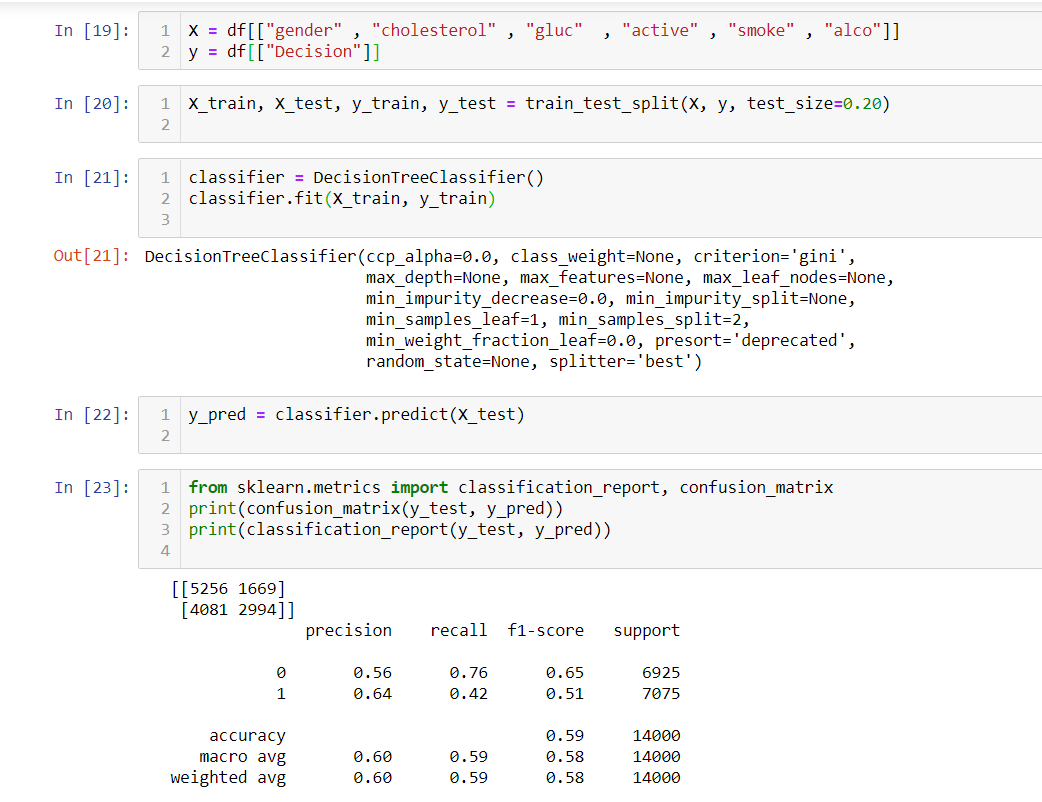


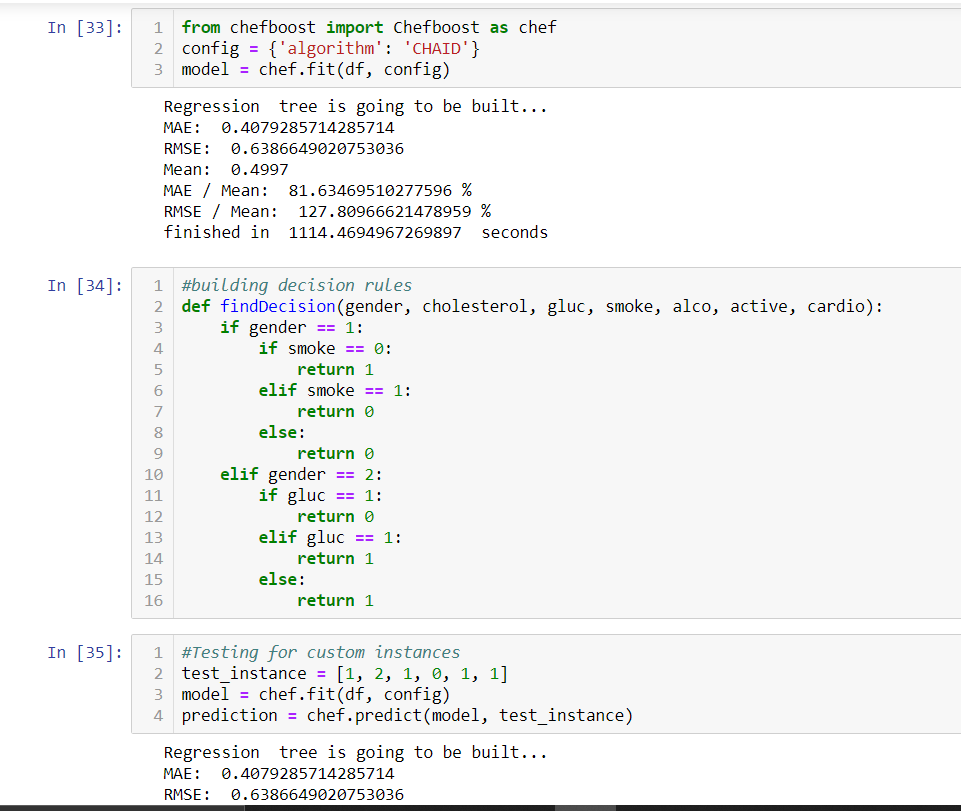
1. Model development (Model 1)

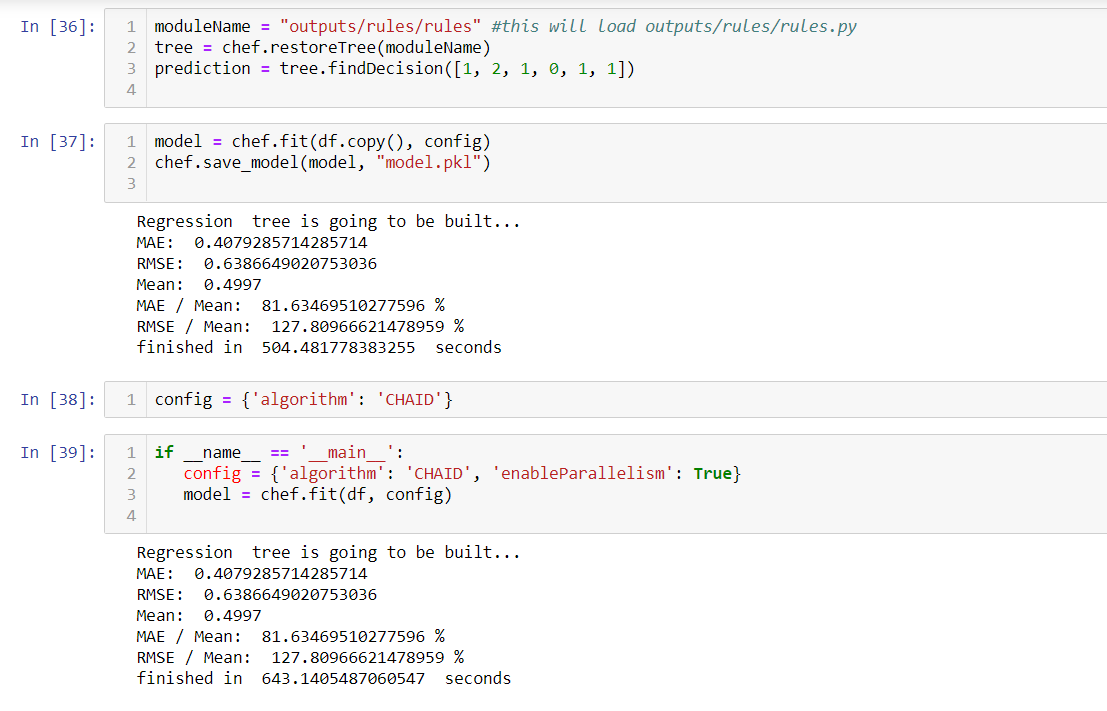


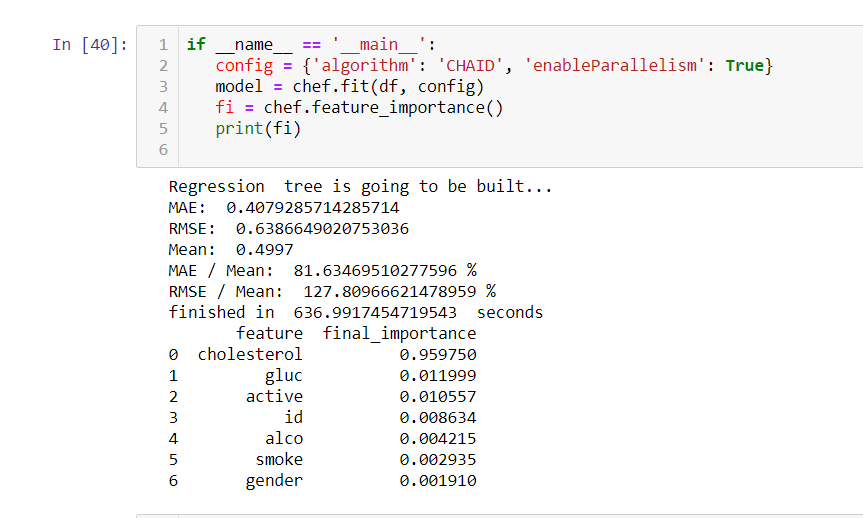


Model Development(Model 2)









1. Checking accuracy-

