

**IS 733: Data Mining**



**Group 7**

**Heart Disease Risk Prediction**



Submitted to

Dr. James Foulds

By:

Anil Umesh

Milind Pawar

Rohit Bhosale

Vishal Sarvagod

Sambhaw Sharma

M.S. Information Systems

Department of Information Systems

University of Maryland Baltimore County

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr no.** | **Section** | **Page number** |
| **1** | **Abstract** | **3** |
| **2** | **Background and related work** | **4** |
| **3** | **Introduction** | **4-5** |
| **4** | **Exploratory Data Analysis** | **5-10** |
| **5** | **Data Preprocessing** | **11** |
| **6** | **Resampling** | **11** |
| **7** | **Logistic Regression** | **11-12** |
| **8** | **Decision Tree** | **13** |
| **9** | **Support vector machine** | **13-14** |
| **10** | **Random Forests** | **14-15** |
| **11** | **Naive Bayes** | **15** |
| **12** | **Neural Network** | **15-16** |
| **13** | **Conclusion** | **16** |
| **14** | **Future Enhancements** | **17** |
| **15** | **References** | **17** |

**1.Abstract:**

Over the last few decades, heart-related disorders or cardiovascular diseases (CVDs) have been the primary cause of a large number of deaths worldwide and have emerged as the most life-threatening illness, not just in the USA, but worldwide. There is also a need for a reliable, specific and viable method to detect these diseases in time for proper care. There must be a fast and reliable screening technique in order to minimize the number of deaths from heart diseases. The Decision Tree is one of the most powerful techniques of data mining to date. Decision Tree, Naïve Byes, Support vector machine (SVM), Logistic regression, Random Forests, and Neural networks are the algorithms used in this project. Heart disease describes many clinical issues linked to the heart, which are vast in nature.

**2. Background and Related Works:**

Machine learning (ML) has been shown to be successful in helping to make choices and predictions from the vast volume of data generated by the healthcare industry, according to Senthil Kumar (Effective Heart Disease Prediction using Hybrid Machine Learning Techniques 2019). Techniques for ML Latest innovations in various fields of the Internet of Things are being included in (IoT). With ML methods, different studies only offer an insight into forecasting heart disease. The writers recommended a new approach aimed at discovering Important characteristics through the application of machine learning methods to enhance the precision of cardiovascular disease prediction. With numerous combinations of characteristics and many known classification techniques, the prediction model is implemented. Via the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

Early identification of heart diseases and constant screening will reduce the mortality rate, according to Ed-Daoudy (Real-time machine learning for early detection of heart disease using big data approach 2019). The rapid development of data from numerous outlets, such as wearable sensor devices used in health tracking, broadcasting and other Internet of Things, has produced an immense amount of data on an ongoing basis. The convergence of streaming big data analytics and machine learning is a revolutionary technology that can have a major effect on the field of healthcare, especially early heart disease detection. It can be more efficient and less costly with this technology. Based on Apache Spark, a real-time heart disease prediction method has been suggested. There are two major sub-components of the framework, namely streaming processing and data storage and visualization. The fifth one With Spark streaming, Spark MLlib uses and applies the data case classification model to forecast heart disease. The second uses Apache Cassandra to store the large amount of data that is generated.

Chala Beyene (Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques 2018), proposed the prediction and analysis of Heart Disease Usage Strategies for Data Mining. The main goal is to predict the incidence of heart failure with an automated early diagnosis of the disease within a short amount of time. The recommended solution is also important in the health sector for practitioners who do not have experience and skills. It utilizes different medical characteristics, including blood sugar and heart rate, age, sex, some of the Characteristics for deciding whether you have heart disease. To assess the output of data sets, WEKA program is used.

**3.Introduction:**

One of the human body's most significant organs is the heart. Heart problems are one of the most common heart diseases in the USA. The heart, through the body's circulatory system, pumps blood. Oxygen is distributed through the circulatory system of the body in the body component of the blood, and if the heart does not operate correctly, the entire human blood system will fail. So if the heart doesn't operate well, it may also lead to suicide, due to a severe health problem.

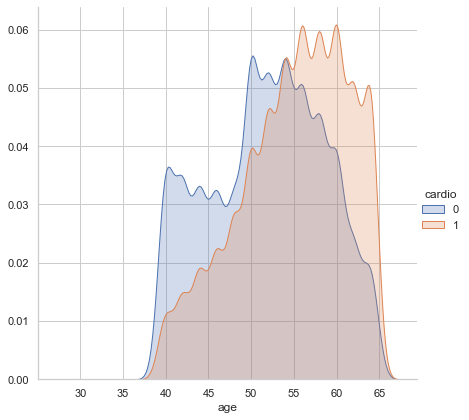
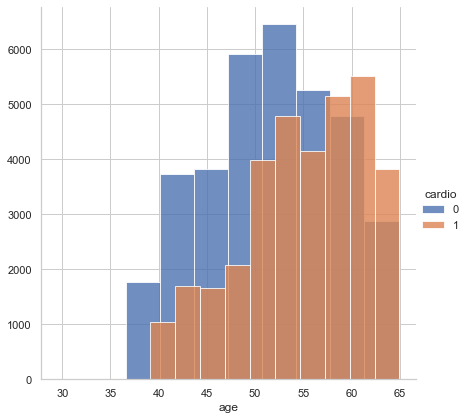
Medical associations compile data on numerous health-related conditions all over the world. Such data can be used to extract practical observations using different machine learning techniques. The data gathered is very huge, though and this data can also be very noisy. Using different machine learning methods, these datasets, which are too overwhelming for human minds to grasp, can be quickly explored. Thus, these algorithms have recently been very helpful in correctly forecasting the occurrence or absence of heart-related diseases.

**4. Exploratory Data Analysis:**

**Exploratory data** analysis is used to identify prevalent patterns, to detect anomalies, to provide context for further data analysis and to assist in checking data set assumptions. The Heart Diseasedataset's exploratory analysis of data is as follows:

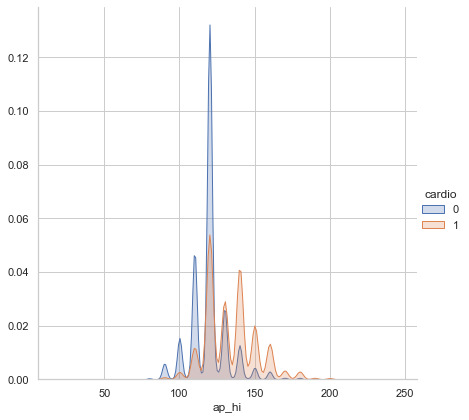
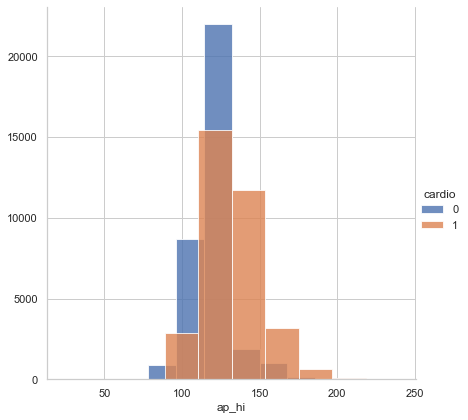
**Data Visualization**: The digital representation of information and data is the Visualization of data. Using graphic elements such as pie charts and data analysis applications for graphs offers an accessible way to see and interpret data trends, outliers, and patterns. Visualization of data is an important way of portraying the data output and the process done on it. For numerous types of visual components, Python libraries such as matplotlib and seaborn are used.

**Age –** Exploratory Data Analysis for effect of age on heart diseases suggests that as the age is increasing we are seeing more percentages of patients with heart disease similarly Kernel Density Estimate is used for visualizing the Probability Density of heart disease with respect to age and we can see steep rise in Probability Density of heart disease from age 35 to Age 55 and from there it somewhat Remains constant, plotting histogram shows similar kind of results that number of patients in sample size are rapidly increasing as age is increasing from 40 to 60 where in age group 40 less than 50% of patients had heart related problems compare to patients with healthy heart but age 60 group number of patients with heart diseases are almost equal to the number of patients with healthy heart and after 60 number of patients without heart disease decreased compared to the people with heart related problems, this analysis suggests that the probability of having heart problems increases with age and Age is an important predictor variable in this study.

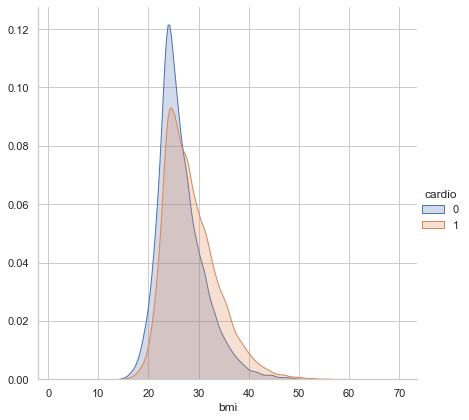
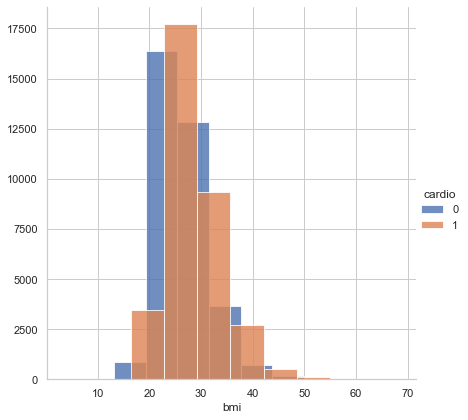
Kernel Density Estimate Plot Histogram

**Systolic Blood Pressure(ap\_hi) -** Exploratory Data Analysis for effect of Systolic Blood Pressure on heart diseases suggests that as the Systolic Blood Pressure is increasing from 120 we are seeing more percentages of patients with heart disease similarly Kernel Density Estimate is used for visualizing the Probability Density of heart disease with respect to Systolic Blood Pressure and we can see rise in Probability Density of patient heart disease where blood pressure in above 120 as we can see Area under Orange curve is more than the area under the blue curve where ap\_hi > 120and when ap\_hi < 120 Area under Orange curve is less than the area under the blue curve, plotting histogram shows similar kind of results that number of patients in sample size are rapidly increasing as Systolic Blood Pressure is increasing, this analysis suggests that the probability of having heart problems increases with Systolic Blood Pressure and Systolic Blood Pressure is an important predictor variable in this study.

 ****

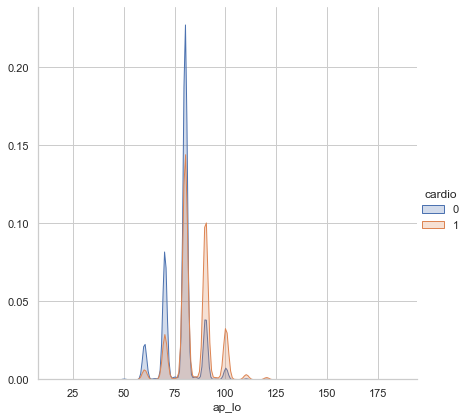
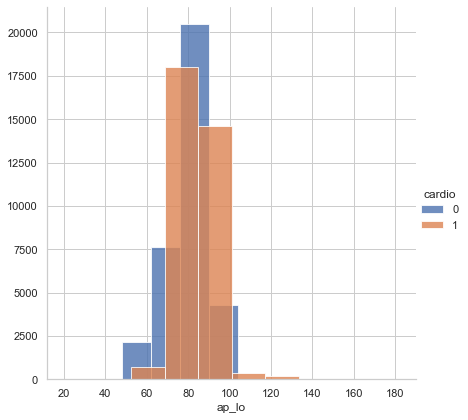
Kernel Density Estimate Plot Histogram

**Body Mass Index (BMI) -** Exploratory Data Analysis for effect of Body Mass Index on heart diseases suggests that as the Systolic Blood Pressure is increasing from 24 we are seeing more percentages of patients with heart disease similarly Kernel Density Estimate is used for visualizing the Probability Density of heart disease with respect to Body Mass Index and we can see rise in Probability Density of patient heart disease where Body Mass Index in above 25 as we can see Area under Orange curve is more than the area under the blue curve where BMI > 25and when BMI < 25 Area under Orange curve is less than the area under the blue curve, plotting histogram shows similar kind of results that number of patients in sample size are increasing as Body Mass Index is increasing, this analysis suggests that the probability of having heart problems increases with Body Mass Index and Body Mass Index is an important predictor variable in this study.

** **

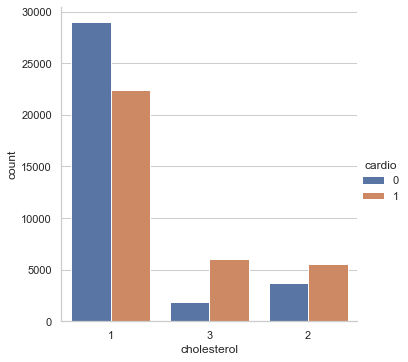
Kernel Density Estimate Plot Histogram

**Diastolic Blood Pressure(ap\_lo)** Exploratory Data Analysis for effect of Diastolic Blood Pressure on heart diseases suggests that as the Diastolic Blood Pressure is increasing from 80 we are seeing more percentages of patients with heart disease similarly Kernel Density Estimate is used for visualizing the Probability Density of heart disease with respect to Diastolic Blood Pressure and we can see rise in Probability Density of patient with heart disease where blood pressure in above 80 as we can see Area under Orange curve is more than the area under the blue curve where ap\_hi > 80and when ap\_hi < 80 Area under Orange curve is less than the area under the blue curve, plotting histogram shows similar kind of results that number of patients in sample size are increasing as Diastolic Blood Pressure is increasing, this analysis suggests that the probability of having heart problems increases with Diastolic Blood Pressure and Diastolic Blood Pressure is an important predictor variable in this study.

** **

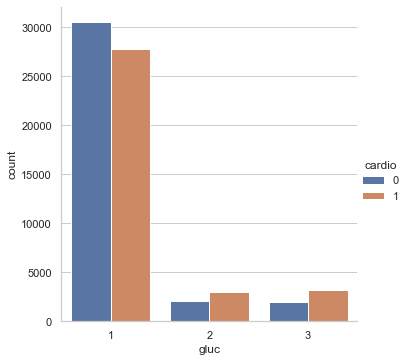
Kernel Density Estimate Plot Histogram

**Cholesterol –** Cholesterol is a categorical value is in this dataset and it has three categories 1 represents normal cholesterol, 2 above normal, 3 very high as we can see at normal cholesterol level there are less chance of having heart disease, similarly for category 2, Above normal, number of heart patients are higher than patients with healthy heart and for category 3 high cholesterol level number of heart patients are very high compared to patients with healthy heart hence we can say that Cholesterol is an important predictor variable to predict heart disease.

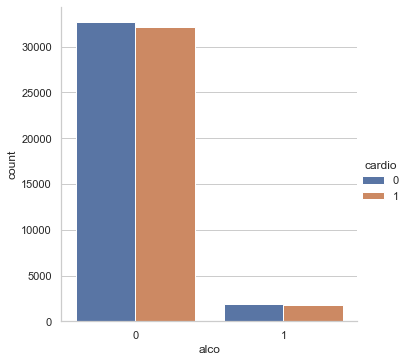
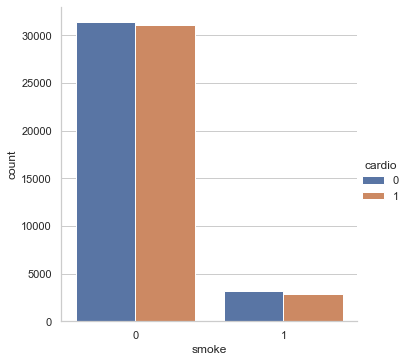


**Bar chart**

**Glucose –** Glucose is a categorical value is in this dataset and it has three categories 1 represents normal Glucose, 2 above normal, 3 very high as we can see at normal Glucose level there are less chance of having heart disease, similarly for category 2, Above normal, number of heart patients are higher than patients with healthy heart and for category 3 high Glucose level number of heart patients are very high compared to patients with healthy heart hence we can say that Glucose is an important predictor variable to predict heart disease.

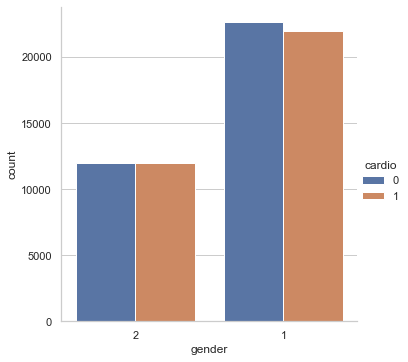
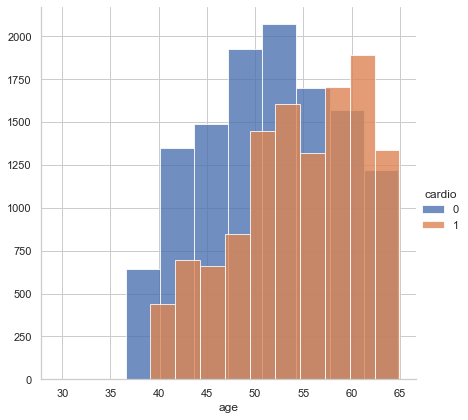
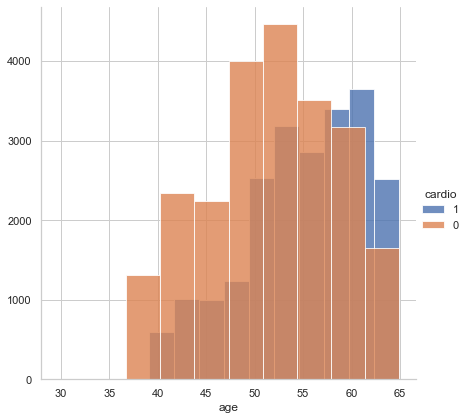
****

**Alcohol and Smoke -** These two categories represent the smoking and alcohol consumption habit of patients as the Exploratory Data Analysis suggests the in case of if person does not take alcohol number of patient with heart problems are almost equal to number of people with healthy heart and we can say same for the people who take alcohol and for categorical value smoke follows the similar pattern as alcohol, in case of if person does not smoke number of patient with heart problems are almost equal to number of people with healthy heart and we can say same for the people who smoke.

** **

Bar plot Alcohol Bar plot Smoke

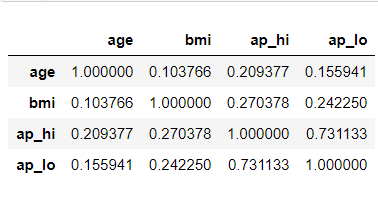
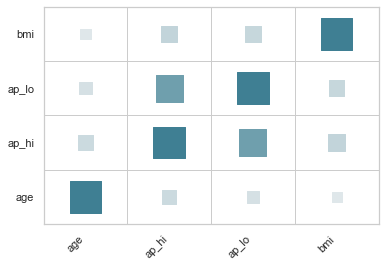
**Gender –** Gender is an categorical value in this study where gender = 1 means men and gender = 2 means women here as we can see in total cases vas gender graph number of cases in case of both man and women are almost equally split 50% have heart problems and 50% don’t but further data analysis suggests that men tend to develop heart disease earlier in life according to the first histogram, male vs Age where histogram female vs Age suggests that females develop heart disease at later ages.

****

male (Gender =1) vs Age female (Gender = 2) vs Age Total Case vs gender

**Correlation matrix –** is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables here as the table suggests there is strong positive correlation between Diastolic Blood Pressure and systolic Blood Pressure of 0.73 and weak positive correlation between BMI and Diastolic Blood Pressure and systolic Blood Pressure.

* Systolic blood pressure and Diastolic blood pressure
* Systolic blood pressure and Age
* Systolic blood pressure and BMI
* Diastolic Blood Pressure and BMI

**Correlation matrix Correlation matrix**

**5. Data Preprocessing:**

Data Preprocessing in data mining is an essential phase in the processing of raw data into a usable format that allows one to implement different methods of data mining that we can do in the future.

Data cleaning and Preprocessing is an important step in this study as the data set has mixed variables both continuous as well as categorical as well as contained few NA values and outliers hence, we cleaned data and the final dataset used had total 68544 rows. This study required conversion of numerical variables into categorical so we can apply various algorithm like Naïve Bayes or decision tree hence using blood pressure chart (Diastole vs. Systole:, 2020) converting Ap\_hi and Ap\_lo variables into categorical variable using binning, we also created new variable BMI using Height and weight variable, using Body Mass Index chart (Calculate your BMI, 2020) and binning, created various weight categories similar for Age Variable hence converting all continuous variable into Categorical variable.

**6. Resampling:**

Resampling is one of the key concepts or steps in machine learning. Resampling is done for the dataset under area of interest to obtain the performance of the various models involved. In this methodology there are various steps involved. They are:

1. Continuously drawing a sample from the training data.
2. Refitting the model of interest with each new sample.
3. Examining all the refitted models and then drawing appropriate conclusions.

There are two major resampling techniques:

1. Cross-Validation: generally used to estimate the error (model assessment) associated with a given learning model and / or to select the appropriate learning model (model selection).

2. The Bootstrap: most used to provide a measure of accuracy of a parameter estimate or of a given learning method.

In this project we have done a 5-fold cross validation and validation data-set approach. Performance on the validation set is a random variable depending on the data-split into training and validation sets. In the validation dataset approach, we have divided the dataset into two parts.50% of the data used for the training and the other 50% is used for testing of the trained models.

K-Fold Cross-Validation proceeds as follows:

1. Partition the entire dataset into K equal-sized components.

2. For k = 1, . . ., K, fit the model to the other K − 1 components and calculate the prediction error on the Kth component.

3. Combine the K estimates of the prediction error to obtain an average prediction error

**7. Machine Learning Algorithms**

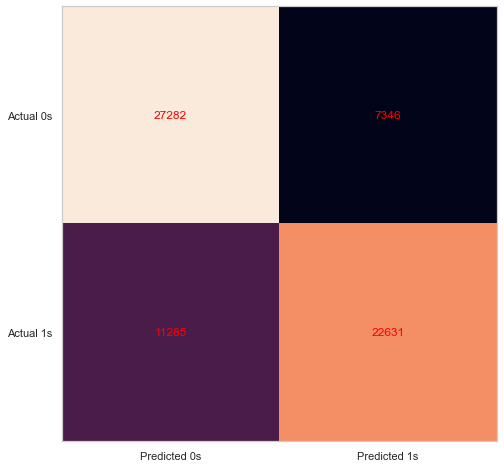
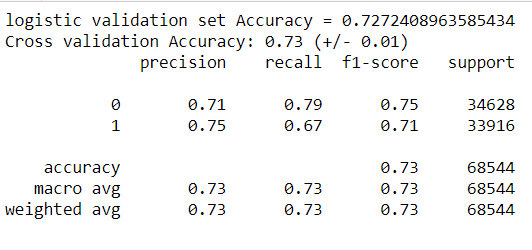
**Logistic Regression:**

Logistic regression is a classification algorithm for supervised learning that is used to estimate the likelihood of a target variable. The existence of the target or dependent variable is binary, indicating that only two classes are possible.

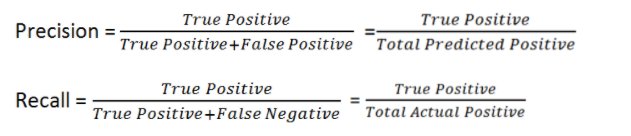
The response variable is conditional in nature, with this dataset coded 1 which stands for yes patients have heart disease or 0 which stands for patients who don’t have heart disease. Mathematically, as a function of X, a logistic regression model predicts P(Y=1). It is one of the easiest machine learning algorithms that can be used for different problems of classification. Hence Applying this algorithm in this data set to calculate accuracy and precision.

Logistic regression with validation set approach provided results given below we received accuracy of 72.72% with validation set approach and 73 % accuracy with 1% variance using

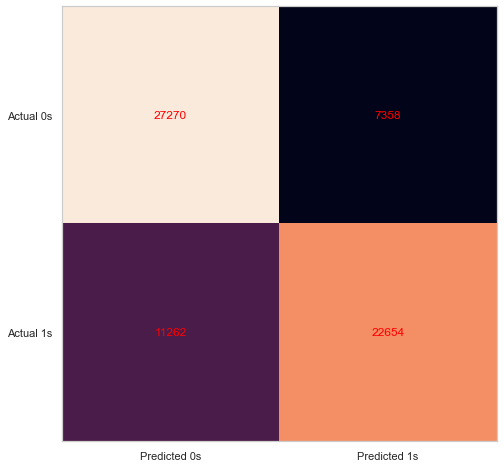
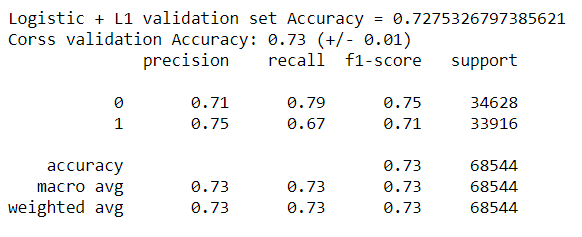
5 Fold cross validation approach.



As we can see, the logistic model is better at predicting true positive cases compared to true negative cases when we look at the precision value but recall value for class 1 is poor resulting in a lower f1 score.



Applying L1 regularization to see any improvement in f1 value and accuracy



As we can see there is a slight difference in accuracy of the validation set approach other than that there is no significant change in any other value such as precision, recall or F1- score.

**Decision Tree:**

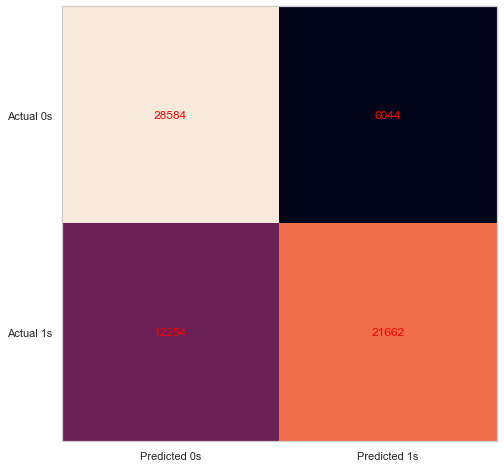
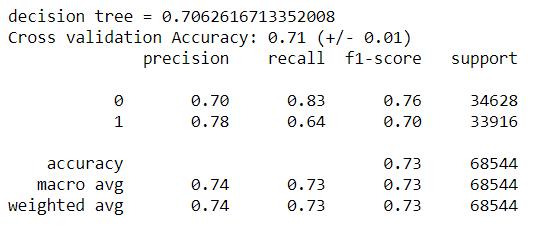
Decision tree analysis is a method for statistical modeling that can be used in many ways. Decision trees can be built by an algorithmic method that, depending on different parameters, can segment the dataset in different ways. The most efficient algorithms that come within the supervised algorithms group are Decisions Trees.

Diagram

Description automatically generated

Root of the decision tree with depth = 4

In the above decision tree, decision nodes are the problem, and leaves are the results.

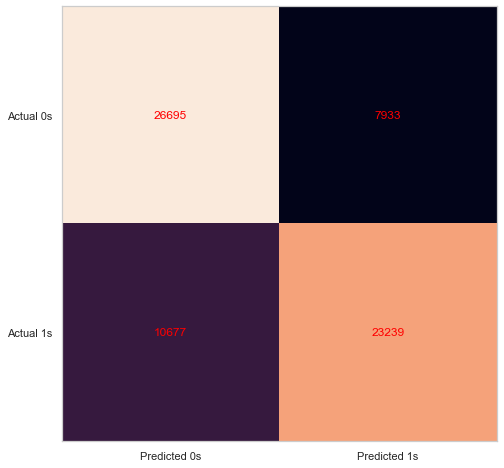
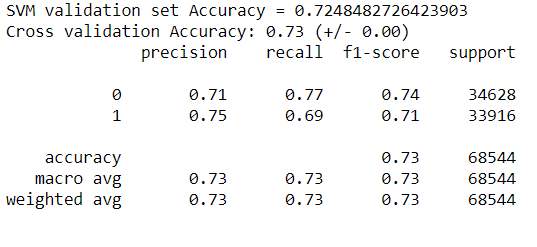


As we applied decision tree we received following results Decision tree with validation set approach provided results given below we received accuracy of 70.62% with validation set approach and 71% accuracy with 1% variance using 5 Fold cross validation approach in this case we received even better f1 score for class 0 but, overall accuracy has decreased but visualization decision tree can be helpful to find out importance of variable as we can see rooth of the tree is Ap\_hi means systolic blood pressure hence we can say systolic blood pressure is most important factor behind a heart disease.

**Support vector machine**

Support Vector Machine (SVM) is a fast and memory efficient Machine Learning algorithm which tries to find a linear hyperplane that could help us to separate the data into classes where each class would be on either side of the hyperplane present in a multidimensional space. It is mainly used for classification, regression and outlier detection problems.

It draws a decision boundary also known as hyperplane separate the extreme points in the dataset. The main goal while drawing the hyperplane is to separate the training data with the intention to maximize the between the hyperplane and the closest data points. Now applying Support vector machine classifier into out Dataset we received following results

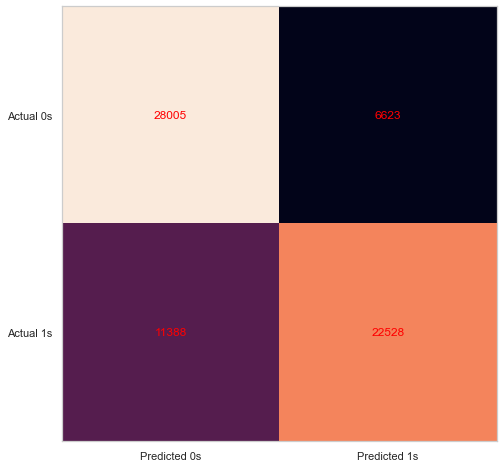
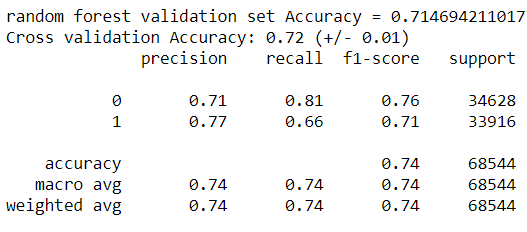


Support vector machine classifier with validation set approach provided results given below we received accuracy of 72.48% with validation set approach and 73% accuracy with less then 1% variance using 5 Fold cross validation approach in this case we received better f1 score for class 0 and 1 compare to decision tree, overall accuracy has also better compare to decision tree and when comparing all precision, Recall and F1 score as well as variance Support vector machine classifier provided better results compare to both Logistic model and decision tree.

**Random Forests:**

A supervised learning algorithm that is used for both classification and regression is Random Forest. But it is, still, primarily used for problems with grouping. As we know, a forest is made up of plants, and more trees make the forest better. Similarly, the random forest algorithm produces decision trees on data samples and then gets the estimate from each of them and chooses the best solution by voting. It is an ensemble strategy that is stronger than a single decision tree and by integrating the result, it decreases the over-fitting.

when applying random forest algorithm we received following results

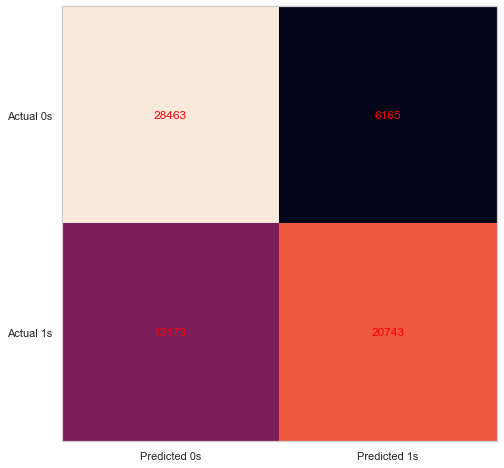
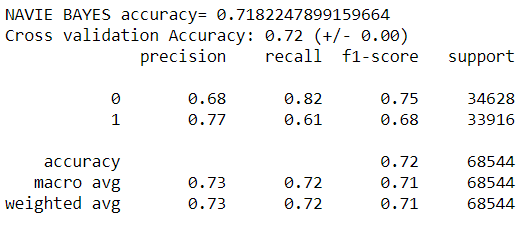


Random Forest classifier with validation set approach provided results with accuracy of 71.46% with validation set approach and 72% accuracy with 1% variance using 5 Fold cross validation as we can see the accuracy of Random Forest classifier on training data set is highest 74% so we can say that Random Forest classifier is overfitting, using this approach we received better f1 score for class 0 similar to decision tree, overall accuracy has also better compare to decision tree but when comparing all precision, Recall and F1 score as well as variance Support vector machine classifier provided better results compare to Random Forest classifier in this case.

**Naïve Bayes**

A supervised learning algorithm based on the Bayes theorem and used to solve classification problems is the Naïve Bayes algorithm. It is used primarily in text classification, which requires a dataset of high-dimensional training. It is one of the simplest and most powerful classification algorithms that helps construct fast machine learning models that can make predictions quickly. It is a probabilistic classifier, which implies, on the basis of an object's likelihood, it predicts.

The theorem of Bayes is also known as Bayes' Rule or the law of Bayes, which is used with prior knowledge to determine the likelihood of a hypothesis. The conditional probability depends on it. When we applied the Naïve Bayes classifier, we received the following results.

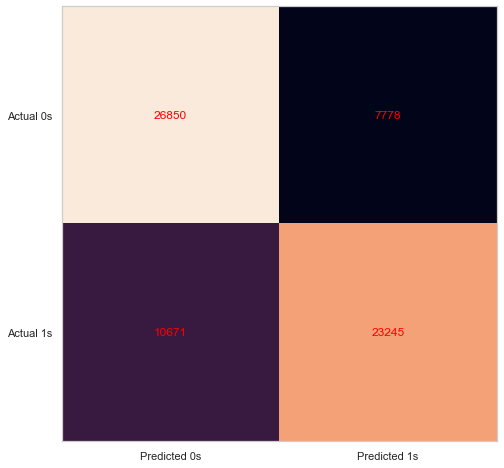
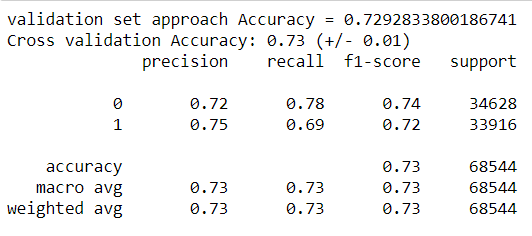


Naïve Bayes classifier with validation set approach provided results with accuracy of 71.82% with validation set approach and 72% accuracy with less 1% variance using 5 Fold cross validation as we can see the accuracy of Naïve Bayes classifier is less the SVM and logistic regression and when comparing all precision, Recall and F1 score as well as variance Support vector machine classifier provided better results compare to Naïve Bayes classifier in this case.

**Neural Network-**

It is a multi-layer neural network that we use to identify things, render forecasts, make predictions etc. Neural nets are a means of machine learning in which a computer learns by analyzing training examples to perform some task. The examples have typically been hand-labeled in advance. For example, an object recognition system could be fed thousands of labeled images of cars, houses, coffee cups, and so on and in the images, it would find visual patterns that correlate consistently with specific labels.

when applying the Neural Network MLP Classifier we received following results



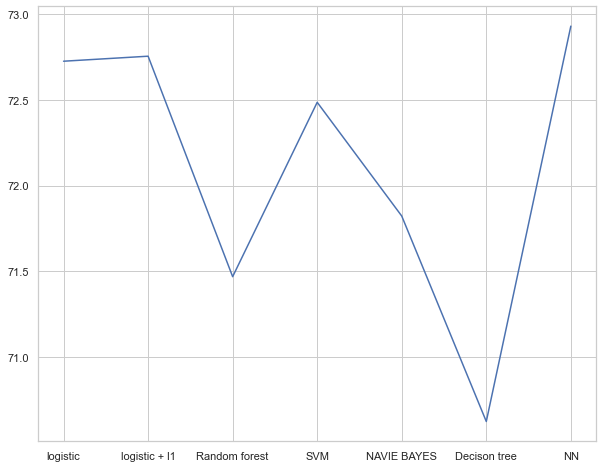
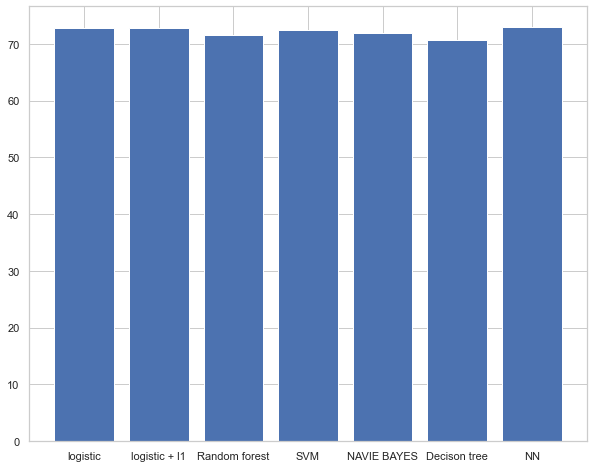
Neural Network MLP Classifier with validation set approach provided results with accuracy of 72.92% with validation set approach and 73% accuracy with less 1% variance using 5 Fold cross validation as we can see the accuracy of Neural Network MLP Classifier is higher than the SVM and logistic regression and when comparing all precision, Recall and F1 score as well as variance Neural Network MLP Classifier provides better results compare to any other machine learning algorithms in this case.

**Conclusion:**

As per our analysis after doing a 5-fold cross validation and validation dataset approach it has been found that the most important predictor variable is ‘ap\_hi’ and the next most important is ‘Age’ and the least significant is ‘gender’. Therefore, a person is with the higher chances of having cardiovascular problems if he/she is having following characteristics. They are:

1. Has a high blood pressure level.
2. Has a higher age
3. Has higher cholesterol level
4. Has obesity

Among the classification methods used, the Neural network has yielded in higher accuracy and better F1 score. Therefore, Neural networks could be used as an ideal method to predict cardiovascular diseases.

****

**Future Enhancements:**

1. If there were more health related variables in the dataset that could have helped us achieve better accuracy.
2. If the Dataset contains the human characteristics variable aging from 20-90 years then we can achieve even better accuracy of predicting the cardiovascular related health issues.
3. If the number of observations were more we could have drawn better predictions about the cardiovascular features of the Humans.
4. If the dataset contained many geographical area individuals data we can narrate the prediction for the cardiovascular characteristics of humans across all geographical locations.
5. if ECG and sleeping cycle habits data were involved in the dataset the prediction could have been more accurate.

**References**

[1] Mohan, Senthilkumar & Thirumalai, Chandra Segar & Srivastava, Gautam. (2019). Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2923707

[2] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5. doi: 10.1109/WITS.2019.8723839.

[3] Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018

[4]*Calculate your BMI*. (2020, 12 10). Retrieved from www.bassmedicalgroup.com: https://www.bassmedicalgroup.com/bariatric-surgery/body-mass-index-bmi-calculator

[5] Diastole vs. Systole*:*. (2020, 12 8). Retrieved from [www.webmd.com](http://www.webmd.com):

<https://www.webmd.com/hypertension-high-blood-pressure/guide/diastolic-and-systolic-blood-pressure-know-your-numbers#1>

[6]Resampling Methods

<https://uc-r.github.io/resampling_methods>

[7]Resampling Methods

<http://www.columbia.edu/~mh2078/MachineLearningORFE/ResamplingMethods.pdf>