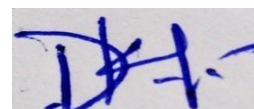


Project Summary

Batch details	PGPDSE-FT Hyderabad Apr21-Group-1
Team members	D.Pavan (00010KMHNZ), Doddapaneni Kundana (ODBVY7GU8Y), Y. Aravind Kumar Reddy (DEENCEAF60), G.Mahesh Kumar (HZXXIM48OK), E.Sai (BD5C2P1DP3)
Domain of Project	HEALTH CARE
Proposed project title	Prediction of cardiovascular disease
Group Number	5
Team Leader	Doddapaneni Kundana (ODBVY7GU8Y),
Mentor Name	ANIMESH TIWARI

Date: 14/10/2021



Signature of the Mentor

Signature of the Team Leader

1. INTRODUCTION:	3
1.1 OVERVIEW	3
1.2 CARDIOVASCULAR DISEASE RISK FACTORS	4
1.3 TOPIC SURVEY IN BRIEF	4
2. METHODOLOGY:	5
2.1 BUSINESS UNDERSTANDING	5
2.2 DATA UNDERSTANDING	5
2.3 DATA PREPARATION	5
2.4 MODELING	5
2.5 EVALUATION	6
2.6 DEPLOYMENT	6
3. DATASET AND DOMAIN:	6
3.1 DATASET SOURCE	6
3.2 DATASET OVERVIEW	6
3.3 DATA DICTIONARY	7-9
3.4 VARIABLE CATEGORIZATION	9
3.5 PROBLEM STATEMENT	10
3.6 TOOLS USED	10
4. DATA PREPARATION:	11
4.1 DATASET INFORMATION	11
4.2 DATA TYPE CONVERSIONS	11
4.3 DISTRIBUTION OF VARIABLES	11
4.4 FEATURE ENGINEERING	12
4.5 UNIVARIATE ANALYSIS	13-15
4.6 BIVARIATE ANALYSIS	15
4.7 MULTIVARIATE ANALYSIS	16
4.8 MISSING VALUES	17
4.9 HANDLING CATEGORICAL DATA AND PREPARING THE DATA	17
5. MODEL BUILDING PREREQUISITES:	18
5.1 SCALING THE DATA	18
5.2 TRAIN-TEST-SPLIT	18
6. MODELS FOR CLASSIFICATION:	19-24
7. IDENTIFICATION OF BEST MODEL	25
8. BUSINESS CASE	25
9. CONCLUSION	26
10. REFERENCES	27

1. INTRODUCTION:

1.1 Overview:

Cardiovascular disease (CVD) is one of the most dangerous or fearful diseases which is taking more than 17.3 million lives every year all over the world and numbers are expected to grow by 23.6 million deaths per year by 2030. It is said that heart diseases are more prevalent in men than women. Compared to different continents people in Asia are more likely to succumb to death due to CVD compared to others. According to WHO it has been estimated that 24% of deaths in India are due to CVD.

CVD is an umbrella term under which there are groups of disorders that are related to heart and blood vessels. Many of these disorders are conditions called atherosclerosis. In this condition walls of arteries build up fatty plaques which causes narrowed or blocked blood vessels that may lead up to different Cardiovascular diseases.

The prevalence of its associated risk factors has been found to exist increasingly in the population. With such a fast pace of increasing incidence, a number of epidemiological studies have been carried out in India to trace the prevalence of CVD over time. Some of them have forecasted the future incidence and prevalence of CVD in India.

With the increase in cases numbers of epidemiological studies have been carried out in India to trace the prevalence of CVD over time. It is the first among the top 5 causes of death in the Indian population. In 2000, there were an estimated 29.8 million people with CVD in India out of a total estimated population of 1.03 billion or a nearly 3% overall prevalence.

CVD comprises many different types of conditions. Some of these might develop at the same time or slowly develop into other conditions or diseases within the group.

Diseases and conditions that affect the heart include:

Angina: A type of chest pain that occurs due to decreased blood flow into the heart

Arrhythmia: or an irregular heartbeat or heart rhythm

Congenital heart disease: in which a problem with heart function or structure is present from birth

Coronary artery disease: slows blood flow to your heart muscle, so it doesn't get the oxygen it needs.

Heart attack: A sudden blockage to the heart's blood flow and oxygen supply

Heart failure: wherein the heart cannot contract or relax normally etc.,

1.2 CARDIOVASCULAR DISEASE RISK FACTORS:



Fig-1.2.1: Risk Factors of Cardiovascular Disease

With a complex collection of diseases under CVD, there are many risk factors involved in this. These factors individually and by interacting with one another will lead to different kinds of diseases. Researchers reported in the journal JAMA that the lifetime risk of CVD is more than 50% for both men and women. The risk factors that will lead to CVD are age, sex, family history, smoking, high blood pressure, poor diet, high blood cholesterol levels, diabetes, physical inactivity, stress, and poor dental health. There are some risk factors that come through hereditary like diabetes mellitus, high blood pressure which can cause CVD, and some daily lifestyle habits such as eating unhygienic food, lack of physical activity, and obesity.

In recent times, Heart Disease prediction is one of the most complicated tasks in the medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amounts of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance.

1.3 TOPIC SURVEY IN BRIEF

Problem understanding: Based on the given data we are trying to predict the patients with cardiovascular disease and the factors that are influencing this prediction. This might help in generating apps/devices that can easily give a warning to the person with high readings to seek medical help.

Current solution to the problem: Currently we have many risk factors that lead to disease but we intend to find the major risk factors. In the real world, there are devices that just give a few readings but not all that help to predict the CVD disease. It will also help the common man to know about the disease.

Proposed solution to the problem: Using ML classification techniques and algorithms and with effective data preparation we analyze and predict the factors that are high-risk factors to the disease and take necessary precautions.

2. METHODOLOGY

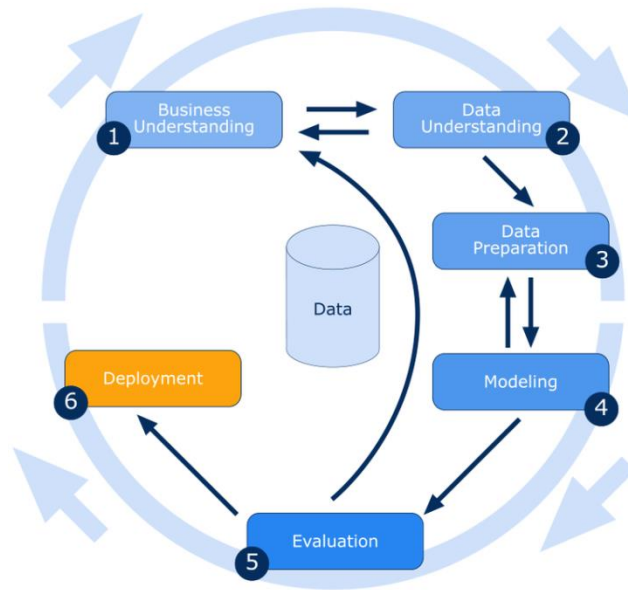


Fig-2: Methodology

2.1 BUSINESS UNDERSTANDING:

We have the data about the Cardiovascular Disease which includes the features are 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'cardio', 'BMI'. These are the features of the data of cardiovascular disease. The aim is to predict the disease and whether the person has the disease of cardiovascular disease

2.2 DATA UNDERSTANDING:

This stage describes from where the data is collected and also gives a detailed description of the data. Explanatory Analysis takes place to prepare the data for modeling so that we get accurate results

2.3 DATA PREPARATION:

In this stage, the dataset is polished up so that it will be easy for modeling and also give a better recall score. Explanatory analyses like Data Imputation, Feature Engineering, Data cleaning are performed on the raw data and the final dataset is constructed for modeling.

2.4 MODELING:

This phase involves finding the model that captures the solution to the business problem using available data. Here, we have tried multiple models and go forth and forth between data preparation and modeling to choose the correct model with the highest recall score.

2.5 EVALUATION:

We can check the model performance on data. Using that data we can compare the predicted and the actual values. We mainly look into recall scores and the model with the highest recall score is chosen.

2.5 DEPLOYMENT:

Deployment refers to moving an object to a place where some action can be performed on it. Here in this project, we have no need for this.

3. DATASET AND DOMAIN:

3.1 DATASET SOURCE: The dataset is extracted from Kaggle and is used to predict the presence or absence of Cardio-vascular disease.

3.2 DATASET OVERVIEW

Variable	Type	Definition
Age	integer	Age of the patient
height	integer	Height of a person in cm
weight	float	Weight of a patient in Kg
gender	object	gender of the patient (1-women,2-men)
ap_hi	integer	Systolic blood pressure
ap_lo	integer	Diastolic blood pressure
cholesterol	object	cholesterol levels (1: normal, 2: above normal, 3: well above normal)
gluc	object	glucose levels 1: normal, 2: above normal, 3: well above normal
smoke	object	Smoking: whether a patient smokes or not (binary) 1-smokes,2-doesn't smoke
alco	object	Alcohol intake: whether a patient consumes alcohol or not (binary) 1-alcoholic,2-non-alcoholic
active	object	Physical activity: whether the patient actively participates in physical activity or not (binary) 1-does physical activity,2-does not do physical activity
cardio	object	Presence or absence of cardiovascular disease: Target variable(binary) 1-Presence of the disease,2-Absence of the disease

3.3 DATA DICTIONARY:

Systolic blood pressure (Ap_hi): Systolic pressure – the pressure when your heart pushes blood out.
diastolic pressure – the pressure when your heart rests between beats.

- During a heartbeat, the heart is pushing blood out into the arteries. Doctors call this "systole," and that's why it's called systolic blood pressure. It's the pressure during a heartbeat and the highest pressure measured.

- When the reading is 120 mmHg or a little below while a person is sitting quietly at rest, the systolic blood pressure is considered normal.
- In each age group, the lowest risk for cardiovascular disease was in people with a systolic blood pressure of 90–114 mm Hg

Diastolic Blood Pressure(Ap_lo) :

Diastolic pressure – the pressure when your heart rests between beats.

The heart rests between beats so it can refill with blood. Doctors call this pause between beats "diastole." Your diastolic blood pressure is the measurement during this pause before the next heartbeat.

- Normal diastolic blood pressure during quiet rest is 80 mmHg or a little below.¹ If you have high blood pressure, the diastolic number is often higher even during quiet rest.
- Diastolic blood pressure of 60–74 mm Hg, with no evidence of a J-shaped increased risk at lower blood pressures.

Smoking :

Smoking has long been recognized as a major risk factor in cardiovascular disease, the risk being greater the more one smokes. As previously discussed, the carbon monoxide present in cigarette smoke binds to hemoglobin in the blood, making fewer molecules available for oxygen transport. In addition, coronary blood flow is reduced, forcing the heart to work harder to deliver oxygen to the body. Such strain places smokers at significantly greater risk for myocardial infarction, or heart attack, and stroke. There are, however, regional and sex differences in the incidence of smoking-related cardiovascular disease. In China, for example, where about 53 percent of adult males smoke (as opposed to about 2.4 percent of adult females)

Cholesterol :

Cholesterol is an organic molecule. It is a sterol, a type of lipid. Cholesterol is biosynthesized by all animal cells and is an essential structural component of animal cell membranes. It is a yellowish crystalline solid.

- Cholesterol values are measure in mg/dL (milligrams per deciliter)
- The total cholesterol/HDL ratio is an indicator of your potential for developing blockages in the arteries of your heart. A ratio greater than 4.5 is considered high risk for coronary heart disease. The ratio may be decreased by increasing your good (HDL) cholesterol and/or decreasing your bad (LDL) cholesterol.

Glucose :

Glucose is the main type of sugar in the blood and is the major source of energy for the body's cells. Glucose comes from the foods we eat or the body can make it from other substances. Glucose is carried to the cells through the bloodstream. Several hormones, including insulin, control glucose levels in the blood.

- Fasting blood glucose is an important determinant of CVD burden, with the considerable potential benefit of usual blood-glucose-lowering down to levels of at least 4.9 mmol/l.
- Patients affected by diabetes show an increased risk of cardiovascular disease (CVD) and mortality that reduces their life expectancy by 5–15 years (depending on the age at diagnosis).

Alcoholic :

An alcoholic drink is a drink that contains ethanol, a type of alcohol produced by fermentation of grains, fruits, or other sources of sugar that acts as a drug

- Having more than 1 alcoholic drink a day for women or more than 2 drinks a day for men may: Contribute to high blood pressure, which is a risk factor for coronary artery disease. Increase your risk of stroke. Directly damage heart muscle (alcoholic cardiomyopathy), which may weaken the heart, leading to heart failure.

Physical Activity :

Physical activity is any form of exercise or movement of the body that uses energy. Some of your daily life activities.

- Less active, fewer fit persons have a 30-50 percent greater risk of developing high blood pressure. Physical inactivity is a significant risk factor for CVD itself. It ranks similarly to cigarette smoking, high blood pressure, and elevated cholesterol.

Age :

- The aging and elderly populations are particularly susceptible to cardiovascular disease. Age is an independent risk factor for cardiovascular disease (CVD) in adults, but these risks are compounded by additional factors, including frailty, obesity, and diabetes.
- Adults aged 65 and older are more likely than younger people to suffer from cardiovascular disease, which is problems with the heart, blood vessels, or both. Aging can cause changes in the heart and blood vessels that may increase a person's risk of developing cardiovascular disease.

Gender :

Either of the two sexes (male and female), especially when considered with reference to social and cultural differences rather than biological ones.

- Women live longer than men and develop cardiovascular disease (CVD) at an older age. Metabolic syndrome represents a major risk factor for the development of CVD, and gender differences in this syndrome may contribute to gender differences in CVD.
- In recent years, metabolic syndrome has been more prevalent in men than in women. Prevalence is increasing and this increase has been steeper in women, particularly in young women, during the last decade. The contributions of the different components of metabolic syndrome differ between genders and in different countries.

Height :

The height of a person or thing is their size or length from the bottom to the top.

- Height has recently been linked to risks of both CVD and cancer, but its direction is the opposite: shorter people are at greater risk for CVD but lower risk of cancer.

Weight :

- Obesity has consistently been associated with an increased risk for metabolic diseases and cardiovascular disease.
- A 10 kg higher body weight is associated with a 3.0 mm Hg higher systolic and 2.3 mm higher diastolic blood pressure; this increase estimates a 12% increase in coronary heart disease and

a 24% increased risk for stroke.

BMI (Body mass index) :

Body mass index obtained from Height and Weight

- The increased risk of all-cause, CVD, and cancer mortality associated with an elevated BMI was significant at levels above 30 kg/m²; however, overweight individuals (BMI 25-29.9 kg/m²) also had an approximately 60% higher risk of CVD mortality.
- Body Mass Index is a simple calculation using a person's height and weight. The formula is BMI = kg/m² where kg is a person's weight in kilograms and m² is their height in meters squared.
- A BMI of 25.0 or more is overweight, while the healthy range is 18.5 to 24.9.

3.4 VARIABLE CATEGORIZATION

We have 12 variables in the data. In this we two types of variables

Numerical Variables: We have 5 numerical variables. We consider "age", "height", "weight", "ap_hi", "ap_low" as numerical variables from the data

Categorical Variables: We have 6 categorical variables. We consider "gender", "cholesterol", "glucose", "smoke", "alco", "active" as categorical variables from the data.

Target Variable:

We take "**cardio**" as the target variable. This variable will decide the presence or absence of cardiovascular disease (CVD).

3.5 PROBLEM STATEMENT:

The dataset contains risk factors of CVD and using these we find whether the patients have cardiovascular disease or not.

3.6 TOOLS USED:

Programming Language: Python

4. DATA PREPARATION:

Exploratory Data Analysis

4.1 DATA INFORMATION:

The data set consists of 70000 rows and 13 columns.

Additional column BMI has been added as it is very important to understand the problem.

In total, we are working with 14 columns, of which 7 are categorical and 6 are numerical.

The one that is left is the “ID” column which is not required for the analysis part. Therefore, it is dropped.

4.2 DATA TYPE CONVERSION:

7 categorical variables which are supposed to be in “Object” data type were in numerical data type as the data was replaced with 0s and 1’s(encoded).

Hence, we have converted them to “Object” data type.

We have not converted the target variable.

4.3 DISTRIBUTION OF VARIABLES:

Checking the presence of outliers of the continuous variables 'age','height','weight','ap_hi','ap_lo':

The boxplots show that the variables 'ap_hi' and 'ap_lo' are not normally distributed and the other variables are near normally distributed. Also, it can be easily seen that all the variables have outliers.

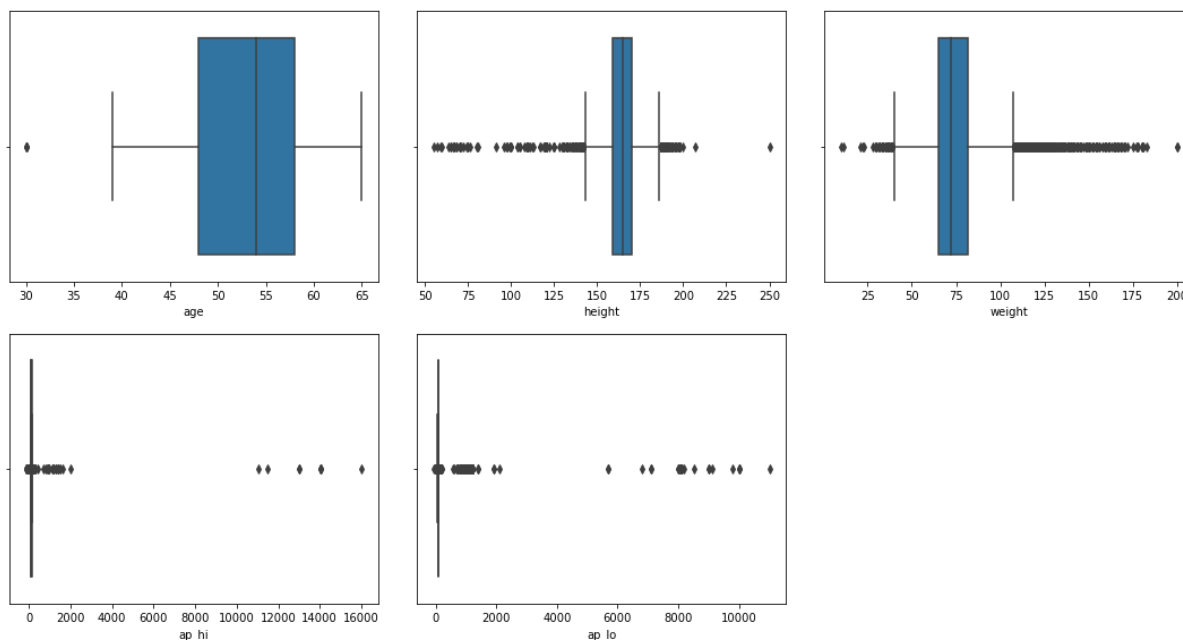


Fig-4.3.1: Distribution of Continuous variables of the data

Count plots of the categorical variables 'gender','cholesterol','gluc','smoke','alco','active':

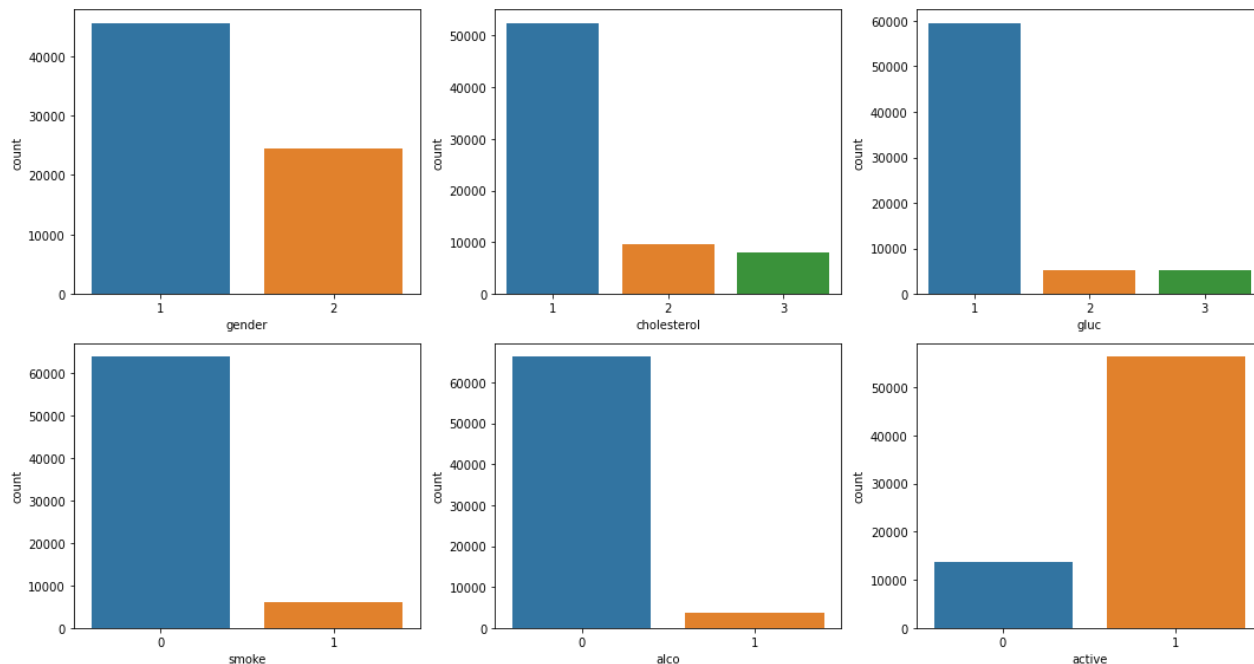


Fig-4.3.2: Distribution of Categorical variables of the data

1. The gender column has 65% of male and 35% of female patients.
2. 74% of patients have normal cholesterol levels and only 11% have high cholesterol levels.
3. 84% of patients have normal glucose levels and only 7% have high glucose levels
4. 91% of patients don't smoke and only 8% of persons smoke.
5. 94% of patients are not alcoholic and only 5% are alcoholic.
6. 80% of patients are doing physical activity and 19% are not.

Count plot of the target variable "Cardio":

The target variable cardio has an almost equal count.

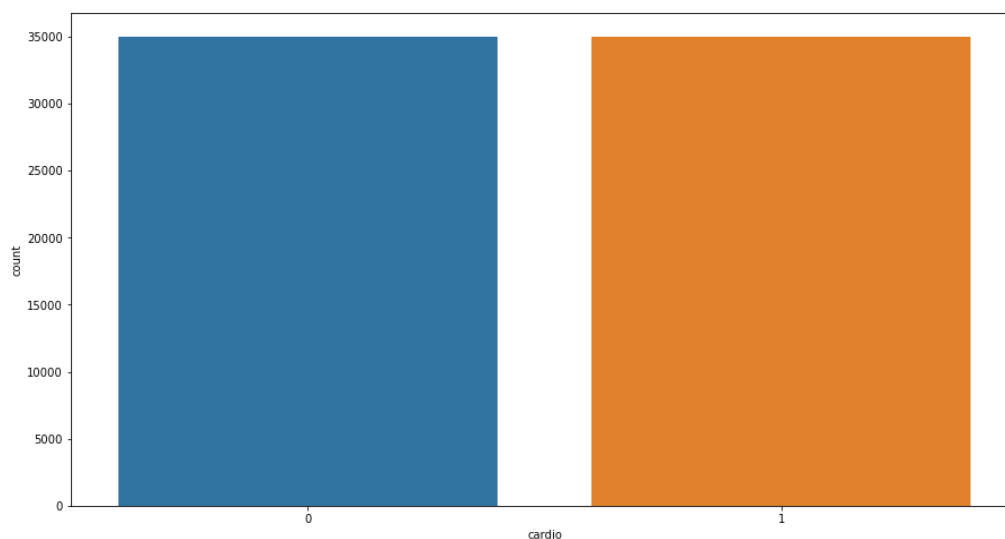


Fig-4.3.3: Distribution of Target Variable

4.4 FEATURE ENGINEERING:

New column "BMI"-Body Mass Index is created with height and weight variables and gender columns values are turned to 0 and 1.

4.5 UNIVARIATE ANALYSIS:

Age:

The skewness of the age is -0.30 which is close to 0. Hence, we can say that the data is normally distributed.

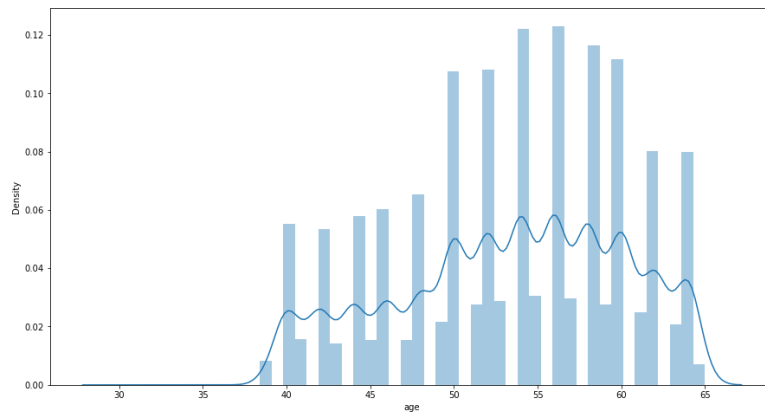


Fig-4.5.1: Distribution of Age

Height:

The skewness of the "height" variable before outlier treatment is -0.6.

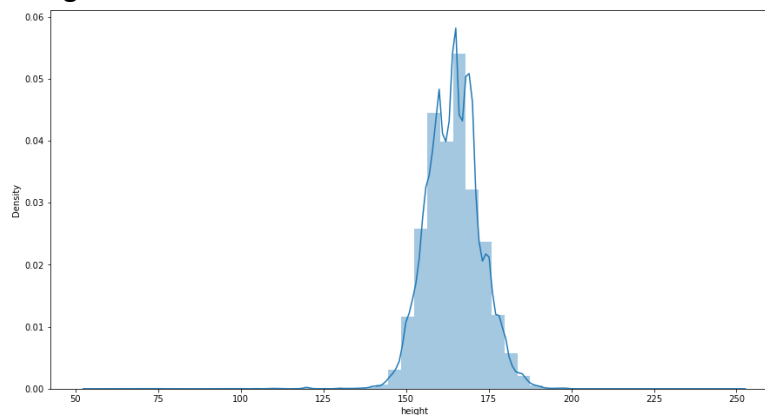


Fig-4.5.2: Distribution of Height before outlier Treatment

After outlier treatment:

Outliers above and below $1.5 \times \text{IQR}$ have been removed.

The height variable is slightly negatively skewed and has a lot of outliers. After removing the outliers the variable seems to be normal as 523 values have been removed.

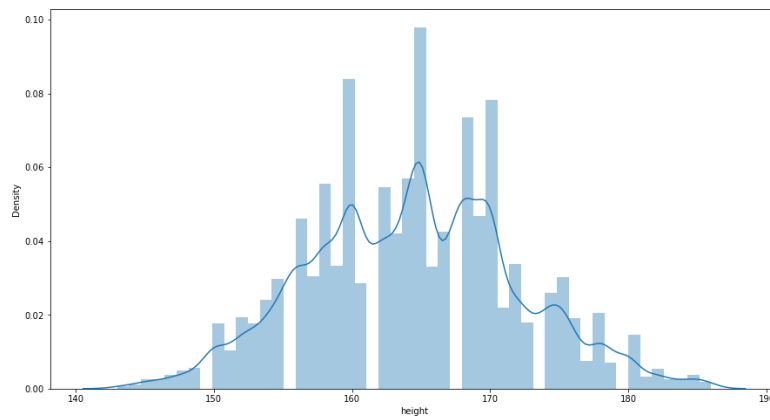


Fig-4.5.3: Distribution of Height after outlier Treatment

Weight:

The weight variable is positively skewed and has a lot of outliers. The upper limit of weight is important for the analysis so only the lower limit values are dropped and 45 outlier values have been removed.

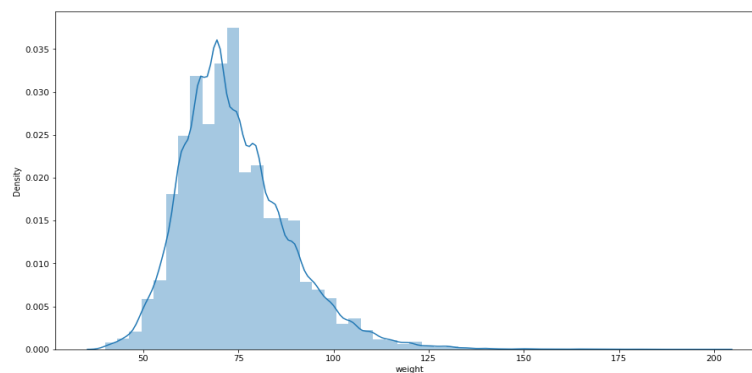


Fig-4.5.4: Distribution of Weight

ap_hi:

Before transformation and outlier treatment:

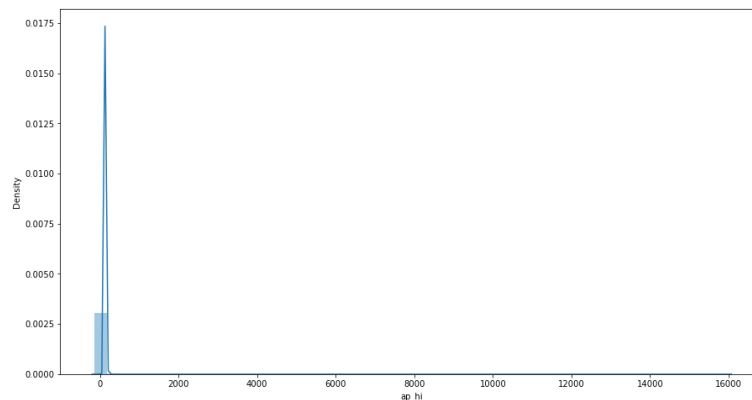


Fig-4.5.5: Distribution of ap_hi before outlier Treatment

After transformation and outlier treatment:

The systolic blood pressure variable is highly positively skewed and has a lot of outliers. After the outlier treatment, the skewness has drastically reduced and became near normal.

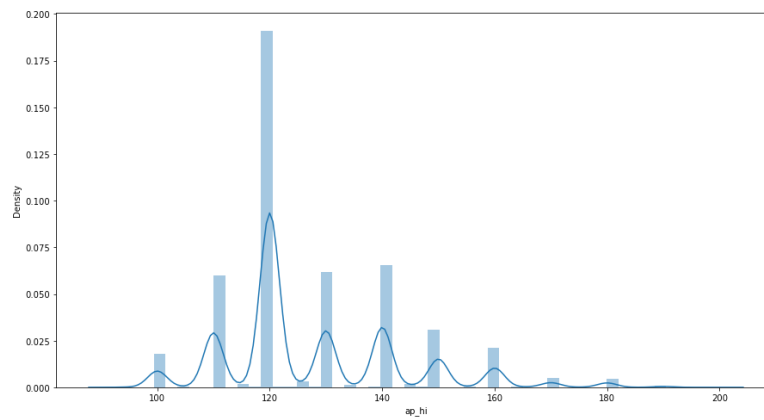


Fig-5.5.6: Distribution of ap_hi after outlier Treatment

ap_lo:

The diastolic blood pressure variable is highly positively skewed and has a lot of outliers. After the outlier treatment, the skewness has drastically reduced and became near normal.

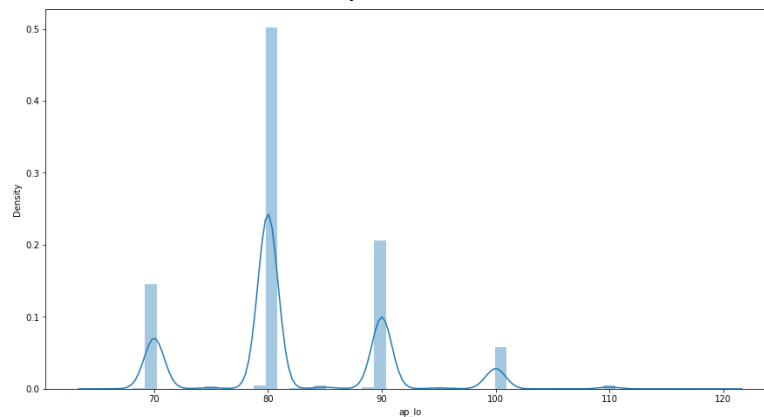


Fig-5.5.7: Distribution of ap_lo after outlier Treatment

BMI:

The BMI variable is positively skewed and has outliers. After the outlier treatment, the skewness has drastically reduced and became near normal.

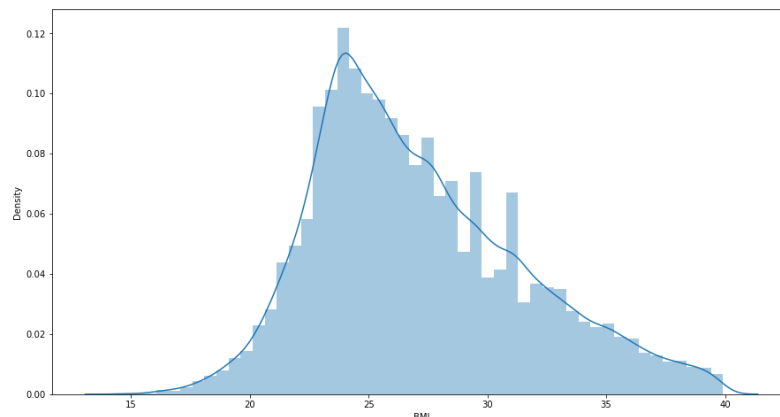


Fig-5.5.8: Distribution of BMI

4.6 Bivariate Analysis:

Plotting variables 'age', 'height', 'weight', 'ap_hi', 'ap_lo' with cardio:

The boxplot of the continuous variables explains their relation with the target variable 'Cardio'.

- The patients with the cardio disease are of the average age of 57.
- The height doesn't show any relationship with cardio as both have the same height.
- The patient with the disease has a higher weight.
- Patients with the disease have higher systolic blood pressure and Diastolic blood pressure.

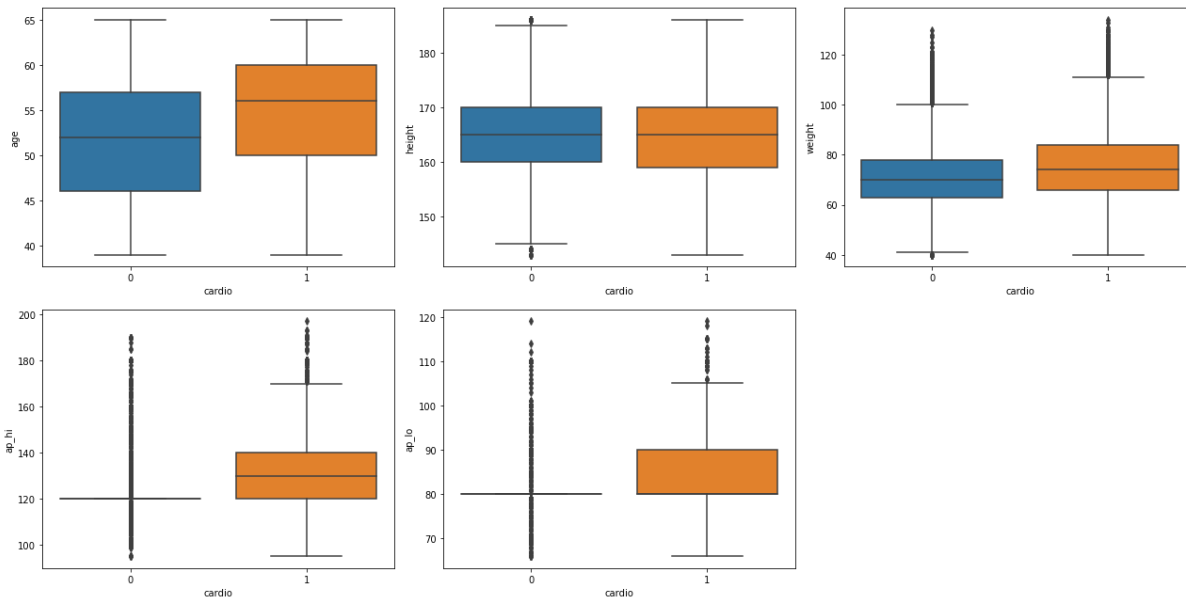


Fig-4.6.1: Relationship between continuous variables and target variable

Plotting variables 'gender','cholesterol','gluc','smoke','alco','active' with cardio:

The bar plot of the categorical variables explains their relationship with the target variable 'Cardio'.

- There is a complete balance in the gender of the people with and without the disease.
- Patients with different levels of cholesterol are prone to the disease.
- The count of people with the disease and smokers is very less.
- Very few people who are alcoholics are prone to the disease.
- Patients with cardio are less active comparatively.

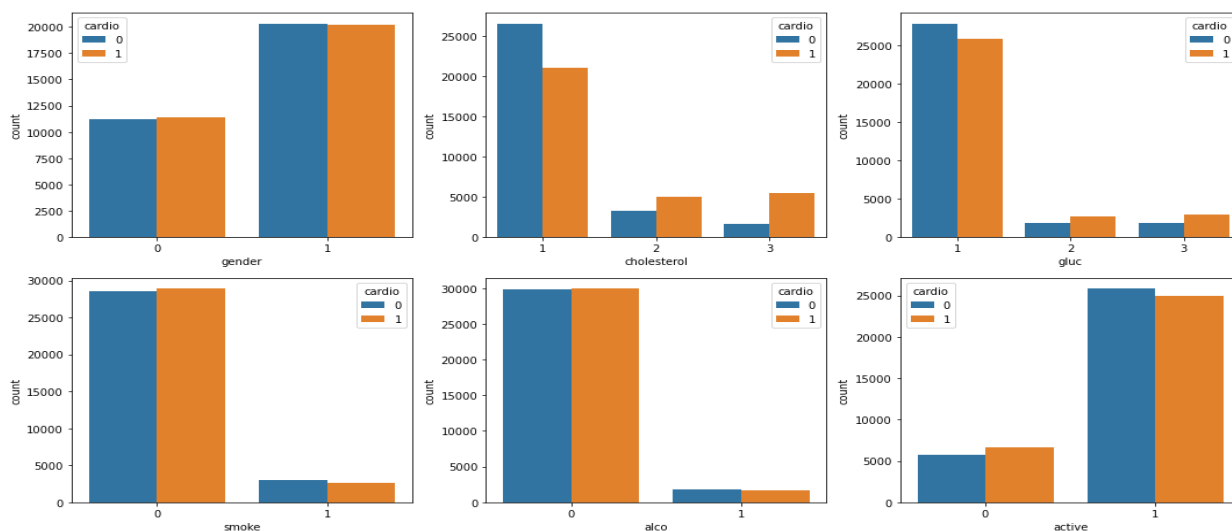


Fig-4.6.2: Relationship between categorical variables and the target variable

4.7 Multivariate analysis:

A pair plot has been used to understand the relationship among all the variables.

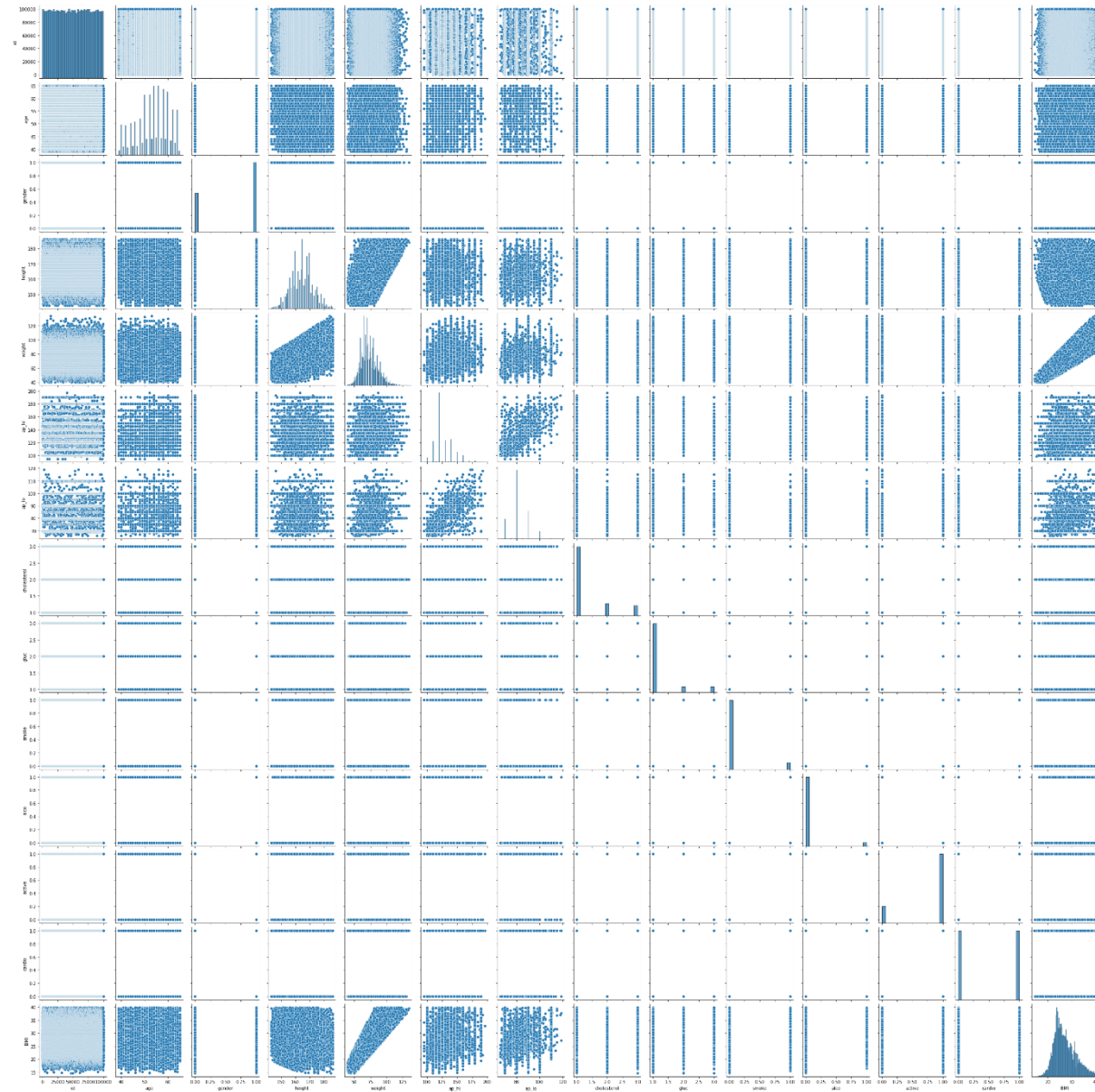


Fig-4.7: Multivariate analysis

4.8 Missing values:

The heat map shows no missing values.

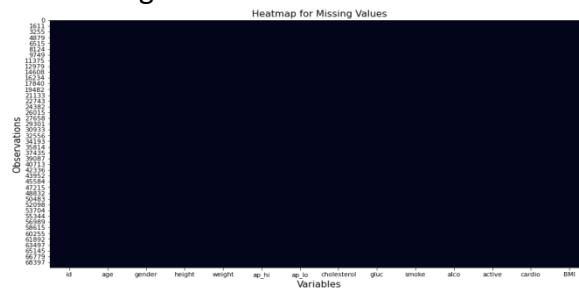


Fig-4.8: Heat map for missing values

4.9 HANDLING CATEGORICAL AND PREPARING THE DATA:

The dataset is split into two based on the data type- Numerical and Categorical. All the categorical variables are dummy encoded dropping the first column formed of the labels in a variable.

5. MODEL BUILDING PREREQUISITES:

5.1 SCALING THE DATA:

All the numerical variables separated from the data are scaled using StandardScaler to bring all the values of the variables to the same scale. After scaling the dummy encode categorical variables and scaled data are concatenated together as a single data frame and this is our final dataset.

5.2 TRAIN-TEST-SPLIT:

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

- **Train Dataset:** Used to fit the machine learning model.
- **Test Dataset:** Used to evaluate the fit machine learning model.

The final dataset after all the modifications is now split to train and test. For the model performance evaluation, we use a 70:30 split where 70% of the data is used for training the model and 30% of the data is used for testing the model.

6. MODELS FOR CLASSIFICATION:

LOGISTIC REGRESSION:

Logistic Regression is a Machine learning Algorithm, which works as predictive analysis. It is a process of modeling the probability of a discrete outcome given an input feature. Logistic Regression is one of the basic and popular algorithms to solve a classification type of problem in Machine Learning. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical value. It can be either {Yes , No} or {0 or 1},{ true or False},etc.

In Logistic regression, we fit an “S” shaped logistic function instead of fitting a regression line, which predicts two maximum values(0 or 1). Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. It can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

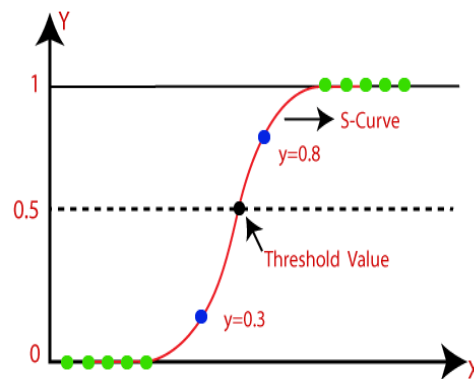


Fig-6.1: Logistic Regression sigmoid graph

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It is also known as the Logistic function.

Assumptions for Logistic Regression:

- (1) The dependent variable must be categorical in nature.
- (2) The Independent variable should not have Multi-Collinearity.

Steps in Logistic Regression: To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

- Data Pre-processing step
- Fitting Logistic Regression to the Training
- Predicting the Test Result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result

DECISION TREE ALGORITHM:

A Decision Tree Algorithm is a Machine Learning Algorithm. This algorithm can be used for the both Classification and Regression types of the Data. It is a tree-structured classified data and where internal nodes represent the features of a data set.

Branches represent the decision rules and the Leaf node represents the Outcome of Data.

In a Decision tree, there are two nodes 'Decision Node' and 'Leaf Node'. The decisions are taken on the performance of a feature of a given data. It starts with a root node and which expands on further branches and constructs a tree-like structure.it built from top to bottom. A decision tree is built based on the simple answer{Yes, No}, which further split the tree into sub-trees. A decision tree can contain categorical data as well as numerical data. The below diagram shows the general structure of a Decision Tree.

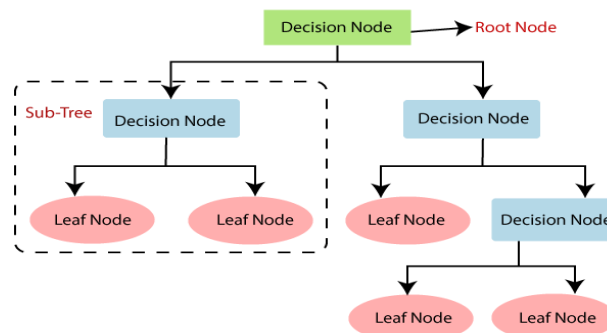


Fig-6.2 Decision Tree split

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from training data.

Types of Decision Trees:

Types of decision trees are based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** A decision Tree has a categorical target variable then it is called a “Categorical variable decision tree”.
2. **Continuous Variable Decision Tree:** The decision Tree has a continuous target variable then it is called” Continuous Variable Decision Tree”.

Assumptions for a Decision Tree:

- Firstly, the whole training set is considered as a Root.
- Feature values are preferred to be categorical. If the values are continuous then they are discredited prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

The decision of making strategic splits heavily affects a tree’s accuracy. The decision criteria are different for classification and regression trees. The algorithm selection is also based on the type of target variables. Let us look at some algorithms used in Decision Trees:

RANDOM FOREST ALGORITHM:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex Problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions and it predicts the final output. The Greater the number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest Algorithm:

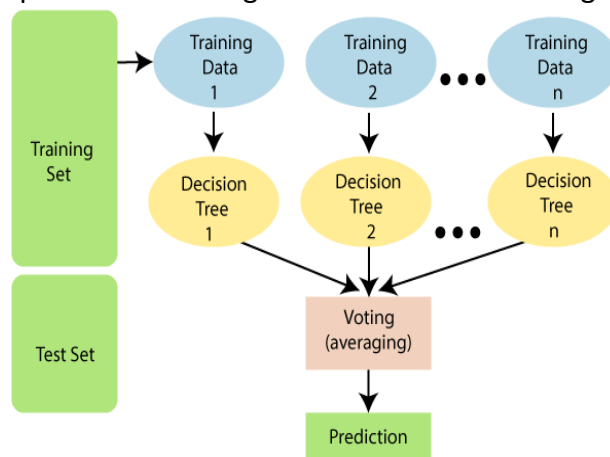


Fig-6.3 Random Forest

Assumptions for the Random Forest :

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

K-NEAREST NEIGHBOR ALGORITHM (KNN ALGORITHM) :

K-Nearest Neighbors(KNN) is one of the simplest algorithms used in Machine Learning for both regression and classification problems. KNN algorithms use data and classify new data points based on similarity measures .classification is done by a majority vote to its neighbors. The number of nearest neighbors to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure. 'K'in KNN is a parameter that refers to the number of nearest neighbors to include in the majority of the voting process.

Working of KNN Algorithm :

1. For implementing any algorithm, we need a dataset. So during the first step of KNN, we must load the training as well as test data.
2. Select the number K of the neighbors
3. Calculate the Euclidean distance of K number of neighbors
4. Take the K nearest neighbors as per the calculated Euclidean distance.
5. Among these k neighbors, count the number of the data points in each category.
6. Assign the new data points to that category for which the number of the neighbor is maximum.
7. And then, our model is Done.

The below figure shows the working of the KNN Algorithm:

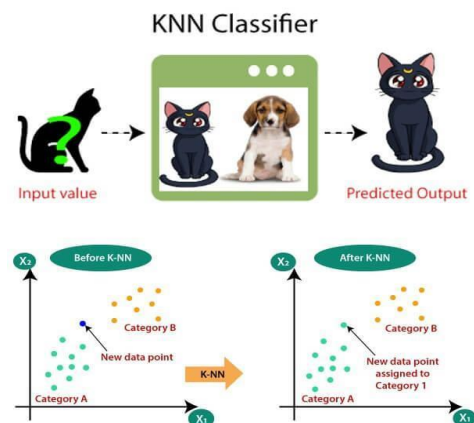


Fig-6.4 K-Nearest Neighbor Classifier

NAIVE BAYES ALGORITHM

The Naive Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and is used for classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which help in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of the Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Working of a Naive Bayes Algorithm:

Steps to implement the Naive Bayes Algorithm in Python,

- Data Pre-processing step
- Fitting Naive Bayes to the Training set
- Predicting the test result
- Test accuracy of the result(creation of Confusion matrix)
- Visualizing the test set result.

Types of Naive Bayes Algorithm:

There are 3 types of the Naive Bayes Algorithm Model. They are:

1. Gaussian
2. Multinomial
3. Bernoulli

BOOSTING ALGORITHMS:

Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model, and then trained sequentially- that is, each model tries to compensate for the weakness of its predecessor.

The term 'Boosting' refers to a family of algorithms that converts weak learners to strong learners.

There are three types of Boosting Algorithms which are as follows:

1. **AdaBoost algorithm**
2. **Gradient Boosting algorithm**
3. **XG Boost algorithm**

Adaptive Boosting Algorithm:

AdaBoost Algorithm is a Boosting Technique used as an Ensemble method in Machine Learning.

Adaboost helps you combine multiple "weak classifiers" into a single "strong classifier". AdaBoost Algorithm can be used for the both Classification and Regression Problem. AdaBoost can be used to boost the performance of any machine learning algorithm. It is used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost is decision trees with one level. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is

grown from previously grown learners.

Working process of Adaptive Boosting Algorithm:

- Importing the dataset
- Splitting the dataset into training and test samples
- Classifying the predictors and Target
- Initializing the Adaboost classifier and fitting the training data.
- Predicting the classes for the test set.
- Attaching the predictions to test set for comparing.

The below figure explains the model of Ada Boost algorithm:

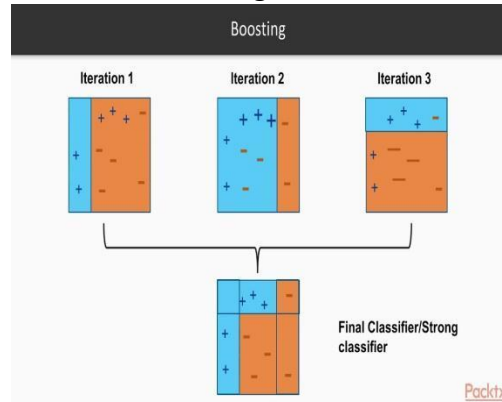


Fig-6.5 Adaptive Boosting Ensemble technique

GRADIENT BOOSTING IN ALGORITHM:

Gradient Boosting Algorithm is a Boosting Technique used as an Ensemble method in Machine Learning. Gradient boosting is a machine learning technique for regression, classification, and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient Boosting algorithm is one of the most powerful algorithms in the field of machine learning. We know that the errors in Machine Learning algorithms are broadly classified into two categories i.e., Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms, it is used to minimize the bias error of the model.

Gradient boosting algorithm can be used for predicting not only continuous target variable but also categorical target variable. When it is used as a regressor, the cost function is Mean Square Error (MSE). And when it is used as a classifier then the cost function is Log loss.

Working steps of the Gradient Boosting Algorithm:

- Calculate the average of the target Label.
- Calculate the residuals
- Construct a decision tree
- Predict the target label using all of the trees within the ensemble
- Compute the new residuals.

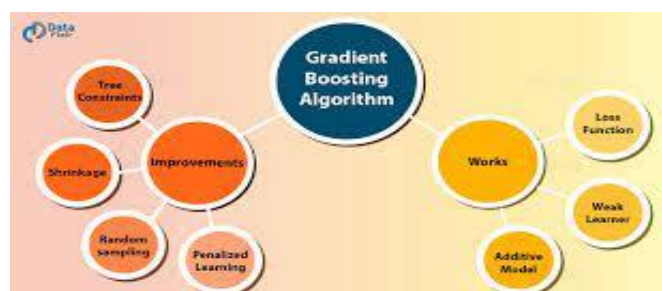


Fig-6.6 Gradient Boosting Algorithm- Ensemble technique

XG BOOSTING ALGORITHM:

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. We use the XGBoost for only two reasons :

- (1) Execution of speed
- (2) And Model Performance.

Gradient Boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

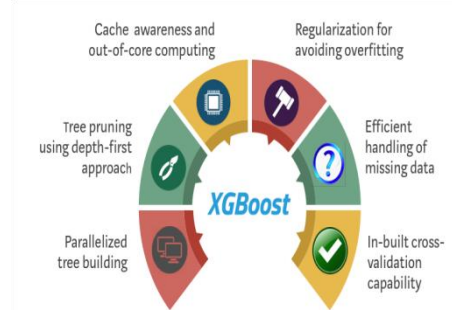


Fig-6.7 XGBoost Algorithm features

Working of XG Boosting Algorithm:

- Load the Libraries of an Xgboost
- Load the Dataset
- Data Cleaning and Feature Engineering.
- Tune and Run the model.
- These are the simple steps for a data problem to solve the boost Algorithm.

7. IDENTIFICATION OF BEST MODEL:

As per the dataset, we have performed these classification models. The aim was to find the best recall score at a threshold of 0.3. We have found that at 0.3 cut-off value, Logistic Regression is providing better results compared to the other classification models.

MODEL	Train Dataset				Test Dataset			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.596	0.907	0.72	0.65	0.59	0.91	0.71	0.67
Decision Tree	0.94	0.99	0.96	0.97	0.61	0.64	0.63	0.62
After tuning	0.63	0.87	0.73	0.68	0.62	0.84	0.72	0.67
Random Forest	0.88	0.99	0.94	0.93	0.61	0.85	0.71	0.65
After tuning	0.63	0.89	0.73	0.68	0.62	0.88	0.72	0.66
K-Nearest Neighbors	0.64	0.99	0.78	0.72	0.56	0.88	0.69	0.61
After tuning	0.62	0.90	0.73	0.68	0.60	0.88	0.71	0.65
Naïve Bayes	0.71	0.67	0.69	0.70	0.72	0.68	0.70	0.71
Adaptive Boosting	0.50	1.00	0.67	0.50	0.49	1.0	0.49	0.49
Gradient Boosting	0.74	0.94	0.83	0.80	0.62	0.84	0.72	0.67
XG Boosting	0.69	0.92	0.79	0.75	0.62	0.86	0.72	0.67
Support Vector Machine	0.77	0.67	0.72	0.73	0.76	0.68	0.72	0.73

7.1 IDENTIFICATION OF IMPORTANT FEATURES:

As Logistic Regression has given the best recall score it is identified as the best fit model to the data. The key features that are highly contributing to the prediction of data are Cholesterol_3, ap_hi(systolic blood pressure), Cholesterol_2, age and etc.,

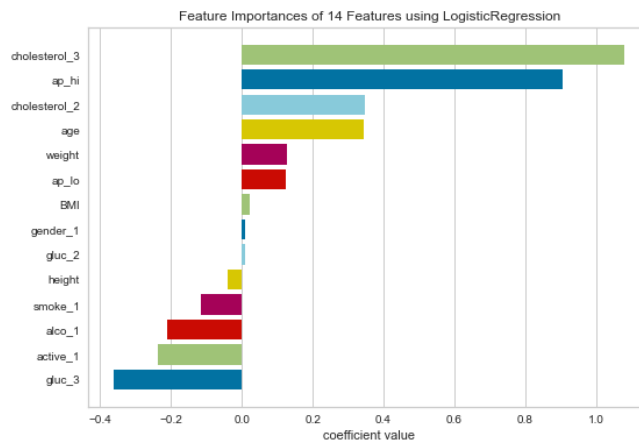


Fig-7.1.1 Feature Importance

8. BUSINESS CASE:

Based on Logistic Regression the confusion matrix is as below:

	Actual:0	
	Predicted:0	Predicted:1
Actual:1	2430	3887
	572	5732

Fig-8 Confusion Matrix

- Total data points are 63104.
- 5732 patients with the disease are actually predicted as having Cardiovascular disease
- 2430 patients not having the disease are predicted as not having the disease
- Model recall score is 0.91% that is 91% of patients who actually had CVD are diagnosed by the algorithm as having CVD.

9. CONCLUSION:

Key takeaway points

From the dataset:

- The patients at an average age of 53 are highly prone to get the disease.
- The average systolic and diastolic blood pressure is 127/82.
- There are 51:49 ratios for the target variable i.e. 51% are given to have the disease and 49% are not having the disease.

From the Model:

- Logistic Regression has given the highest Recall score of 0.91%
- 8160 correct predictions were made from the dataset.
- As predicted by the doctors, cholesterol, systolic blood pressure, age, weight, diastolic blood pressure is identified as the important feature for prediction of the disease.
- Glucose level-3 has a negative impact on the prediction.

Recommendations to the Industry:

- There have been a lot of outliers as in wrong readings of the patients. It is recommended to take exact body readings so that the disease is correctly predicted.
- Recommending the patients to get health check-ups frequently.
- Estimating the signs of the disease in the early stage will be helpful for the patients.
- Educating the people about the risk factors of CVD and also treatment to be done before reaching the hospital to avoid serious conditions.

References:

Dataset Reference links:

- <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

Reference links for articles/journals/ Dataset Resources:

- <https://reader.elsevier.com/reader/sd/pii/S235291481830217X?token=4AEED587D1649E4A585217292252D6D95913B42ED62090562D80D739FF68C2D8E9A7DF68161A5AE6CA0890F9005592E8>
- <https://link.springer.com/article/10.1023/A:1009715923555>
- <https://ieeexplore.ieee.org/document/8544333>
- <https://www.sciencedirect.com/science/article/pii/S1532046415001999>
- <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=>
- <https://www.bmj.com/content/353/bmj.i2416.long>
- <https://link.springer.com/article/10.1007/s10916-015-0290-7>
- [https://linhttps://journals.lww.com/psychosomaticmedicine/Abstract/2005/09000/Depressive Symptoms, Coronary Heart Disease, and.4.aspxk.springer.com/article/10.1186/s12911-019-0918-5](https://linhttps://journals.lww.com/psychosomaticmedicine/Abstract/2005/09000/Depressive_Symptoms,_Coronary_Heart_Disease,_and.4.aspxk.springer.com/article/10.1186/s12911-019-0918-5)
- [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>
- <https://www.sciencedirect.com/science/article/pii/S235291481830217X>
- [https://www.researchgate.net/publication/326733163 Prediction of Heart Disease Using Machine Learning Algorithms](https://www.researchgate.net/publication/326733163_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms)