# Abstract

The number of death due to heart attack is increasing day by day worldwide. This is the one of the major reasons which is affecting individual medical health all over the world. Health care industry is facing huge problems due to this and they are doing research to overcome this situation. Therefore, early diagnosis of disease is very important and intricate job in medicine. But, predicting the person might get heart attack in future is difficult task so, it should be done correctly and efficiently. The identification of heart disease is frequently accompanied by impulsive effects and some assumptions are made. This research study focusses to reduce the efforts and time required by doctor by automating the risk prediction with the help of machine learning and using the classification algorithms. Due to progress made in machine learning, there are various methods that can be successfully adopted to predict the heart disease of an individual. Before training a model using particular method, a dataset is downloaded from Kaggle which contain records of 330 patients with 14 different attributes. As it difficult to collect data of each and every patient, the only way possible is to download the data from repository present on google. The classification algorithm used to classify a patient is decision tree, random forest, Artificial Neural Network, Support Vector Machine, Boosting Algorithms and many more. All this algorithm is used to develop an application which helps in binary classification. The best algorithm is selected on the basis of comparison and the algorithm which gives good accuracy is used for prediction. Best accuracy was given by support vector machine which was 91% and Artificial Neural Network gave the accuracy of 90.89%. The classification report such as confusion matrix, ROC, AUC curve, precision, recall, F1 score was obtained for each and every implemented algorithm. The algorithm which gave the best accuracy for heart attack possibility prediction is 90.16%.

**Key words used in this research**

Machine learning techniques, Human Heart Attack, Adaboost algorithm, Features of dataset, Classification Report, Accuracy, Confusion Matrix

# Acknowledgement

Doing master and writing a dissertation is tedious job which cannot be done singlehandedly. So, I will gravelly thanks to God for his blessings throughout my dissertation work to complete it successfully.

It is genuine pleasure to express my deep sense of thanks and guidance to my professor, without his help this dissertation would not have been possible. His timely advice and scholarly advice have helped me to a very great extent to accomplish this task. I will also thank my professor for providing me the best topic, which would be helpful for my future. I am highly obliged in taking the opportunity to sincerely thanks to all members of my department. Last and foremost, I will thank my university to provide all the resources whenever it was necessary and needed.

I am extremely thankful to my all friends who provided me necessary technical suggestion

during my thesis work.

I am also thankful to my colleague who motivated me throughout the thesis work.

**Name of the Student**

**Table of Contents**

[Abstract 1](#_Toc80264439)

[Acknowledgement 2](#_Toc80264440)

[1. Introduction 5](#_Toc80264441)

[2. Literature Review 7](#_Toc80264442)

[2.1. Heart Anatomy: 7](#_Toc80264443)

[2.2. What is Heart Attack? 9](#_Toc80264444)

[2.3. Factors affecting heart attack: 11](#_Toc80264445)

[2.4. Symptoms of heart attack: 14](#_Toc80264445)

[2.3. Tips for reducing heart attack 15](#_Toc80264445)

[2.6. Importance of data science in heart attack possibility prediction: 16](#_Toc80264446)

[3. Problem Definition: 17](#_Toc80264447)

[4. Aims and objectives: 18](#_Toc80264448)

[5. Research Questions: 18](#_Toc80264449)

[6. Expected Outcomes: 18](#_Toc80264450)

[7. Everything about data: 19](#_Toc80264451)

[7.1. Data Source: 19](#_Toc80264452)

[7.2. Data Description: 19](#_Toc80264453)

[8. Methodology: 22](#_Toc80264454)

[8.1. Literature study: 22](#_Toc80264455)

[8.2 Data Analysis 23](#_Toc80264456)

[9. Result: 33](#_Toc80264457)

[10. Discussion and conclusion: 40](#_Toc80264458)

[11. References 43](#_Toc80264459)

[Figure 1: Anatomy of Heart 8](#_Toc77499849)

[Figure 2: Blocked cardiovascular artery 10](#_Toc77499850)

[Figure 3: Showing image of renaming the column for better understanding 22](#_Toc77499851)

[Figure 4: Showing image of plotting the age versus serum cholesterol level with target variable 23](#_Toc77499852)

[Figure 5: Showing image of plotting the age versus resting blood sugar with target variable 24](#_Toc77499853)

[Figure 6: Showing image of confusion matrix 30](#_Toc77499854)

[Figure 7: Showing image of confusion matrix of Adaboost implemented on test data 38](#_Toc77499855)

[Figure 8: Showing image of AUC curve of Adaboost implemented on test data 39](#_Toc77499856)

# 1. Introduction

The heart plays a vital role to pump blood to all parts of the body. However, malfunction in heart may lead to different heart disease or hear attack. Heart attack occurs mainly due to two reasons, first when the tissues or cell of heart is not functioning well or are destroyed due to factors such as smoking, drugs, age, stress and many more. Secondly, when there is blockage in coronary arteries due to high cholesterol in individual’s body or due to inappropriate diet in daily life. The risk of heart attack may increase due to unhealthy diet or oily food, smoking, high blood pressure, high cholesterol level, no exercise etc. The traditional approach was not found useful and was very complex. Due to unavailability of expert doctor and medical diagnosing tools specifically in undeveloped countries, diagnose and prevent heart attack is very complex. However, early prediction of heart attack may prevent from more damage. Heart attack is increasing in both economically developed and undeveloped countries. According to the data, the 30% global deaths occur mainly due to heart attack.

Conventional methods to diagnose takes more time such as angiography or ECG which may prove fatal and is not that useful for early prediction. These methods such as angiography is very costly and can be done by only cardiologist. However, the advancement in medical field in past decades enables in early identification of highly risk factors which may contribute to heart attack. But this conventional method is not efficient due to human mistakes. To overcome the disadvantage of conventional methods for the prediction of heart attack, researchers attempted to develop smart healthcare system using machine learning techniques. Machine learning is a part of artificial intelligence, in which machine is trained on the data that is collected from patient medical history. The training is done to predict the test data or more precisely to predict the data which might be available in future. The main aim of machine learning algorithm is to find a pattern in data and make prediction. The development of this technique has led to decrease the number of death due to heart attack. The research using the same technique is done by many researchers and published in different articles; by the help of the same the dissertation work is done and improvement is done in some area to get good score and efficiency.

Hassan Shaban Hassan et al [5] applied different data mining techniques to predict the person may suffer with heart attack in future. The attributes that he took under consideration such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart attack. The data which gave pattern having value greater than a threshold were chosen for the heart prediction was chosen for further process.

Dr. V. Reenvisage et al [10] proposed artificial neural network method to predict the person is affected by heart disease. These ensemble-based methods lead to creation of new models which was combination of the predicted values from multiple predecessor models. They considered 3 independent neural networks models were used to construct the ensemble neural network model. But when the number of neural networks were increased there was no improvement seen in the model accuracy.

For the classification of the human emotions that data will be collected through the use of signals that are collected using techniques like and based on the data the interpretation of heart attack will be done. Firstly, in this research the data will be collected and then the data will be processed so that any inconsistency in the data can be removed like reduction in components, handling missing data etc. and then the final dataset will be obtained. This dataset will then be analysed using machine learning techniques. The results that will be obtained will help in the classification of heart attack data. This classification system can be applied to the machine learning in the future by which the expert will be able to analysed and classify the heart attack may occur or not and take necessary decision. The analysis of data is necessary for the future advancement because in the future there will be more applications in the field of health care industry. and with the wide range of applications the automation that are used currently are not autonomous and needs the assistance of human beings when there is interaction of machine and human beings. So, there is a need that the technique that can understand the situation of heart attack so that the interactions can be more flexible and the machine can be more independent in taking decisions in the real time conditions and this issue will be focused in this research. In this research paper, the data is downloaded from the Kaggle which contain the record of 330 patients and 14 attributes were taken into consideration. In this there are two types of people, the one those who can get heart attack and other one who have less chance of heart attack. The classification of this is done using the machine learning classification algorithm which are Naïve Bayes, Support Vector Machine (SVM), Decision tree, Random Forest. Using this algorithm, the accuracy, confusion matrix, ROC, AUC is obtained for each and every algorithm. The algorithm which gives best result is selected for research. Before the selection of algorithm, the data is analysed and pre-processed and feature engineering technique has been applied if necessary. To improve the accuracy of model, the boosting techniques such as XG boost is used but due to this also the accuracy was not able to increased. There was only a marginal increase in the accuracy as compared to another algorithm. For improvement Artificial Neural Network (ANN) was used which uses the concept of backward propagation, which lead to improvement of accuracy by 5 percent. Therefore, the best model which was obtained was using the artificial neural network and the support vector machine. This research detail explanation and how the outcome was is explained further

# 2. Literature Review

## 2.1. Heart Anatomy:

Heart is muscular organ and has conical form which is enclosed in pericardium. It weighs about 300 grams. It is located above and between sternums. The unique features of heart are that cells or tissues of heart does not multiply but only grows in size. So, the number of cells of heart at is same throughout the lifespan of human beings. The heart pumps blood using arteries and veins which is known as cardiovascular system. The heart is divided into two parts inter atrial and inter ventricle septa. It is further divided into four chambers which are as follows:

* Right atrium

It receives blood from veins and it pumps to the below ventricle i.e., right ventricle.

* Right ventricle

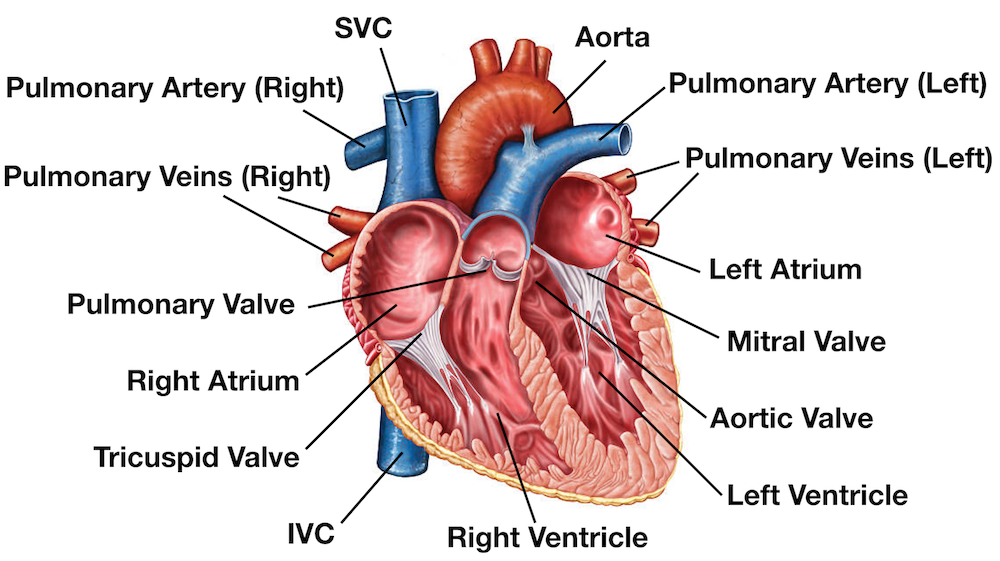
It receives blood from the right atrium and it pumps to the lungs where the blood is oxygenated.

* Left atrium

The blood from lungs which is oxygenated is received by left atrium, so it contains oxygenated blood.

* Left ventricle

It receives blood from the left atrium and it pumps blood to all parts of the body using arteries, so this also oxygenated blood.



[Figure 1: Anatomy of Heart. Image downloaded from https://www.ezmedlearning.com/blog/heart-anatomy-labeled-diagram]

So, the left part of heart contains the deoxygenated blood which is pumped to lungs for oxygenation which is under low pressure whereas right part contains oxygenated blood which is pumped to all parts of body. There is web of nerve network surrounding the heart which helps in relaxation and contraction of heart.

There are four valves in heart which prevents backward flow of blood in heart. The valve which separates right atrium and right ventricle is known as tricuspid valve whereas the separation between left atrium and ventricle is known as mitral value. It is also known as bicuspid valve. The valve which is located in outflow tract of ventricles is known as aortic valve for left side whereas for the right side it is known as pulmonary valve.

* Tricuspid valve

It helps in regulation of deoxygenated blood between right atrium and right ventricle.

* Mitral valve / Bicuspid valve

It helps in regulation of oxygenated blood between left atrium and left ventricle.

* Pulmonary valve

It carries blood from right ventricle to lungs for oxygenation where blood is mixed with oxygen.

* Aortic valve

The oxygenated blood is then pass to aorta, the largest artery of body. The opening of this aorta is aortic valve.

The heart is surrounded by sac called pericardium which is further divided into three parts which are as follows:

* Epicardium

It is the outermost layer of heart which is formed by visceral layer of the serous pericardium.

* Myocardium

It is the most important layer of heart which helps in conducting. It is made up of muscular tissue.

* Endocardium

It is the innermost layer which forms the outer layer for atrium and ventricle of heart.

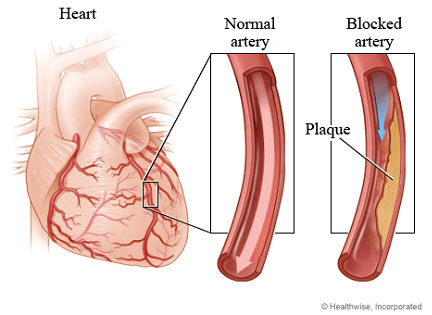
The conducting system of heart is unique which is done due to electrical signals generated by sinoatrial node or SA node. It is located at the top of right atrium. It is also known as natural pacemaker. When the electrical signal is generated by natural pacemaker it is passed to muscle fibres of heart which helps in contraction. Although, this contraction may increase or decrease on behalf of various factors such as stress, physical demands or hormonal level.

The blood vessels of heart are divided mainly in two parts arteries which carries blood from heart to different parts of the body and veins which takes the blood from the different parts of the blood to the heart for oxygenation.

## 2.2. What is Heart Attack?

Heart Attack is also known as Myocardial infarction. As each and very organs of body need blood for proper function and regulation. In the similar manner, heart is supplied by blood with the help of artery. If the blood vessels which is responsible to provide blood to the muscles of heart is blocked, due to which the muscles are damaged and are not able to function properly which leads to heart attack.

The heart attack occurs when the coronary arteries that supply blood to the muscles of heart is narrowed due to fats, cholesterol and many other substances which is altogether known as plaque. This process is known as [atherosclerosis](https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis). If due to healthy lifestyle this plaque is broke but this might spread to other parts of heart which may block other blood vessels of heart.



[Figure 2: Blocked cardiovascular artery. Image downloaded fromhttp://www.nulifewellnesscentre.com/chelation-therapy-heart-blockage]

Before a heart attack, one of the plaques is burst which causes a blood clot at the side of rupture. This clot may trigger the chances of major heart attack. The CHD is the major condition of the blocked blood vessels.

The starvation of heart muscles due to oxygen and nutrients is known as Ischemia. When this ischemia increases or causes more damage to the muscles which leads to heart attack or myocardial infarction.

There are no symptoms of atherosclerosis. Whenever the coronary artery which supply blood to heart is narrowed or blocked, the nearby vessels compensate by expanding themselves and supply blood to heart. This expanded blood vessels is also known as collateral circulation which prevents heart attack; this may also developed after heart attack to help heart muscle to recover.

The damage which is caused by heart attack depends upon the number of factors such as size of the area of blood vessels which is blocked and the time between injury and treatment of heart attack. The damaged heart muscles are recovered by formation scar tissue. The time taken to heal heart muscles depends upon the extent of injury and the own rate of healing. Though, the heart muscles are damaged it works properly but sometimes cannot fulfil the needs of blood supply to body. According to the research, it takes eight weeks for heart muscles to heal. The scar tissue which is formed at damaged cannot contract.

There is mostly no chest pain but there is a common kind of chest pain known as angina. Angina occurs mainly when there is not adequate supply of nutrients or blood to heart muscles. There is a difference between angina and heart attack, angina does not permanently affect the heart muscles.

The symptoms of heart attack that can prevent heart attack which are as follows:

* Chest Pain

It often occurs at top or left of chest. It might last few weeks or hours; it may sometimes go away and comeback. The heart can feel like fullness, pressure or pain and sometimes there is indigestion and heartburn also. The chest pain is sometimes mistaken with indigestion.

* Shortness of breath

It occurs when we are sleeping or doing some kind of physical activities. This is mainly due to chest discomfort.

* Feeling of anxiety.
* Sometimes the person feels weakness.

## 2.3. Factors affecting heart attack:

There are many factors which may lead to heart attack but some factors cannot be changed or controlled but some factors can be modified and treated as well as to prevent heart attack. The factors affecting are as follows:

* Increasing Age

This factor is not in human being’s control because as years pass the age goes on increasing, the older people of age 65 or above is at high risk. The majority of death due to heart attack is seen in old age group only. The health muscles cannot contract and relax efficiently during old age which leads to heart attack.

* Gender

The man is at higher risk of heart attack as compared to women according to research done at USA. The heart attack at early age is commonly seen in men whereas women are at high risk when they reach the age of menopause; this is when the death rate of women increases.

* Inheritance / Heredity

The parents who have heart disease is more likely to pass this heredity characteristics to their children. Like the African American are at high risk of blood pressure which may also cause heart attack, this is passes to their children also.

All the factors above which is mentioned is not in human being control, just as the person cannot control his age, gender and family inheritance which is passes from ancestors.

* Smoking

The cigarettes contain harmful substance such as nicotine which is harmful to individual’s lungs. As the function of lungs is to supply blood to heart after blood oxygenation, if lungs are damaged due to smoking it can indirectly affect heart and causing heart attack. The carbon monoxide which is present in the cigarettes decrease the amount of oxygen carried by blood through lungs. This not only affects to the active smoker but also to the passive smoker which increases their chance of heart attack.

* High blood cholesterol

The body need some amount of cholesterol to build cells but if it increases then it is harmful for heart because it is waxy like substance which creates blockage in blood vessels. The body contains different types of cholesterol which are Low Density Lipo-protein cholesterol (LDL) which is also known as bad cholesterol. The LDL shown be minimum in body but according to American Heart Association this should not be a main factor to consider a treatment of heart. Furthermore, the other type is High Density Lipo-protein cholesterol (HDL) which should be high in body. If the person HDL is low it may put at high risk of heart attack. The last type is triglyceride which is most common type of fat found in body. The high triglycerides with low HDL or high LDL may put the person in high risk. The total cholesterol is calculated as HDL+LDL+20 % of triglycerides level. Daily exercise or yoga would help to reduce bad cholesterol and increase good cholesterol. The total cholesterol level should less than 200 mg/dl, good cholesterol level (HDL) should be more than 40 mg/dl and bad cholesterol level (LDL) should be less than 130 mg/dl.

* High blood pressure

The high blood pressure not only causes damage to heart but also to other organs of the body which might lead to organ failure. This increases the workload of heart making the heart muscle stiff and thick. These effects the heart functionality and due to abnormal functioning, there is high risk of heart attack. These is one of the most common symptoms of heart attack. It should be controlled by reducing the salt in food, managing weight, daily exercise and if possible, doing medications.

* Obesity

This factor not only give rise to heart related ailments but also different problems such as high blood pressure, high cholesterol, diabetes and many more. The excess fat in body requires more blood supply, which cannot be pump by heart which causes heart attack. According to research, the person those has excess fast in waist are at higher risk than others. If the weight loss is done then all the risk factors related to obesity is directly reduced. By eating food which contain less fat, more protein, doing daily exercise would help to lose weight and decrease the chance of heart attack. The body mass index (BMI) should be not greater than 25.

* Stress

Due to modern lifestyle or workload in office, may affects person mental health which creates a stressful scenario in person’s life. The person who is in stress condition starts overeating, smoking etc. If the socio-economic status of person is not good which may lead to stress in daily life and indirectly there is high risk of heart attack. Anger and poorly controlled stress can sometimes lead to heart attack and stroke, it should be avoided by managing time, doing relaxation exercise, yoga or stress therapy.

* Physical Activity

The daily routine exercise helps to reduce the heart related problems. It also helps to reduce blood pressure, high cholesterol, obesity whereas sometimes it also helps to decrease blood pressure in few people. Many people live sedentary life, they don’t do exercise or does infrequently. Those who don’t do exercise at all have higher chances of heart attack as compared to others. Individual should do exercise at least 30 minutes a day.

* Drinking Alcohol

According to the research by the England Heart Association, drinking alcohol increases the blood pressure of individual, this may increase the chances of heart attack. It is suggested that the men should not take a drink more than 1 in a day whereas the women should not take more than 2 drink a day.

* Diabetes

If it not controlled then I would damage heart muscles and reduce the pumping capacity which would eventually lead to heart attack. The diabetes can be controlled by reducing sugar in daily food, doing exercise, maintain healthy weight and if possible, doing medication as prescribed by doctor. Diabetes not only affects heart but the organ such as kidney and the flexibility of blood vessels is reduced.

2.4. Symptoms of heart attack:

For research to find the symptoms of heart attack approximately 11,000 people were taken into consideration who belong to different countries such as Australia, United States of America, United Kingdom, India, Sri Lanka, China, Germany, Africa and Russia. Only 50% of participants were able to find chest pain as one of the symptoms of heart attack when they were provided with seven varieties of conditions. Moreover, there were almost 8% of participants who were not able to find any symptoms of heart attack. The participants were divided into two group old age group and young age group. The old age was aware of heart attack symptoms as compared to young age. Thus, by increasing the awareness among young age people would help to decrease the mortality rate due to heart attack and reduce delay treatment. Insufficient supply of oxygen rich blood supply to the working heart muscles, causes a pain in heart muscle known as angina pectoris. This pain is accompanied by different symptoms such as fullness at the centre of the chest, and a feeling of pressure on the chest. This lasts for a few minutes and repeats itself in relevant periods of time. There are other symptoms other than chest pain such as pain in the left shoulder, arm, jaw and sometimes back can be felt as a symptom of heart attack. There are some rare symptoms which are not usually seen such as shortness of breath, cold, sweating, breathing difficulty, vomiting or feeling sick as per World Health Organisation. It is important to note that the symptoms are not common for everyone who may suffer from heart attack which is the main problem of increasing mortality rate. The patient who has been suffering from diabetes for a long time may not feel any kind of chest pain. While the symptoms such as back pain, sickness, vomiting, nausea, shortness of breath are more commonly seen in women as compared to men. Heart attack is the only most common problem which can cause sudden death of a person. There are complications which can lead to heart attack such as cardiogenic shock which leads to death in 85% cases. The other one is cardiac arrhythmias which causes a sudden type of cardiac arrest where the patient's heart is not able to function at all. The last is heart failure which occurs due to the damage caused to the heart valve or problem related to ligaments of the heart. In some patient’s heart attack is sudden without any symptoms or early awareness. While in others there are severe symptoms such as severe chest pain, pain in the left body part. If there are any warning symptoms or sign it may be like pressure on chest or chest pain when a person is taking rest. So, to save the life of a person who has suffered a heart attack, the pulse rate should be checked to find out if he is breathing. If the person is not able to breath, then the chest should be pushed hard about 100 to 120 times per minute more specifically, he should try CPR.

2.5. Tips for reducing heart attack:

People those who are smoking or taking tobacco, following unhealthy lifestyle, no physical activity are at high risk of heart attack. Tobacco not only causes problems related to heart but also causes cancer, emphysema. The active and passive smoker both are at high risk of heart attack. The nicotine present in the tobacco are absorbed in blood. This absorption led to stimulation of hormone named epinephrine. The epinephrine is produced by adrenal gland, which causes an increase in blood pressure, heart rate or pulse rate and also breathing. To avoid problems of heart disease people should say not to tobacco or smoking. The government of each country in the world is running an awareness about the effects of smoking of individuals health.

Other than tobacco or smoking, the unhealthy lifestyle which is followed by the people in this current world are at high risk of heart attack. The unhealthy diet includes food rich in fats, sugar, less fruits or vegetables, and salt. This unhealthy diet causes obesity or person becomes overweight. With the overweight another problem arises which are high blood pressure, sometimes diabetes and increase level of body fat in blood. The healthy lifestyle should be followed to avoid the issues of different disease caused due to them.

Due to lack of physical exercise in their daily routine or lifestyle give rise to problems related to heart. The inactivity in lifestyle causes increase in level of cholesterol and obesity. If doing daily exercise will reduce the level of overweight, decrease in weight. According to World Health Organisation, doing exercise will directly reduce the blood pressure, bad cholesterol level, also improve the blood oxygen level. People should pump his heart by going to gym, cycling which improves the blood circulation, doing yoga which helps to improve the inherited problem of heart.

The high blood pressure or hyper tension damages the muscles or wall of blood vessels. The friction and damage caused due to pressure and tension causes plaques of cholesterol causing blockage in the arteries. This blockage can cause sudden death of patient. The blood vessels is narrowed due clot of substance such as blood and increase cholesterol. This clot in blood vessels causes blockage and restrict the flow of blood to various organs of blood which eventually leads to heart attack. Furthermore, the glucose which is present in blood due to lack of insulin released by pancreas. The increase blood sugar in patient those who are suffering from diabetes from long time causes blockage in blood vessels of arteries which restricts the flow of blood efficiently. Due to this atherosclerosis is build and increases the chances of heart attack.

Living a healthy lifestyle such as following a good diet which is rich in protein and carbohydrates instead of fats, sugar will directly reduce the chance of heart attack. The person should stop using tobacco products or smoking first, which will remove the amount of nicotine level in the body and harmful effects caused due to this substance. The person should do at least exercise 30 minutes per day which reduce the factors affecting heart attack such as reduce weight, improve blood pressure, decrease the level of cholesterol, sugar and salt. The person should eat raw vegetables, fruits to gain it's nutrition without cooking it. Last but not the least, the educated member of family should motivate each family members to do daily exercise, follow healthy life, sent a family member to rehabilitation centre if he is unable to quit smoking and provide proper medics as prescribed by doctors to reduce the chance of heart attack.

## 2.6. Importance of data science in heart attack possibility prediction:

The detection of heart attack at every early stage was not possible with the conventional practise. Any disease or problem does not occur all of a sudden but it is a continuous process. The disease occurs due to result of unhealthy lifestyle for a long period of time and it sometimes gives a common symptom which is not detectable. After few years, it would eventually lead to disease such as heart attack which is life threatening. Heart attack is preventable if we modified our lifestyle such as reducing alcohol, eating healthy food, etc. When it is coupled with the early treatment it would improve its prognosis. Although, it becomes very difficult when the patient suffers from high blood pressure, high cholesterol because they are at high risk of heart attack. To overcome this scientist and doctors thought of using machine learning techniques which is part of data science. So before applying machine learning, they collected the data of patients such as their blood pressure, gender, cholesterol level, etc. The analysis was done on the data to get useful pattern such as who is more likely to get heart attack male or female. Few such research was done by using this technique, but in this the improvement of that is done wherever it was possible.

Data Science helps to build screening tools, its superiority to recognise pattern in dataset and classify it; Doctors and Scientist are trying to use this. This has been effective to predict and assist in making decision when large amount of data is provided by healthcare industry. When the traditional approach is not able to solve the problems related to medical history of patients, this is when this new approach is able to rescue.

In Paolo Melillo, Nicola De Luca, Marcello Bracelet and Leandro et al., [45] the risk assessment checking in patients suffering from heart attack is done by automatic classifier which does the classification directly. It uses a term named as Long-term heart rate variability; this classifier classifies the person those who are at low risk from the person those who are at high risk automatically. Classification and Regression tree (CART) is used to classify the patients is at high or low risk. The fluctuation of heart rate is also one of the important factors in finding the high risk of heart attack in patient.

Stephanie C. Lemon et al [20] presented a project which was to detect the heart disease using the fuzzy rule model. This model is used to give optimistic results which was necessary. Another research paper done by Choi, J et al [12] introduced clustering algorithm which is known as K-means that is used to extract the data appropriate to heart attack from the data warehouse where lot of data is stored. The important data of the heart attack and give important updates of the same.

# 3. Problem Definition:

Earlier, the doctors and health professionals used to predict the chances of heart attack using their prior knowledge and experience. The knowledge of each disease they acquire during their education but sometimes it is insufficient due to lack of practise. Furthermore, the expert doctor can only efficiently predict that patient may suffer heart attack or not. The number of such professional doctors is quite less. There is only one cardiologist for 10,000 citizens of country like India as per studies done. On the other side, the number of physicians is more but are not specialist of heart related ailments.

To overcome this problem, there was a need of system which can help to find the health of heart. This system would be revolutionary in medical industry, assist doctor and young physicians as tool to improve the quality of healthcare for patient. There is lot of medical record of patients present in warehouse of health care industry who were diagnosed with heart problems. Therefore, these medical records would be helpful for new patients who do not have any medical records currently.

The unused dataset of patients would be used in machine learning technique. The different classification algorithms will be using the dataset to recognise the pattern to predict the chances of heart attack. Each and every algorithm is implemented and model is chosen on behalf of accuracy, precision, recall, F1 score. If person may suffer with heart attack it will be represented as 1 else 0.

# 4. Aims and objectives:

The aim is to classify the patient have the chances of heart attack or not. This classification is done using different machine learning techniques using the medical record of patients. The main objective of thesis is to make prototype of heart attack prediction system that would make intelligent medical decision which traditional techniques cannot. Moreover, it would reduce the cost of medical tests and also help to learn the hidden knowledge present in historical medical dataset.

# 5. Research Questions:

1. How does increasing age affects heart attack?
2. Does gender affect heart attack?
3. What should be level of cholesterol to reduce the chance of heart attack?
4. What is minimum and maximum blood pressure that causes heart attack?
5. To find at which range of heart rate does heart attack occurs?
6. Which type of chest pain or angina is more severe that can lead to heart attack?
7. How does exercise help to reduce the chances of heart attack?
8. What should be the blood sugar level in body to reduce the chance of heart attack?
9. What other factors affects heart attack other?

# 6. Expected Outcomes:

The expected outcome of this research is that through the application of the appropriate machine learning techniques the features can be extracted from the heart attack dataset such that the features will be helpful in the classification of the heart attack possibility accurately. Presentation of the performance report of the heart attack so that accuracy of the prediction can be identified.

# 7. Everything about data:

## 7.1. Data Source:

Total 303 patient’s data have been collected keeping in mind their medical history and problems related to heart. The patient of different age group data has been collected and mostly the patients are above 30. The number of death due to heart related ailments is seen greatest in middle age group according to World Health Organization (WHO). There are 14 attributes which is taken into consideration and it is considered as the most important factors of people prone to heart attack. The dataset contains mixture of data of both people who have chance of heart attack or not. This dataset is downloaded from Kaggle free of cost. The dataset is uploaded from UCI repository to Kaggle, where is it collected by group of medical organization related to field of healthcare industry. There were many attributes which was collected but only few was most common problem in heart patients. The dataset was collected by organization in such a way it contains both patient who have chance of heart attack or not. Although, the number of datasets collected of patient is quite less to train a machine learning models which require huge amount of data. The form is filled up for each patient and it is stored in excel or csv file. All 13 columns are useful for training and last column is target one.

## 7.2. Data Description:

There are 14 attributes which are collected of 303 patients which are as follows:

* Age

The different age group people data has been collected. The prefer age group is 30 and above because according to research done they are more prone heart related ailments. The datatype of this column is integer type.

* Sex

According to studies, the chances of heart attack is more in male as compared to female. Therefore, this feature is also considered and the data is represented as 1 for male and 0 for female. The number of male patients in dataset is 206 and for female is 96. The datatype of this column is integer type.

* Chest Pain Type

The early symptoms of each heart related problem are chest pain which is also known as angina. The different types of chest pain are typical angina which is represented as 0, atypical angina which is represented as 1, non-anginal pain which is represented as 2, asymptomatic which is represented as 3. The number of typical anginas is 143, atypical angina is 87, non-anginal pain is 50 and asymptomatic is 23.

* Resting Blood Pressure

The normal blood pressure (BP) is less than 120 mm Hg for systolic pressure and less than 80 mm Hg for diastolic pressure. This is the most common factor affecting heart attack. It is distributed in a range of low blood pressure to high blood pressure, the lowest blood pressure is 129 mm Hg and highest blood pressure is 200 mm Hg.

* Serum Cholesterol

The total cholesterol level is stored in csv file of 303 patient. The normal total cholesterol level is less than 200 mg/dl. The total cholesterol level is calculated to store in sheet using High Density Lipo-Protein (HDL), Low Density Lipo-Protein (LDL) and triglycerides. The high density lipo-protein should be more than low density lipo-protein, triglycerides. This column is of integer datatype.

* Fasting Blood Sugar

The normal fasting blood pressure is less than 100 mg/dl. Sugar in blood affects heart muscles functionality if it increases more than marginal level. The column of this data is of integer datatype.

* Resting Electrocardiographic Results

This value is stored using Electro Cardio Graphic (ECG) and classified into 3 different types which are normal represented as 0, ST-T wave abnormally represented as 1, probable or definite left ventricular hypertrophy represented as 2. The count of type 0 (normal) is 152, type 1 (ST-T wave abnormally) is 147, type 2 (probable or definite left ventricular hypertrophy) is 4. This column is of integer datatype.

* Maximum Heart Rate Achieved

The heart rate is measured of each patient, the normal heart rate ranges between 60 to 100 beats per minute. The maximum heart rate of patient in dataset is 202 beats per minute whereas for person with minimum heart rate is 139 beats per minute. The heart rate is second most factors affecting heart attack. The column is of integer datatype.

* Exercise Induced Angina

During excessive exercise sometimes the heart is not able to pump efficient blood. The inability leads to chest pain which is represented if chest pain occurs as 1 and if it does not then it is represented as 0. The number of 1 is 204 whereas for 0 is 99. The column is integer datatype.

* Depression Induced by Exercise

This attribute represents the ST depression value when person is doing excessive exercise. The ST depression means the interval between depolarisation and repolarisation of ventricles. The value is found using ECG (Electro Cardio Graph) of each person after they have done exercise. It is a continuous value, the ST depression up to 0.1mV is considered to be normal. The column is of integer datatype.

* Slope of the Peak Exercise

The slope value of ST wave is stored in database, it is calculated using ECG which is electronic machine and capture the signals released by pace maker. The slope of ST segment is of 3 type which is up sloping that is represented by 0, flat that is represented by 1, down sloping that is represented as 2. The count of type 0 (up sloping) is 21, type 1 (flat) is 140 and type 2 (down sloping) is 142 in dataset. The datatype of this column is integer.

* Number Of Major Vessels

The number of vessels that is used to supply blood to different organs or parts of the body. To find which vessels is blocked this factor is taken into consideration. There are in total 5 major blood vessels considered which range from 0-4. The blocked is present in blood vessels can be obtained by flour spy. The count of blood vessels for type 0 is 175, type 1 is 65, type 2 is 38, type 3 is 20, type 4 is 5.

* Thalassemia

It is a categorical variable which represents few numbers. The value of 3 is considered as normal thalassemia, 6 is considered to be fixed defect whereas the 7 is reversible defect. Thalassemia directly affects the heart that’s the reason this factor is considered

* Target Variable

This is the only dependent variable in the dataset. This is almost a balance dataset; the classification of the same column is to be done. The person may suffer heart attack that count is 165 whereas the person that may not suffer heart attack is 138. The target value is output column for classification.

|  |  |
| --- | --- |
| Features | Values |
| Age | It is continuous data |
| Sex | 1 – Male  0 - Female |
| Chest pain | 0 - typical angina  1 - atypical angina  2 - non-anginal pain  3 - asymptomatic |
| Resting blood pressure | It is continuous value measured in mm Hg |
| Cholesterol level | It is continuous value measured in mg/dl |
| Fasting blood sugar | It is continuous value measured in mg/dl |
| Resting electrocardiographic results | 0 – Normal  1 - ST-T wave abnormality  2 - definite left ventricular hypertrophy |
| Maximum heart rate achieved | It is continuous value |
| Exercise induced angina | 1 – Yes  0 - No |
| ST depression induced by exercise | It is continuous value |
| Slope of the peak exercise ST segment | 0 – up sloping  1 – flat  2 – down sloping |
| Number of major vessels | It is categorical variable |
| Thalassemia | It is categorical variable |
| Target | It is categorical variable to predict person may suffer a heart attack or not |

# 8. Methodology:

## 8.1. Literature study:

The method that is utilized in this research is the constructive research as the result of this research can be easily validated with accuracy (Adegoke, 2011). For conducting this research, the information was collected from a variety of literary sources for getting the theoretical knowledge for conducting the study. Firstly, literature was studied so that the theory can be developed for finding the answers to the research questions. The literature was studied for finding the question for the classification of the heart attack dataset for classification of possibility of the subject evoked when the stimulus is induced while heart attack prediction data collection. Through the analysis of the research work the heart attack possibility were identified in the individuals as the chances can be grouped into may suffer heart attack or may not and through this data the classification can be done for the identification of these groups.

# 8.2 Data Analysis

Data Analysis is a process of to find useful information, making conclusion and make the decision by cleaning dataset, transforming the dataset, modelling dataset when it is required. This is the most important step before training any data using machine learning algorithms. The business-related tactics or decision is made by analysing the data in today’s world. It also helps to increase the progress of business efficiently. The different steps in data analysis are importing data, processing data, cleaning data, exploring data (Exploratory Data Analysis), Modelling and Algorithms and many more as required by the client or to increase the accuracy of model. The different process required for this is explained below

* Importing Data

The downloaded data from Kaggle data is in row and column form. The data is stored in csv file. This csv file has to imported using the library of python named pandas. The data is imported as it is row and column form. Due to this the analysis can be done easily and useful information can be obtained. The function that is used to import is read\_csv (). So, the imported data contain 303 rows and 14 columns.

* Data pre-processing

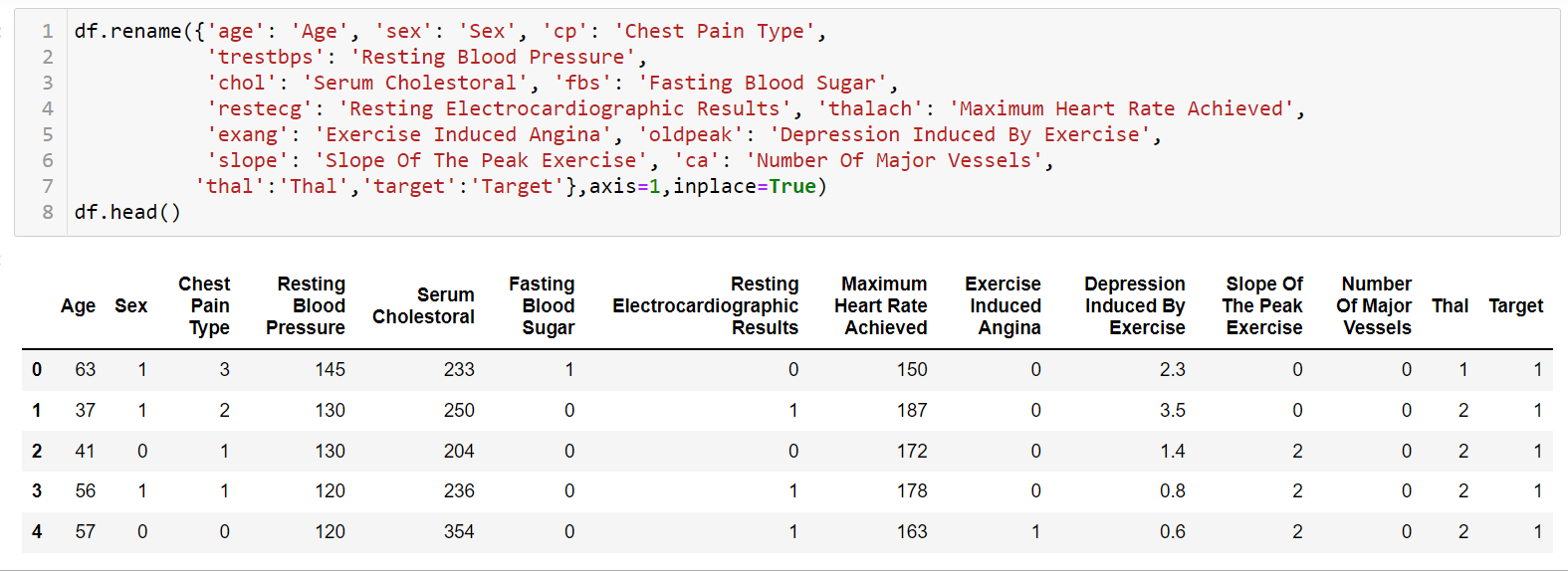
To understand the data more efficiently it is pre-processed before cleaning. If the imported data is not in table format / structured data it is to be converted into that format that would be useful for analysis. The first row of each data represents column name or more specifically attribute of data.

* Data Cleaning

This is the most time tedious task but important one. The data may contain some duplicates, missing values or some error. Due to human error some values would be wrongly entered and stored in database. This type of data has to be cleaned before it is provided to model for training. The common steps are identifying inaccuracy in data, finding duplicates, column segmentation. The all process is done using variety of analytical techniques. The data use for this thesis is cleaned and does not contain any missing values. For each and every medical data of 303 patient is appropriate as shown in below image.

* Data transformation

After that before going towards modelling the data, the dataset was firstly normalized and standardized. In the process of normalization, the input variables are normalized in the range between 0 and 1. Through this way the difference between the values can be reduced and this is done to improve the precision of the model. In the process of standardization each input variable is scaled such that the mean value is subtracted from the input value and then the value thus obtained is divided by the standard deviation so that the distribution can be shifted such that the mean value can be obtained as 0 and the standard deviation value can be obtained as 1. The standardization method that is used in this study is Minmax scalar by which the data can be scaled and transformed in the values ranging from 0 and 1 such that the minimum value obtained is 0 and the maximum value that is obtained is 1. This is done because if the model measures the values at different scales to model will not be able to indiscriminately predict the correct results, so the variable is transformed such that they fit in the model at equal scale without any kind of bias in the data. But for these projects, we have not transformed the data using any standardization technique. Only rename of the column is done to understand the features very well. The image of the same is shown of renaming column



[Figure 3: Showing image of renaming the column for better understanding. Image is used from the Jupiter notebook]

* Exploratory Data Analysis

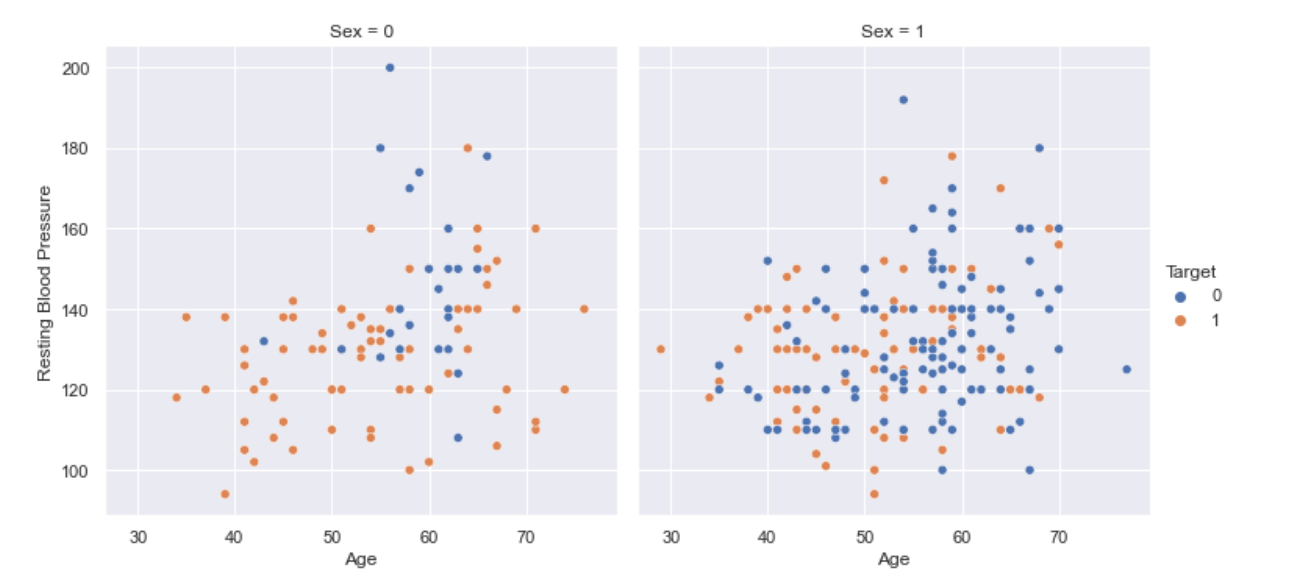
To understand the data in table form or spreadsheet is quite difficult and complex. Without visualization, to see the data is a boring, time tedious task. It is an approach to visualise important features or attributes of a dataset. By visualization, the data can be summarised and can be learned what data is trying to say to us. Before training the dataset, the dataset should be understood by EDA (Exploratory Data Analysis). Sometimes, it also helps data scientists how to manipulate the data, how to remove the outlier, how to find any kind of anomalies. It can also help to apply which statistical techniques, hypothesis test, which algorithm of machine learning to be applied. The dataset provided by the business client, to explain them into detail the data scientist or analyst uses the concept of EDA. It helps to find the standard deviation, type of variable, confidence interval of the dataset. The correlation between features can be found out and feature selection can also be done using this. The need of EDA is to understand data before making any type of assumption. The visualization can be helped to show how the model has performed to business clients to grow their business. The stakeholders are also benefited due to this because the question which they are thinking about can be confirmed by using this technique. Finding a useful relationship between features was possible due to exploratory data analysis. The tools of exploratory data analysis is clustering or dimensionality reduction which helps to visualise high dimensional data which is not understood easily. The K -means clustering is performed to visualise the dataset, the dataset is divided into k clusters on the basis of distance between centroid of each cluster.



[Figure 4: Showing image of plotting the age versus serum cholesterol level with target variable. Image is used from the jupyter notebook]

As the age increases the cholesterol level of patient also goes on increasing this does not depend upon the sex of person. But as the age increases the chances of heart attack increases in male. The death caused due to heart attack by increasing cholesterol level is more in women as compared to men. If comparison is made on the age, resting blood pressure and gender. The blood pressure goes on increasing with age but the death due to heart attack is more in men similar to women in cholesterol level. The death due to heart attack is more in men as compared to women as plotted in notebook file. The number of death due to chest pain type of non-typical angina is more as compared to others. While the death due to asymptomatic chest pain is least whereas for the person with type 0 that is typical angina and type 1 that is atypical angina the death due to this is almost same. The plot of electro cardio graphic result with targets shows that the death due to ST wave abnormality is more which is of type 1, whereas the type 2 that is ventricular hypertrophy is least. On the other hand, the death due to type 0 that is wave which is normal is more than the type 2. The death due to fasting blood pressure due to a increase in certain level of mg/dl that is type 1 is more as compared to the type 1. So, it can be said that if the level of blood sugar increases the chances to suffer heart attack is also increased.

The exploratory data analysis is mainly divided into univariate analysis and multivariate analysis. The univariate analysis is the simplest one which deals with only one feature. Due to a single feature, there is no need to find the relationship. The type of univariate analysis is boxplot which plots 5 different characters of single feature which are median, maximum, minimum, first quartile, second quartile. The other type is a histogram which helps to find the distribution of the dataset. The multivariate is to plot data points for more than one feature. The example of multivariate is scatter plot, bar chart which considers at least two parameters to plot the data in the X and Y axis. There are different libraries which are used to plot data such as matplotlib which is the most basic library. The other one which is known to me is seaborn, plotly and cuffins which is an advanced version of matplotlib. This library requires very less code to plot complex graph as compared to matplotlib. It is also helpful for the data cleaning and transformation process because it tells how to handle the missing values, and how to deal with the noisy data in the dataset. There is also other graph which is difficult to plot in matplotlib but can be easily plotted in seaborn such as heat map, bubble chart, count plot and many more.



[Figure 5: Showing image of plotting the age versus resting blood sugar with target variable. Image is used from the jupyter notebook]

* Feature Selection

Feature selection is the process by which only those features are selected that contributes to the most in the prediction of the output value. This is done because if there are irrelevant values in the model for the prediction of the output the accuracy of the model decreases. So, only those features are selected that links the most to the prediction of the output value. This technique is very helpful as the over fitting of the data is reduced by feature selection procedure as the redundancy in the data is reduced. When there is less data that is misleading from the result then the accuracy of the model increases significantly. Through this way the model becomes more efficient as the value can be predicted through the use of model in lesser time because the mis leading data is removed through feature selection procedure. In this research for the purpose of feature extraction the methodology that is used is Univariate selection by which the features having the strongest relation with the output variable that is the 'label' variable are selected. The feature selection is not done for this heart attack dataset because the number of features is already less to train and the no two column are highly correlated to drop anyone. Only those features are taken in the sample that is selected through feature extraction process. This new data frame will be used for the further data analysis by building the classification model for the classification of the heart attack.

After making the new data frame the same is converted into the NumPy array. The sample data frame is converted to the NumPy array named as X and the variable "label" from the original df3 data frame is also converted into the NumPy array and is stored into the y array variable. After that the train test split is performed and through this way the performance of the ML algorithm is compared with the model that is prepared. This method is very useful for the analysis of large dataset as in case of the heart attack dataset that is very large. For using the train test split procedure, the scikit learn library is used for the evaluation of the performance of the machine learning algorithm. In this process the dataset is divided into two sections from which the first section is used as the training dataset in which the model is fit and the second section is used as the training dataset for the training of the model. After that the predictions are made by the model and the results of the predicted values by the model are compared with the actual results. This way the model can be fit using the training dataset and the model can be tested using the testing dataset and this way the performance of the model can be identified.

* Modelling and Algorithms

There are different algorithms in machine learning which have certain procedures for training data to create a model of machine learning. As it is commonly said that data is fit on an algorithm to create a model. So, the model is the output of an algorithm that is run on data. The trained model can be saved for future prediction on unseen data. The model represents what has been learned from data by using a particular algorithm. To differentiate between model and algorithm, model includes both a procedure or process for predicting a new data and algorithm is a process of how to train data. The model is selected if it predicts efficiently and correctly for the unseen data. However, machine learning engineer wants a model and don't care much about the algorithm but it should be known how algorithms works. The algorithm is classified into three main categories which are regression, classification and clustering. The classification algorithm is used to classify between two categorical values such as for this thesis using different clinical attributes of the patient, the prediction is made that he might suffer a heart attack or not. If there is any chance the patient may suffer heart attack it will predict 1 or else it will give output as 0. There are different classification algorithms used which are logistic regression, support vector machine (SVM), naive bayes, decision tree, random forest, k nearest neighbour (KNN). The detailed explanation of how this algorithm works and helps to train data is shown below whereas the model created after training data and its results is explained in the next section.

1. Logistics Regression

It is a type of supervised machine learning algorithm. It is used to predict the target dependent variable when there are different independent variables given. The classification of the dataset is done on the basis of a probabilistic function that lies between 0 and 1. The classification is done on the basis of S type curve which is also known as sigmoid curve. If the value lies above a certain value or it gives likelihood whether the patient will suffer a heart attack or not. The threshold value is declared, for example 0.5 if the probability value is greater than threshold value the target value is classified that person may suffer heart attack, if it less than the output value would be 0.As the probability value is not more than 1, the value of sigmoid curve or 'S' shape always lies between 0 and 1. The assumption of logistics regression is that dependent variable or which is target column in heart prediction dataset should be categorical in nature as it classification algorithm. Moreover, the independent variables such as age, sex, cholesterol level, type of chest pain, blood pressure should not be multi-colinear in nature. The probability is calculated as y/1-y but this value may range from minus infinity to plus infinity, that's the reason the log value of y/1-y is taken into consideration. The different types of logistics regression are binary, multinomial and ordinal. For this thesis binary logistics regression is used as we have to predict the person will suffer heart attack or not. The dataset is split into training and testing dataset where training dataset is fit into logistics regression algorithm. The model is trained using logistics regression and the result of the same is explained further in next section. To increase the accuracy hypermeter tuning can be done where the variety of parameter is changed and which parameters provided good accuracy is used for business strategy.

1. Decision tree

Unlike, logistics regression which can be used only for regression problem decision tree can be used for both classification and regression problems. It is a tree-based classifier, the node represents the attribute of the dataset, and the leaf node represents the outcome of that node. The main or decision node does not contain any branches which are useful to make decisions and leaf nodes are the outcome of that decision which is not further divided into branches. It takes help of classification and regression algorithms to build a decision tree. It simply takes a decision by asking a question of Yes / No. This algorithm is easy to understand as it mimics human thinking. The main problem arises is how to select best features, this problem is solved by attribute selective measure. Therefore, for this two techniques information gain and Gini index are used. Information gain is the measure of changes in entropy for a given feature to build a decision tree. Gini index is the measure of impurity present for a given attribute.

1. Random Forest

It uses the concept of ensemble learning where it uses more than one algorithm to solve complex problems. Random Forest is a classification and regression algorithm, which takes help from the number of decision tree and takes the average of all to improve the accuracy for prediction. More the number of decision tree the higher is the accuracy and reduction in the problem of overfitting. The prediction of each decision tree is taken into consideration and the average accuracy is considered as the final output. Taking the average of each decision tree is also known as voting. There are few assumptions by random forest that there should be some actual value present in dependent features instead of guessed. Secondly, the correlation between each prediction of the decision tree should be very less. The advantage of random forest over other algorithm is that it gives good accuracy although there may be missing values in dataset, training time for large dataset is less as compared to others, most of the time the accuracy is high. The only disadvantage of random forest is that though it may be used for both classification and regression but it is not useful for regression. It is used in healthcare industry, marketing, banking and many more. For this thesis, there are both categorical variables and continuous variables. The categorical variable such as age, sex, chest pain type is divided into two branches by asking question yes or no. While for continuous variables such as cholesterol level, blood pressure level the dataset is divided into two parts by considering a threshold value. On the basis of that the question is asked for continuous features and a decision is made at the parent node to get output in the leaf node.

1. K Nearest Neighbour

K Nearest Neighbour is a supervised classification algorithm. It checks the similarity between the data points which are present with the new data point. The newest ones are then classified into the most similar categories. As compared to other algorithms which make assumptions, this is a non - parametric algorithm which does not assume anything on the given dataset. It stores each and every data point instead of training immediately. That's the reason it is also called a lazy learner algorithm. When whole data is stored at that moment it starts classification. So, it can be said that initially it just stores and when new data is added it starts to classify. This algorithm is also implemented for heart attack prediction the output of the same is explained in detail in next section. The advantage of this algorithm is that it is easy to understand, it can handle outliers efficiently, to train large data it takes very less time. The most difficult part is to determine k value because there is no appropriate method to find it, trial and error method is mostly used. Furthermore, the cost to calculate the distance between the points takes too much time when the dataset is huge. The heart attack dataset is classified using KNN where the hyperplane or straight line is drawn to classify the data points. Initially the k is assumed and the distance between the k nearest point or neighbour is calculated using the Euclidean distance. The Euclidean distance is calculated as the square root of the sum of squares of x points and y points. The target column is classified using this algorithm and the output of it is explained in the result section of the thesis.

5. Naive Bayes

It is a probability-based algorithm which make predictions on basis of occurrence of particular value in dataset. This algorithm is based on Bayes theorem which uses concept of probability to classify the feature. It mostly used for text classification, spam filtering, sentimental analysis. Naive Bayes is made up of two individual concept one naive which makes an assumption that the appearance of some feature is independent of others. Other one is Bayes because it uses the principal of the same. The advantages of using this algorithm is that it performs very well for multiclass classification when compared with others, it is easy and fast algorithm which is mostly used for text classification. The disadvantage of this algorithm is that it does not consider the relationship between the attribute of dataset or more technical it does not consider correlation. The Bayes theorem can be gives as P(B|A) P(A) / P(B). Here P(B|A) means the probability of occurrence of B when probability of A is known. There are different types of naive Bayes which are taken into consideration on the basis of dataset which are Gaussian Naive Bayes which considers that the continuous values are under normal distribution, Multinomial Naive Bayes is used to classify large amount of data or huge categorical data which uses the frequency of word present in particular sentence. The third type is Bernoulli Naive Bayes which is almost similar to Multinomial but in this it assumes the target variable is independent variables. The heart attack prediction uses Gaussian Naive Bayes to classify person may suffer heart attack or not and detail of the same is explained in next section.

6. Support Vector Machine

Support vector machine is specifically used for classification problems though it may be helpful for regression also. The aim of this algorithm is to make a line for separating the n dimension classes which is also called a hyperplane. This is the decision boundary which will help to put a data point in particular partition for the unseen dataset and enable them to classify. The extreme points are used to make hyperplane that classify the data point which is termed as support vector. When the distance between two margin or gap is more it is known as optimal hyperplane. The support vector machine is of two main types on the basis of the number of classes. Firstly, if there are only two classes or the data can be classified just using a single line such as in heart attack prediction it is known as Linear SVM. Secondly, when there are more than 2 classes to classify where 3-D planes are used it is referred as Non-Linear SVM. The Hyperplane dimensionality depends upon the number of features to be classified if there are 3 then 3D hyperplane is used. Sometimes it becomes difficult to visualise the hyperplane with more than 2 target variables. In these projects, the person will suffer a heart attack or not; to classify points a straight line is drawn by taking the input of 13 attributes as mentioned in the dataset. Then algorithm is implemented using the sklearn library which imports the SVM from that library for training data using this algorithm.

7. Adaboost

The better way to improve the accuracy of model is to implement first decision tree, random forest and then move on to the boosting technique such as Adaboost. The same procedure is followed for this topic. Adaboost stands for Adaptive boosting which can be implemented for regression and classification both but it is preferred for classification in most cases. The weight assigning technique makes this algorithm different from other. It assigns the higher weight for each incorrectly classified weights at every sequence. In these correctly classified weights are not provided further to model but only the incorrect one is fed for training. Unlike random forest which makes lots of node for prediction. This algorithm makes only two branch of leaf nodes which is known as stump. The error found in first stump is moved on to second stump for minimization of error. The prediction on test data is done similar to random forest where the output of first stump is predicted. While passing to the next stump, the voting is taken into consideration where the majority of output is taken into considered. The adaboost algorithm can be explained as take random weights but it should be equal for each data points, use input as weight and predict the wrongly classified points, for the wrongly classified points increase the weight. If the weights are classified correctly the points are classified successful or else same step has to be repeated till the end. The advantage of adaboost is that hypermeter tuning is not required as the parameters are not jointly optimised, the problem of overfitting also does not occur, it is the best algorithm to increase the accuracy of weak classifier. The disadvantage of adaboost is that it takes lots of time to train a model, it is also not robust toward outlier or noisy data which is to be handle carefully before applying adaboost. For this thesis, the adaboost gives the best accuracy, precision, recall, F1 - score which is explained in detail in next section. It was initially created for only binary classification which is type of ensemble method. It learns from its previous mistake to improve the accuracy and the predict the unseen data efficiently. The number of weak classifiers can be given to adaboost to increase its accuracy whenever required for the business growth as directed by the client.

# 9. Result:

The current work is a contribution to society to build an intelligent clinical decision system for the prediction of heart attack. The dataset is trained using the different algorithm as mentioned in modelling and algorithms under data analysis section. For training the sklearn library of python is used. The processing of data is done using Anaconda IDE, the language used was Python which was the main tool for implementing all the classifiers. The main library other than sklearn (sci-kit learn) which were used are pandas for analysis, seaborn for visualisation. This section of thesis discusses the experimental results of various classification algorithms such as logistics regression, naïve bayes, support vector machine (SVM), decision tree, random forest, boosting technique such as ad boost. The best algorithm is chosen on the basis of performance evaluation matrices more specifically accuracy, selected feature space is also taken into consideration for evaluating performance of algorithms. The other evaluating feature taken into consideration is confusion matrix, F1- score, precision, recall which is mentioned in detail below.



[Figure 6: Showing image of confusion matrix. Image downloaded from https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html]

* Confusion Matrix

In the analysis of the confusion matrix the values that are correctly predicted by the model are called as True positives. This is because the actual value and the predicted value are same so the model is able to predict the correct value. The value is called the true negative when the model is able to correctly negatively predict the value if the value is not present it is correctly predicted as that. The situation is called as the false positive when the value is not present in the actual data but it is predicted by the model and lastly the value is false negative when the value is present in the actual data but is predicted as not present as shown in the figure below. The confusion matrix is n\*n matrix where n represents the number of categorical values in target or dependent column. As this classification is binary classification, the confusion matrix for this research work is 2\*2. It provides a holistic view of how our model is classifying the unseen dataset whether it is True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). The False Positive (FP) is also known as Type 1 error whereas False Positive is known as Type 2 error.

* Accuracy

Accuracy is the measure by which it can be determined that how well the model is able to correctly predict the true values. Accuracy is the ratio of True positive and true negative and all the observations are true positive, false positive, false negative and true negative. Accuracy is the fraction of number of correct predictions to total number of predictions. Only accuracy cannot be considered when the dataset is imbalance where there is number of positive and negative value differ largely.

Accuracy = True Positive (TP) + True Negative (TN) / (True Positive (TP) + False Positive (FP) + True Negative (TN) + False Negative (FN))

* Precision

The measure of precision shows that the percentage of the values that are correct are correctly predicted by the model. So, precision is the ratio of positive observations and the total observations that are positive. This way the correctness of the model in predicting the correct value can be determined. The Precision is taken into consideration when the False Positive is higher than False Negative of dataset. Precision is used mostly in recommender system or e-commerce website, because wrong results would lead to loss in business or the customer churn.

Precision = True Positive (TP) / (True Positive (TP) + False Positive (FP))

* Recall

Recall is the measure for the determination of the probability by which the model determines the positive value. This measure is also known as sensitivity of the model and it is the ratio of the observations that are predicted correctly to the total observations, so false negative observations are also included in this case. The metric recall is the most appropriate measure for finding the performance of the model as in this case the false negative values are not ignored when the prediction is made. Recall is used mostly in healthcare industry where it is necessary to detect actual positive is not remained undetected.

Recall = True Positive (TP) / (True Positive (TP) + False Negative (FN))

* F1 Score

F1 score is the measure by which the percentage of the correctness of the predictions that are made can be predicted. Through this score a balance can be evaluated between the precision and recall. The value of F1 score is between 0 and 1 and in case the F1 score is 1 is shows that there is perfect balance between precision and recall. F1 score is also said to be the harmonic mean which is the combination of both the matrices.

F1 score = 2\* Recall (R) \* Precision (P) / Recall (R) + Precision (P)

* ROC- AUC Curve

ROC stands for receiver operating curve which is plotted to find the performance of the classification algorithm trained on a specific dataset, it is plotted at each and every threshold value. It basically considers two parameters which are True Positive (TP), False Positive (FP). By lowering the value of threshold would increase the value of both true positive and false positive. To evaluate the model this curve concept was insufficient, which led to the need of AUC. The AUC stands for area under the roc curve. The value of AUC can be calculated by considering all the classification thresholds. AUC assumes that the probability of a positive random sample is more as compared to a negative random sample. The AUC is considered as utmost important because it ranks the prediction instead of absolute value which is termed as scale-invariant. But this cannot always be considered because sometimes we don't need the probability. Secondly, it does not consider the classification threshold value but chooses the quality of model prediction which is termed as classification threshold-invariant. The classification threshold invariant cannot be considered when we want to improve the type 1 or type 2. But for this thesis, I have plotted AUC curve after implementing the adaboost algorithm which is explained at last part.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 - Score |
| Logistics Regression | 83.60 percent | 0 - 84 percent  1 – 83 percent | 0 -78 percent  1 -88 percent | 0 - 81 percent  1 – 86 percent |
| Decision Tree | 78.68 percent | 0 – 73 percent  1 – 84 percent | 0 – 81 percent  1 – 76 percent | 0 – 77 percent  1 - 80 percent |
| Random Forest | 88.52 percent | 0 – 86 percent  1 – 91 percent | 0 - 89 percent  1 – 88 percent | 0 – 87 percent  1 – 90 percent |
| Ada boost | 90.16 percent | 0 – 86 percent  1 – 94 percent | 0 – 93 percent  1 -88 percent | 0 – 89 percent  1 – 91 percent |
| Naïve bayes | 85.24 percent | 0 – 88 percent  1 – 84 percent | 0 -78 percent  1 – 91 percent | 0 – 82 percent  1 – 87 percent |
| K Neighbours | 85.24 percent | 0 -85 percent  1 -86 percent | 0 – 81 percent  1 – 88 percent | 0 – 83 percent  1 – 87 percent |
| Support Vector | 86.88 percent | 0 – 88 percent  1 -86 percent | 0 – 81 percent  1 - 91 percent | 0 – 85 percent  1 - 89 percent |

As mentioned in the above table, the various algorithms such as logistics regression, decision tree, random forest, naive Bayes, Ada boost, support vector machine (SVM), K nearest neighbour (KNN) have been implemented. The classification report of the same such as precision, recall, F1 - Score, confusion matrix has been plotted. The first algorithm which was trained was logistics regression; the accuracy obtained was 83. 60%. The confusion matrix with wrongly predicted output as type 1 error (True Negative) was 6 and type 2 error (False Positive) was 4. The precision and recall that person may not suffer a heart attack was 84% and 78% respectively. On the other hand, precision was 83% and recall was 88% for people who may suffer heart attacks. The mean harmonic average i.e., F1 score was 81% and 86%. The value of weighted average and macro average is almost the same for each and every parameter of the classification report. To improve the accuracy, a decision tree was implemented but the accuracy was dropped by approximately 5% as compared to logistics regression. The error rate was also increased instead of improvement such as type 2 error was increased to 8% which was a major concern for medical dataset. The precision that a person may suffer a heart attack was increased by just 1% but for the other class it was decreased to 73%. Recall acted almost opposite. The F1 score was reduced in both scenarios when compared to logistics Regression. Therefore, after implementation of the decision tree it was unable to fulfil what was needed. The confusion matrix, precision, recall, F1 score was almost better of logistics regression as compared to decision tree.

Logistic Regression

The accuracy score of the model is: 83.60655737704919%

Confusion matrix

[[22 5]

[ 8 26]]

precision recall f1-score support

0 0.73 0.81 0.77 27

1 0.84 0.76 0.80 34

accuracy 0.79 61

macro avg 0.79 0.79 0.79 61

weighted avg 0.79 0.79 0.79 61

According to the implementation done in different research paper, to increase the accuracy of model after decision tree the random forest is implemented which is a type of ensemble method. The accuracy sky rocketed to 88% directly with major increment of 10% as compared to use of one decision tree. The precision, recall, F1 score was also improved by few percentages. All the parameters were the best which was obtained through random forest. The precision that there is any chance of heart attack was 91% and for no chance it was 86%. The recall was almost similar for both 88% with a minor difference of 1%. As the precision and recall was good which directly affects the F1 score, the F1 score for person was 87% and 90% for person may suffer heart attack or not respectively. The type 1 error was reduced as compared to logistics regression which was just 3 misjudged predictions. The improvement in type 1 error led to increased True Positive and False Positive. The type 2 error was same as logistics regression that is 4 but better than decision tree which was had worst type 2 error. This was best model that was obtained until now but more improvement was needed thus different algorithms was implemented other than this.

Random Forest Classifier

The accuracy score of the model is: 88.52459016393442%

Confusion matrix

[[24 3]

[ 4 30]]

precision recall f1-score support

0 0.86 0.89 0.87 27

1 0.91 0.88 0.90 34

accuracy 0.89 61

macro avg 0.88 0.89 0.88 61

weighted avg 0.89 0.89 0.89 61

After random forest, K nearest neighbour (KNN), Support vector machine (SVM), Naive Bayes was implemented, the accuracy for all was somewhat similar that is 86%. There was dropped of 2% accuracy if compared to random forest. The average precision of support vector machine was better than k nearest neighbour and naive bayes that is appropriately 87% for person may suffer heart attack or not. The recall for all the three that there is no chance of heart attack was almost similar 81% but for other class percentage was different. The person may suffer heart attack was 91% for all three k nearest neighbour, support vector machine, naive bayes. If F1 score is compared the support vector machine had good score 85% and 89% for person may suffer heart attack or not. For others such as KNN and naive Bayes was approx. 83% that the chances of heart attack is less whereas no chance of heart attack had 87%. The confusion matrix was not good than random forest the type 1 error for all was more. The true negative was approx. 6 for incorrectly predicted class. Shockingly, the type 2 error of naive Bayes was by better than random forest. As it predicted 3 classes wrong instead of 4 so, it would be considered in medical cases but there was huge difference in accuracy. For KNN and SVM the type 2 error was worst and cannot be considered for further development.

Support Vector Machine

The accuracy score of the model is: 86.88524590163934%

Confusion matrix

[[22 5]

[ 3 31]]

precision recall f1-score support

0 0.88 0.81 0.85 27

1 0.86 0.91 0.89 34

accuracy 0.87 61

macro avg 0.87 0.86 0.87 61

weighted avg 0.87 0.87 0.87 61

To gain the accuracy of 90% or more than that, the boosting technique was implemented. The ada boost (Adaptive boosting) was implemented which gave better accuracy, precision, recall, F1 score and reduced the type 1 error and type 2 error. The accuracy was increased to 90% which was much better than all the algorithms. The type 1 error (True Negative) was almost reduced from 4 to 2. So, there was only two data of patients which was misclassified. But for the type 2 error which is mostly concern in clinical diagnosis was not improved, the number of incorrectly classified class was 4 similar to logistics regression. The true positive and false positive showed a increment was good sign. The precision that there is heart attack possibly was 94% and no possibility of heart attack was 86% only which was huge difference in both classes. The recall for no chances of heart attack was 93% and there was major difference of 5% between both the classes recall. The F1 score which is totally dependent on precision and recall, shown an improvement of 89% and 91% for the heart attack prediction possibility. Lastly, it can be said that this algorithm which assigns new weight at each sequence was much better if compared to classification report and accuracy of model. This model should be chosen for prediction of heart attack possibility which would be useful to save lives of human beings by early prediction. The harm that is caused due to heart attack is almost reduced due to use of this technique to predict heart attack on basis of medical record of patients.

Adaboost Classifier

The accuracy score of the model is: 90.1639344262295%

Confusion Matrix

[[25 2]

[ 4 30]]

precision recall f1-score support

0 0.86 0.93 0.89 27

1 0.94 0.88 0.91 34

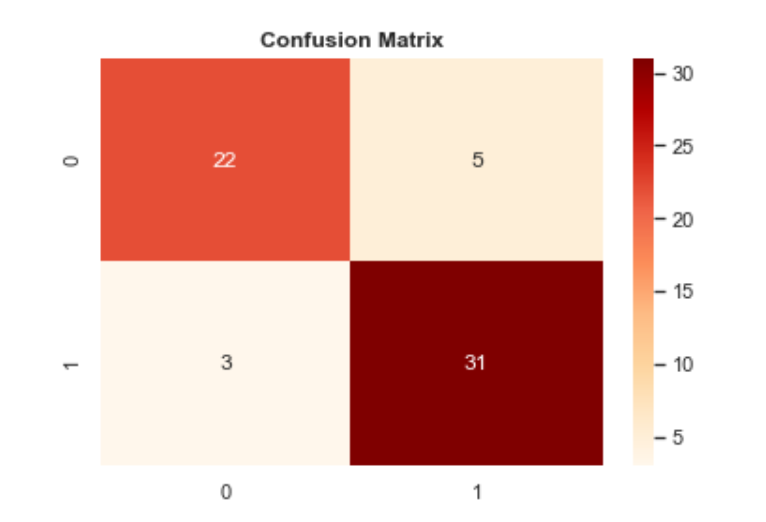
accuracy 0.90 61

macro avg 0.90 0.90 0.90 61

weighted avg 0.90 0.90 0.90 61

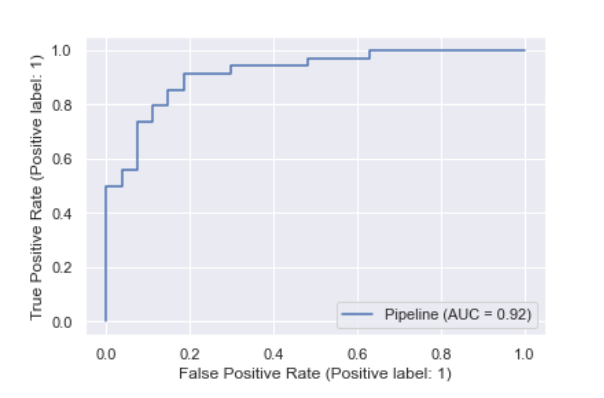
# 10. Discussion and conclusion:

Heart attack is increasing rapidly in both developed and undeveloped countries. It is considered to be a fatal chronic disease. This thesis tries to build a system which can efficiently predict the heart attack at an early stage. Therefore, early prediction and proper treatment would reduce the harm caused to heart attack. More the effect of heart attack, the more is the recovery time. It is essential for patients to understand the after effects of heart attack. For instance, the people above age group 65 takes at least 8 weeks to recover due to age related complications and ailments. While the recovery of younger people is fast as compared to younger age due to their activeness. Although, there might be some patients who would have to be hospitalised more than 8 weeks or be in the home healthcare system or in nursing care. Moreover, the caretaker should provide medicine prescribed by the cardiologist at the proper time for good recovery. However, some people are unable to understand the importance of the same. People should bring changes in their lifestyle such as eating less oily stuff, eating vegetables, and fruits. More specifically they should follow a balanced diet, their food should contain a fish daily, if possible, which is rich in omega 3 fatty acid. Patient should limit the amount of saturated fat and salt in their diet. It should be the responsibility of not only the patient but also the caretaker and family members to follow a heart healthy diet. Undoubtedly, the patient should do regular exercise which will benefit not only young but also old age people. Doing regular exercise will help to fight obesity, diabetes, stress, decrease blood pressure which will indirectly reduce the chances of heart attack. The level of exercise depends upon the patient's health condition. The patients who have already suffered a heart attack should do moderate exercise as suggested by physiotherapist. They should not try risky exercise which would harm their heart muscles. Furthermore, physical activity should be such that there is a minor risk of injury such as walking, jogging. There is a program called cardiac rehabilitation in each hospital which tries to speed up the patient's recovery time. If the patient notices early symptoms of heart attack such as chest pain, they should not panic and inform their closed ones as soon as possible. The chest pain (angina) sometimes is severe which cannot be treated using medication. To decrease the chances of heart, attack the person should quit alcohol, quit smoking, follow a healthy diet and try to control weight according to their height. The patient must be given mental strength and support so that he does not go into a stress or depression condition which is more harmful to the heart. If they are suffering from any mental issue, they should consult a mental health specialist. The patient should discharge from hospital if and only if there is no chest pain, he is able to do exercise without any pain or stress, the anxiety or fear is reduced. The death due to heart attack can be reduced by educating people and making them aware about symptoms of heart attack, following healthy lifestyle. If the patient is admitted to hospital there are several tests taken by doctors to diagnose the reason behind heart attack. The treatment of heart attack is angioplasty, electro cardio graphs (ECG), clot busting medicine, all these techniques are costly and time tedious. According to the report of Canada Public Health, the diagnose of heart disease led to decrease in the death rate by 5% in 2018. The heart attack can be reduced by anti-platelet agents, beta- blockers and statin. The emergency case arises in the patient who have suffer major heart attack due to blockage in many blood vessels which is passing to and fro from heart. To decrease the death due to heart attack, the public should be trained how to not panic after any person have suffer heart attack. How to provide first aid to the patient should be taught to each and every one though it may or may not be from health care industry.



[Figure 7: Showing image of confusion matrix of Adaboost implemented on test data. Image used from implemented Jupiter notebook]

To avoid heart, attack a medical history of patients such as age, sex, cholesterol level, blood pressure level is taken and the pattern is recognised using machine learning. This technique would not only help patients but also doctors to cure patient. The different machine learning algorithms which is taken into consideration are as follows support vector machine, Naive Bayes, K Nearest Neighbour, decision tree, random forest, logistics regression, adaboost. From all these the accuracy of adaboost was much higher approximately 90.16%. The classification report obtained due to this was much better than other algorithms. There may me many computers aided techniques which can reduce the cost of clinical test and help to detect, recover faster. The confusion matrix of Adaboost as seen in image below, the type 1 error and type 2 error were quite less as compared to others. The number of correct prediction such as true positive and false negative was high. After implementation of different algorithms adaboost classified the data more efficiently and precisely. Using this concept different device can be designed which would be helpful when there is not availability of cardiologist or expert doctors. In this thesis, 7 different algorithms have been implemented to find the accuracy and the help of the research paper is taken to implement and improve the accuracy of algorithm if possible. As the number of patients take into consideration was very less and the count of features affecting the heart was not sufficient. So, in future large dataset should be taken with many features. The future plan is to explore more and more dataset by contacting the different patient from hospital. The artificial neural network should be also implemented to improve the accuracy of the data if needed. If the count of people or doctors using this technology will increase the death rate due to heart related ailments will automatically decrease.



[Figure 8: Showing image of AUC curve of Adaboost implemented on test data. Image used from implemented Jupiter notebook]

# 11. References

1. On the Singapore Resident, the Knowledge of the Signs and Symptoms of which Heart Attack and Stroke among the country in the year 2014.

2. [County health different factors associated which can be avoidable deaths from cardiovascular disease in the country such as United States (US), at the year 2006–2010.](https://journals.sagepub.com/doi/10.1177/003335491613100310)

3. [Awareness about the reason of heart attack symptoms and response among adults—United States, 2008, 2014, and 2017](https://www.cdc.gov/mmwr/volumes/68/wr/mm6805a2.htm).

4. Harshit Jindal1, Sarthak Agrawal1, Rishabh Khera1, Rachna Jain2 and Preeti Nagrath2 in the year 2018,” The Heart disease prediction using machine learning algorithms which includes classification algorithm”.

5. Early and accurate detection and not only diagnosis of heart disease using the different intelligent computational model. Yar Mohammad, Mohammad Tahir, Maqsood Tayar, Kil to Chong in the year 2020.

6. Analyzing heart disease prediction based on ensemble methods and improving the same if needed, Xiao-Yan Gao,1 Adelgid Amin Ali,2 Hassan Shaban Hassan,2 and Emani M. Anwar in the year 2020.

7. Heart Attack Detection Using the Android Smartphone by the students [Md Ashrafuzzaman,](https://www.researchgate.net/scientific-contributions/Md-Ashrafuzzaman-2069051393) [Md Masaharu Huq](https://www.researchgate.net/profile/Md-Mazaharul-Huq), [Chandan Chakraborty](https://www.researchgate.net/scientific-contributions/Chandan-Chakraborty-2069050843) in the year September 2013.

8. The Effective Solving The Heart Disease Prediction Using Hybrid Machine Learning different algorithm , [Gautam Srivastava](https://www.researchgate.net/profile/Gautam-Srivastava), [Senthil Kumar Mohan](https://www.researchgate.net/profile/Senthilkumar-Mohan), [Chandra Segar Tirumala](https://www.researchgate.net/profile/Chandra-Segar-Thirumalai) in the year June 2019.

9. Animas Harz, 2Subrata Kumar Mandal, Amit Gupta, Sarcomata Mukherjee and Amitay Mukherjee, West Bengal in the year 2017, Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques.

10. Heart Attack Prediction System using Neural Network which is Cascaded, R. Chitra and Dir. Reenvisage in the year 2013.

11. Contributed by Mrs. K. Uma Maheshwari, Ms. J. Jasmine on the Neural Network based Heart Disease Prediction, at the International Journal of Technology (IJERT) in the year 2017.

12. Done by the Sushmita Manikandan, Heart Attack Prediction System, International using the Conference Communication, Data Analytics and Soft Computing (ICECDS) in the year 2017.

13. Prediction of Heart Disease Using Machine Learning. Aditi Givhan, Goutham Kokkola, Ishi Pandya, Prof. Kailas Divakar (PhD) in the year 2018.

14. Evat in the year 2018 done the research on, A heart attack in elderly people at the Lahti University Degree Programme in Nursing which was awarded.

15. Rohith Ravi, Ganesan R, Kalaallit Buddha Bhargav published the paper at the famous publication of International Research Journal of Engineering and Technology (IRJET)Efficient Heart Disease Prediction for Large Volume of data using Artificial Neural Network in April 2021.

16. The Wearable heart attack detector, Ashwini K, Koala K , IJERT, Volume [NCESC – 2018 (Volume 6 – Issue 13)](https://www.ijert.org/ncesc-2018-volume-6-issue-13) in the year 2018.

17. Neural Network based Heart Disease Prediction, Mrs. K. Uma Maheshwari, Ms. J. Jasmine, Volume, [RTICCT – 2017 (Volume 5 – Issue 17)](https://www.ijert.org/rticct-2017-volume-5-issue-17) in the year 2018.

18. Vol. 10, No. 6 in the year 2019, Fahd Saleh Alotaibi, done on the Implementation of Machine Learning Concept to Predict Heart Attack which may occur to patients, (IJACSA),

19. [Mohammad Ayoub Khan](https://arxiv.org/search/cs?searchtype=author&query=Khan%2C+M+A) in the year December 2020, the research was done on, An IoT Framework for Heart Disease Prediction based on MDCNN which is a type of Classifier.

20. J.E. Malasada, M.L. Gevalia, N.D. Gohil and S.C. Jamali this researcher, Impelling Heart Attack Prediction System using Data Mining and Artificial Neural Network, International Journal of Engineering on April 2020.

21. [Kriti Gupta](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Julianna F. Lamphouses](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Bernold Birkdale](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Parva MoD](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Ruins Chen](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Vivek T. Kulkarni](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202), [Kumar Dharmaraja](https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.112.970202) did a research on heart disease in Woman, this was most important research.

22. Prevention of Cardiovascular Disease, World Health Organisation (WHO) in the year 2007.

23. [P. Gayathri](https://www.researchgate.net/scientific-contributions/P-Gayathri-2037093597) , [N. Jai Shankar](https://www.researchgate.net/profile/N-Jaisankar), in the year June 2013 did a research on the Heart Disease Diagnosis Using Data Mining, Soft Computing Techniques.

24. L. Rogers, in the year 2016 did a research on the heart disease in 20th century: a historical review, Heather.

25. Bhalla, K. Jyoti (October – 2012) An Analysis of Heart ISSN: 2278-0181 Vol. 1 Issue 8 Disease Prediction using Different Data Mining Techniques.

26. Indo Asian News Service on May 2021, with the help of few experts did a project, why the risk of death in men is more as compared to women.

27. Sanga Ware, Shaun k Rakesh, Bharat Choudhary, Formalistic Heart Attack Prediction by using learning the machine Techniques, (IJRTE), in the year January 2020.

28. Dresswear Sedale, Kumari Deepika, “Predative analytics to prevent and control chronic disease”, https://www.researchgate.net/punlication/316530782, January 2016.

29. Nouri Ajam, “Heart Disease Diagnoses using Artificial Neural Network”, The International Institute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11

30. Paranjape, S. & Awang, R. In 2008 IEEE/ACS International Conference on ICCCA 108–115 (IEEE New York).

31. Phan, N. 2018. Bachelor thesis. A heart attack in elderly people. A guidebook for relatives and health care professionals. In Theseus.fi.

32. The American Psychological Association (APA) in the year 2017. Older adults: Health and age- related changes [referenced 18.10.2017]. Available at the website: <http://www.apa.org/pi/aging/resources/guides/older.aspx>.

33. Choi, J. 2017. Heart attack. Causes, symptoms, and treatments [referenced 14.9.2017]. In Medical situation which occurs of New Today Online Available at the website: <http://www.medicalnewstoday.com/articles/151444.php>.

34. Chitrali, S. Pate (June 2012), Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, journal of computer science and application*,* vol 47-no10, pp 0975-888.

35. Shanta Kumar (2009), Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European journal of scientific research, vol.3.1 no.4, pp 642-656.

36. Sonali. B. Mind, Priyanka Wacker The research paper was developed on the Artificial Neural Network (ANN)", International Journal (IJRITCC), Vol. 2, No. 1, January 2014, pp. 96-100.

37. Stephanie C. Lemon, Jason Roy, and Melissa A. Clark, Peter D. Friedmann, William Rakowski, "Using the methodology of classification with the help of logistics regression and using classification and regression tree analysis for the same", The Society of Behavioural Medicine, Vol. 26, No. 3, 2003, pp. 172-181.

38. Syed Umar Amin, Kavita Agarwal, Rd. Rizwan Beg "Using the genetic data to determine the risk of heart attack in individual", Proceedings of IEEE Conference on Technologies, 2013.

39. Y Wei, T Liu, R Valdez, M Gwinn and M.J Khoury,” To develop result of heart attack on the diabetic and the non-diabetic patient in efficient manner which is done efficiently ,” BMC Medical Making, vol. 10, 2010.

40. M. Salian, W. Hussain, N. A. Saqib and M. A. Khan, identifying the risk factor of heart attack by unstructured data of the cardiac arrest people in the year 2016 International Parallel (ICPPW), Philadelphia, PA, 2016, pp. 426-431. Doi: 10.1109/ICPPW.2016.66.

41. M. Hertzog and B Peel, “Different clustering analysis technique to get the patient will get heart failure”, Journal of Nursing, vol. 25, pp. 273-283, July/August 2010.

42. Ahmed Fawzi Odom , Emad E. Abdallah , Yousef Kalani , Ahmed Kef aye and Mohammad Ashour(2015) the effective Diagnosis and Monitoring of person getting the Heart Disease ISSN: 1738-9984 IJSEIA.

43. Raihan M, Mondal S, More A, Sagar M O F, Skidder G, Majumder M A & Ghosh K (2016, December). For Ischemia the heart prediction is done using the smart phone of patient to predict the heart failure of particular person. In 2016 19th (ICCIT) (pp. 299-303). IEEE.

44. Wolgast G, Ehrenburg C, Israel son A, Hollander J, Johansson E & Mane fjord H

(2016). Heart attack detection using the wireless connected to the body of patient [Education Corner].

45. S Nouri Ajam, “Using The Artificial Neural Network for the earlier prediction of patient heart failure”, The International Institute of Science, vol.5, No.4, 2015, pp.7-11

46. Adam, P. Asma (2012), Prediction System For Heart Disease using Naïve Bayes, International journal using mathematical science and advance computing, vol 3, pp 290-294.

47. Rohith Ravi, Ganesan R, Kalaallit Buddha Bhargav published the paper at the famous publication of (IRJET)Efficient Heart Disease Prediction for Large Volume of data using Artificial Neural Network in April 2021.

48. Bashir S, Qamar U & Jaded M Y (2014, November). For heart diagnosis using the intelligent framework that is ensemble methods. In International Conference on IT Society (Society 2014) (pp. 259-64). IEEE.

49. Tutu Harahan, Rozita Ghazali, Nazari MoD Naqi, Mustafa Mat Doris, "Using the recent advancement in data mining technique", The Second International Conference (SCDM-2016).

50. Stephanie C. Lemon, Jason Roy, and Melissa A. Clark, Peter D. Friedmann, William Rakowski, "Using the CART Algorithm to do analysis of public health dataset to predict the heart failure", Vol. 26, No. 3, 2003, pp. 172-181.