Diagnosing Heart Problem using Machine Learning

Pathan Juberkhan

10516983

MSc in Data Analytics August 2020

DECLARATION:

I, Pathan Juberkhan, State that this thesis is my original work and that it never has been applied for qualification or admission to another college or university. Furthermore, I have properly applied to both the information and materials available, and the thesis complies completely with the scholarly integrity principle of the Dublin Business School.

ACKNOWLEDGEMENT

I would like first to thank Dr. Marian Rocha, a research guide to the Dublin Business School, for supervising Data Analytics. The door to the office of Mariana Mam was still accessible when I went to an uncomfortable location or asked regarding my inquiry or composition. She has faithfully rendered this paper my own job, however she has guided me privileged everything she felt it was. The work could not have been successfully done without her enthusiastic involvement and knowledge.

Author – Pathan Juberkhan

ABSTRACT

The main part of the metabolism in the case of the human body is the requirement of the oxygenated blood so that the process can be complete. In the result of the metabolism, the wastes and the deoxygenated blood needs to be pumped out from the different organs present in the body so that the human body can sustain the process of living. Heart being the pumping organ of the body, not only supplies the oxygenated blood to the different parts of the body but also removes the wastes and the deoxygenated blood. Hence, to take the proper care of the heart the regular check up is very important. There can be many reasons which can be due to genetic history and some of the acquired habits that can result in an adverse effect to the heart. There has been a subsequent amount of research works that takes into the account the prediction of the well being of the human heart.

In this research we have analysed around 1095 patient cases and tried to identify the important risk factors that can be a primary reason for the heart problems. This research work tries to present a work that can be an immediate step to find a probability score for the heart problem. The different risks factors are those which can be the primary reason for the occurrence of the heart problem. In this research we have analysed different classification models to identify the heart problem. The data collected is from five different cities in India and from different age groups. The primary aim of such activity is to present a transparent solution which can help the patient to know statistically whether there is any probability of happening of any heart problem.

This solution is in fact not a replacement for a medical practitioner but rather aiding some help in the diagnostic process of any doctor. This in order creates a transparency in the treatment between a doctor and a patient. The model is checked with the number of False Positives and the False Negative that the model performed and based on that the best algorithm is selected for the prediction.

Keywords: Heart Problem, Machine Learning, Data Analysis, Recommendation Engine, Artificial Neural Networks, Random Forest, XGBoost, Data Mining, Visualization3

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The sustenance of the human body needs the circulation of the important substances to be provided to it that can make it possible to not only function properly but also help the human body to live. One of the most important part of the human system as a whole is the circulation of the blood which carries important ingredients for its own sustenance. The main organ hence is the heart which is the pumping system that creates the blood to flow towards each part of the human body. Hence the monitoring the heart become of the important part of the human body condition monitoring. In order to do a proper diagnostic of the heart, the data forms one of the important factors. In order to use machine learning to predict the condition of the heart, the researchers are having an important research in that particular direction. To analyse this particular research, the main objectives that we will be solving are as below:

1. To analyse the different machine learning algorithms and to find the best algorithm that can predict the heart problem at higher probability and less false predictions
2. To find the important risk factors or the factors that contribute to the heart problem and heart disease
3. To condition monitor the heart and to foresee the heart failure

One of the primary reasons of natural death in the Globe is due to the heart failure which is caused due to various reasons happening to the heart problem. Some of the reason can be the family genetic history and some may be raised due to the different factors that the person takes into the account in his life. This asks for a regular check up of the heart for the well-being of it.

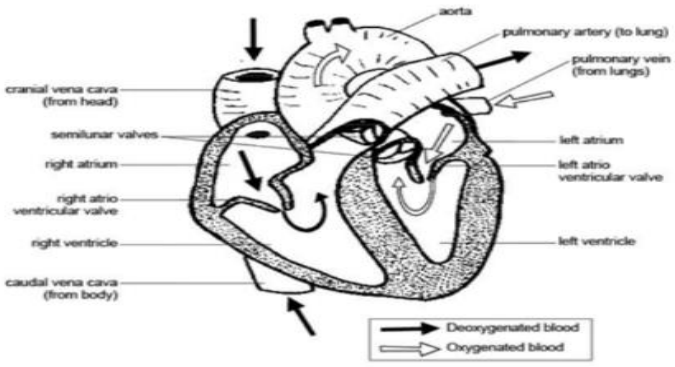


Figure 1: Image of the Human Heart. Black arrow pointing to the flow of the Deoxygenated blood in terms of removing the waste from the different parts of the body and the white arrow points to the flow of the oxygenated blood.[4]

The above is the figure of the human heart. In this we can see that there are four chambers which has the different function associated with it. The primary function of the heart is to carry the oxygenated blood to the different parts of the body which helps in the carrying out the metabolism and then carry

the de-oxygenated blood back to the heart. The main objective of this project is to present a solution for which the doctor can check with the confidence about the possibility of having a heart problem in a patient. The coronary tomography which is done in order to check if there is a heart problem or not exposes the human body to a high frequency radiation [10][11]. This radiation will be having some adverse effects for a normal human body.

Also, the cost for the coronary tomography is very high. In this research, we will be considering different risk factors such as Family History, Fasting Glucose, Smoking Habits, Hypertension, Dyslipidaemia, Obesity, Sedentary Lifestyle, CABG and High Serum to model a solution which can give a probability of the heart problem happening or not. Apart from these risk factors, some of the features known as the demographic details of the patients can be taken into the account for the modelling. Such features like Gender, Location and age etc. can be taken into account. The heart being the pumping organ undergoes through different kinds of the blockage that can lead to the improper flow of the blood through the blood vessels.

According to a recent WHO organization survey 17.5 million people die every year. It will rise to 75 million a year 2030 [33]. Medical professionals working in the area of cardiac disease have their own limits and can estimate a risk of a heart attack of up to 67%. Precise [34], with the new epidemic situation, doctors need a support network for forecasting heart disease more accurately [35]. Using the Machine Algorithm and deep learning open new door opportunities for predicting heart attack accurately. Journals offer plenty of state-of-the-art information about Methods of computing and deep learning in computers. An empirical comparison was given to aid the work of new research in this area. There are different kinds of the heart diseases and problems that has been taken into the account. Some of the problems are mentioned as below.

1. *Atherosclerosis:* In this kind of the heart problem the vessels become hard and stiff due to the fatty deposits and causes plaques [12]. In a research this has been found that smoking is a power risk factor that causes the problems in the heart. There is a strong relationship between the active and the passive smokers on the progression of this kind of the disease [13]. In this the research is done on the basis of the data collected for the determination of the progression of the atherosclerosis. A total of 10,914 participants data was collected to analyse the cardiovascular risk factors associate with it. In a result it was found that 50% of the current smokers were in the zone of increase of the progression of the atherosclerosis.
2. *Cerebrovascular Diseases (CVD):* This can be associated due to the blockage of the flow of the blood through the blood vessels which causes extreme force to circulate the same amount

of the blood [14]. In a research it was found that Asthma increases the CVD (cardiovascular disease) risk [15]. To extend the relation between different factors the authors associated with the relationship with the sex, concurrent allergy and different asthma medications to check with the play in the enhancing this problem of the heart. In a separate yet powerful research it was seen that the cholesterol concentrations in the blood (low level) is strongly associated to the cardiovascular and coronary diseases [16].

3. *Ischemic Heart Disease:* The most important effect of having the deposit of cholesterol in the walls of the arteries [17]. The figure below depicts the causes of this effect and the deposits of plaques in the inner side of the walls of the arteries causes immense pressure on the heart to pump the required amount blood to the organs.

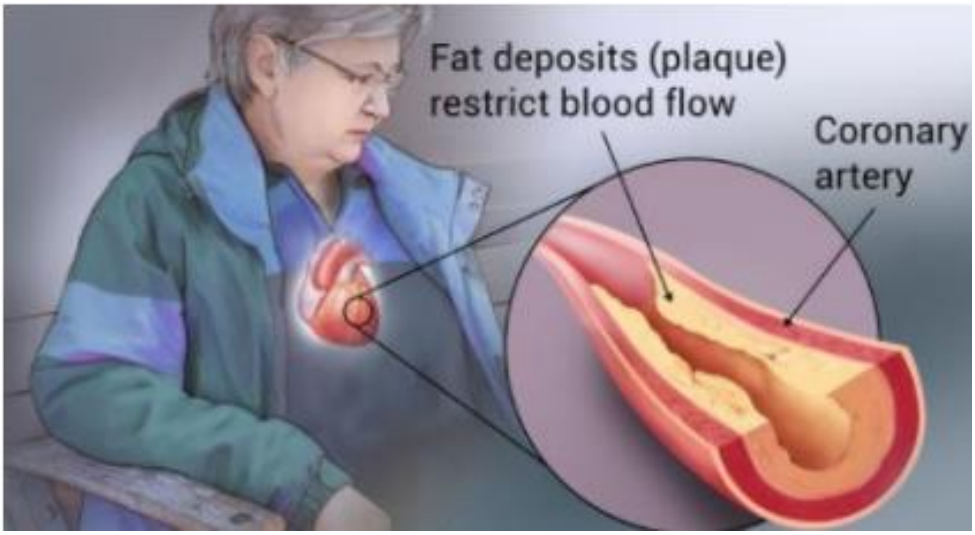


Figure 2: Deposits of Plaques in the inner side of the arteries causing the volume of blood flow to decrease and increase the pressure for the pumping organ heart to pump out the required amount of blood [14]

Research shows that the presence of the albumin excretion in the urine is directly correlated to this kind of the heart problem [18]. In this research, it was found prospectively whether a slight increase of urinary albumin excretion, i.e. microalbuminuria, adds to the increased risk of ischemic heart disease among hypertensive subjects.

4. *Hypertensive Heart Disease:* The risk factor hypertension is one of the major problems in the human beings as of now. This has the adverse effect on the morbidity and the mortality [19].

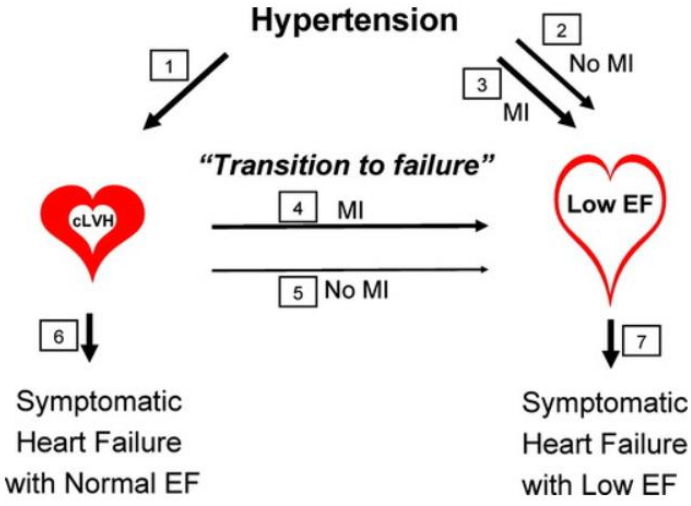


Figure 3: The seven paths for the progression of the Hypertension to the heart problem [9]

Hypertensive heart disease is a constellation of abnormalities that includes left ventricular hypertrophy (LVH), systolic and diastolic dysfunction, and their clinical manifestations including arrhythmias and symptomatic heart failure. The classic paradigm of hypertensive heart disease is that the left ventricular (LV) wall thickens in response to elevated blood pressure as a compensatory mechanism to minimize wall stress. Subsequently, after a series of poorly characterized events ("transition to failure"), the left ventricle dilates, and the LV ejection fraction (EF) declines (defined herein as "dilated cardiac failure"). In other words, the HHD is due to the hypertension in the body that causes extra pressure on the blood flow and hence the effect on the heart [20].

In this research I have taken the binary classification problem for the determination of the probability of the happening of the heart problem in the body. All the different kinds of the heart related problems are classified as 1 and heathy situation as 0. In this research we will be considering the analysis on the data that has been collected to check on the different hypothesis on the studies done. The following sections proceed as the discussion on the existing research works that has been going on and has been done on the fields towards the determination of the heart disease prediction. This will be followed by the methods and the concepts discussion on the system development of the problem of detecting the heart disease. It will be followed by the actual implementation of the system design that has been done in this research.

I will discuss the results and the data analysis that I have done in this project followed with the conclusion of the whole idea of building the model.

CHAPTER 2: LITERATURE REVIEW

Healthcare is one of the fields where there is enormous amount of the data. Hence using the features and the hidden insights within the data and to predict with certain probability about the heart attack becomes one of the primary aims in the field of data science. In one of the research areas, some of the datasets used are classified in the terms of the medical parameters [1]. The algorithms that are used for the predicting the level of accuracy in the case of the heart problem prediction are Decision Trees and Naïve Bayes. In most of the research works, Naïve Bayes has been used as the primary algorithm to take care of the prediction of the heart attack [2][3]. In another research for the heart problem prediction system [4], many of the risk factors are analysed to see the correlation of the individual feature with the patient's heart condition and then the classification algorithms are analysed to see which one performs the better task of the prediction of the heart problem with minimum false negative rate.

In this research a new metric known as the selection value is taken and defined which not only takes the accuracy into the factor rather it also takes the false negative rate into the factor for the identification of the best algorithm in this perspective. Random Forest Classifier performed the best and followed by the Support Vector Machines with Radial Basis function as the kernel. A normal diagnostic which includes ECG, CT Scan to confidently tell whether the patient is suffering from any kind of heart problem or not is actually time consuming and costly [3]. On the top of it if a normal human body is exposed to high frequency radioactive waves, this might have some adverse effect on the body as well [4].

In another research work [5], ensemble methods are used in terms of improve the prediction of the heart disease risk. The whole idea behind this approach is not only to improve the accuracy of the predication ability of the model rather also to see whether the machine learning models can be deployed to the medical environment that can make early predictions for the severity of the human body. An accuracy raise of 7% was achieved using the Boosting and Bagging methodologies which was further enhanced with the implementation of the feature selections in order to show significant improvement. In another research many of the risk features such as gender, age, cholesterol level, smoking, hypertension, eclampsia etc. are used in order to determine the risk with the help of the fuzzy logics and the decisions are made on the basis of the weight fuzzy rules.

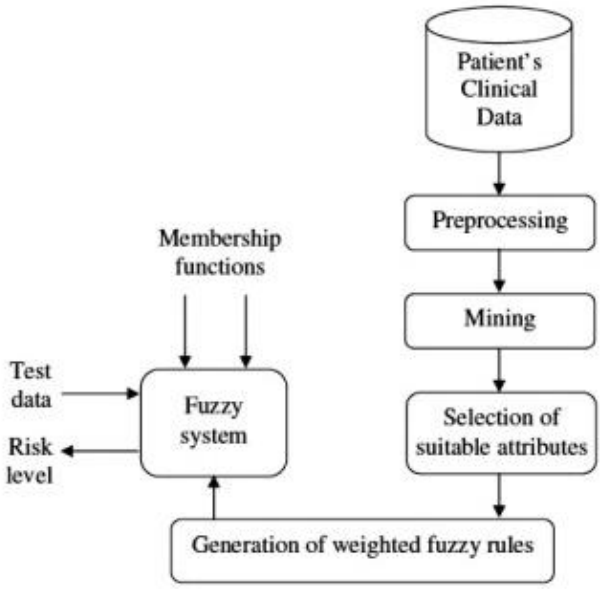


Figure 4: Architecture used [5]

Some non-invasive methods have been employed which termed to be very efficient in terms of the diagnosis of the heart failure and in order for its prevention [6]. An extensive analysis is done in order to have a proper analysis for the several machine learning algorithms that are being employed and several feature reductions and extraction techniques that has been used in this particular research. In order to give some kinds of alerts the features such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking etc. are used [7]. Since the above risks can't help in predicting the heart failures rather a conditional monitoring can be done in order to have preventive measures.

The whole system predicted with an accuracy of 89%. Even some of the research used neural networks to predict the severity of the heart problem [8]. In a research done on Endovascular aneurysm repair (EVAR) that is an innovative minimally invasive surgical procedure that helps to minimize the recovery time, postoperative morbidity and mortality for patients [21]. This research proposes a model of the ensemble for predicting postoperative morbidity after EVAR. A training collection of consecutive patients who underwent EVAR between 2000 and 2009 was used to build the ensemble model. All data needed for modelling predictions, including patient demographics, preoperative, co-morbidity, and complication as outcome variables, were prospectively collected and entered into a clinical database. To categorize numerical values into informative feature space, a discretization approach was employed.

In this research the Bayesian Network (BN), Artificial Neural Network (ANN), and Support Vector Machine (SVM) have been introduced as base models and several stacking models combined. The study findings consisted of an ensemble model for predicting postoperative morbidity after EVAR,

the frequency of prospectively reported postoperative complications, and BNs' knowledge of causal effect with Markov's blanket principle. In a research in which the work is a blurry one about the growth of Decision Support System for coronary artery diagnosis an evidentiary disorder [22]. Info on coronary artery disease University of California Irvine (UCI) sets are included. This takes the information base of the Fuzzy decision support program By using the Rough Set Theory-based Rules extraction tool. The rules are then picked and blended based on knowledge From discrete numerical attributes. Fuzzy weight rules Is proposed to use the details from the extracted help Regulations. Sets of data obtained from U.S. heart disease UCI, Switzerland and Hungary, Ipoh Specialist Hospital info Malaysia is to have the proposed program tested. Evidence of display that the device is capable of giving coronary percentage of the blocking of arteries more than cardiologists and angiography.

The results of the program proposed were checked and validated by three cardiologists, and others are called to be simple and useful. The framework is constructed using incomplete sets of CAD data as the training collection. This set of training is imputed using RST on ANN. The imputed collection of training is then is created. This training collection contains 358 items (patients). RST rule creation with the use of ROSETTA software results in 3881 rules [23]. Only 27 rules can be selected through the proposed RST based rule selection method. Table 2 reveals only a few rules out of 27. This set of rules is evaluated using full collection of CAD data and contrasted with other approaches as shown in the Table below [24].

Selection Methods	Accuracy	Coverage	Number of rules
Proposed Method	0.852	0.937	27
Support Based (Training Data)	0.847	0.799	29
Support Based (Testing Data)	0.844	0.868	27
Michalski μ=0.5	0.845	0.785	27
Torgo	0.845	0.785	27
Brazdil	0.845	0.785	27
Pearson	0.845	0.785	27
Cohen	0.863	0.65	29

Table 1: Summary of the Rules that evaluates the full correlation of the CAD Data [5]

The selection approach proposed in this research provides improved accuracy and coverage efficiency.

Rule No.	Rules
1.	oldpeak ([0.3,*)) AND slope (2) AND thal (7) => num (1)
2.	fbs (0) AND thalach ([33,*)) AND slope ((1) AND ca ([*,1)) AND thal (3) => num (1)
3.	fbs (0) AND cal([1,*)) AND thal (7) =>(1)
4.	sex 1) AND fbs (0) AND thalach ([33,*)) AND excang (0) AND ca ([*,1)) AND thal (3) => num (1)
5.	sex (1) AND fbs (0) AND restecg (0)) AND oldpeak ([0.3,*)) AND thal (7) => num (1)

27.	Age ([53,*)) AND tresbps ([129,*)) AND restecg (0) AND excang (0) AND ca ([*,1)) => num (1)

Figure 5: Rules used in the research for the Analysis of the Model Performances [5]

In a research towards the determination of the impact of Hypertension into the heart problem [25], evidence that artificial intelligence (AI) is useful for predicting and administering risk factors for hypertension. But I am far from harnessing the ground breaking AI methods to predict these hypertension risk factors and apply them to personalized management. This study summarizes recent advancements in the field of computer science and medicine, highlighting the ground breaking AI method for future detection of hypertension in early stages. Furthermore, with an eye to personalized medicine, I have reviewed current research and future implications of AI in hypertension management and clinical trials. Although recent studies show that AI is feasible and possibly useful in research into hypertension, AI-informed care has yet to transform control of blood pressure (BP).

This is partly due to the lack of data on the consistency, accuracy, and reliability of the AI in the BP sphere. Nevertheless, several factors lead to poorly regulated BP including issues related to genetics, the climate, and lifestyle. AI offers insight into extrapolating data analytics to warn prescribers and patients about particular factors that could influence their control over BP.

To date, AI has been primarily used to examine risk factors for hypertension, but due to the limitations of study design and physician participation in computer science literature, it has not yet been used to control hypertensions. AI 's future of more flexible architecture using multi-omics methods and wearable technologies will undoubtedly be an important resource for integrating biological, behavioural, and environmental variables into the decision-making process for effective drug use to monitor BP.

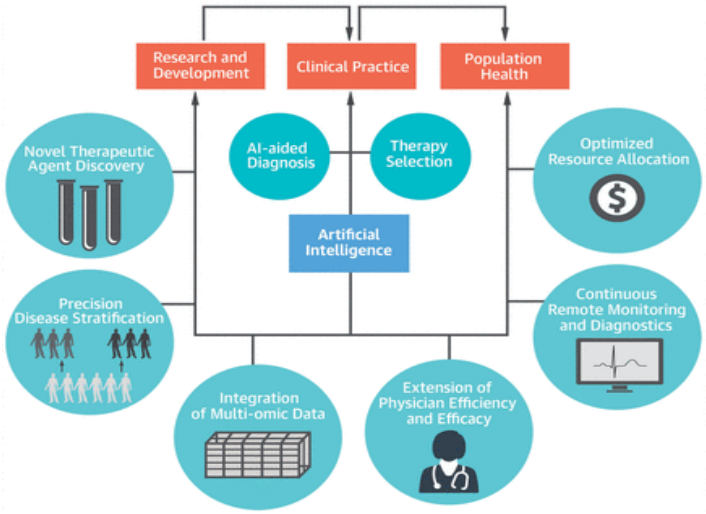


Figure 6: Monitoring of the Patients for the Heart Disease Identification [6]

Artificial intelligence and machine learning are capable of affecting almost any area of human life, and cardiology is no exception [26]. In this paper provides clinicians with a guide on specific aspects of artificial intelligence and machine learning, reviews selected implementations of these approaches in cardiology to date, and discusses how artificial intelligence could be integrated into cardiovascular medicine in future. The paper reviews, in particular, predictive modelling concepts relevant to cardiology such as selection of features and frequent pitfalls such as improper dichotomisation. Second, common algorithms used in supervised learning are discussed and selected implementations are reviewed in cardiology and related disciplines.

Third, it describes the emergence of deep learning and related methods collectively known as unsupervised learning, offers contextual examples in general medicine as well as in cardiovascular medicine, and then explains how these methods may be applied to allow cardiac precision and improve patient outcomes.

In a similar research, cardiovascular medicine, AI techniques have been applied to explore novel genotypes and phenotypes in existing diseases, improve the quality of patient care, enable cost-effectiveness and reduce readmission and mortality rates [27].

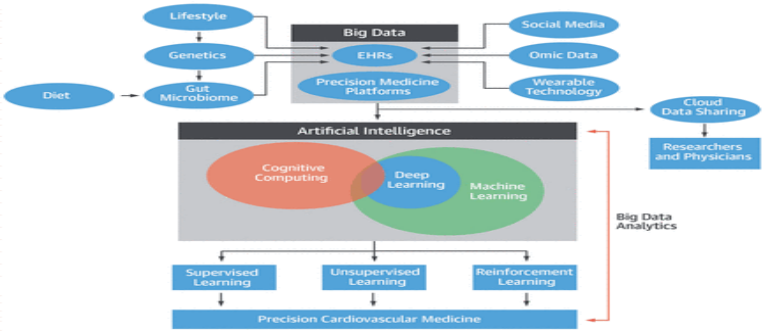


Figure 7: Big Data Analytics Application for the Patient Monitoring and Prediction Detection[9]

Several machine-learning techniques were used for the diagnosis and prediction of cardiovascular diseases. To apply the optimal machine-learning algorithm, each problem requires some degree of understanding of the problem, both in terms of cardiovascular medicine and statistics. AI can contribute to a paradigm change towards cardiovascular precision medicine in the near future. AI 's potential is tremendous in cardiovascular medicine; however, ignorance of the challenges can overshadow its potential clinical effects. This paper provides an overview of the application of AI in cardiovascular clinical care, and discusses its potential role in facilitating cardiovascular medicine with precision. In a study a comparison of different approaches with the Cardiovascular Health Study (CHS) data set approach was done for stroke prediction [28]. Here, the decision tree algorithm is used for the process of selection of features, the main component analysis algorithm is used to reduce the dimension, and the neural network classification algorithm adopted back propagation is used to construct a classification model.

This work has the optimum predictive model for the stroke disease with 97.7 per cent accuracy after analysing and comparing classification efficiencies with different methods and variation models accuracy. Clinical diagnosis is mostly done through the expertise of a doctor and patients were asked to take no diagnostic tests [29]. But not all of the research can lead to successful disease diagnosis. Selection of the subset of features is a pre-processing step used to reduce the dimensionality and eliminate irrelevant data. In this paper we present a classification method that uses ANN and selection of subset features to classify heart disease. PCA is used to pre-process and to decrease no. Of attributes which indirectly reduce the number of diagnostic tests that a patient needs to take. We applied our approach to the database on heart disease in Andhra Pradesh. Our experimental findings indicate that the accuracy has improved over conventional methods for classification.

For the treatment of heart disease this method is feasible and faster and more precise. In a research, Artificial Neural Networks (ANN) algorithms have been developed to diagnose heart disease from Phonocardiogram (PCG) signals using an artificial intelligence system [30]. Four new featured signal characteristics, namely operation, intensity, mobility and spectral peaks from the power spectral density plots are used as inputs into the neural network. In this analysis, 94 PCG signals were used to test the accuracy of the neural networks for three heart diseases. After filtering the signals and extracting the feature properties, the features are fed to the neural networks. In this, classification is carried out using the techniques of the Radial Base Function (RBF) network and the Back Propagation Network (BPN). The Operating Characteristic of the receiver (ROC) is determined to measure the

consistency of both systems. The results show that , compared with 90.8 percent for BPN, RBF received 98 percent accuracy in predicting the disease.

The developed artificial intelligence algorithm has proven to be a effective technique for automated heart disease diagnosis using PCG signals. Since, Cardiac disease is the most severe condition. This disease is very popular now a few days we have used various attributes that can well apply to these heart diseases to find the right way to predict and in this research has also used predictive algorithms [31]. Naive Bayes, algorithm based on risk factors is evaluated on dataset. In this paper, researchers used decision trees and combination of algorithms to predict heart disease, based on the attributes listed above. The results showed that the correct results are given when the dataset is small naive Bayes algorithm and when the dataset is large decision trees give the same results. The proposed model and algorithm is as shown below:

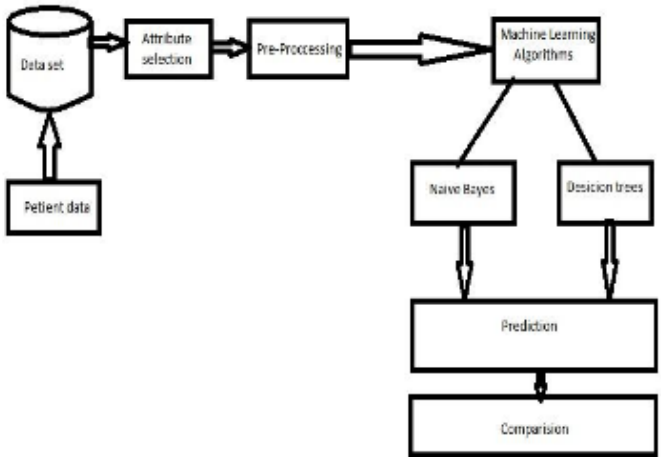


Figure 8: Hybrid of Naïve Bayes and Decision Tree Classifier for the Heart Detection System [12]

In a research proposing a mixed solution, suggested applies to Data set for heart disease; findings indicate the effectiveness and robustness of the hybrid approach proposed in Data analysis of different forms for the diagnosis of heart disease [36]. This thesis investigates therefore the different machines Training algorithms and comparing results using different performance metrics, i.e. precision , accuracy, recall, rate f1 etc. Total classification accuracy of 99.65 per cent using the optimized FCBF, PSO model and instead ACO. The results show that the performance of the system proposed is greater than that of the classification Technique discussed above.

CHAPTER 3: METHODOLOGY

3.1 Steps

In this section I will be discussing about the methods that can be applied for the achieving the solution for the predicting the heart problem in a human body. The whole of the approach can be done as based on the following flow chart:

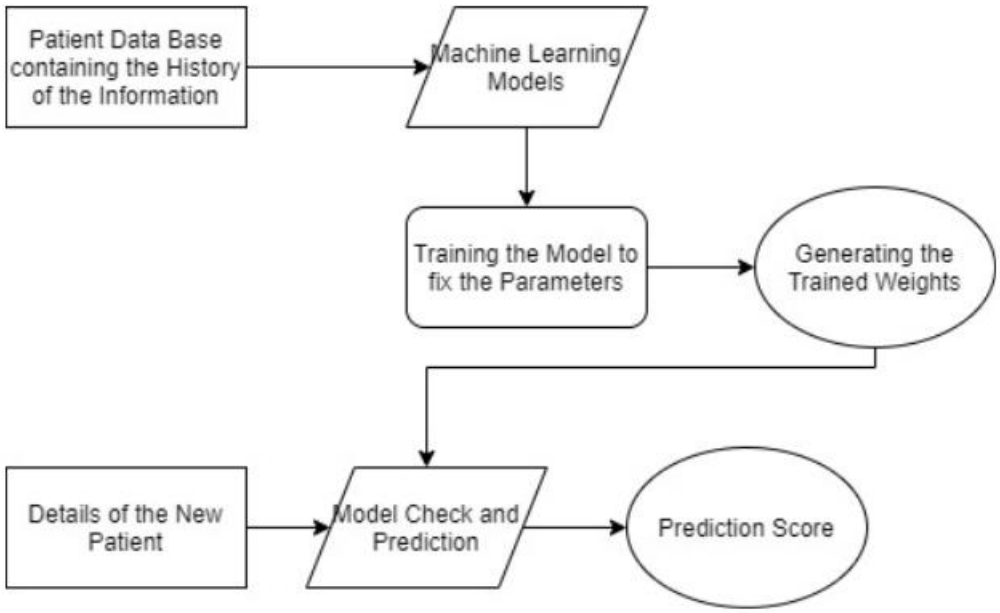


Figure 3: Proposed Architecture for the determination of the heart problem in a patient based on the history of the data available.[24]

3.2 Algorithms Description

Step 1: The patient History data is stored in some kind of database which can be accessible enough for the training the model and getting the best results out of it

Step 2: In order to predict in the healthcare system, there cannot be any secondary thoughts in term of the model performance. Hence the best performing model needs to be hired for the predication. Hence, to fix the model, a lot of algorithm needs to be passed in through this framework to validate the model performance and check

Step 3: The trained weights are then stored which can be used for the check of the performance of the model during the real time analysis period

Step 4: Real time analysis of the real patient data. The error that is generated should be passed on to the training part to re-tune the model parameters in order to have a better accurate results.

I will be using the following classification algorithms to train and test the data.

1. Logistic Regression: In the presence of more than one explaining variable, logistic regression is used for obtaining an odds ratio. The method is somewhat similar to a multiple linear regression, except for the binomial response parameter. The result is the impact on the odds ratio of the event of interest observed by each variable. The main advantage is that all variables are analysed together to prevent confusing effects. We present in this post, use illustrations to clarify the logistic regression method. If the procedure is established, it stresses the fundamental understanding of the findings and several particular problems are then addressed.

2. KNN (K-Nearest Neighbours): KNN can be used both for statistical problems of classification and regression. But in classification problems in the industry, it is more widely used.

3. Naïve Bayes : It is simple but very powerful algorithm for predictive modelling. It can be states as P(h|d) = (P(d|h) * P(h)) / P(d). Where P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability. P(d|h) is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

4. Linear SVC (Linear Support Vector Classifier): The aim of a linear SVC is to conform with the information you provide and return a hyperplane "best match" that divides or classifies your data. You can then feed your classifier some features after you have obtained the hyperplane to see what the "predicted" class is.

5. SGD (Stochastic Gradient Descent): The term 'stochastic' implies a random probability-linked mechanism or operation. Therefore, a few samples are selected alone in Stochastic Gradient Descent instead of the entire data set for each iteration. There is a term called "batch" in Gradient Descent, which indicates the total number of data samples used in each iteration to calculate the gradient. The sample is the whole dataset of traditional downward optimization of the gradient, including Sample Gradient Descent. However it is very useful to use the entire dataset to achieve the minimum in a less noisy and uncertain way but when our datasets get larger, the problem arises.

$$for\ i\ in\ range\ (m):$$
$$\theta_j = \theta_j - \alpha\ (\hat{y}^i - y^i)\ x_j^i$$

6. Decision Tree Classifier: A decision tree is a tree structure with a fluctuation chart where the internal node is the function(s) and the branch is a decision rule, with each node representing the outcome. The decision tree The highest node is considered the core node of a decision tree. You can use the attribute value to partition. It recurringly partitions the tree to call recursive partitioning. This fluid diagram helps you make choices.

7. Gradient boost Algorithm: Gradient boosting is a regression and classification technique which produces a prediction model, usually decision trees in the form of a collection of weak prediction models. It constructs this model in a manner close to other boosting approaches and generalizes it by requiring an undefined differentiable loss function to be optimised.

3.2 Dataset

The data that will be used is shared by a start-up company located in Mumbai known as ‘A’ Labs. The data set consist of details of patients history data with different risk factors such as different demographic details of the patients such as gender, age, blood pressure and different risk factors such as hyper tension, fasting glucose, dyslipidaemia. This data will be considered as the out of sample data and the whole system to be fixed based on some of the open source data [9].

3.3 Data Pre Processing Techniques

Various data prepossessing techniques and data imputation techniques needs to be hired to make the dataset a proper one to have it input to the modelling and hence to carry out the research work. Some of the common techniques are missing values imputation, outliers treatment, data scaling, polynomial features generation, One hot encoding etc. In case of the data imputation, since the dataset is the real time data and of healthcare domain, the important part to be is consulting a domain expert to help in case of imputation rather than imputing it in terms of the statistical measures.

3.4 Modelling

For the analysis and getting the predictions right some of the best methods in the modelling like the SVM, Decision Trees, Logistic Regression, KNN, Guassian Naïve Bayse, Linear SVC, SGD, Decision Tree Classifier, Gradient Boosting Trees were used and get the best methods out the packs. In this case, the model with different kinds of the hyper parameters setting were put into the hyper parameters optimisation algorithms like the Grid Search and Random Search Algorithms. The best algorithm setting is done with the help of the selected hyper parameters through the best hyper parameter optimization setting. Among the whole process, the best setting models are then go through the prediction process and using the results the best model is selected. The best model with the hyperparameters is saved and put into the real test where the out of the samples data are passed through them to check the accuracy and the performance in the real time. Also the analysis and the recommendation models help the doctors to diagnosis better. Unlike many of the papers and researches online, we will not be competing the doctor rather we will be helping the doctor to have a proper diagnosis with the decreasing the turnaround time for each of the patients diagnosis.

CHAPTER 4: IMPLEMENTATION

4.1 The Data Set

The data set that I am using this research paper is the patient's data set from different location of India with count of 1095 records. During this study, about 1095 patients were examined and the key factors which may be the root cause of the heart attacks were identified .. This study aims to pose a job that can be an imminent move towards reaching a possibility of the heart issue. The specific risk factors are the key causes why the cardiac condition exists. Various classification frameworks have been tested to describe the central issue in this study. Data obtained was gathered from five separate cities and classes of age in India.

Risk parameters which are considered in this dataset are Family history, Smoking, Hypertension, Fasting Glucose, Obesity, Lifestyle, CABG, High Serum. The records are in binary where 1=Yes and 0=No. Along with this there are other three data columns Age and Sex and location.

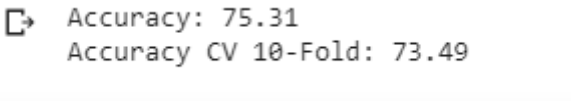
4.2 Data Processing

For this study, I used various methods for data analysis to clean and explain the results. Specific methods like pre-possession and data imputation would be employed in order to have the data collection the correct one to be included in the modelling and also for the inquiry. Some popular techniques involve imputing the missing values, processing outliers, scaling of results, generation of polynomial functions, one hot encoding. Once the data is imputted, provided that the dataset includes the real-time data and health care context, it's essential to consult a domain authority instead of imputing it on mathematical methods to support you imput it. In Data Processing data frames are now classified in features X and Y. The location column has been converted to categorical data. I also converted Sex column to a categorical data.

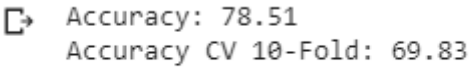
4.3 Model

I worked on multiple algorithms to find out which is a more accurate to detect a heart problem providing an available data. Algorithms such as logistic regression, KNN, GNB, Linear SVC, Stochastic Gradient Descent, Decision Tree Classifier, Gradient Boosting Tree, XGBoost, Grid search CV. The intention was to find out which is more accurate reason being heart is one of the most important body part of human body and less accurate prediction about this can leads to the death of

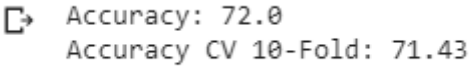
the patient. In the first segment of model coding I found a correlation between the features. I have also implemented a function which return the accuracy of the model which helps us to differentiate between different algorithms. In the implementation of logistic algorithm, The accuracy is 75.31 and accuracy CV 10-Fold is 73. 49. Output of logistic regression is as follows.



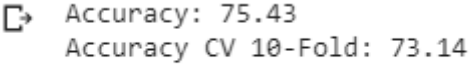
The second algorithm which I have implemented is KNN. As explained KNN in methodology segment, This algorithm returned 78.51 accuracy and CV 10-Fold accuracy is 69.83 a shown in below output image.



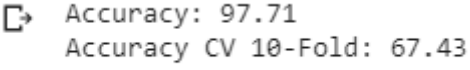
Gaussian naïve bayes model is one of the classification model which I have used to find if this is more accurate. After implementation I found that accuracy rate of this algorithm is 72.0 and Cv-Fold is 71.43 as per below image from output window.



Linear SVC is another algorithm which I have implemented which is with accuracy of 75.43 and with CV-10 Fold accuracy is 73.14. Please see below snapshot from output from google colab.



Stochastic Gradient Descent is one of the algorithm which I have implemented with accuracy and CV 10-Fold accuracy 69.94 and 70.63 respectively. Please check below output image.



I have also implemented Decision Tree Classifier which showed a best results with accuracy of 97.71 but CV 10-Fold accuracy is not something which we can consider which is 67.43 as below image.

```
Accuracy: 97.71
Accuracy CV 10-Fold: 67.43
```

Gradient Boosting Tree is our algorithm. During my implementation I found it more accurate than other algorithms with 81.37 accuracy rate and with 75.2 CV 10-Fold accuracy.

```
Accuracy: 81.37
Accuracy CV 10-Fold: 75.2
```

XG Boost is another algorithm of boosting regression family which I have implemented to find accuracy which turned to be 73.51.

```
Accuracy = 0.7351598173515982
```

I also implemented some optional code for hyper parameter tuning for XG Boost using GridSearchCV. Which also turned to be same as XgBoost with same 73.51. See below output from google colab.

```
Accuracy = 0.7351598173515982
```

4.3 Hardware and Software

In implementation of this research paper I used google colab which is platform which provides virtual GPU and CPU to perform your coding in python which more faster than out legacy platform. This is completely cloud based platform which needs no setup to perform your activity. Developer no need to install any thing to write a code.

Compared to the other local platforms I found this quit fast and easy to use and import thing that I worked on my project without my own laptop as your code with stored in cloud you can use it wherever you are whenever you need it. Google Colab : https://colab.research.google.com/

CHAPTER 5: RESULTS AND ANALYSIS

In order to do the analysis of the different machine learning models and algorithms, various metrices can be used to analyse better. Some of them are ROC curve, Confusion metrices, accuracy etc. The algorithm which is stable and performing best in this process can be used for the model fixing.

Results of different data processing code segment and algorithm is as follows.

5.1 Data Processing

First thing first removal of irrelevant data. I dropped a column ID which is Irrelevant.

```
[ ] df.drop('ID',axis=1,inplace=True)
```

```
df.head()
```

	Age	Sex	location	RF1(Family History)	RF2(Smoking)	RF3(Hypertension)	RF4(Dyslipedimia)	RF5(Fasting Glucose)	RF6(Obesity)	RF7(Li
0	48	Male	Delhi	1	1	0	0	0	0	
1	61	Male	Delhi	0	0	0	0	0	0	
2	54	Male	Delhi	0	0	1	0	1	0	
3	40	Female	Delhi	0	0	1	0	0	0	
4	58	Male	Delhi	1	0	1	1	1	1	

Fig. Output 5.1.1

We removed column ID as we do not need it in research for further observation.Another step which I did is a removal of ambiguity such small letter and capital letter are same for instance delhi and Delhi (City in India) and for that I performed below.

```
[ ] df["location"].replace({"delhi": "Delhi"}, inplace=True)
```

```
[ ] df.location.value_counts()
```

```
Chennai      599
Kolkata      302
Delhi        125
Hyderabad     35
Bangalore     33
Name: location, dtype: int64
```

Fig. Output 5.1.2

I also performed a visualization based on location of and resulted output looks as follows.

```
[ ]  plt.figure(figsize=(10,7))
     df['location'].value_counts().plot(kind='barh',cmap='RdYlBu');
```

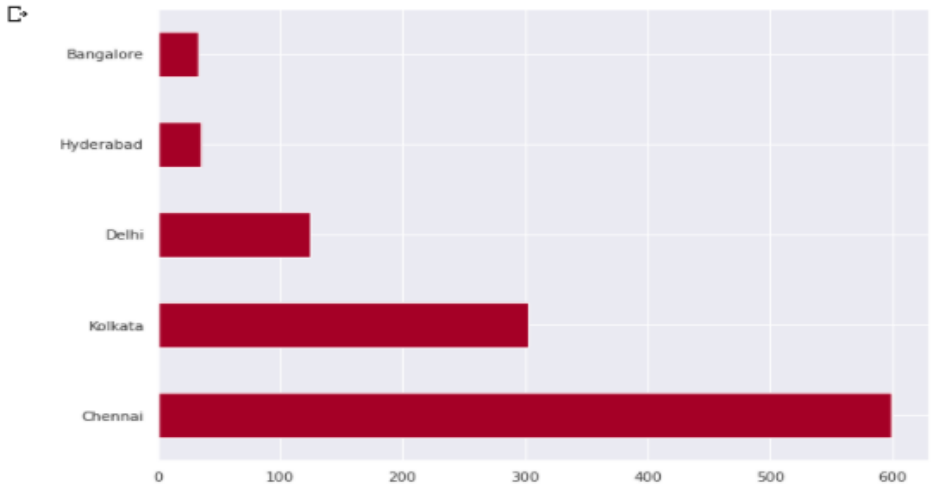


Fig. Output 5.1.3

I did a renaming of columns for better understanding of data. I have added RF a abbreviation for risk factor as shown in below output snapshot.

E	F	G	H	I	J	K	L	M
RF1(Family Histc	RF2(Smoking)	RF3(Hyperter	RF4(Dyslipec	RF5(Fast	RF6(Obesity)	RF7(Lifest	RF8(CAB	RF9(High Seru

Fig. output 5.1.4

I also tried to find a missing value from data which and there was no any missing value which can be shown by a graph as per below output image.

```
missingno.matrix(df)
<matplotlib.axes._subplots.AxesSubplot at 0x7f25daffe748>
```

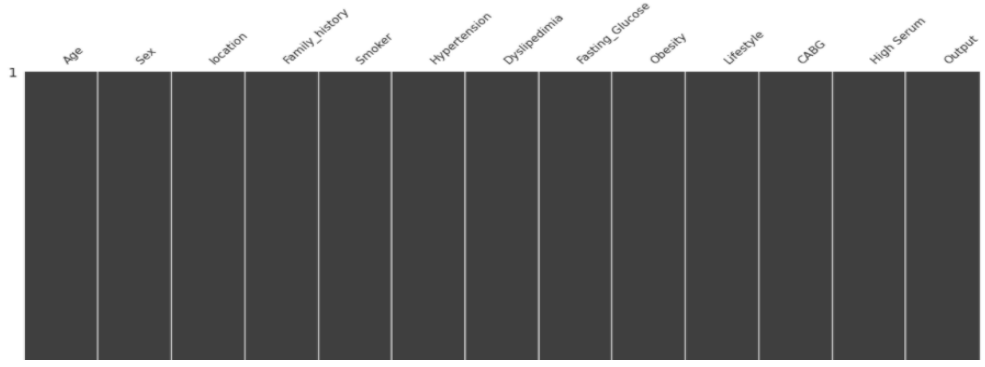


Fig. Output 5.1.5

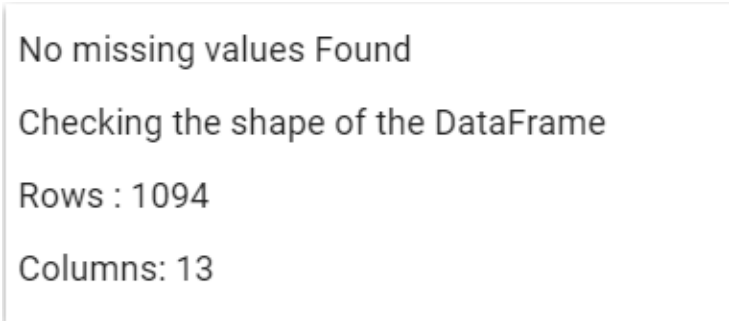


Fig. Output 5.1.6

5.2 Variable Identification

On this section I split the dataframe into the features X and Y.

	Age	Sex	location	Family_history	Smoker	Hypertension	Dyslipedimia	Fasting_Glucose	Obesity	Lifestyle	CABG	High Serum
0	48	Male	Delhi	1	1	0	0	0	0	1	0	1
1	61	Male	Delhi	0	0	0	0	0	0	1	0	1
2	54	Male	Delhi	0	0	1	0	1	0	1	0	1
3	40	Female	Delhi	0	0	1	0	0	0	1	0	1
4	58	Male	Delhi	1	0	1	1	1	1	1	0	1

Fig. Output 5.2.1

After that I converted location column to categorical data using one hot encoding.

	Age	Sex	location	Family_history	Smoker	Hypertension	Dyslipedimia	Fasting_Glucose	Obesity	Lifestyle	CABG	High Serum	0	1	2	3	4
0	48	Male	Delhi	1	1	0	0	0	0	1	0	1	0.0	0.0	1.0	0.0	0.0
1	61	Male	Delhi	0	0	0	0	0	0	1	0	1	0.0	0.0	1.0	0.0	0.0
2	54	Male	Delhi	0	0	1	0	1	0	1	0	1	0.0	0.0	1.0	0.0	0.0
3	40	Female	Delhi	0	0	1	0	0	0	1	0	1	0.0	0.0	1.0	0.0	0.0
4	58	Male	Delhi	1	0	1	1	1	1	1	0	1	0.0	0.0	1.0	0.0	0.0

Fig. Output 5.2.2

After that we have column called SEX which I have converted to categorical data and results are as follows.

```
[ ]  X.Sex.value_counts()

     Male      735
     Female    359
     Name: Sex, dtype: int64
```

Fig. Output 5.2.3

Along with that I also tried visualize distribution of SEX using bar graph to identify how many females and Male we have in out data set. Please find below output image.

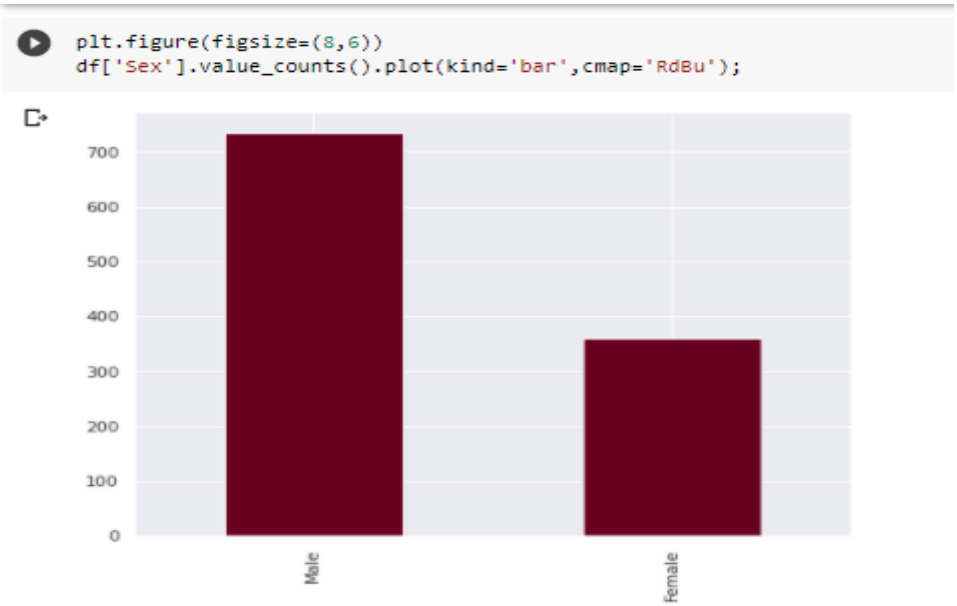


Fig. Output 5.2.4

Along with that I also distributed Age which can be seens in below output image.

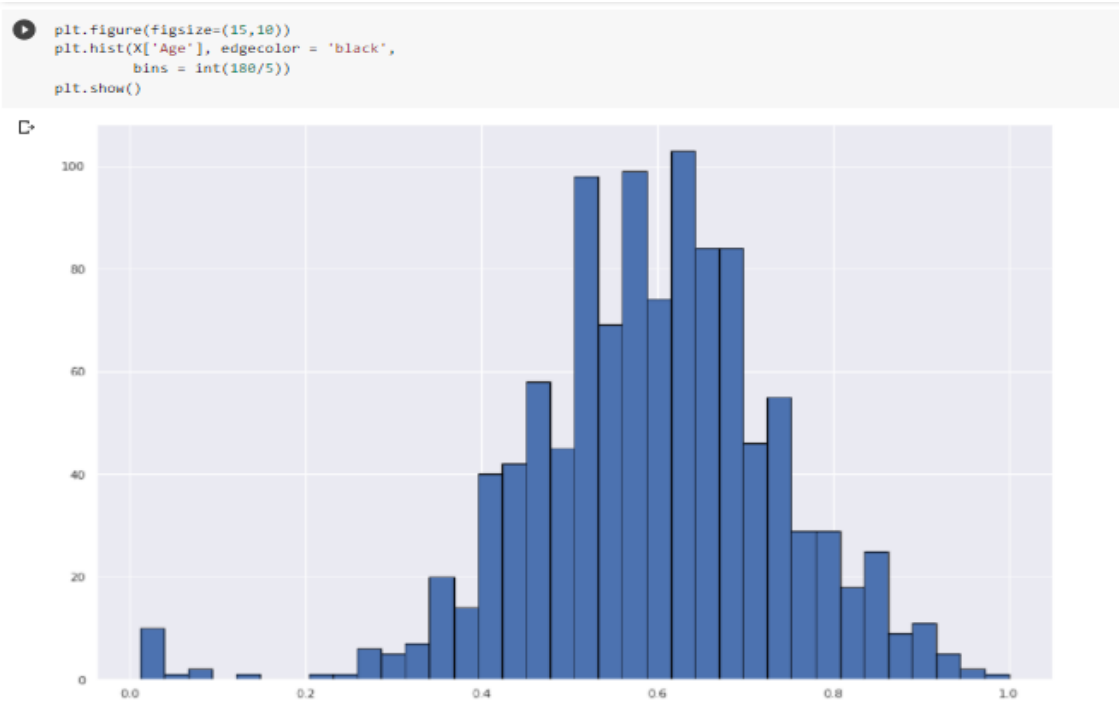


Fig. Output 5.2.5

Distribution based on family history as per below bar graph.

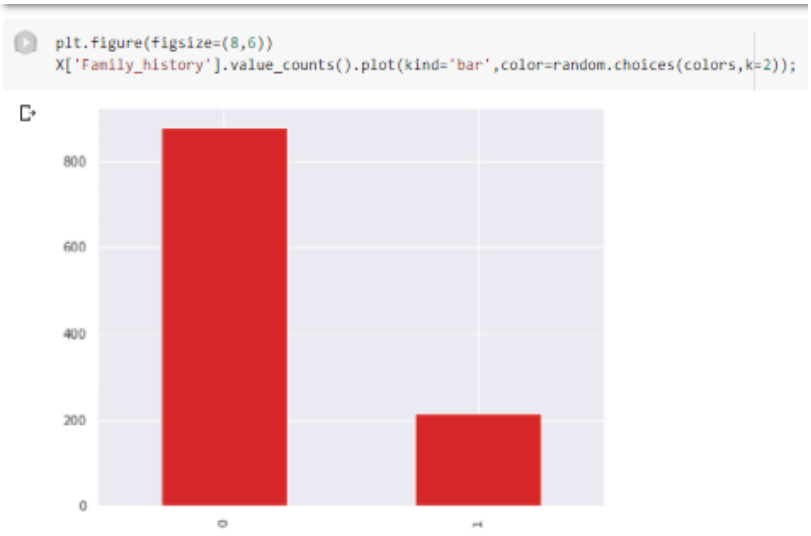


Fig. Output 5.2.6

Distribution based on Smoker Risk factor.

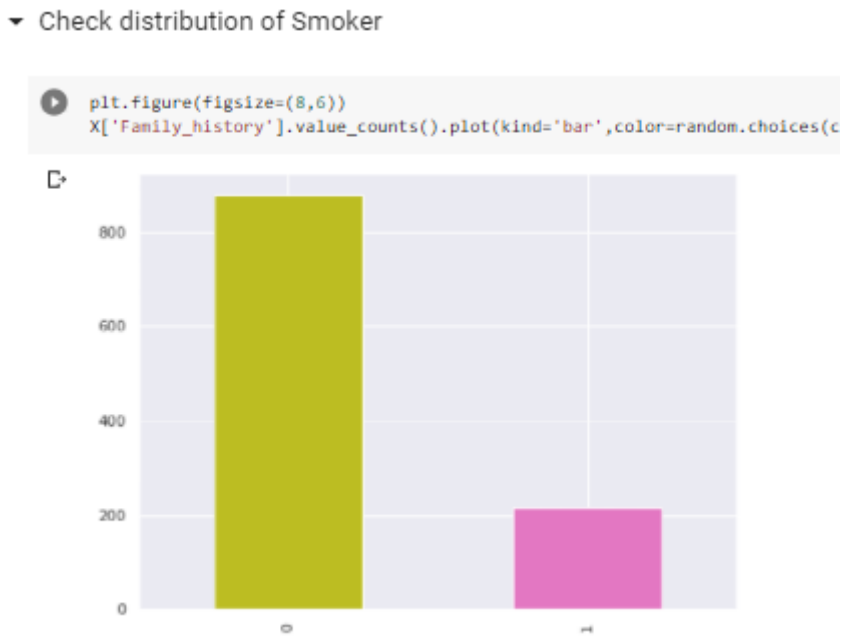


Fig. Output 5.2.7

Hypertension is one of the another risk factor which can be considered as important factor for the heart failure. I also tried to show distribution based on hypertension which is as follows.

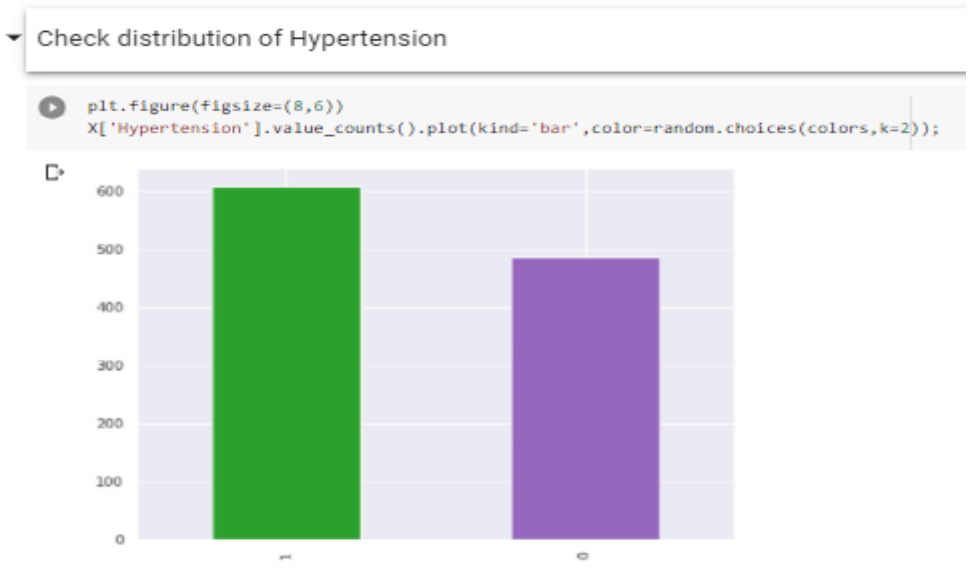


Fig. Output 5.2.8

Dyslipedimia is one of the factor which I considered for this research and the distribustion for this is as follows.

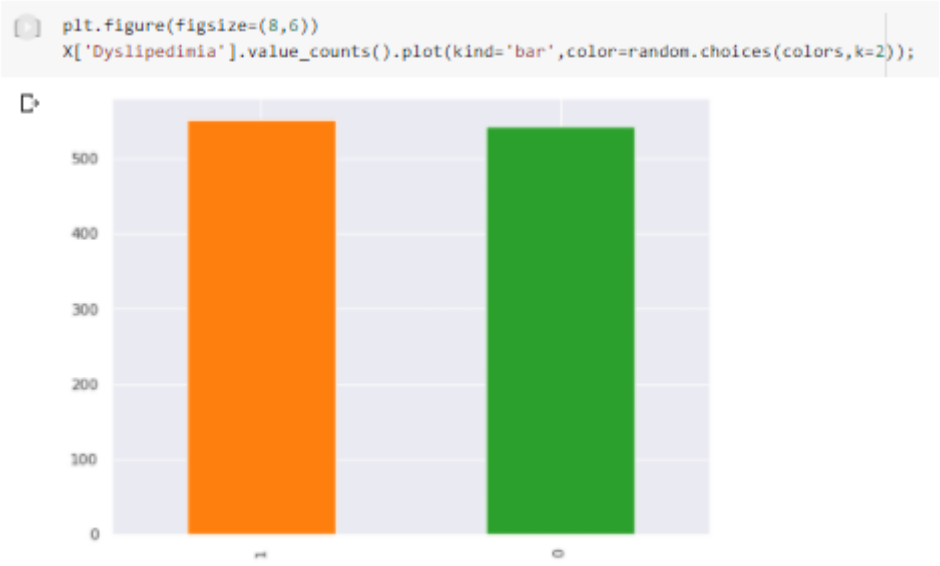


Fig. Output 5.2.9

Another distribustion is for fasting glucose which can be seen in below snapshot.

Fig. Output 5.2.10

Life style is another risk factor which can be considered to predict a heart problem and distribution for that is as follows.

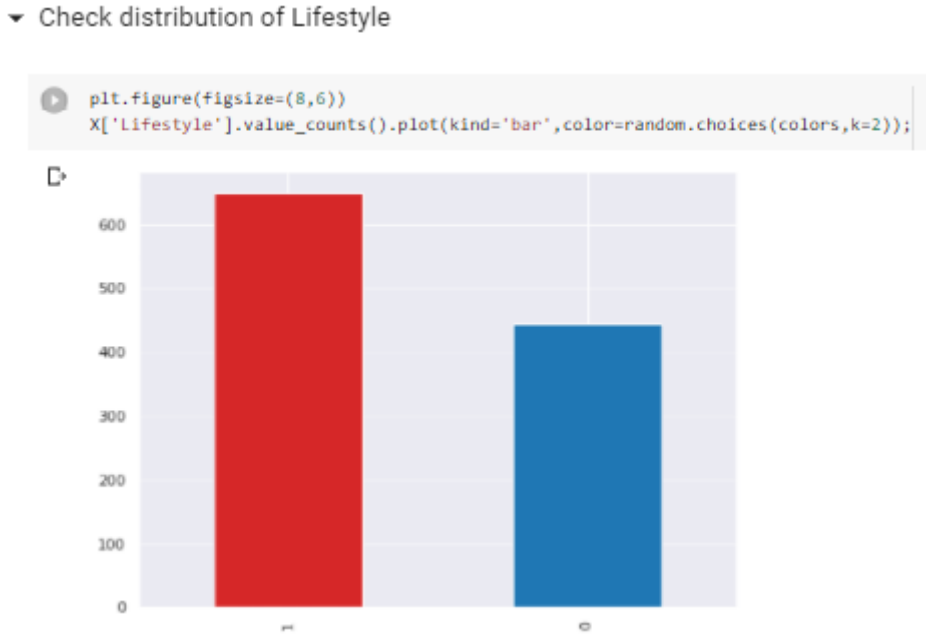


Fig. Output 5.2.11

Obisity is another risk factor with distribustion in bar graph as follows

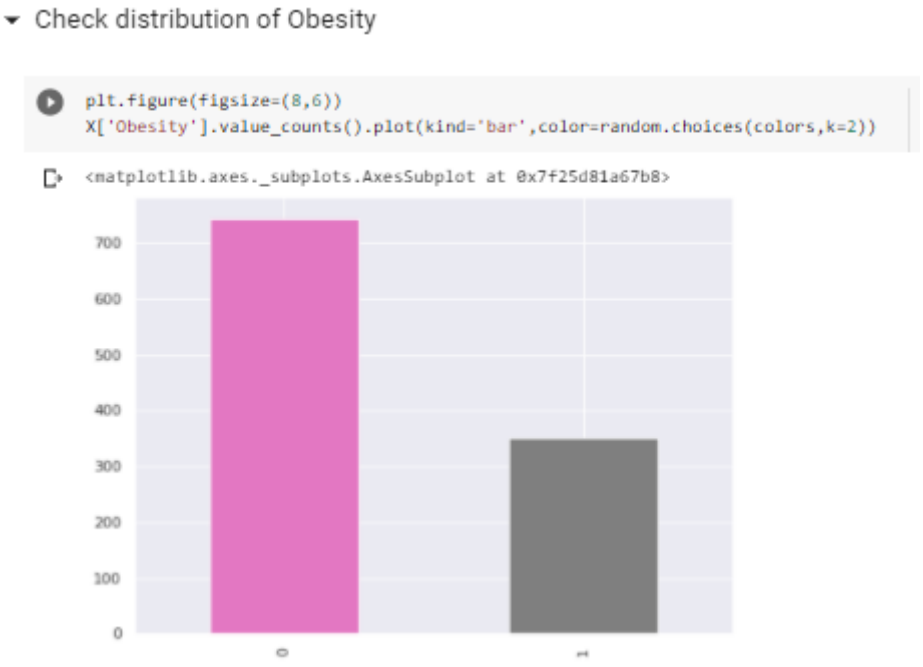


Fig. Output 5.2.12

Distribution for CABG is as follows.

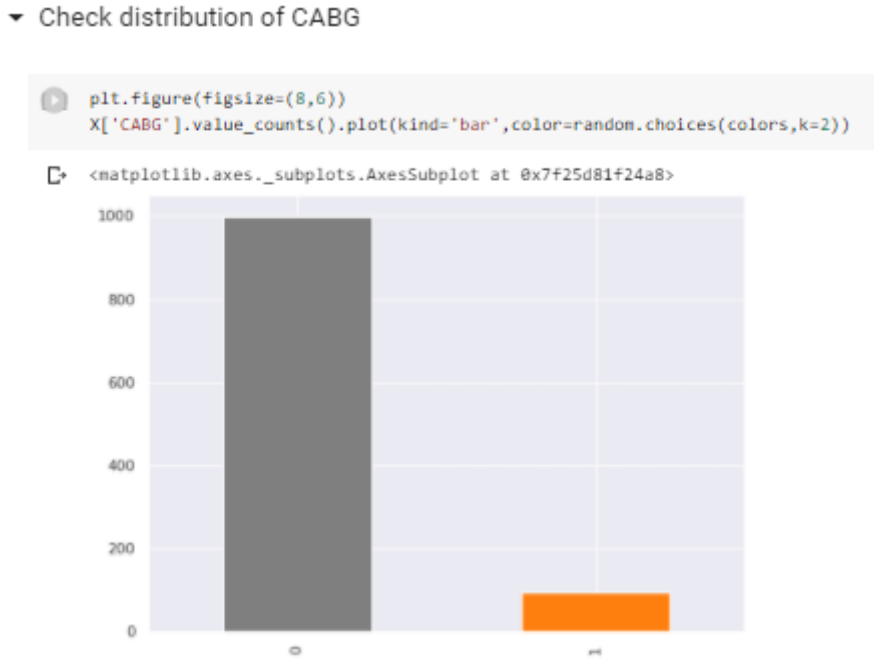


Fig. Output 5.2.13

The last risk factor for my research is high serum and distribustion is as follows.

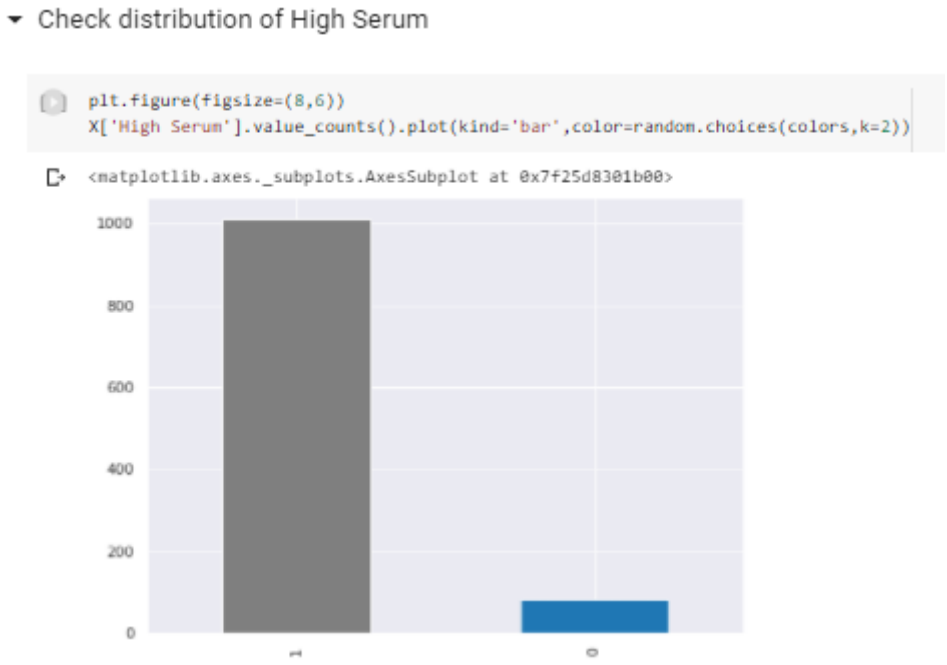


Fig. Output 5.2.14

The last thing which I did for the data processing is splitting data into train and testing data. And correlation can be seen in below visualization.

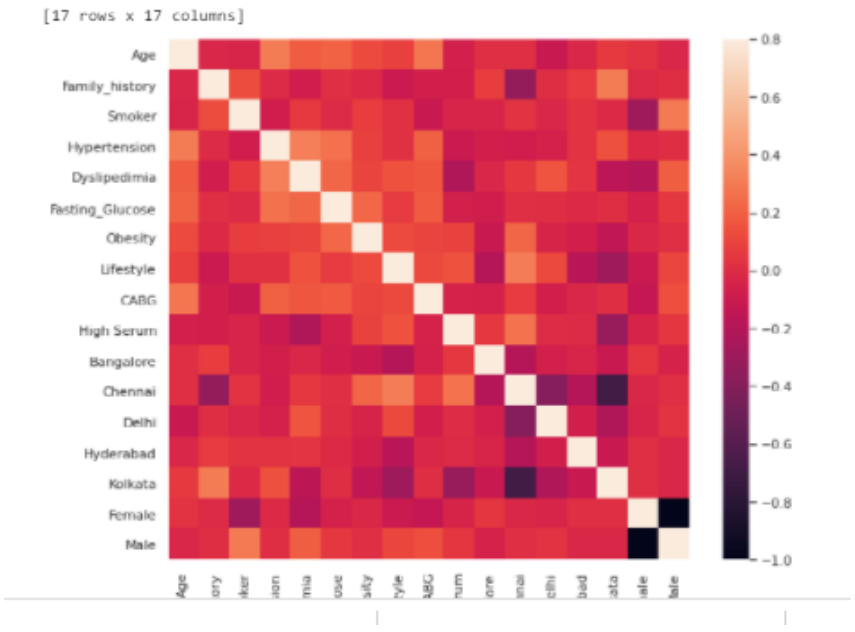


Fig. Output 5.2.15

5.3 Implemented Algorithms

As already described in methodology which are the algorithms which I have used to do this research outputs of that algorithms are described in this section.

1) Logistic regression: Logistic regression is used to produce an likelihood ratio in the case of more than one explanatory variable. The approach is much like a multi linear regression, but for the parameter binomial answer. The effect is the influence of each vector on the likelihood ratio of the occurrence of concern. The biggest value is that the study of all factors is performed together to eliminate confounding results. Throughout this essay, we describe the logistic regression mechanism using diagrams. Throughout the development of the method, it stresses a basic interpretation of the findings and solves some particular issues.

Fig. Output 5.3.1

2) KNN : I also applied the second method, KNN. This algorithm has 78.51 accuracy and CV 10-fold accuracy, as seen in the following output graphic, as KNN discussed in the methodology section.

Fig. Output 5.3.2

3) GNB : Gaussian naïve bayes model is one of the classification model which I have used to find if this is more accurate. After implementation I found that accuracy rate of this algorithm is 72.0 and Cv-Fold is 71.43 as per below image from output window.

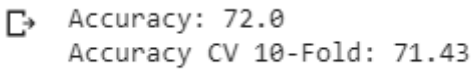


Fig. Output 5.3.3

4) Linear SVC :

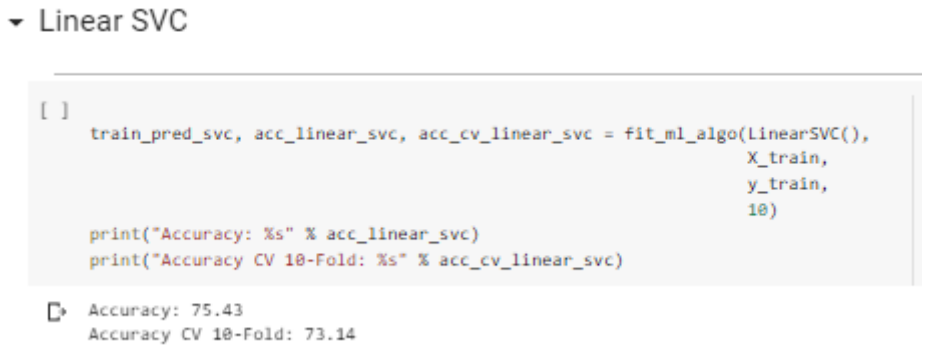


Fig. Output 5.3.4

5) Stochastic Gradient Descent

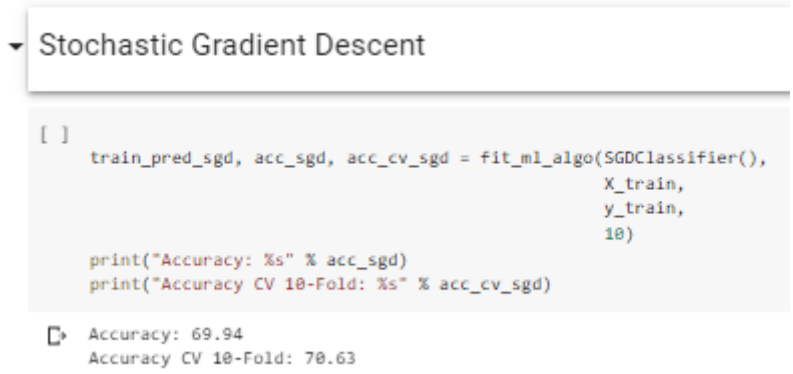


Fig. Output 5.3.5

6) Decision Tree Classifier

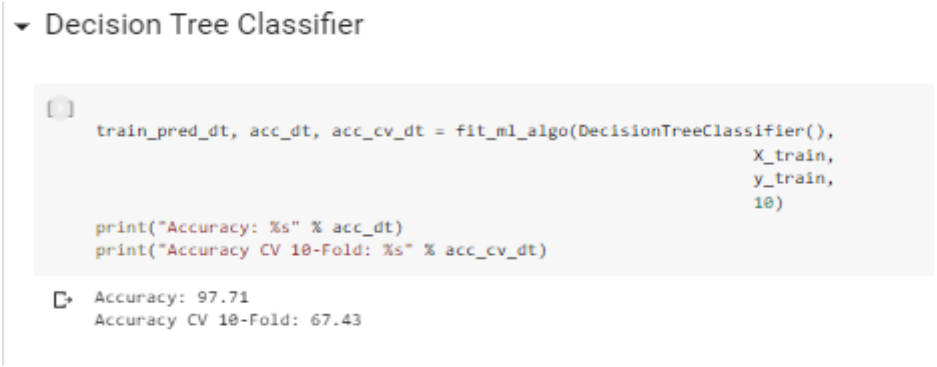


Fig. Output 5.3.5

7) Gradient Boosting Tree

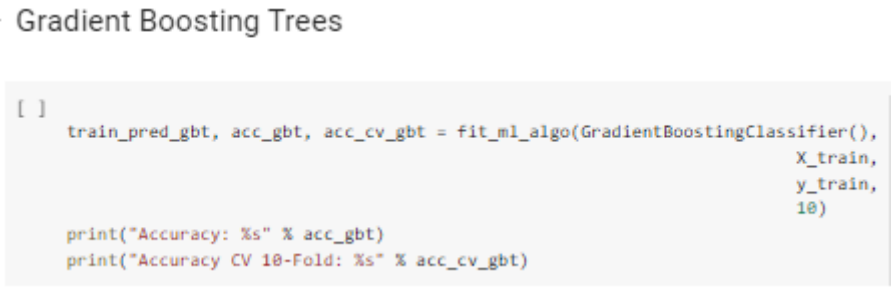


Fig. Output 5.3.6

8) XG Boost

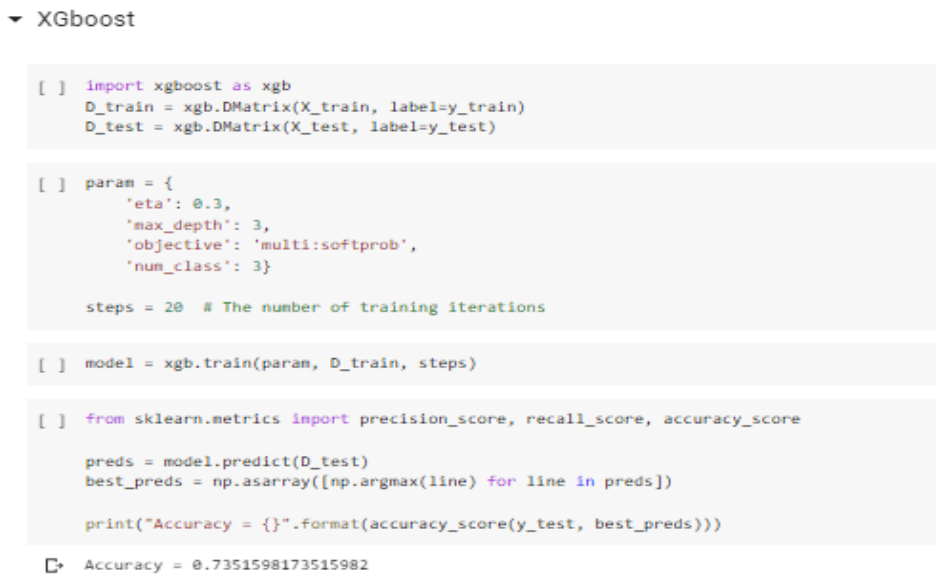


Fig. Output 5.3.7

CHAPTER 6: CONCLUSION

6.1 Conclusion:

Heart problem is one of the major reason od death. As per World Health Organization 17.9 million people die every year in which 80% deaths are due to Heart problems. The intension of this research is use technology like machine learning which can help us to find a heart related issue in prior stage which can be be treated and a person may get another chance to live.In a technical perspective machine learning helped me to find which is the best suited algorithm can be used in pharmaceutical industry which can help to develop a devise or software system which detect heart related issue on early stage. I observed that XGBoost is one of the algorithm which can used to test a human body for heart related issue depending on the provided risk factors.

I found XGboost a much more accurate algorithm in my research. It doesn't mean that other algorithms are not useful, Indeed they help me to understand the probability of whether to use them or not and health is something that we cannot negotiate with it so using a correct algorithm or a mechanism which show most correct result is need of hour.

On the basis of my research I want to conclude that XGBoost algorthim or model can be used to detect a heart related issue based on the given parameters.

6.2 Future Work:

I believe this research can be extended to the further level where we can used deep learning algorthims which can be used to find more accurate results.

References

[1] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, CHENNAI, India, 2019, pp. 1-5.

[2] Nichenametla, Rajesh & Maneesha, T. & Hafeez, Shaik & Krishna, Hari. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering and Technology(UAE). 7. 363-366. 10.14419/ijet.v7i2.32.15714.

[3] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2018, pp. 1275-1278.

[4] Adhikari, Nimai Chand Das. "Prevention of heart problem using artificial intelligence." *International Journal of Artificial Intelligence and Applications (IJAIA)* 9.2 (2018).

[5] Anooj, P. K. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules." *Journal of King Saud University-Computer and Information Sciences* 24.1 (2012): 27-40.

[6] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems, 2018*.

[7] S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," *2013 IEEE Conference on Information & Communication Technologies*, Thuckalay, Tamil Nadu, India, 2013, pp. 1227-1231.

[8] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2018, pp. 1275-1278.

[9] https://www.kaggle.com/ronitf/heart-disease-uci

[10] Kolmannskog, S., Moe, P.J. and Anke, I.M., 1979. Computed tomographic findings of the brain in children with acute lymphocytic leukemia after central nervous system prophylaxis without cranial irradiation. *Acta Pædiatrica*, *68*(6), pp.875-877.

[11] Ochs, J.J., Berger, P., Brecher, M.L., Sinks, L.F., Kinkel, W. and Freeman, A.I., 1980. Computed tomography brain scans in children with acute lymphocytic leukemia receiving methotrexate alone as central nervous system prophylaxis. *Cancer*, *45*(9), pp.2274-2278.

[12] Diaz, Marco N., et al. "Antioxidants and atherosclerotic heart disease." New England Journal of Medicine 337.6 (1997): 408-416.

[13] Howard, G., Wagenknecht, L.E., Burke, G.L., Diez-Roux, A., Evans, G.W., McGovern, P., Nieto, F.J., Tell, G.S. and ARIC investigators, 1998. Cigarette smoking and progression of atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) Study. *Jama*, *279*(2), pp.119-124.

[14] Rodgers, Anthony, et al. "Blood pressure and risk of stroke in patients with cerebrovascular disease." Bmj 313.7050 (1996): 147.

[15] Iribarren, C., Tolstykh, I.V., Miller, M.K., Sobel, E. and Eisner, M.D., 2012. Adult asthma and risk of coronary heart disease, cerebrovascular disease, and heart failure: a prospective study of 2 matched cohorts. *American journal of epidemiology*, *176*(11), pp.1014-1024.

[16] Heart Protection Study Collaborative Group, 2004. Effects of cholesterol-lowering with simvastatin on stroke and other major vascular events in 20 536 people with cerebrovascular disease or other high-risk conditions. *The Lancet*, *363*(9411), pp.757-767.

[17] Gertler, Menard M., et al. ”Ischemic heart disease.” Circulation46.1 (1972): 103-111.

[18] Jensen, J.S., Feldt-Rasmussen, B., Strandgaard, S., Schroll, M. and Borch-Johnsen, K., 2000. Arterial hypertension, microalbuminuria, and risk of ischemic heart disease. *Hypertension*, *35*(4), pp.898-903.

[19] Drazner, M.H., 2011. The progression of hypertensive heart disease. *Circulation*, *123*(3), pp.327-334.

[20] Diamond, Joseph A., and Robert A. Phillips. ”Hypertensive heart disease.” Hypertension research 28.3 (2005): 191-202.

[21] Hsieh, N.C., Hung, L.P., Shih, C.C., Keh, H.C. and Chan, C.H., 2012. Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. *Journal of medical systems*, *36*(3), pp.1809-1820.

[22] Setiawan, N.A., Venkatachalam, P.A. and Hani, A.F.M., 2020. Diagnosis of coronary artery disease using artificial intelligence based decision support system. *arXiv preprint arXiv:2007.02854*.

[23] A. Ohrn, "Discernibility and rough sets in medicine: tools and applications," in Department of computer and information science. Trondheim: Norwegian University of Science and Technology, 1999, pp. 223.

[24] T. Agotnes, "Filtering large propositional rule sets while retaining classifier performance," in Department of Computer and Information Science: Norwegian Universtiy of Science and Technology, 1999, pp. 143.

[25] Krittanawong, C., Bomback, A.S., Baber, U., Bangalore, S., Messerli, F.H. and Tang, W.W., 2018. Future direction for using artificial intelligence to predict and manage hypertension. *Current hypertension reports*, *20*(9), p.75.

[26] Johnson, K.W., Soto, J.T., Glicksberg, B.S., Shameer, K., Miotto, R., Ali, M., Ashley, E. and Dudley, J.T., 2018. Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, *71*(23), pp.2668-2679.

[27] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M. and Kitai, T., 2017. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, *69*(21), pp.2657-2664.

[28] Singh, M.S. and Choudhary, P., 2017, August. Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 158-161). IEEE.

[29] Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2013. Classification of heart disease using artificial neural network and feature subset selection. *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, *13*(3), pp.4-8.

[30] Abdel-Motaleb, I. and Akula, R., 2012, May. Artificial intelligence algorithm for heart disease diagnosis using phonocardiogram signals. In *2012 IEEE International Conference on Electro/Information Technology* (pp. 1-6). IEEE.

[31] Rajesh, N., Maneesha, T., Hafeez, S. and Krishna, H., 2018. Prediction of Heart Disease Using Machine Learning Algorithms. *International Journal of Engineering & Technology*, *7*(2), pp.p363-366.

[33] William Carroll; G. Edward Miller, “Disease among Elderly Americans : Estimates for the US civilian non institutionalized population, 2010,” Med. Expend. Panel Surv., no. June, pp. 1–8, 2013.

[34] V. Kirubha and S. M. Priya, “Survey on Data Mining Algorithms in Disease Prediction,” vol. 38, no. 3, pp. 124–128, 2016.

[35] Sharma, H. and Rizvi, M.A., 2017. Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), pp.99-104.

[36] Khourdifi, Y. and Bahaj, M., 2019. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *Int. J. Intell. Eng. Syst.*, *12*(1), pp.242-252.