

CERTIFICATE

It is certified that the work contained in this thesis entitled “**HEART DISEASE PROGNOSIS TOOL USING MACHINE LEARNING**”, by **Abhinav Dubey** (Roll No. 1200449001), for the award of **Master of Technology** from Babu Banarasi Das University has been carried out under our supervision and that this work has not been submitted elsewhere for a degree.

Signature: _____

Ms. Hirdesh Varshney

Assistant Professor

Department of CSE

School of Engineering BBDU

Lucknow

Date: _____

Signature: _____

Mr. Neeraj Baishwar

Associate Professor

Department of CSE

School of Engineering

Lucknow

Date: _____

ABSTRACT

One of the leading causes of death around the world is heart disease. Medical practitioners cannot simply forecast it because it is a challenging process that necessitates experience and knowledge. As per a recent study by WHO, heart-related diseases are increasing. 17.9 million people die every year due to this. Data mining is a great developing technique that revolves around exploring and digging out significant information from the massive collection of data which can be further beneficial in examining and drawing out patterns for making business-related decisions. Talking about the Medical domain, implementation of data mining in this field can yield in discovering and withdrawing valuable patterns and information which can prove beneficial in performing clinical diagnosis. Machine Learning techniques have accelerated the health sector through multiple types of research. Various algorithms in machine learning will be employed to solve this problem. The approaches are k-Nearest Neighbor, Support vector classifier, Stochastic gradient descent, decision tree classifier, logistic regression, Random Forest, XGBoost, Gradient boosting machine, Adaboost, and Multi-layer perceptron classifier. The following variables are extracted from medical profiles to forecast the likelihood of heart disease in a patient: blood pressure, sex, blood sugar, chest discomfort, cholesterol level, age of the individual, and some other attributes. Given's heart disease prediction technology improves medical care while lowering costs. This investigation has provided us with valuable information that can aid in the prediction of heart disease patients.

The world has witnessed an exploding spread of cardiovascular diseases (CVD) and it has been contemplated as one of the major causes of death. Further, the lack of awareness about various influential factors of CVD limits its early diagnosis and treatment. Therefore, the American Heart Association provides and recommends the guidelines for effective prediction of CVD based on hypertension, cholesterol, diabetes, age, and smoking. Furthermore, the machine learning (ML) models have proved their effectiveness in identifying the hidden patterns of data and therefore, many reported literature works have employed ML techniques for the prediction of CVD.

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to Assistant Prof. Hirdesh Varshney, Dept. of Computer Science and Engineering, Babu Banarasi Das University, Lucknow, India for his generous guidance, help, and useful suggestions.

I express my sincere gratitude to Assistant Prof. Neeraj Baishwar, Dept. of Computer Science and Engineering, Babu Banarasi Das University, Lucknow, India, for her stimulating guidance, continuous encouragement, and supervision throughout the present work.

I also wish to extend my thanks to Dr. Manuj Darbari, Dr. Gaurav Kumar Srivastava, Mrs. Akanksha Singh, Miss Hina Rabbani, and other faculty members, and colleagues for attending my seminars and for their insightful comments and constructive suggestions to improve the quality of this research work.

I am extremely thankful to Dr. Praveen Kumar Shukla, HOD, Department of Computer Science and Engineering, Babu Banarasi Das University, Lucknow, India for providing me with infrastructural facilities to work in, without which this work would not have been possible.

ABHINAV DUBEY

BABU BANARASI DAS UNIVERSITY

TABLE OF CONTENTS

	Page No.
Candidate's Declaration	i
Abstract	ii
Acknowledgment	iii
List of Figures	vi
List of Tables	viii
Nomenclature	ix
Chapter 1: INTRODUCTION	1-3
Chapter 2: LITERATURE REVIEW	4-19
Chapter 3: PRESENT WORK	20-35
3.1 Random Forest (RF)	26
3.2 Multilayer perceptron	26
3.3 K-Nearest Neighbor	27
3.4 Extra tree Classifier	28
3.5 XG-Boost	29
3.6 Support Vector Classifier	29
3.7 Stochastic Gradient Descent	30
3.8 Adaboost Classifier	31
3.9 Decision Tree Classifier (CART)	31
3.10 Gradient Boosting Machine	32
3.11 Mathew Correlation coefficient (MCC)	34
3.12 Log Loss	34
3.13 F1 Score	35
Chapter 4: RESULT AND DISCUSSION	36-48
4.1 Pearson correlation	41
4.2 Chi Selector	41
4.3 Recursive Feature Elimination	42

4.4	Light GBM	43
4.5	LR Model	44
4.6	RF for Feature Selection	45
Chapter 5: CONCLUSION AND FUTURE SCOPE		49
REFERENCES		50-58

LIST OF FIGURES

Fig. No.	Figure Name	Page No.
1.1	A flowchart of a supervised machine learning model	2
1.2	Heart Disease Prediction Model	3
3 (a)	Exploratory Data Analysis	20
3 (b)	Distribution of heart disease	21
3 (c)	Gender and age-wise distribution	21
3 (d)	Distribution of chest pain type	22
3 (e)	Distribution of Rest ECG	22
3 (f)	Electrical signals recorded by an electrocardiogram	23
3 (g)	Outliers representation plot in terms of Cholesterol	24
3 (h)	Outliers representation plot in terms of age	24
3 (i)	Distribution of Numerical features	25
3.1	Random Forest diagram	26
3.2	Multi-layer perceptron diagram	27
3.3	K-NN diagram	28
3.4	Receiver operating characteristics curve of an extra tree classifier	28
3.5	The general architecture of XGBoost	29
3.6	Support Vector Classifier	30
3.7	Stochastic Gradient Descent figure	30
3.8	Adaboost Classifier figure	31
3.9	A simple decision tree classifier with 4 features	32
3.10	Gradient boosting flow diagram	33

3.11	Graphical Representation of Mathew Correlation Coefficient	34
3.12	Logg Loss Graph	35
4 (a)	Distribution of Numerical features	36
4 (b)	Flow Chart for determining modified z-scores of outliers	37
4 (c)	Correlation with Response Variable class plot	37
4 (d)	Train Test Split distribution diagram	38
4 (e)	Cross-Validation Diagram	38
4 (f)	ROC Curve	40
4 (g)	Precision-Recall Curve	40
4.1	Pearson Correlation map for feature selection	41
4.2	Feature Selection by Chi-Square	42
4.3	Recursive Feature Elimination Flow	43
4.4	Light GBM Architecture	44
4.5	Selection of the parameter in the LASSO model by 10-fold cross-validation based on minimum criteria.	44
4.6 (a)	Random Forest for Feature Selection	45
4.6 (b)	Distribution of most contributing feature	46
4.6 (c)	Hard vs soft voting classifier	46
4.6 (d)	ROC Curve	48
4.6 (e)	Precision-Recall Curve	48

LIST OF TABLES

Table No.	Table Name	Page No.
3 (a)	Heart disease patient's count	23
4 (a)	Comparison of Heart Disease Classification Techniques	39
4 (b)	Analysis of Machine Learning Techniques with Soft Voting	47

NOMENCLATURE

Abbreviation	Full-Form
ran	Random Forest Classifier
mlp	Multi-Layer Perceptron
et	Extra Trees Classifier
ada	Ada Boost Classifier
dec	Decision Trees Classifier
svc	Support Vector Classifier
cm	Confusion Matrix
acc	Accuracy
prec	Precision
rec	Sensitivity
F1	F1 Score

CHAPTER 1: INTRODUCTION

During the last decade, cardiovascular disease remains the major foundation of loss of life internationally. According to the sector fitness employer, over 17.9 million individuals bite the dust consistently because of cardiovascular infection, with coronary vein sickness and cerebral stroke representing 80% of these deaths. Countless deaths are predominant in low and middle-income countries. Many predisposing factors, including non-public and expert habits and genetic predisposition, account for coronary heart sickness. Heart disease is caused by a combination of common risk factors such as smoking, excessive alcohol use, strain, and inoperativeness, and physiological elements such as for overweight, high blood pressure, excessive cholesterol, and prevailing coronary heart disease. The importance of early clinical diagnosis of coronary heart disease in implementing preventative steps to avoid death cannot be overstated. System learning is a rapidly developing domain of synthetic intelligence. Those algorithms can monitor vast facts from different fields, one such significant region is the clinical field. it is a trade for the repetitive expectation demonstrating technique utilizing a pc to acquire information on confounded and non-straight connections among various elements involving diminishing the slip-ups in anticipated results. Statistics mining is the procedure of examining large repositories to uncover hidden key decision-making records for future study from a collection of prior repositories. The medical subject comprises super data of sufferers. those records want to be mined through diverse system learning algorithms. Healthcare experts evaluate those records to obtain powerful diagnostic choices through healthcare specialists.

The essential focal point of this examination is on applying AI methods to a conjecture heart infection. Distinct people will experience different heart disease symptoms which will vary as a result. Because myocardial infarction illness is linked to various modifiable risk factors related to lifestyle and intervention, the timing of discovery and diagnostic accuracy is especially important in the therapeutic care of myocardial infarction disease patients.

Data on heart-related concerns are collected by medical organizations and researchers worldwide. This information can be utilized to acquire significant bits of knowledge using an assortment of AI methods. However, the information gathered is gigantic, and it is habitually uproarious. These datasets, which are excessively enormous for human personalities to get a handle on, can be effortlessly broken down with AI models. A lot of studies are being done to discover the risk

factors for heart disease in different people, and different researchers are utilizing different statistical methodologies and data mining algorithms. Gender, age, blood pressure, diabetes, total cholesterol, hypertension, obesity, and lack of exercise are all risk factors for heart disease.

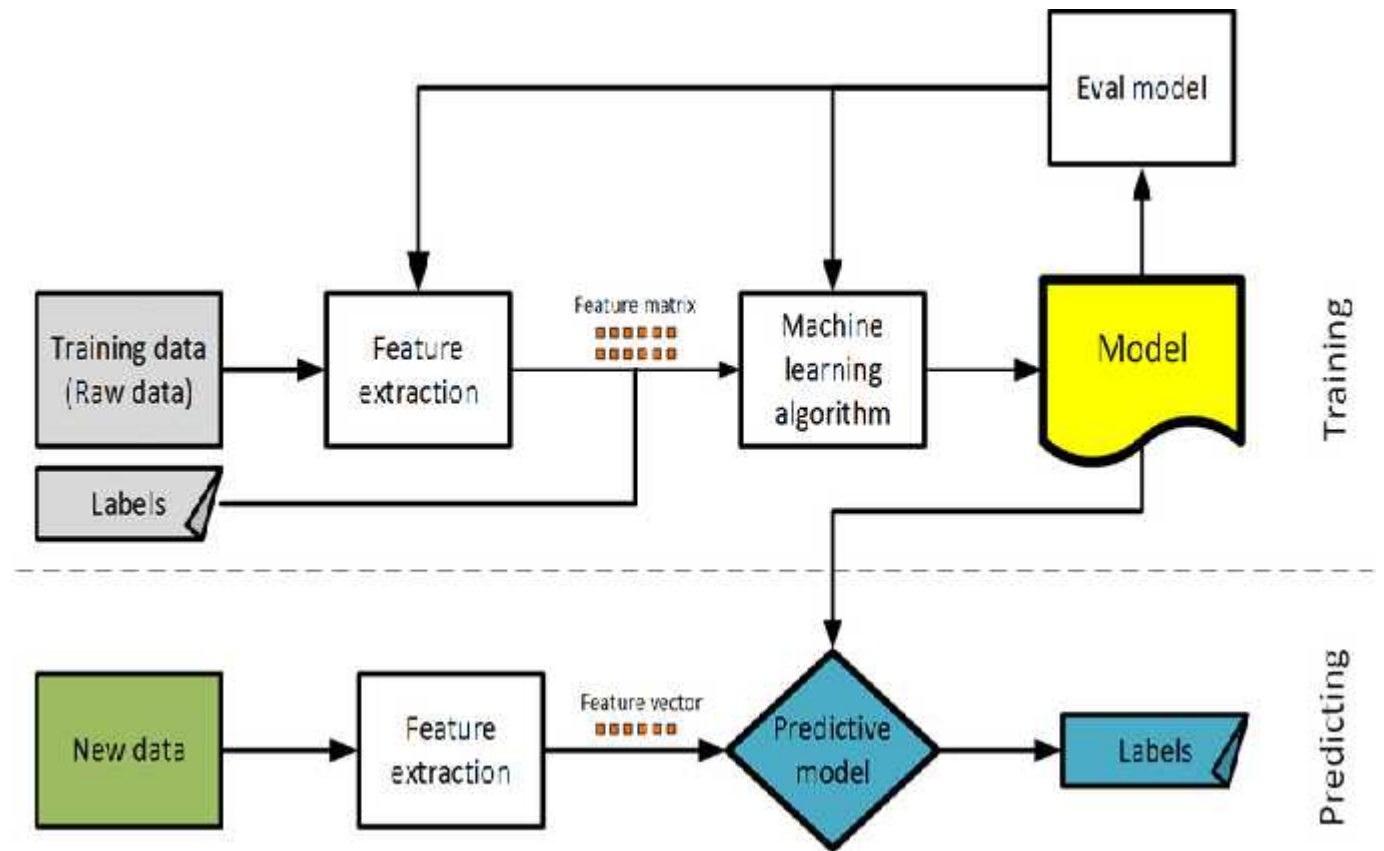


Fig. 1.1: A flowchart of a supervised ML model

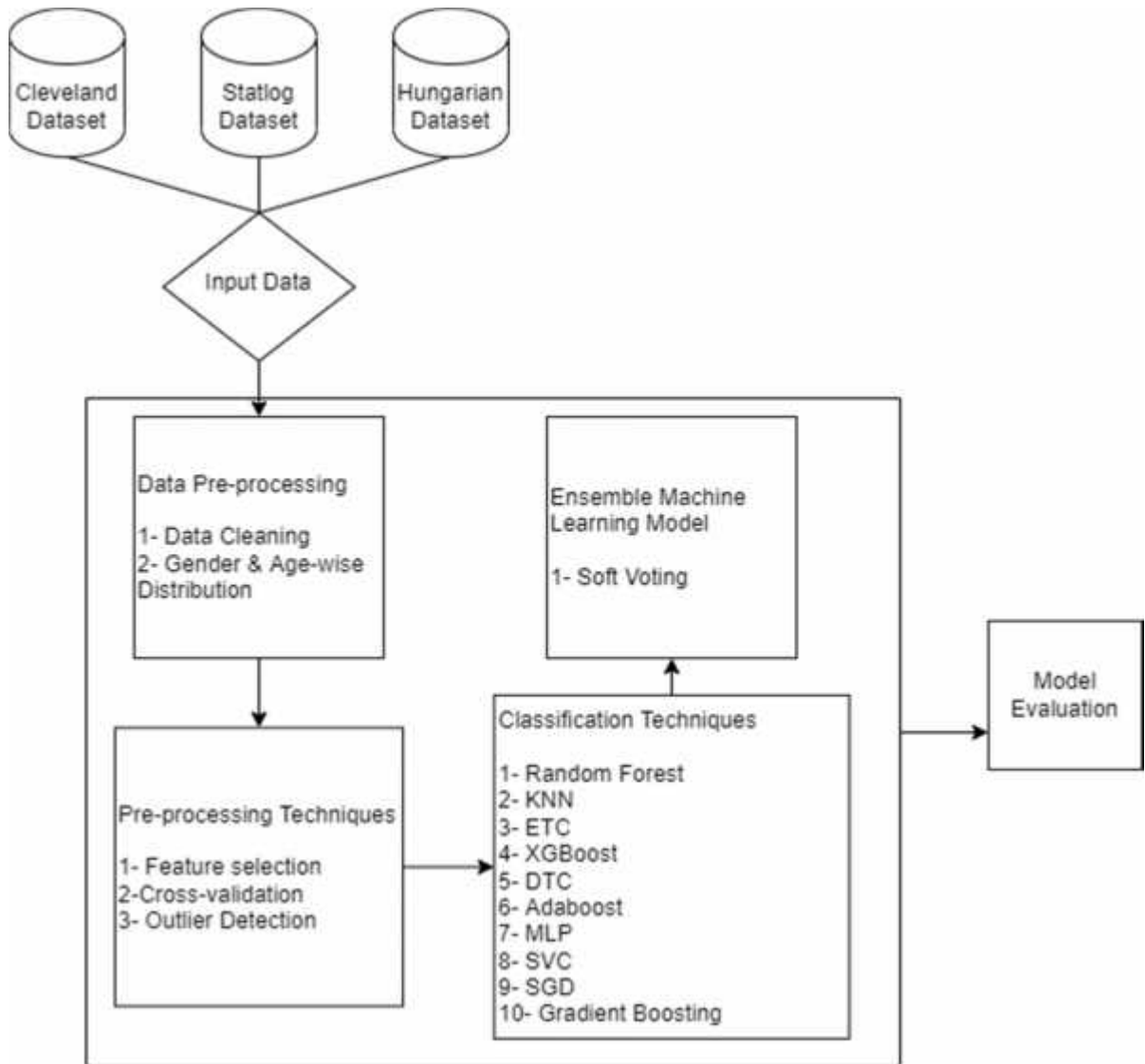


Fig. 1.2: Heart Disease Prediction Model

Under the aegis of the preceding debate, this work utilizes AI (ML) methods to construct a prognostic tool for the correct classification of the heart disease prediction dataset. Two Different open-source Kaggle heart disease datasets are clubbed and used for this purpose. Moreover, in light of the fact that no single AI calculation can perform around ok for all undertakings so in this review, ten famous Machine learning (ML) (Random woods classifier (RF), Multilayer perceptron (MLP), Decision tree classifier(CART), K-NN (K-NN), Extra tree classifier (ETC), Extreme Gradient Boosting (XGB), Support-vector machine (SVM), Stochastic gradient descent (SGD), Adaptive boosting (AdaBoost), Gradient boosting machine (GBM)) procedures are utilized, and their viability was evaluated utilizing an assortment of key execution.

CHAPTER 2: LITERATURE REVIEW

Anjan Nikhil Repaka [1] recommends the method of Naïve Bayes (NB) classification and Advanced Encryption standard to predict the high risk of myocardial infarction in patients. Information has been gathered from various sources. The existing technique surpasses the NB resulting precision of 89.77%. And also, AES provides the most secure performance evaluation in comparison to PHEA. The review inspects how to analyze cardiovascular ailment utilizing past information. For this purpose, NB is employed to gather SHDP (Smart Heart Disease Prediction), for the prediction of risk factors associated with the coronary ailment. Due to the fast development of innovation, versatile well-being innovation has become perhaps the most famous web application. The significant information is collected in a standardized plan. The accompanying credits are separated from clinical profiles to conjecture the probability of coronary illness in a patient: age, circulatory strain, cholesterol, sex, glucose, etc. The NB grouping for anticipating coronary illness involves the gathered characteristics as info. The dataset is separated into two pieces, with 80% being utilized for preparing and the leftover 20% for testing. The proposed technique consolidates the accompanying stages: dataset assortment, application-based client enrollment and login, NB characterization, forecast, and secure information move utilizing AES (Advanced Encryption Standard). Consequently, an outcome is produced. By using information-digging approaches for coronary illness expectation, the exploration expounds and offers different information deliberation techniques. The outcomes show that the ongoing demonstrative methodology is fruitful in anticipating risk factors for cardiovascular problems.

Saleh Alotaibi [2] combined five different algorithms He used Rapid miner as a tool which provided improved accuracy of NB by 3%, Logistic Regression (LR) by 5%, Decision tree (DT) by 11%, and SVM by 8% in contrast with the previous work i.e. Matlab and Weka tool. His work's key restriction is the dataset's modest size. Therefore, various augmentation techniques have been augmented with the dataset. As a result, by merging five different algorithms, this study discussed, suggested, and implemented a ML model. Rapid miner was used in this study, and it was computed with higher precision than Matlab and Weka. In comparison to earlier studies, this study has demonstrated a substantial improvement and great accuracy. When it comes to the UCI dataset, it needs to be expanded. The dataset's unassuming size is the vital imperative in this review. Since the dataset contains a predetermined number of patient records, it was improved by utilizing

important methodologies. Later on, the outcomes demonstrated the way that the methodology could be helpful to specialists and cardiovascular specialists in deciding the probability of a coronary failure in a patient.

Nagraj M. Lutimath [3] used Naive Bayes grouping and SVM for the expectation of myocardial infarction. The dataset utilized in this paper is UCI Cleveland. Mean Absolute Error (MAE), Sum Square Error (SSE), and Root Mean Square Error (RMSE) are accomplished utilizing proper qualities of the dataset utilizing fitting highlights. It is observed that SVM with the outspread piece is offering more recurrence than NB Classification. The UCI Cleveland informational collection is utilized to foresee coronary illness involving Nave Bayes characterization and outspread bit SVM in this examination. MAE, SSE, and MSE are determined based on the appropriate properties and features of the data set in comparison to the radial kernel and the Naive Bayes classifier. SVM classification using a radial kernel is more accurate than Nave Bayes classification. Male and female traits are also examined in the data set. Utilizing the consistency estimations RMSE, we find that females are more impacted by coronary illness than guys. Other AI strategies, for example, profound learning, affiliation rule examination, and transformative calculations, will be researched in store for anticipating exactness with proper execution measures.

T. Nagamani [4] used the MapReduce algorithm when trained persistent fuzzy NNs along with a meta-heuristic approach are compared. The dataset used for predicting myocardial infarction is UCI Cleveland. This study achieved 98.12% accuracy for 45 instances of the testing set. Because of linear scaling and dynamic schema, the proposed method's output accuracy, which is the MapReduce algorithm, is distinctively better. This study has some latency due to batch processing which should be reduced in future studies for better results. According to the World Health Organization (WHO), the cardiovascular illnesses (CVD) are being considered as the main source of global mortality, including in India. A few examinations have been directed to foster models utilizing Data Mining alone or in blend with computational procedures, for example, DTs, Naive Bayes, Meta-heuristics approach, and Trained Neural organizations, Machine knowledge. The testing results demonstrate the way that the proposed technique can arrive at an expectation precision of 98% by and large, which is higher than the conventional intermittent fluffy brain organization. Furthermore, our MapReduce strategy outperformed earlier algorithms with

prediction accuracies of 95–98 percent. These findings show that the MapReduce method could be utilized in the clinic to accurately forecast HD risks.

Rishabh Magar [5] used four algorithms: UCI Dataset has been used for the prediction. This study proposed a Web application-based predictive model. LR is shown to have a maximum accuracy of 82.89 percent. SVM came in second with 81.57 percent, followed by Nave Bayes and DT with 80.43 percent apiece. The four calculations were at first executed. Individual datasets were prepared for every calculation. After that, they were all put to the test. The most efficient method was to be chosen based on several factors. With an accuracy of 82.89 percent, we discovered that the LR technique is the most efficient of the four. SVM had an accuracy of 81.57 percent, whereas DT and NB had an accuracy of 80.43 percent and 80.43 percent, respectively. As a result, a better user interface was used to implement these four algorithms. And using ML methods, this web application creates a robust model for predicting cardiac disease. This would allow end end-users to acquire a basic assessment of their cardiac status. Therefore, considering the role of coronary illness in the global fatality rate, a promising innovation like AI may serve as a lifeline to the society by the early identification.

Apurb Rajdhan [6] used four algorithms: DT, LR, Random Forest (RF), and NB, and compared their accuracy score to predict myocardial infarction. This study used the dataset provided by UCI ML repository. With the rising number of deaths because of coronary illness, it has become important to plan a framework that can really and precisely expect coronary illness. The objective of the exploration was to find the best AI framework for distinguishing cardiovascular issues. Utilizing the UCI AI archive dataset, this study examines the precision scores for anticipating coronary illness. Later on, the review may be value-added by making a web-application in light of the RF strategy and utilizing a bigger dataset than the one utilized in this examination, which would assist with conveying improved results and help wellbeing experts in effectively and proficiently determining cardiovascular illness.

P. Umasankar [7] introduced a new approach that included preprocessing and fuzzy correlation rule mining processes for the decision-making. This suggested study's main focus was on the factors that cause heart attacks in people. The preprocessing phase lowered the size of the dataset.

Using a fuzzy based rule base, a set of rules was constructed to estimate the heart disease based on the selection mechanism. The internal ambiguous set was used to overcome decision-making problems in patients with heart disease who were in a reluctant state.

Senthilkumar Mohan [8] presented an up-to-date mechanism to increase the precision of cardiovascular disease prediction by selecting highly relevant variables using ML techniques. Combining the attributes with well-known categorization approaches created a new forecasting model. Consolidating a coronary illness expectation model with a hybrid RF with a direct model brought about a raised exhibition accuracy level of around 88%. This study can be expanded in the coming time by incorporating additional ML approaches. Moreover, later on, new element determination calculations could be made to accomplish a more extensive view of critical highlights, consequently working on the viability of cardiovascular infection estimating. Coronary illness is the most common cause of death in the cutting-edge world. In the field of clinical information examination, foreseeing cardiovascular illness is a significant trouble. ML can help with decision-making and prediction in the healthcare industry, which generates massive volumes of data. Recent breakthroughs in numerous areas of the Internet of Things have also used ML techniques (MIoT). The use of ML to forecast heart illness has only scratched the surface of study. In this paper, a unique ML strategy has been framed for finding critical attributes. The expectation model is presented utilizing different mixes of qualities and an assortment of notable grouping techniques. We gain an improved exhibition level with an exactness level of 88.7% involving the half-breed irregular woodland with a straight model for coronary illness expectation. (HRFLM)

Aditi Gavhane [9] proposed an application that might use simple symptoms to forecast heart disease vulnerability. The NNs AI calculation was found to have the most noteworthy precision and steadfastness. The proposed method made use of this technique. The users were given a forecast result that included the user's current state, which led to CAD using MLP, a ML technique. ML algorithms have undergone a significant evolution as a result of recent advancements. As a result, the proposed system uses a Multi-layered Perceptron, which has higher efficiency and accuracy (MLP). The proposed calculation gave a practically reliable result in light of the information given by the clients. As the quantity of individuals who utilize such gadgets develops, their familiarity with their ongoing heart condition increments. Subsequently, the number of

people experiencing coronary illness will diminish.

According to Aakash Chauhan [10], the number of people in India who suffer from cardiovascular diseases is increasing. Before very long, coronary sickness is expected to overwhelm cancer disease as the primary driver of death in India. Accordingly, restricting its effect is fundamental. Thus, this study fostered a coronary illness expectation approach for dependably distinguishing the gamble of coronary illness. A novel model for predicting heart disease was developed using data mining techniques. To generate strong association rules, the dataset of patients was subjected to frequent pattern growth association mining. The data could be analyzed, and doctors could reliably anticipate cardiac problems using the proposed method.

Harshit Jindal [11] developed a method that employs three classification algorithms: K-NN, RF, and LR, with a precision of 87.5 percent. In this EHDPS, or powerful myocardial infarction sickness determining framework, LR and K-NN outperform RF, with K-NN offering the best accuracy of the three calculations tried at 88.52 percent. Coronary illness examples are expanding at a disturbing rate, making it basic and alarming to foresee any such afflictions ahead of time. This is a troublesome symptomatic to make, so it should be done accurately and rapidly. The study focuses on identification of patients most susceptible for heart disease depending on several medical characteristics.

Five techniques were used by Ravindhar NV [12] In the exploratory examination of cardiovascular circumstances, a 10-overlap cross-approval system is performed. Backpropagation NNs gave the most noteworthy exactness, with 98.2 percent precision, 87.64 percent review, and 89.65 percent accuracy. Cardiovascular sickness has turned into a worldwide general medical problem, inferable from an absence of wellbeing mindfulness, an unfortunate way of life, and undesirable utilization propensities. With regards to infection finding, experts might have assorted concerns, which prompts differed ends and activities. Then again, even in the occurrence of a typical disease, how much data accessible is huge to such an extent that pursuing exact and trustworthy choices may challenge. With the right arrangement of patient and non-patient clinical imperatives, it's feasible to dependably gauge the probability of an individual creating coronary illness and concentrate valuable information from these frameworks. A mechanized structure for remedial investigation

would likewise boundlessly work on clinical contemplations while radically bringing down costs. In this examination, we fabricated a system for understanding the standards of determining the gamble profile of patients utilizing clinical information qualities. Four AI techniques and one brain network calculation were utilized in this review to assess execution estimations for the identification of cardiovascular problems. To obtain the ability to forecast cardiac attacks, the algorithms were examined in terms of accuracy, recall, precision, and F1 settings. The results demonstrate that our NN system predicted heart illnesses with 98 percent accuracy.

Aravind Akella [13] used six predictive models on the UCI repository and discovered that NNs had the highest precision of 93.03 percent, indicating the little possibility of false negatives and thus an exceptionally precise outcome. A few modifiable gamble factors impact the improvement of coronary corridor sickness (CAD), an exceptionally common infection around the world. Prescient models in view of AI (ML) methods could assist clinicians with recognizing CAD prior and further developing results. Materials and methodology: We utilized six different AI strategies to foresee the presence of CAD among patients in the 'Cleveland dataset.' With a definitive objective of fostering a valuable clinical device for CAD recognizable proof, the subsequent PC code is provided as a functioning open-source arrangement. The precision of each of the six AI calculations was better compared to 80%, with the 'brain network' approach acquiring more prominent than 93%.

N. Saranya [14] proposed a web-based approach for predicting cardiac disease that saves time and money. The RF and K-NN approaches are used in the model. The dataset was acquired from one of Coimbatore's hospitals, and after cleaning and pre-processing, it produced a precision of 100% using RF and 91.36 percent using K-NN. Developing approaches for detecting heart problems earlier and alerting patients to improve their health. We will apply ML approaches to forecast the incidence at an earlier stage to solve this challenge. We will generally expect to utilize specific factors like age, sex, level, weight, case history, smoking, and liquor admission, as well as tests, for example, pulse, cholesterol, diabetes, ECG, and ECHO. There are different calculations in AI that will be utilized to tackle this issue. K-NN, Support vector classifier, choice tree classifier, calculated relapse, and RF classifier are among the methodologies. We should expect in the event

that the patient has coronary illness or not utilizing these boundaries and calculations, and afterward encourage the patient on the most proficient method to work on their wellbeing.

Aadar Pandita [15] made a forecast model that integrates five AI calculations and utilizes the technique with the best accuracy to make a web-empowered application that acknowledges patients' clinical data and predicts whether they have a cardiovascular illness. HTML/CSS and a Flask-based framework are used to create the web application. K-NN had the highest precision of 89.06 percent, while LR had the lowest precision of 84.38 percent. All around the world, heart disease has been the leading cause of death. Heart attacks and strokes account for the majority of deaths connected to cardiovascular disorders. As a result, we must devise strategies to ensure that these figures are kept to a minimum. To diminish the adverse consequences of coronary illness, we should attempt to predict its event at a previous stage. AI calculations can assist us with precisely anticipating such outcomes, which can help specialists and patients identify the beginning of sicknesses, diminishing their effect, or keeping them from creating. We want to foster a framework that can successfully distinguish the presence of heart sickness while likewise being time and practical.

To forecast heart illnesses, Apurv Garg [16] used K-NN and RF ML methods. The balancing of the data was examined once it was obtained and analyzed, and a correlation was discovered in-between numerous variables and their impact on the desired value. The repository used was the UCI repository, which can be found on Kaggle. It was split in half for training and testing in an 80-20 ratio. The desired attribute was found to have a positive link with chest pain and maximum heart rate obtained. When using K-NN, this model had an accuracy of 86.885 percent, and when using RF, it had an accuracy of 81.967 percent. Many individuals experience the ill effects of cardiovascular sicknesses (CVDs), which kill individuals everywhere. AI can be utilized to decide if an individual has a cardiovascular disease in light of specific qualities, for example, chest distress, cholesterol levels, age, and different variables. Cardiovascular illness conclusions can be improved on utilizing order calculations in view of managed learning, a kind of AI.

Chaitanya Suvarna [17] In this study, an intrinsically distributed technique known as Particle Swarm Optimization was employed to solve concerns with interactions between distinct particles,

also known as simple and individual agents. To undertake experimental testing, the study used a well-known data source. The standard protocols were used to rank the predictability of cardiac disease. PSO was combined with a constriction factor, sometimes known as constricted PSO, to create a significantly updated version of this method. The results demonstrated that PSO-based data mining algorithms delivered extremely efficient outputs not just with the progressive method but also with industry-standard algorithms and that such algorithms may be employed very effectively in heart disease prediction depending on the applications.

Rashmi G Saboji (2017) developed a new system for predicting cardiac disease using healthcare data based on specific characteristics [18]. The capacity to forecast the prognosis of cardiovascular disease using a few indicators was the major offering. The RF is utilized to fabricate an estimating arrangement utilizing Apache Spark. For medical care experts, the proposed strategy gave an incredible opportunity to put this investigation on an exceptionally versatile scene for brilliant independent direction. We achieve a level of accuracy of around 98% using this methodology. The wellbeing business, as per ongoing patterns, is gathering a consistently expanding measure of information carefully, making it probably the most datum serious enterprises. The headway of innovation has made it conceivable to handle such tremendous measures of information and precisely gauge wanted results. In this paper, we offer a versatile procedure for anticipating coronary illness in light of specific characteristics utilizing medical care information. Our vital commitment to this work is the capacity to foresee the conclusion of coronary illness utilizing a couple of qualities. Our expectation arrangement depends on Apache Spark's irregular backwoods, which gives a colossal chance to medical care experts to apply this answer for a continually changing, versatile large information climate for informed independent direction. We show the way that up to 98 percent precision can be accomplished utilizing this strategy. We likewise give a correlation with the Nave-Bayes classifier, showing that the irregular timberland approach beats the previous by a huge degree.

C. Sowmiya (2017) advocated evaluating the possibility of nine alternative classification algorithms to predict cardiac disease. Various research studies [19] employed their classification algorithms the most. This study wrapped up on adapting the SVM and apriori algorithms to predict cardiac disease. Medical profiles based on a variety of characteristics were gathered and used in

this study. Patients who were more prone to developing myocardial infarction disease were identified in this study. The medical community was interested in this research to help detect and prevent heart disease. When the proposed strategy was used, the researchers found that it was much more successful and accurate than earlier techniques.

Cardiovascular disease is a leading cause of morbidity and mortality in today's society. Identifying evidence of cardiovascular disease is a critical but difficult task that must be completed meticulously and efficiently, and the correct robotization would be quite appealing. As specialists, each individual cannot be equally skilled. All specialists cannot be equally talented in every subspecialty, and we do not always have gifted and authoritative specialists available. A computerized therapeutic analysis system would improve medical consideration while also lowering expenditures. In this study, Kanak Saxena [20] devised a framework for determining the tenets that may be used to predict a patient's risk level based on a specific health metric. The study's key contribution is to assist non-specialized clinicians in making accurate decisions about heart disease risk levels. Further, various kinds of rules have been proposed in the developed methodology. The structure's presentation is assessed regarding plan accuracy, and the outcomes show that the system has gigantic guarantee for all the more unequivocally expecting coronary disorder risk levels.

The heart and other organs are vital components of the human body. As cardiovascular ailments are the main source of mortality around the world. Stationary ways of life are at fault, as they can prompt heftiness, raised cholesterol levels, hypertension, and hypertension. S. Bhagavathy [21] utilized Various information mining draws near. To extricate information from an enormous number of informational indexes, the SVM and K-NN calculations were utilized. The created reports assist doctors and nurses in identifying diseases and their severity levels, allowing them to better treat patients. In the healthcare industry, text mining is the most widely utilized mining approach. In this research, we evaluate the implementation efficiency of the K-means clustering method and the Map Reduce Algorithm in parallel and distributed systems.

Although a lot of data has been generated by the healthcare sector, this information is not always used to its full potential and is frequently neglected. An illness can be discovered, predicted, and

possibly cured with this massive amount of data. Infections like coronary illness, malignant growth, cancers, and Alzheimer's sickness represent a critical danger to mankind. Muthuvel Marimuthu [22] attempts to zero in on coronary illness forecast in this examination. Coronary illness can be anticipated utilizing AI draws near. Clinical information is utilized as info, for example, pulse, hypertension, diabetes, a few cigarettes smoked each day, etc, and these elements are then displayed for expectation. This model can then be utilized to estimate clinical information later on. K-NN, NB, SVM, and choice tree procedures are utilized. The model's exactness is determined utilizing every one of the calculations. The model for anticipating coronary illness is then picked as the one with the most elevated exactness.

AI changes crude medical services and information into data for better estimation of patient's condition and forecasting. A. Choudhary [23] explains that the goal of this study, dubbed the fine-tune prediction (FTP) model, is to find important variables and include a fusion classifier to increase accuracy. The model obtains a 93.49 percent accuracy rate in experiments. The RF and LM attain just 88.20 percent and 63.60 percent accuracy without FTP. In the result section, we also show the hyperparameter adjustment of the classifier with relevant characteristics.

Cardiovascular sickness is a principal wellspring of mortality among the all-out people. One of the primary subjects in the clinical data assessment division is cardiovascular disease assumption. In the clinical benefits industry, there is a monstrous proportion of data. Data mining changes over a ton of rough clinical benefits data into information that can assist with route and assumption. There have been several studies that have used data mining approaches to predict cardiac disease. Nonetheless, there is research that has focused on the important characteristics that can help predict cardiovascular disease. It is critical to choose the right combination of relevant features to increase the prediction models' performance. Mohammad Shafenoor Amin [24] Proposed this study to find key features and data mining techniques were used to create prediction models: K-NN, DT, NB, LR, SVM, Neural Network (NN), and Vote (a hybrid technique combining NB and LR). Out of the developed models, vote base data mining technique achieved an accuracy of 87.4%.

A data collection of 335 records representing the various 26 variables was used to analyze and forecast coronary artery heart disease [25]. The correlation idea was used to pre-process the data

set. The particle swarm optimization (PSO) technique was used to identify and extract features. Models for NNs, regression, fuzzy logic, and DTs were created. The information was fed into a NN model. The accuracy was found to be 77 percent. It was also used in a regression model. The accuracy was 83.5 percent as a result of this. There were no significant changes in the other fuzzy and DT models.

Heart disease is an important condition that experts are studying to anticipate the patients who may develop it. AI is the craft of making programs that gain for a fact to explicit issues. For characterization of the informational collection, choice trees, brain organizations, Nave Bayes arrangement, SVMs, and hereditary cycles [26] are utilized in AI. Using appropriate medical data, DTs C4.5 and Fast DTs were investigated [27]. DTs were shown to be 69.5 percent accurate, whereas fast DTs were found to be 78.54 percent accurate.

Using a good medical data set, an Apriori procedure employing the Transaction Reduction Method (TRM) was used to diagnose heart disease [28]. The results were matched to some of the more traditional methods. The program produced a result of 93.75 percent accuracy. When SMO was used, an accuracy of 92.09 percent was obtained. When SVM was employed, the accuracy was 89.11 percent. The accuracy of the C4.5 DT was 83.85%, and the accuracy of the Naive Bayes probability classification was 80.15 percent.

R. Subramanian et al., [29], presented the utilization of brain organizations to analyze and estimate coronary illness, circulatory strain, and different highlights. A Deep NN was proposed based on the given infection credits to predict an outcome. Further, 120 hidden layers were employed for ensuring an exact result of having coronary sickness expecting that the model is used for the Test Dataset. The supervised network has been recommended for heart disease diagnosis. At the point when a specialist tried the model with obscure information, the model is utilized and gained from recently scholarly information to foresee the result, permitting the precision of the model to be determined.

[30] introduced a cloud-based choice emotionally supportive network to help heart specialists during the conclusion interaction. For anticipating coronary illness, this framework utilized AI

draws near. The framework was intended to offer help cost-really, with the capacity to interface with existing frameworks. In that review, a solo bunching calculation was used to order the dataset in light of specific gatherings. In [31], the creator utilized a strategy for performing different bunching calculations on a coronary illness dataset to sort out the most effective way for boosting the expectation exactness proportion. The effectiveness of ML approaches in predicting cardiac disease using historical data was further demonstrated in a study utilizing Naive Bayes, DTs, SVM, and other models [32].

Min Chen et al. [33] apply a variety of ML algorithms for excellent chronic disease prediction. For both structured and unstructured data, a multimodal illness risk prediction technique was used. With a faster convergence time, the algorithm's prediction accuracy is better than others. Data mining techniques developed by Tikotikar A et al. [34] are employed in the medical field for clinical diagnosis. An extensive review of medical data is assumed to aid in making well-informed diagnoses and judgments.

The SVM methodology is an effective strategy for predicting cardiac disease by Cincy Raju et al. [35]. By comparing the performance of the DT algorithm to that of the SVM, Praveen Kumar Reddy, M, et al. [36] were able to prove that it was better. The authors Akash et al. used the k-mean technique to combine structured data and patient text data, resulting in improved accuracy. ML approaches were used by Reddy et al. to predict cardiac disease. All of this piqued people's curiosity about using ML to forecast cardiac disease.

Using a data mining approach to predict hypertension from patient medical histories, Feixiang Huang et al. [37] discovered that the J-48 classifier gives better results. M. Amiri et al. [38] created heart sound diagnosis systems. They created classification and regression trees using 116 heart sound signals. M.A. Nishara Banu [39] and colleagues employed a clustering and classification method to predict the patients' risk levels. Theresa Princy et al. examined classification approaches such as NB, NN, K-NN, and DT for predicting a patient's risk level, considering age, gender, pulse rate, blood pressure, and cholesterol.

Ankita Dewan, et al. (2015) The backpropagation (BP) algorithm was employed to update weights

by backpropagating errors. Heart disease prognosis is regarded as the most difficult issue in medical research. As a result, a decision support system for recognizing cardiac disease in patients is required. A heart disease prediction approach has been proposed that combines an efficient evolutionary algorithm with a BP mechanism. Today's medical sector has come a long way in treating patients suffering from numerous ailments. One of the most hazardous is coronary illness, which shouldn't be visible with the unaided eye and strikes abruptly when its cutoff points are reached. A patient's death would result from poor clinical decisions, which no hospital could afford. To obtain an accurate and cost-effective computer-based treatment and support Good decision-making systems can be created. To manage their healthcare or patient data, many hospitals employ hospital information systems which eventually generate massive volumes of information in the form of images, text, graphs, and numbers. Unfortunately, this information is rarely used to support medical decisions. There is a significant amount of unexplored information in this data, which raises the essential question of how to extract usable information from it. As a result, an effective project is required to assist practitioners in predicting cardiac disease before it develops. The major goal of this work is to create a prototype that can extract hidden knowledge (patterns and relationships) from a previous heart disease database. It can answer complex questions about heart disease and thereby help doctors make better clinical judgments, something that older decision support systems couldn't do. It can help to cut treatment expenses by offering efficient treatments.

The HNB classifier was utilized by M. A. Jabbar et al. (2016) to suggest a unique approach to diagnosing heart disease. The HNB was used on the cardiac catalog dataset, and the performance of the suggested technique was tested [41]. According to the experimental data, the proposed HNB model performed better than the existing techniques. The proposed approach expanded the productivity of stowed away Naive Bayes by utilizing discretization and IQR channels. The proposed model had the highest accuracy level when compared to the NB classifier. The HNB model gave reliable DSS for automatically diagnosing the condition. Coronary illness determination is a tedious interaction. The Hidden Nave Bayes (HNB) model, according to our proposed model, can be used to classify cardiac disease (prediction). Our results on a data set of heart disease patients reveal that the HNB is 100 percent accurate and outperforms Naive Bayes.

There was a lack of successful analysis approaches for discovering linkages and trends in health care data, according to Monika Gandhi et al. (2015). Data mining technologies gave better results in this circumstance [42]. As a result, numerous data mining techniques were presented as solutions in this study. This study looked into several knowledge abstraction techniques for predicting heart disease using data mining methodologies. The utilization of information mining-based order strategies to the information disclosure process was likewise researched. Different grouping techniques were utilized to accomplish information arrangement and information extraction, each with its own arrangement of advantages and disadvantages. Moreover, various classifiers were utilized to foster a calculation in medical services associations, which were likewise assessed in this review. It was helpful in breaking down the general execution of different examination studies. Even though medical organizations (healing centers, therapeutic focuses) generate a large amount of data, this data is not adequately utilized. Although the healthcare system is "data-rich," it is "knowledge-poor." There are no effective ways for discovering linkages and patterns in healthcare. In this situation, data mining techniques may be useful. As a result, various data mining approaches can be applied. The purpose of this work is to provide various ways for knowledge abstraction utilizing data mining methodologies that are currently being employed in heart disease prediction research. In this study, algorithms are used to assess data mining approaches such as Naive Bayes, NNs, and DT Algorithms on medical data sets.

K. Prasanna Lakshmi et al. (2015) used experience mining stream association rules to forecast diseases from medical data [43]. A special powerful tree was utilized to oversee the streaming information. In this review, a choice emotionally supportive network called Stream Associative Classification Heart Disease Prediction (SACHDP) was made on the grounds that the worldwide passing rate from heart sicknesses was expanding. Experiments comparing the suggested approach to associative classification techniques revealed that the proposed approach produced better results. The proposed research can be enhanced in the future to lessen the number of generated rules.

According to Jagdeep Singh et al. (2016), multiple relationships and classification algorithms were used to predict heart disease. The association algorithms were utilized to identify the association

rules of cardiac dataset attributes. The correlations among attributes inside the databases were anticipated using classification methods to develop an accurate classifier. The main contribution of this study [44] was achieving excellent prediction accuracy for identifying heart problems. The proposed hybrid associative categorization was implemented using the Weka environment. According to the comparative analysis and results obtained, Apriori associative algorithms produced better outcomes based on gender, age, chest pain type, blood pressure, blood sugar, etc. like factors that can predict early symptoms of heart disease. Apriori, FP-Growth, Naive Bayes, ZeroR, OneR, J48, and K-NN are among the data mining techniques used in this work to predict cardiac illnesses. The construction of a heart disease prediction system is based on the best results, and it uses a hybrid technique for classifying associative rules (CARs) to reach a prediction accuracy of 99.19 percent.

According to N. Priyanka et al. (2017), early detection and treatment were rare. However, because of a lack of resources in medical sectors, this could be an issue, resulting in ineffective cardiac disease prediction [45]. The use of suitable technology support has proven to be extremely advantageous to the medical community and patients. Such problems could be solved using data mining approaches. This investigation yielded highly accurate results even when the patient had a chance of getting heart disease these findings could be used in future initiatives to detect certain types of cardiac disease. As a result, diagnosing and treating heart disease would be simple.

According to J. Thomas et al. (2016). Different data mining approaches and classifiers were discussed in various research. These investigations could lead to a more efficient and precise diagnosis of cardiac disease [46]. As a result, different technologies produced varied precision values based on the number of features. The K-NN was used to detect the risk rate of heart disease. A high level of accuracy was also attained for a variety of qualities. With the assistance of a few additional algorithms, In the future, the number of qualities could be reduced and the accuracy level improved. Medical issues are turning out to be more normal because of present day ways of life and acquired factors. Coronary illness, specifically, has become more broad as of late, jeopardizing individuals' lives. Circulatory strain, cholesterol, and heartbeat rate are totally changed for every individual. In any case, medicinally demonstrated results show that ordinary circulatory strain, cholesterol, and heartbeat rate are each of the 120/90. This report examinations

different order frameworks for deciding an individual's gamble level in light old enough, orientation, circulatory strain, cholesterol, and heartbeat rate. Data mining Medical conditions are turning out to be more normal because of present day ways of life and acquired factors. Coronary illness, specifically, has become more far reaching as of late, seriously endangering individuals' lives. Circulatory strain, cholesterol, and heartbeat rate are totally changed for every individual. In any case, restoratively demonstrated results show that ordinary circulatory strain, cholesterol, and heartbeat rate are each of the 120/90. This record examines different characterization frameworks for deciding an individual's gamble level in light old enough, orientation, circulatory strain, cholesterol, and heartbeat rate.

Purushottam, et al. (2015) found that all doctors cannot be experts in every field and that in certain instances, skilled and specialist doctors are unavailable [47]. Thus, a mechanized clinical determination framework must be worked to work on clinical consideration while likewise bringing down costs. This study attempted to create a system to identify the criteria for forecasting a patient's risk level based on specific health data. The testing results demonstrated that the suggested approach performed well in accurately predicting heart disease risk levels. Cardiovascular disease is a leading cause of morbidity and mortality in today's society. Identifying evidence of cardiovascular disease is a critical but difficult task that must be completed meticulously and efficiently, and the correct robotization would be quite appealing. As specialists, each individual cannot be equally skilled. All specialists cannot be equally talented in every subspecialty, and we do not always have gifted and authoritative specialists available. A computerized therapeutic analysis system would improve medical consideration while also lowering expenditures. In this study, we devised a framework for determining the tenets that may be used to predict a patient's risk level based on a specific health metric. The study's key contribution is to assist non-specialized clinicians in making accurate decisions about heart disease risk levels.

CHAPTER 3: PRESENT WORK

Two different open-source Kaggle heart disease datasets are combined and used for this purpose, each of which is available independently and has 11 similar properties. It includes information from Cleveland, Hungary, Switzerland, and Long Beach VA, and statistics from Statlog. There are 11 similar features in the combined dataset, which contains a total of 1944 records.

Therefore, by analyzing the dataset, it has been revealed that the gathered dataset needs to be modified by converting the categorical features into categorical names to obtain good intuition like ST Slope present in the form of 1, 2, and 3 will be categorized into Upsloping, Flat, and Downsloping. Chest pain types present in the form of 1, 2, 3, and 4 will be turned into typical angina, atypical angina, non- angina, and asymptomatic. Next, to check if there is any missing entry or not `dt.ISNA().sum()` will be passed.

Exploratory data analysis (EDA) and preprocessing are critical components of any ML-driven investigation. It aids in the visualization of the dataset and provides a better understanding of it.

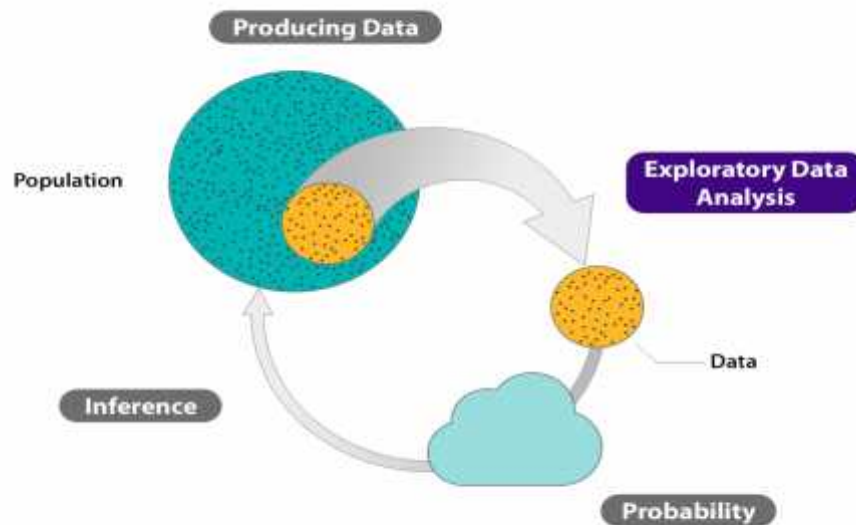


Fig. 3 (a): Exploratory Data Analysis

Fig.3.2 reflects the target variable whether it is a balanced dataset or not because if the dataset is

not balanced then the approach will be different up sampling and down sampling techniques will be applied.

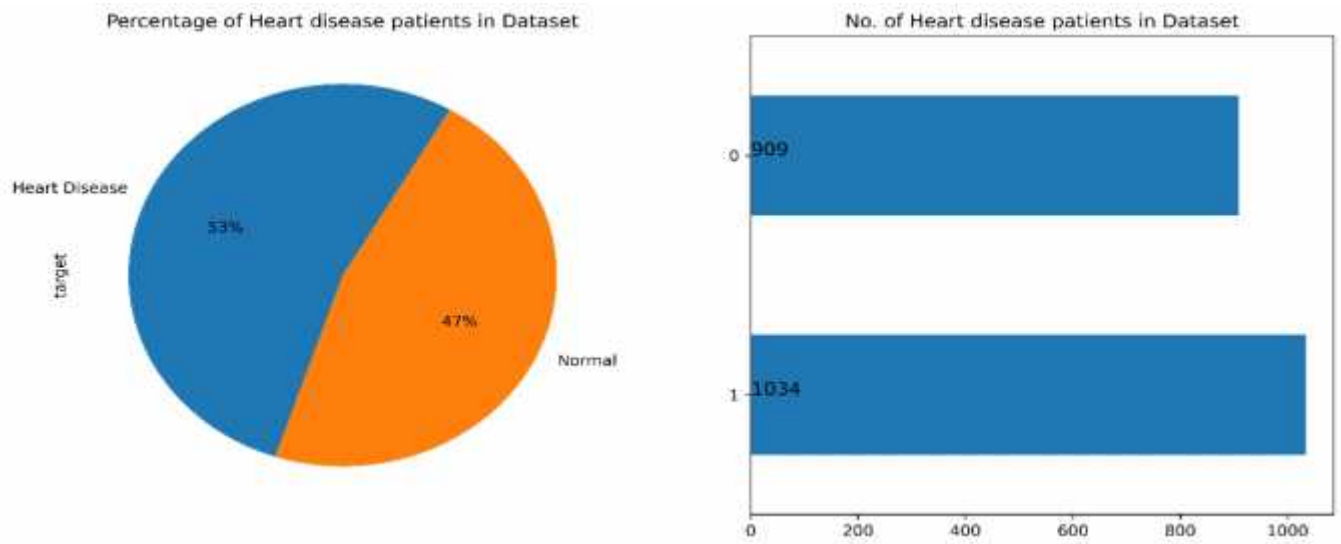


Fig. 3 (b): Distribution of heart disease

The repository is balanced having 1034 myocardial infarction patients and 561 normal patients. The plots in 'figure 3 (c)', 'figure 3 (d)', and 'figure 3 (e)' indicates the Gender and Age-wise Distribution, Chest pain types distribution, and Resting ECG distribution in normal patients and heart patients.

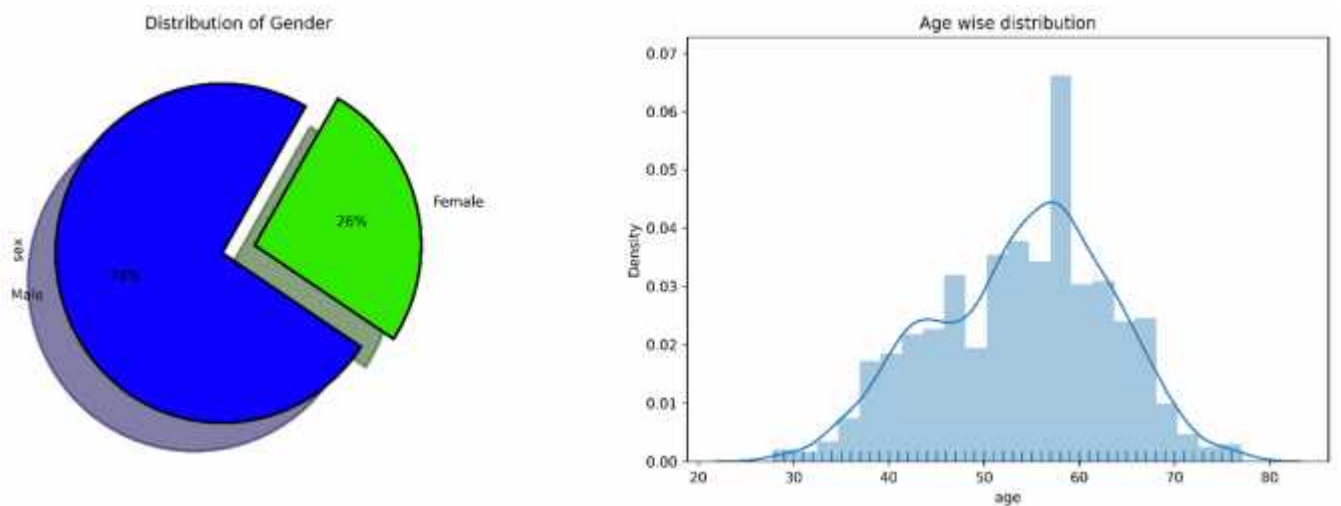


Fig. 3 (c): Gender and Age-wise Distribution

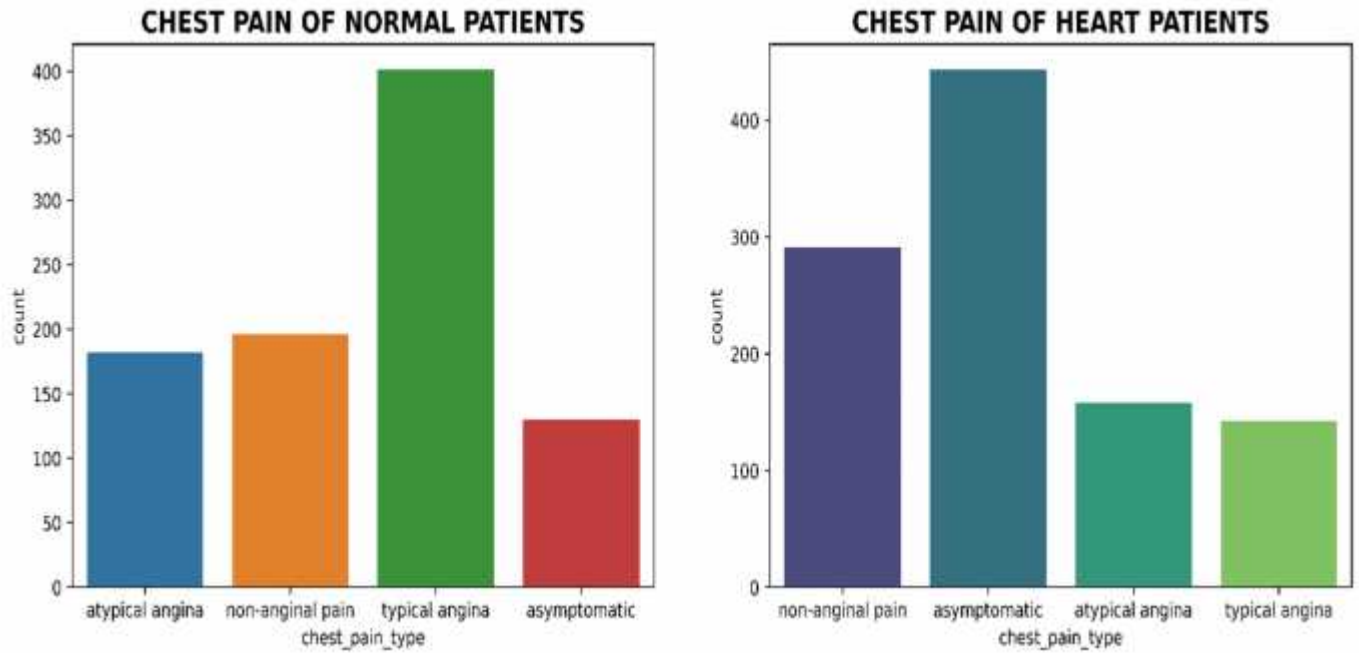


Fig. 3 (d): Distribution of chest pain type

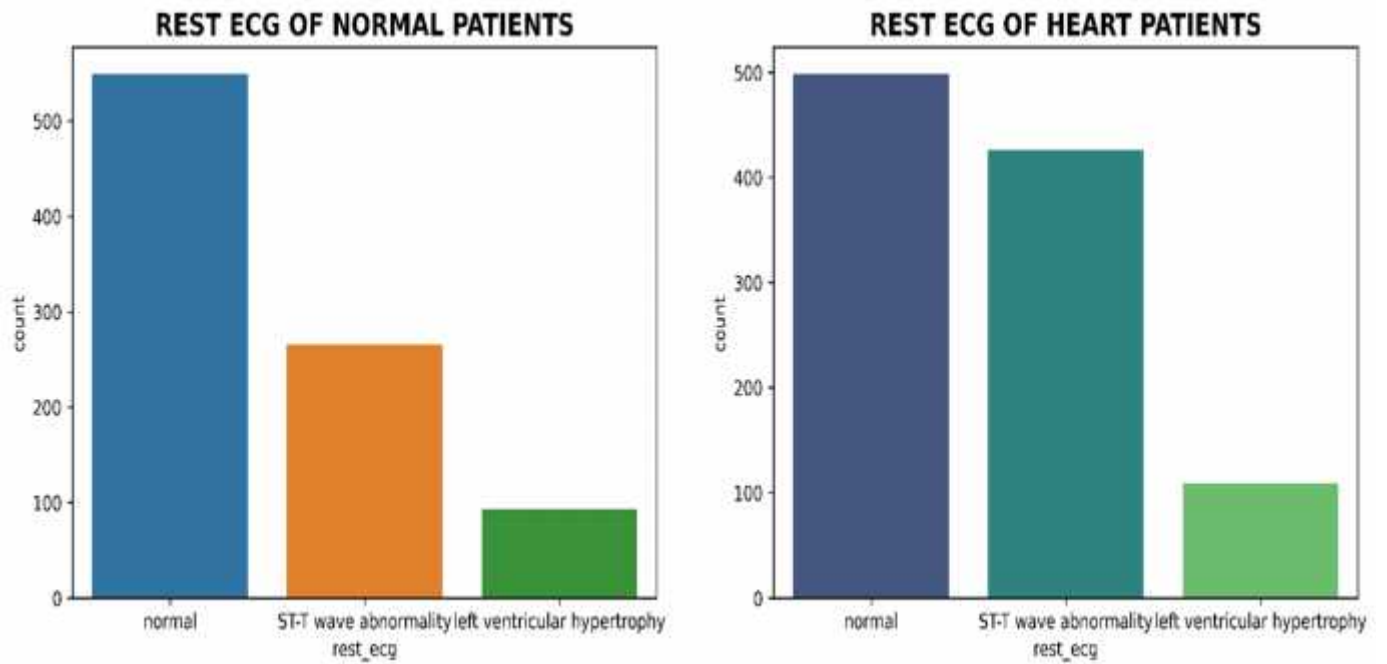


Fig. 3 (e): Distribution of Rest ECG

Table 3 (a): Heart Disease patient's count

Target	0	1
chest_pain_type		
asymptomatic	14.300000	42.840000
atypical angina	20.020000	15.280000
non-anginal pain	21.560000	28.140000
typical angina	44.110000	13.730000

The table shown above depicts a heart disease patient's count. There are different types of chest pain types mentioned in the above table.

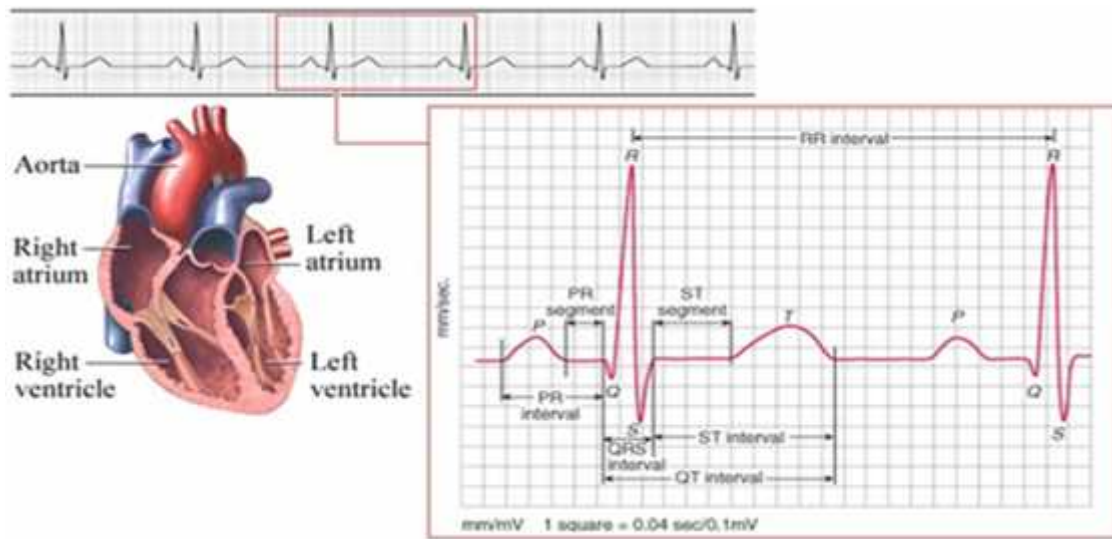


Fig. 3 (f): Electrical signals recorded by an electrocardiogram

An electrocardiogram has been used to analyze the heart on the basis of electrical signals. It's a normal test used to recognize heart issues and screen the heart's status. Electrocardiograms are moreover called ECGs or EKGs. It gauges heartbeat and showing blockages in the conduits doesn't be ensured.

From the underneath plot, we can see anomalies plainly with respect to a portion of the patient's cholesterol is 0 while for one understanding both cholesterol and resting bp is 0 which is perhaps because of missing passages we will channel these exceptions later.

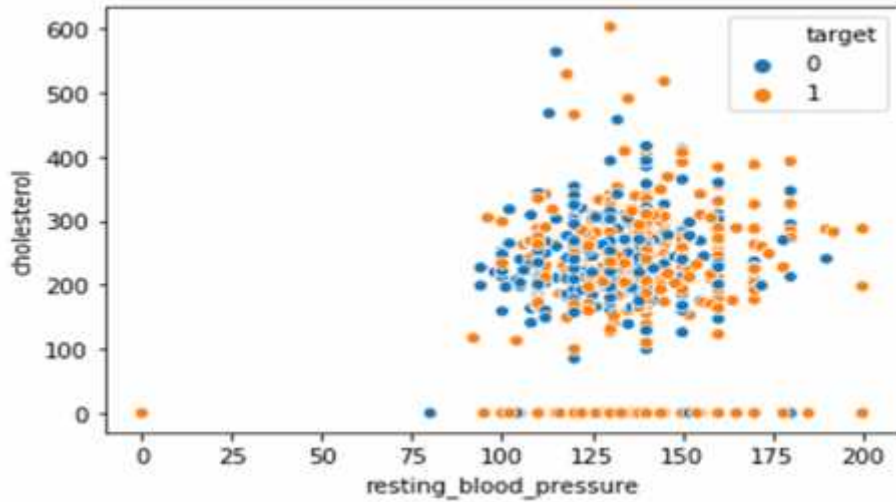


Fig. 3 (g): Outliers representation plot in terms of cholesterol

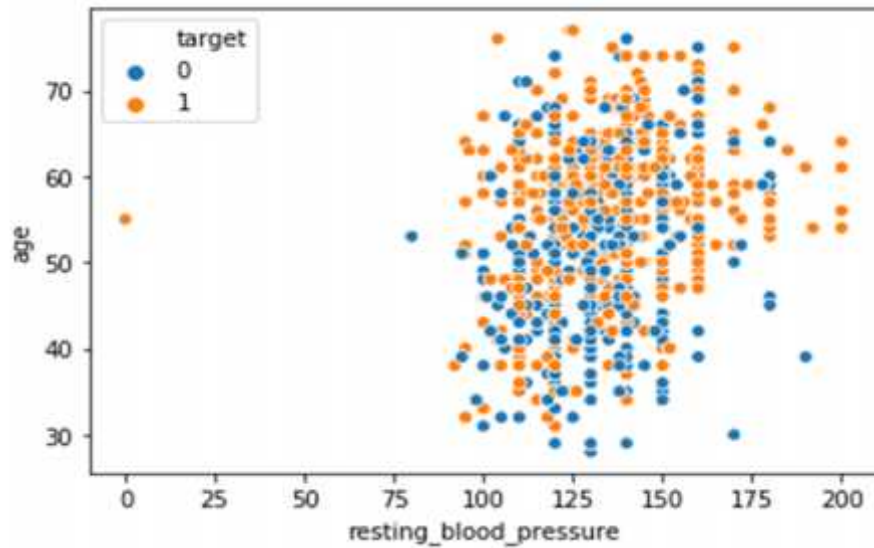


Fig. 3 (h): Outliers representation plot in terms of age

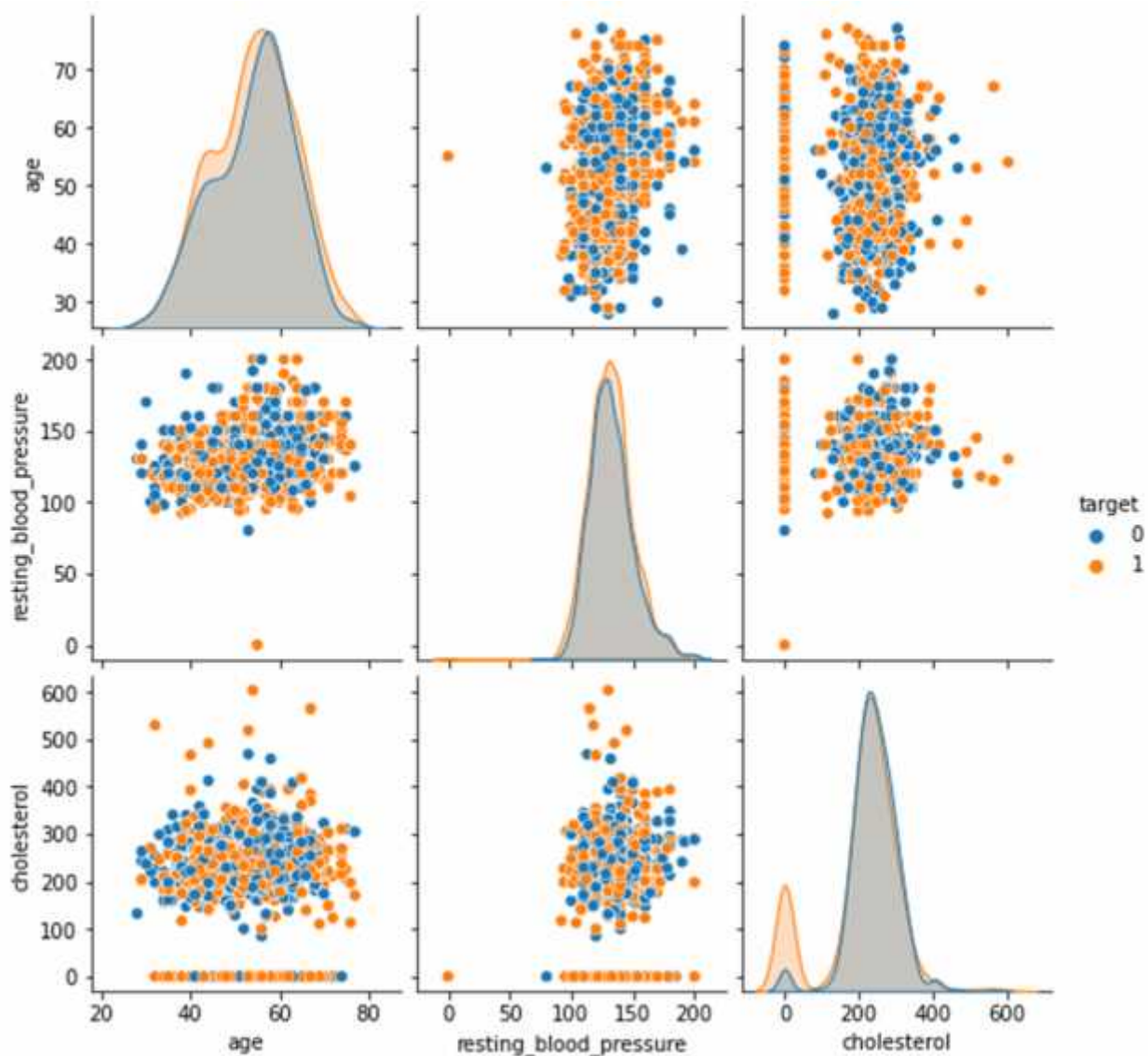


Fig. 3(i): Distribution of Numerical features

From the above plot, it is clear that age is directly proportional to cardiovascular disease. After data cleaning and EDA. Outlier Detection which is a Pre-processing technique is applied. It is the process of detecting and subsequently excluding outliers from a given set of data. Then, at that point, the train-test split is completed. It is a capacity in Sklearn model determination for dividing information clusters into two subsets: preparing information and testing information. With this capacity, you don't have to partition the dataset physically. After Train-test split 10-crease cross approval is applied with the assistance of which a standard model is accomplished further which is contrasted with the underneath referenced AI models.

3.1 RF (RF)

RF is another extensively used supervised ml approach. This method can be applied to both regression and classification problems, but it excels at the latter. As the name implies, the RF technique assesses several DTs before producing output. As a result, it's effectively a DT collection. The concept behind this strategy is that a bigger number of trees will eventually lead to the proper pick. For classification, it uses a voting mechanism and then chooses the class, whereas, for regression, it takes the mean of all DT outputs. It works well with large datasets with several dimensions.

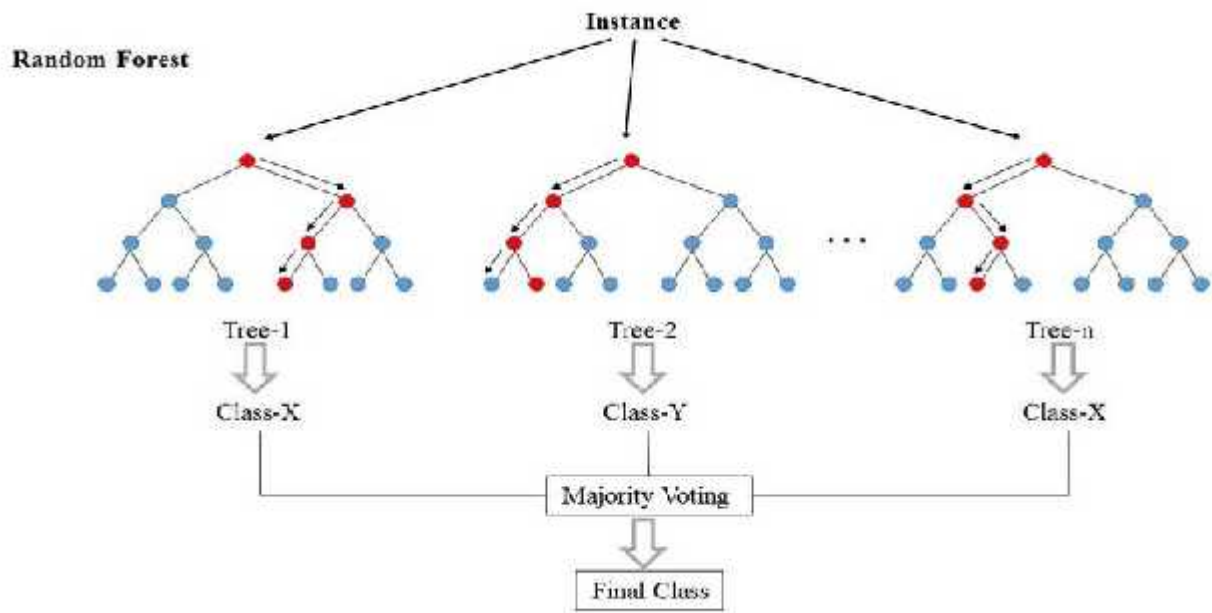


Fig. 3.1: RF diagram

3.2 Multilayer perceptron

It's a kind of completely associated feedforward fake brain organization (ANN). The term MLP is obscure; it can apply to any feedforward ANN or network comprised of many layers of perceptrons. At the point when just a single secret layer is available, multi-facet perceptrons are typically alluded to as "vanilla" brain organizations. An MLP has something like three-center point layers: a data layer, a mystery layer, and an outcome layer. Besides the data centers, each center has a neuron with a nonlinear inception work. MLPs use backpropagation, a controlled learning

approach, during getting ready. The different layers and non-straight institutions perceive MLP from a direct perceptron.

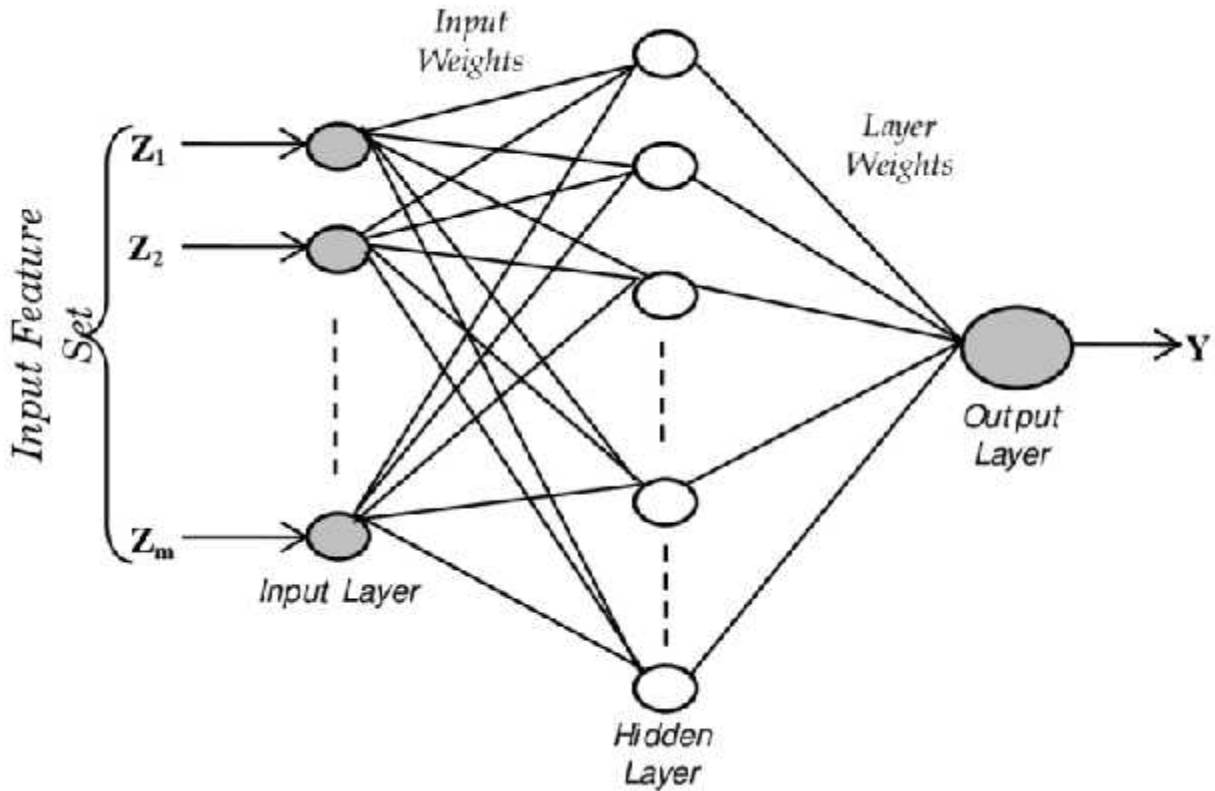


Fig. 3.2: Multi-layer perceptron diagram

3.3 K-NN

Hodges et al. laid out the K-NN rule [13] in 1951, which is a nonparametric example order strategy. The K-NN strategy is perhaps the most fundamental however strong arrangement technique. It makes no doubts about the data and is for the most part consistently used for gathering endeavors where there is basically no prior data on the data scattering available. The motivation behind this methodology is to find the k nearest data of interest in the preparation set to the data of interest for which an objective worth is missing and relegate the typical worth of the found information focuses to it.

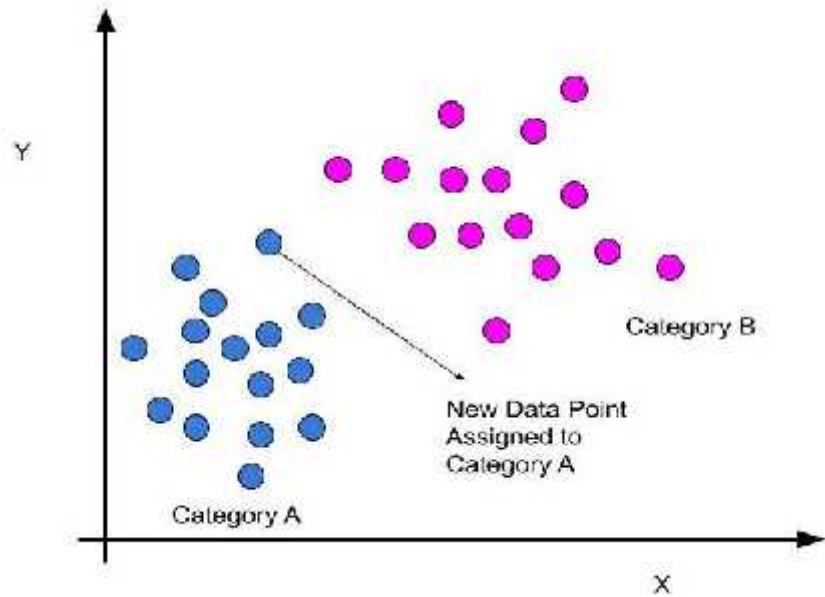


Fig. 3.3: K-NN diagram

3.4 Extra tree Classifier

Very Randomized Trees Classifier (Extra Trees Classifier) is a kind of group learning strategy that produces an order result by joining the consequences of various de-related choice trees assembled in a "woodland." Except for how the choice trees in the backwoods are made, it is hypothetically indistinguishable from an RF Classifier.

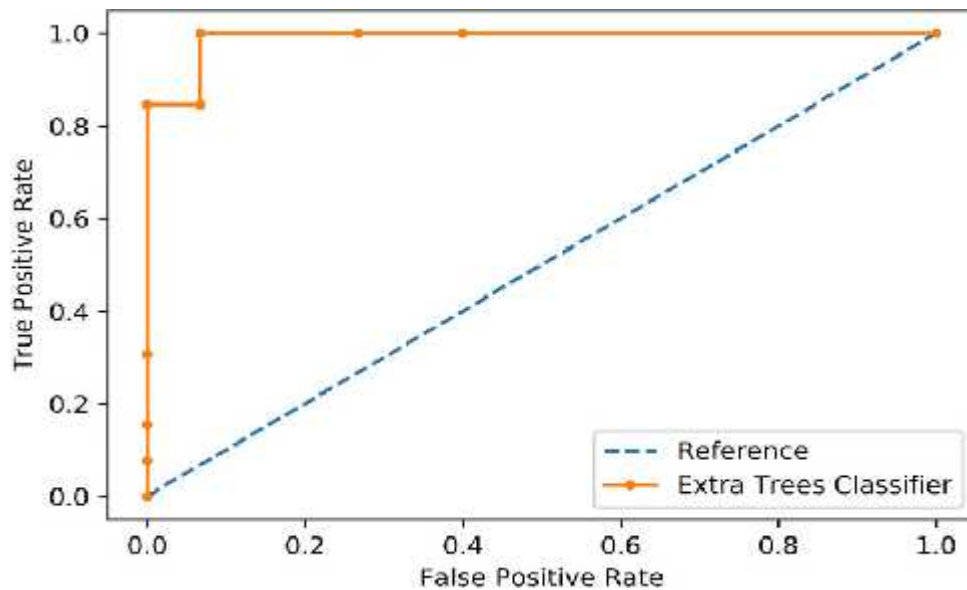


Fig. 3.4: Receiver operating Characteristic curve of an extra tree classifier

3.5 XG-Boost

Gradient boost approaches, which integrate the results of multiple weak learners, are also used by XBG (DT). It's a scalable technology that boosts trees from beginning to conclusion. Furthermore, its execution speed and model performance, particularly for low to middle-level structured data, places it among the most well-regarded and promising ML algorithms.

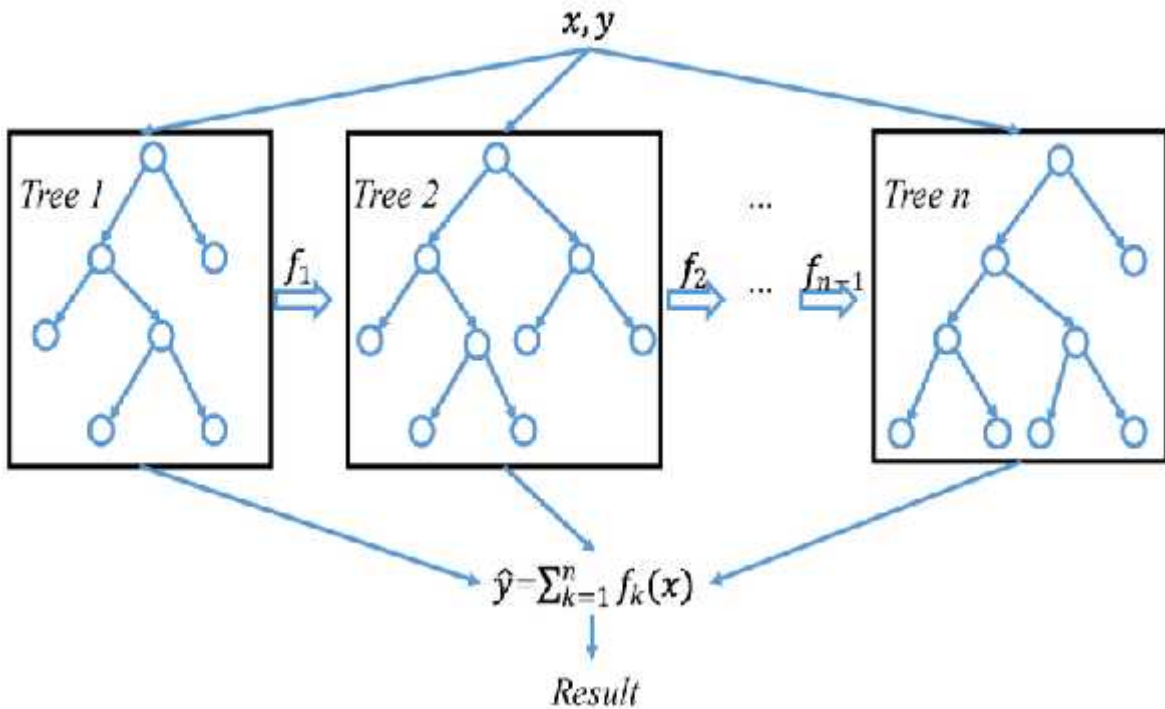


Fig. 3.5: A general architecture of XG-Boost

3.6 Support Vector Classifier

Support vectors focuses on the information that are nearer to the hyperplane. Additionally, these informations should impact the position of the vector along with the heading of the hyperplane. We utilize these help vectors to build the classifier's edge. The place of the hyperplane will be changed by eliminating the help vectors. These are the variables that will impact the development of our SVM.

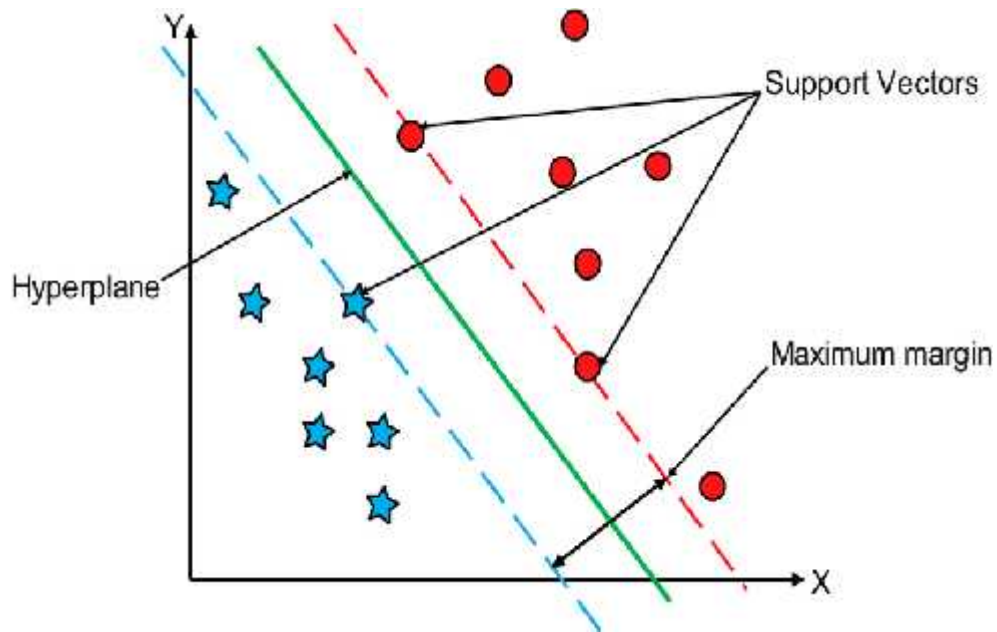


Fig. 3.6: Support Vector Classifier

3.7 Stochastic Gradient Descent

To advance a goal work with adequate perfection models, stochastic slope drop (frequently curtailed SGD) is used in an iterative procedure (for example differentiable or subdifferentiable). It very well may be viewed as a stochastic guess of inclination plunge enhancement since it replaces the genuine slope (taken from the whole informational index) with a gauge (determined from a haphazardly chosen subset of the information). This decreases the very high computational expense, which is particularly significant in high-layered streamlining issues, by considering quicker emphases in return for a lower union rate.

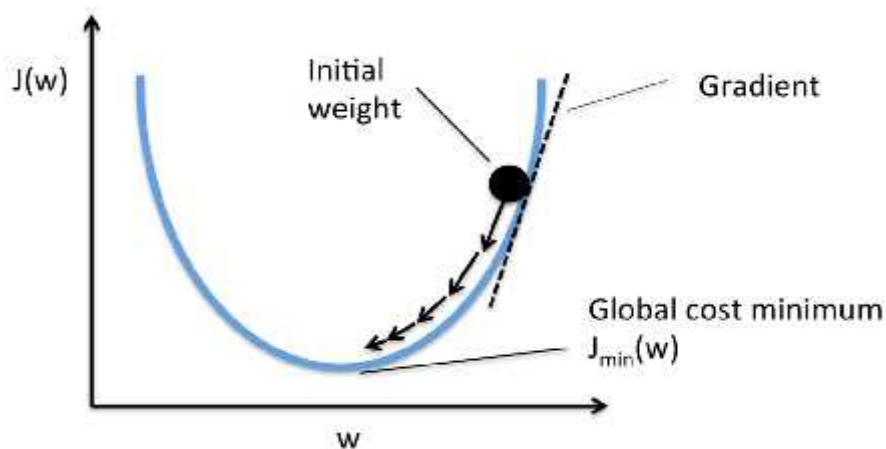


Fig. 3.7: Stochastic Gradient Descent figure

3.8 AdaBoost Classifier

An AdaBoost classifier is a meta-assessor. Firstly, the classifier performs fitting operations on the first dataset, and then, at that point, validates the fitness by employing another set of the equivalent dataset. In this process, continuously updates the weights of misclassified and wrongly estimated samples. Therefore, the succeeding classifiers center around intense cases.

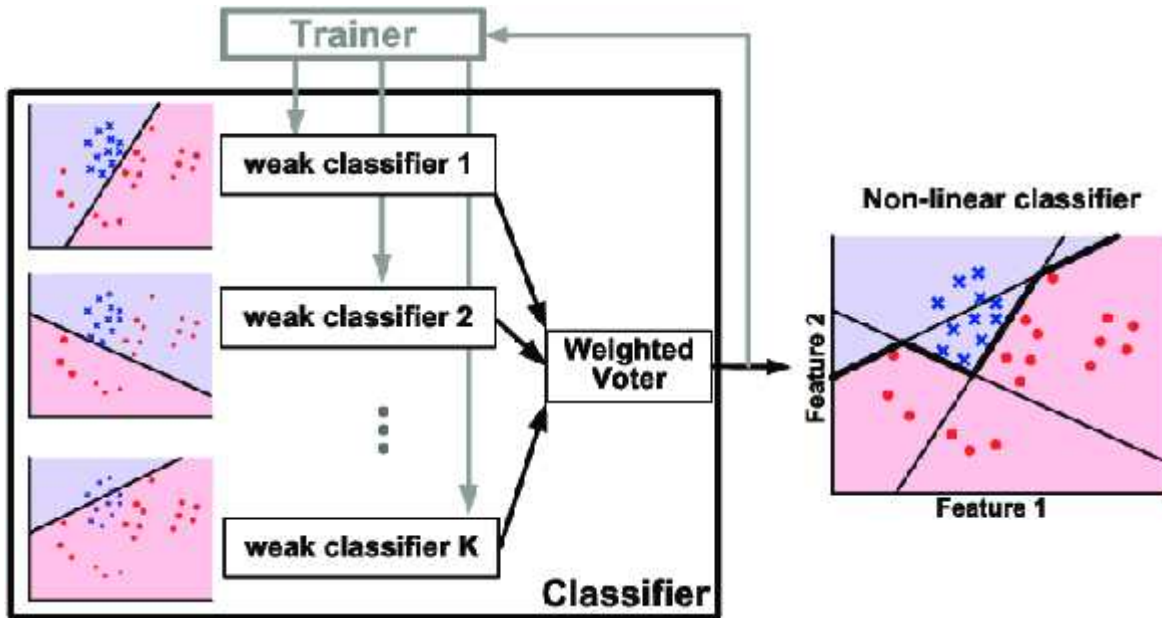


Fig. 3.8: AdaBoost Classifier figure

3.9 DT Classifier (CART)

A choice tree is a regulated learning calculation. This technique is most usually used to tackle order issues. It works brilliantly due to its constant and straight-out highlights. This calculation separates the populace into at least two related bunches in light of the main indicators. The DT calculation ascertains the entropy of every trademark from the beginning. The dataset is then separated into classifications in view of the factors or indicators with the biggest data gain or most minimal entropy.

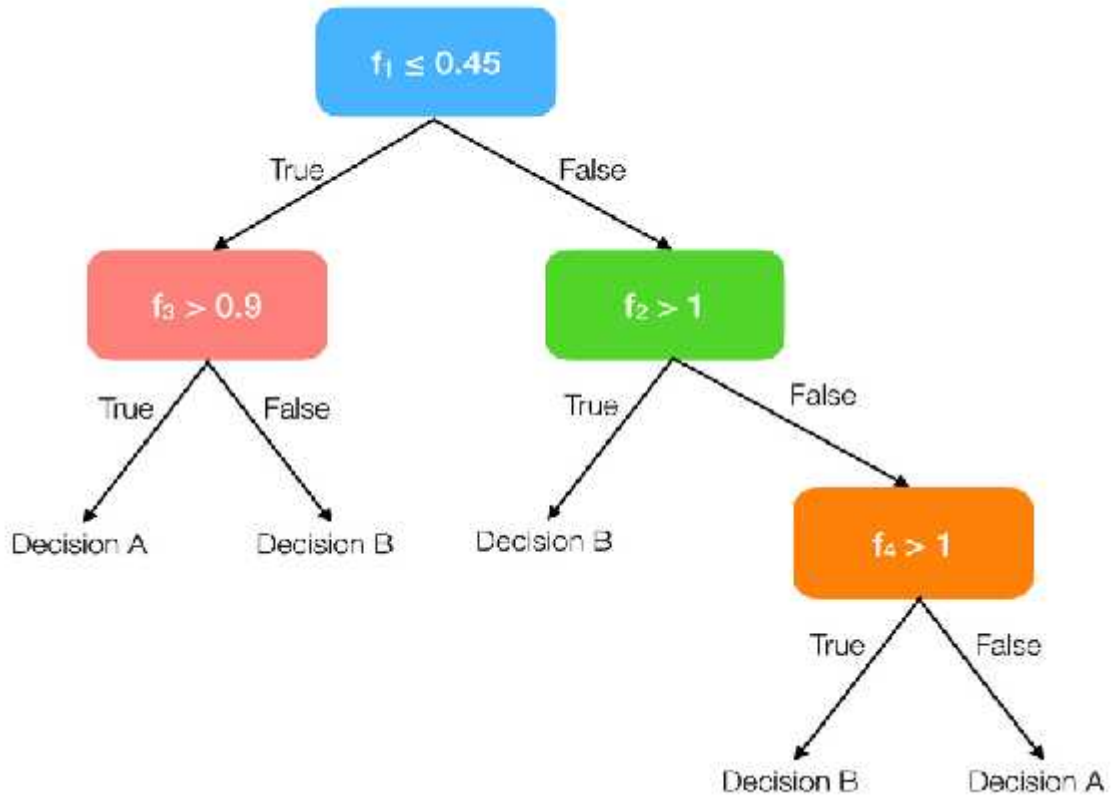


Fig. 3.9: A simple DT classifier with 4 features

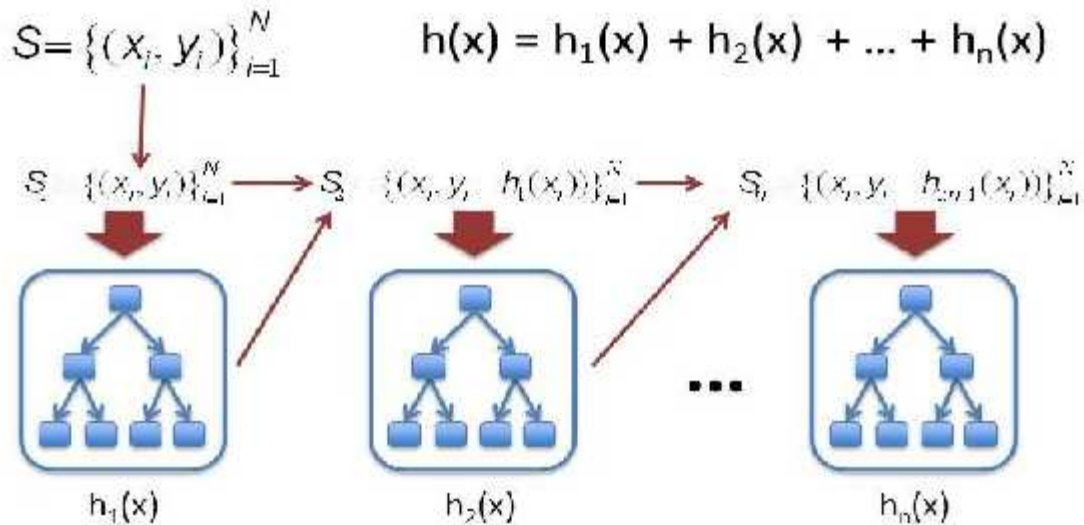
3.10 Gradient boosting machine

Inclination supporting is an AI approach that can be utilized for relapse and characterization, in addition to other things. It returns an expectation model as an assortment of frail forecast models, the most successive of which are choice trees. At the point when a choice tree is an unfortunate student, the outcome is called inclination supported trees, and it ordinarily outflanks an irregular timberland. A slope supported trees model is underlying a similar stage-by-stage way as other helping strategies, however it varies in that it can improve any differentiable misfortune work.

Gradient Boosting (Simple Version)

(Why is it called "gradient"?)
(Answer next slides.)

(For Regression Only)



<http://statweb.stanford.edu/~jbr/frog/trebst.pdf>

24

Fig. 3.10: Gradient boosting flow diagram

The performance of these developed predictive models has also been compared against various crucial parameters such as Accuracy, Precision, Sensitivity, Specificity, F1 Score, ROC, Log Loss, and Mathew correlation coefficient. These parameters can be mathematically calculated using Eq. 1– 6.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \text{-----}(1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \text{-----}(2)$$

$$\text{Sensitivity(recall)} = \text{TP} / (\text{TP} + \text{FN}) \text{-----}(3)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \text{-----}(4)$$

$$\text{MCC} = (\text{TP} * \text{TN} - \text{FP} * \text{FN}) / \sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})} \text{-----}(5)$$

$$F1 \text{ Score} = 2(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \text{-----}(6)$$

3.11 Mathew Correlation coefficient

Instead, the Matthews correlation coefficient is a more solid factual rate that possibly yields a high score assuming the expectation performed well in each of the four disarray lattice classes (genuine up-sides, bogus negatives, genuine negatives, and misleading up-sides), relatively to the size of positive and negative components in the dataset.

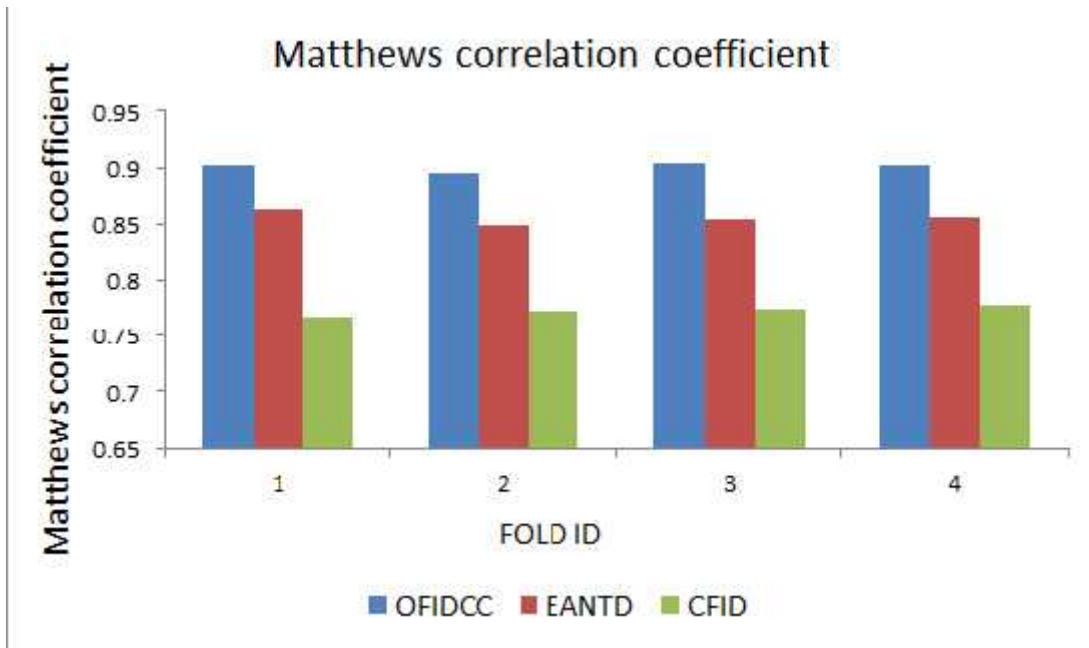


Fig. 3.11: Graphical Representation of Mathew Correlation Coefficient

3.12 Log Loss

The exhibition of an order model with a forecast contribution of a likelihood esteem somewhere in the range of 0 and 1 is estimated utilizing logarithmic misfortune. Our AI calculations plan to diminish this worth however much as could be expected. For a perfect model this loss should be 0. As the anticipated probability diverges, log loss grows. Consequently, large log loss has been witnessed even for a very small predicted probability of 0.12 when the actual observation label is 1.

Given a true observation (is Dog = 1), the graph below shows the log loss, on the other hand, rapidly grows when the expected probability lowers.

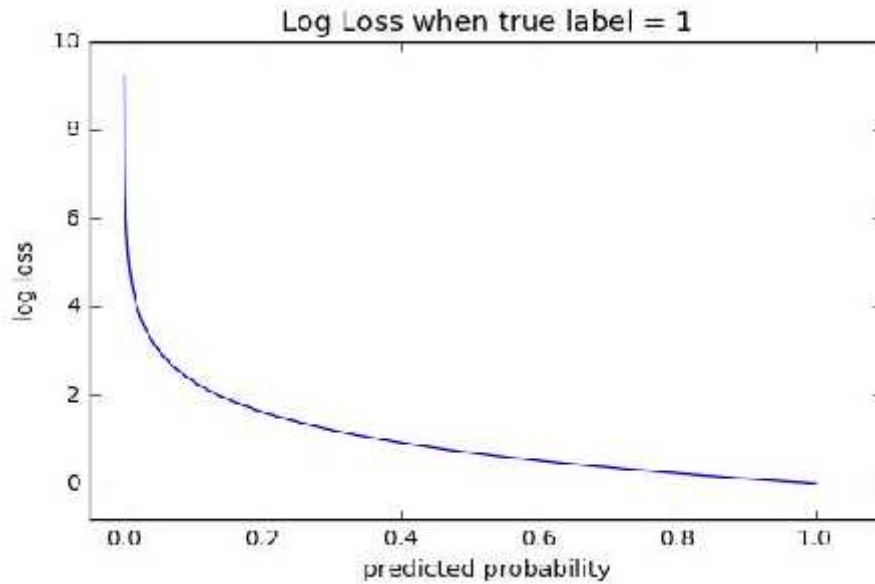


Fig. 3.12: Log Loss Graph

3.13 F1 Score

F1 Score is the weighted typical of Precision and Recall. Naturally, it isn't quite so direct as precision, but F1 is regularly more significant than accuracy, especially if you have a disproportionate class movement. Precision works best in the event that misleading up-sides and negatives have comparative expenses. Assuming the expense of misleading up-sides and bogus negatives are altogether different, it's smarter to check out at both Precision and Recall. For our situation, the F1 score is 0.701.

Where, Genuine Positive (TP) implies that the model correctly anticipates the positive samples whereas, True Negative (TN) represents the correctly predicted negative samples. Further, if the model estimates negative sample against positive sample and positive sample against negative then, it has been represented by False Negative (FN) and False Positive (FP) respectively.

Feature selection is applied to find out which are the most influencing features among the 11 features used in this study. Now top five algorithms will be selected by again applying 10-fold cross-validation by isolating the most consistent, non-redundant, and relevant features after the application of the feature selection process. Then soft voting model which is an ensemble of these top five models will be created and model evaluation is done.

CHAPTER 4: RESULTS AND DISCUSSION

Pre-processing techniques like Outlier Detection, Train test split, Cross-validation, and Feature selection are applied after data cleaning and Exploratory data analysis (EDA) to analyze their impact on the outcome using various ML techniques. Now to detect and subsequently exclude outliers from the dataset Outlier Detection is used.

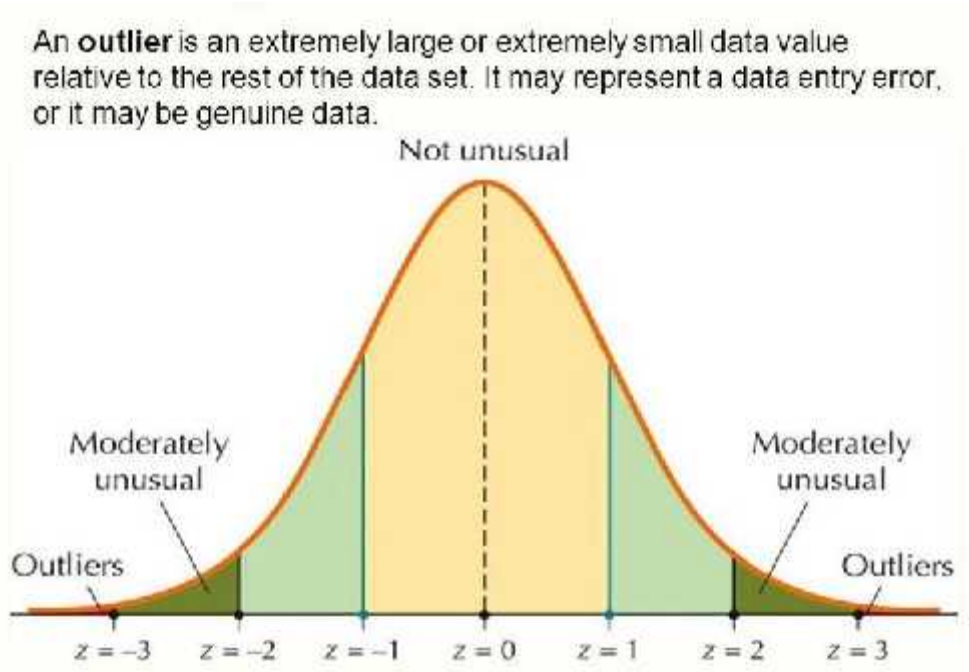


Fig. 4(a): Distribution of Numerical features

In a dataset, an exception is an information point that is phenomenally high or very low in contrast with the closest data of interest and the other close by coinciding qualities. There is a sum of 26 information focuses that are exceptions in this dataset. Exception recognition will be finished utilizing a z-score.

$$\text{Z score} = (x - \text{mean}) / \text{std. deviation} \text{-----}(7)$$

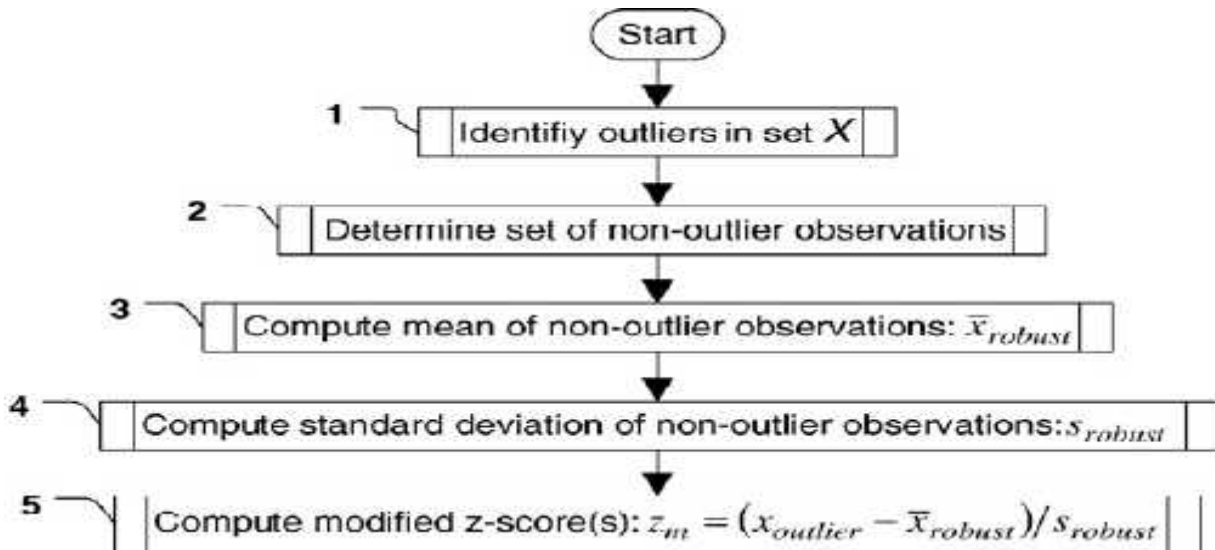


Fig. 4(b): Flow Chart for determining modified z-scores of outliers

Presently Filtering numeric highlights like age, resting bp, cholesterol, and max pulse achieved has exceptions. Subsequent to sifting the numeric highlights in the dataset z-score will be determined. Then the edge will be characterized for sifting anomalies. At last, separating the exceptions that are holding just those information focuses which are beneath the limit. All the 26 outliers are now removed. After outlier detection checking of correlation is done which is shown in the figure below.

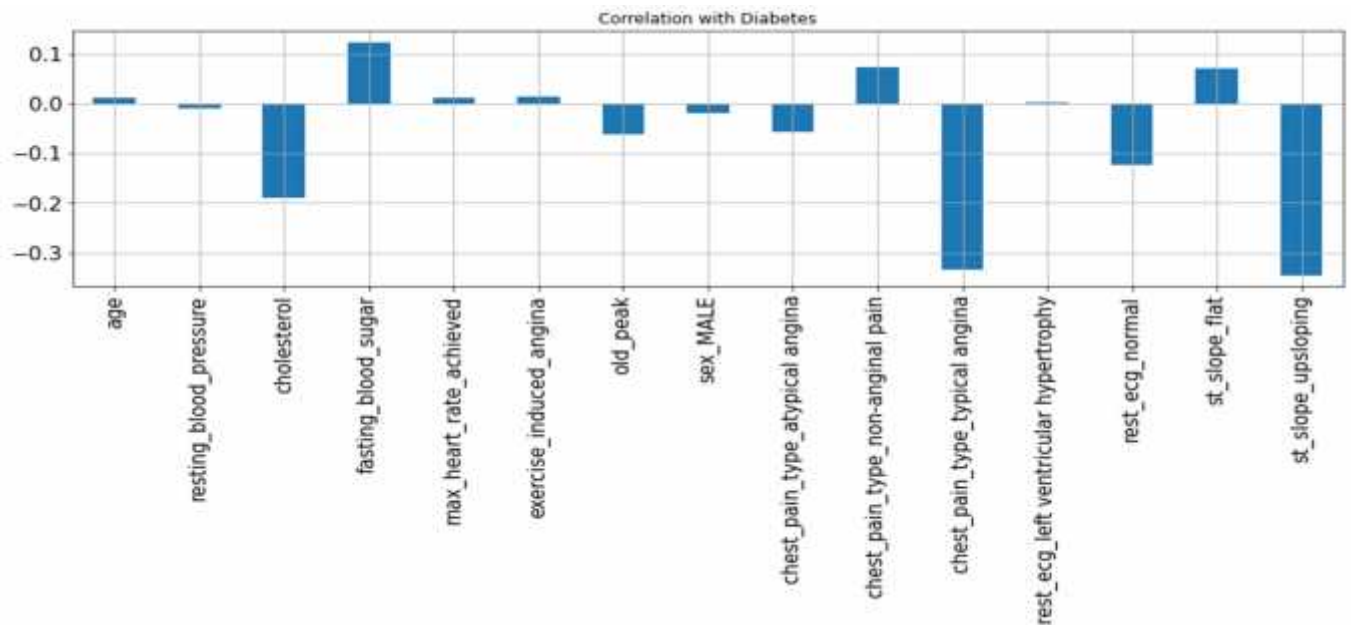


Fig. 4(c): Correlation with Response Variable class plot

Then, the informational collection has been partitioned into two sections, preparing information which is 80% of the entire informational collection, and testing information which is 20% of the entire informational index. Train Test Split is finished utilizing separated inspecting. A readiness dataset is a lot of data used to set up a model. The testing dataset is used to survey the pre-arranged model's show.

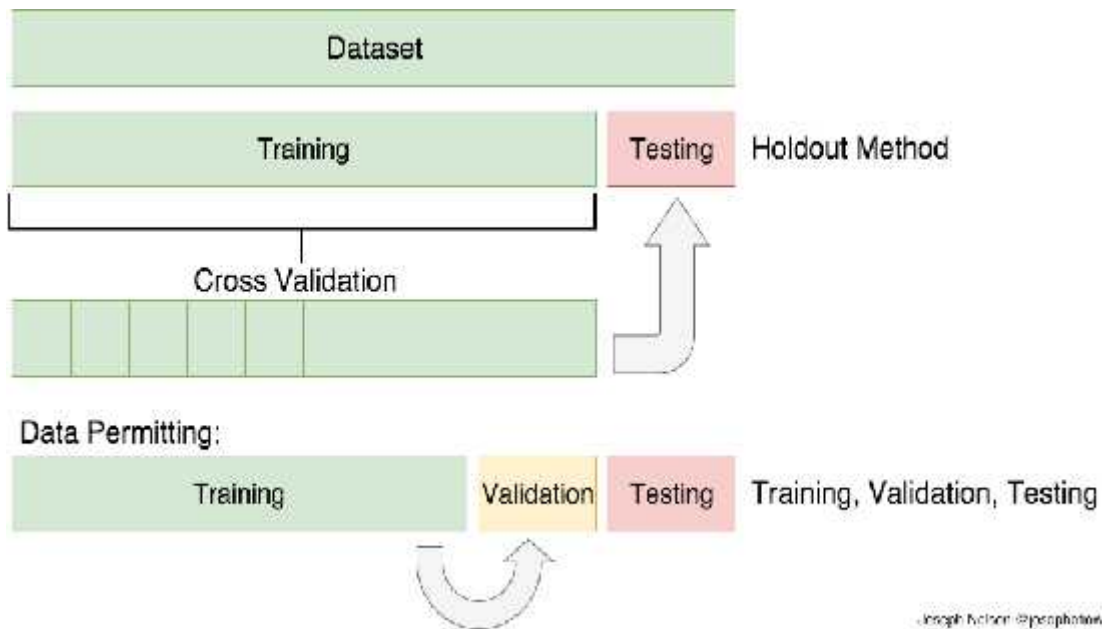


Fig. 4(d): Train Test Split distribution diagram

The method of 10-overlap cross-approval is applied to get the top-performing benchmark model.

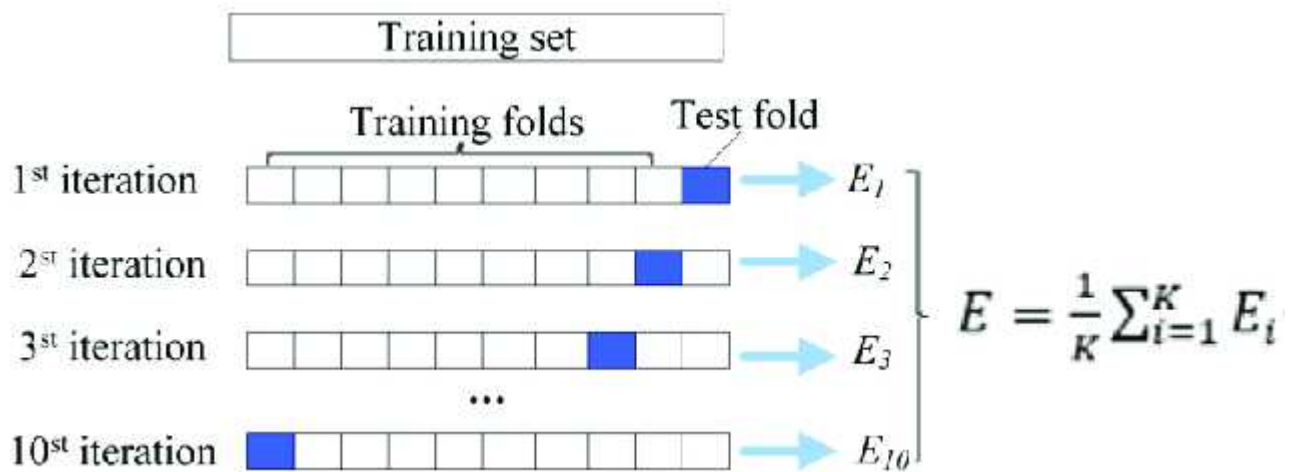


Fig. 4(e): Cross-validation diagram

After the application of cross-validation, it is found that the extra tree is the best baseline model which is compared to all the remaining 9 ML models.

Table 4(a): Comparison of Heart Disease Classification Techniques

Sr. No.	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log Loss	Mathew correlation coefficient
0	Extra Tree	0.914062	0.909091	0.931373	0.894444	0.920097	0.912908	2.968216	0.827452
1	MLP	0.833333	0.815315	0.887255	0.772222	0.849765	0.829739	5.756548	0.666373
2	K-NN	0.854167	0.839450	0.897059	0.805556	0.867299	0.851307	5.036978	0.707760
3	RF classifier	0.901042	0.891509	0.926471	0.872222	0.908654	0.899346	3.417948	0.801492
4	XGB	0.901042	0.895238	0.921569	0.877778	0.908213	0.899673	3.417946	0.801313
5	SVC	0.799479	0.784753	0.857843	0.733333	0.819672	0.795588	6.925844	0.597865
6	SGD	0.684896	0.635179	0.955882	0.377778	0.763209	0.666830	10.883546	0.415854
7	AdaBoost	0.786458	0.777273	0.838235	0.727778	0.806604	0.783007	7.375570	0.571011
8	CART	0.890625	0.889423	0.906863	0.872222	0.898058	0.889542	3.777727	0.780276
9	GBM	0.820312	0.802691	0.877451	0.755556	0.838407	0.816503	6.206278	0.640168



Fig. 4(f): ROC Curve

Obviously the most noteworthy normal region under the bend (AUC) of 0.955 is accomplished by Extra Tree Classifier.

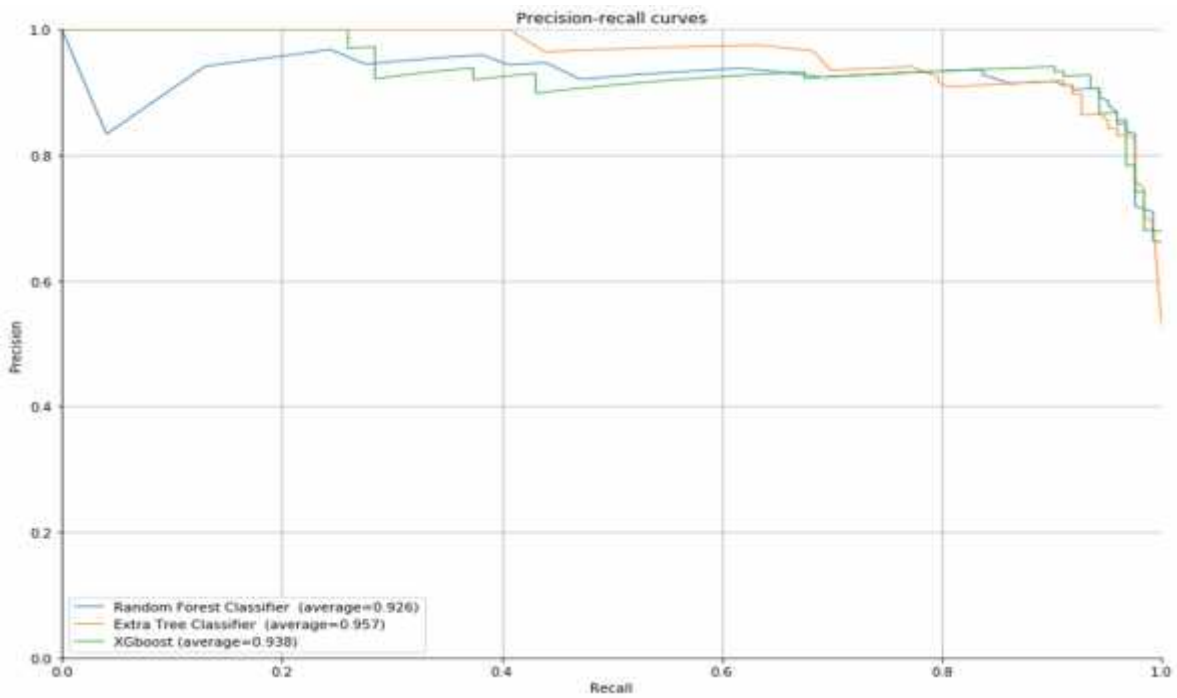


Fig. 4(g): Precision-Recall Curve

A total of 11 features are present in the dataset but all the features are not good enough to provide accurate output so some of the features had to be filtered. So, different feature selection algorithms will be used like Pearson correlation, Chi selector, Recursive Feature Elimination, Light GBM, Logistics, and RF.

4.1 Pearson correlation

The Pearson Correlation Formula is a formula that calculates the correlation between two variables. The expression "connection" alludes to the relationship that exists between two factors. Connection is estimated by the relationship coefficient. This equation is utilized to perceive how the two arrangements of information are connected. The Pearson Correlation coefficient decides the straight association between informational indexes.

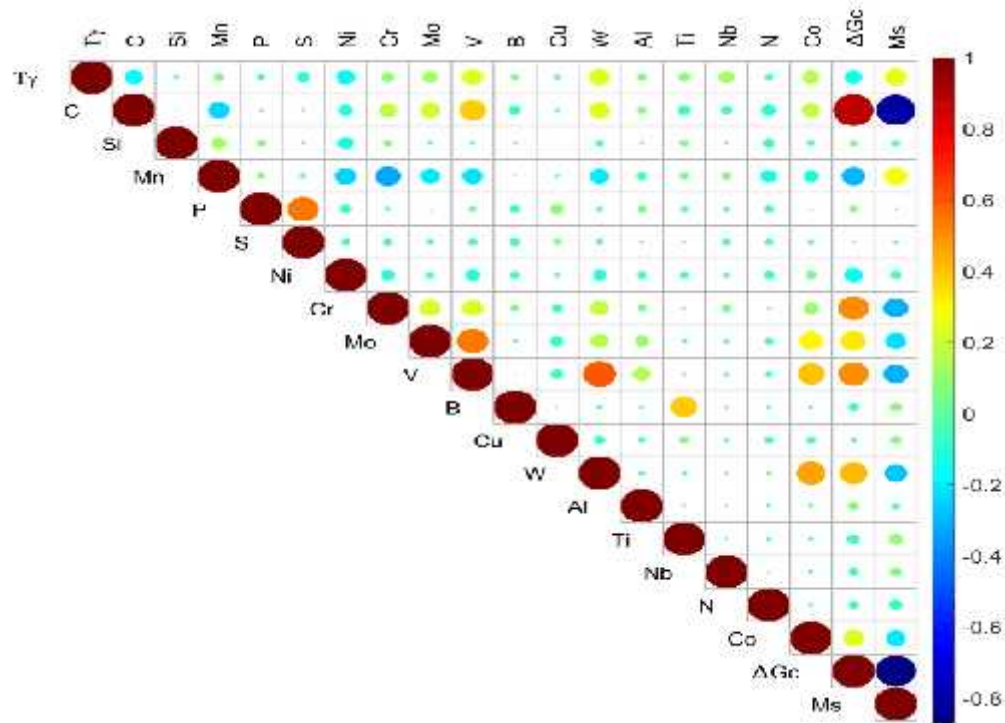


Fig. 4.1: Pearson Correlation map for feature selection

4.2 Chi Selector

We often ask about the Chi-Square test's utility in AI and how it contrasts. Highlight determination is a basic subject in AI, as we will have different elements in line and should pick the best ones to construct the model. By looking at the connection between the highlights, the chi-square test helps with the arrangement of element determination issues.

- a) Chi-Square distribution is one example.
- b) Feature Selection Using the Chi-Square Test
- c) Python-based Chi-Square Test

For straight out highlights in a dataset, the Chi-square test is used. We ascertain the Chi-square between each element and the objective and pick the highlights with the most noteworthy Chi-square scores. It decides whether the example's relationship between two all out factors mirrors their actual relationship in the populace.

Curse of Dimensionality

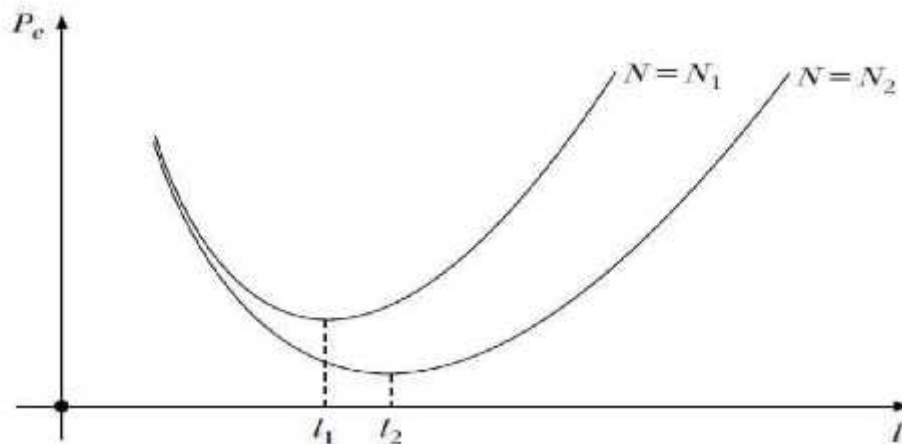


Fig. 4.2: Feature Selection by Chi-Square

4.3 Recursive Feature Elimination

RFE Feature Selection is an element choice strategy that limits the intricacy of a model by choosing significant qualities and disposing of the less significant ones. The determination technique disposes of less significant elements individually until the ideal number is reached to guarantee top execution.

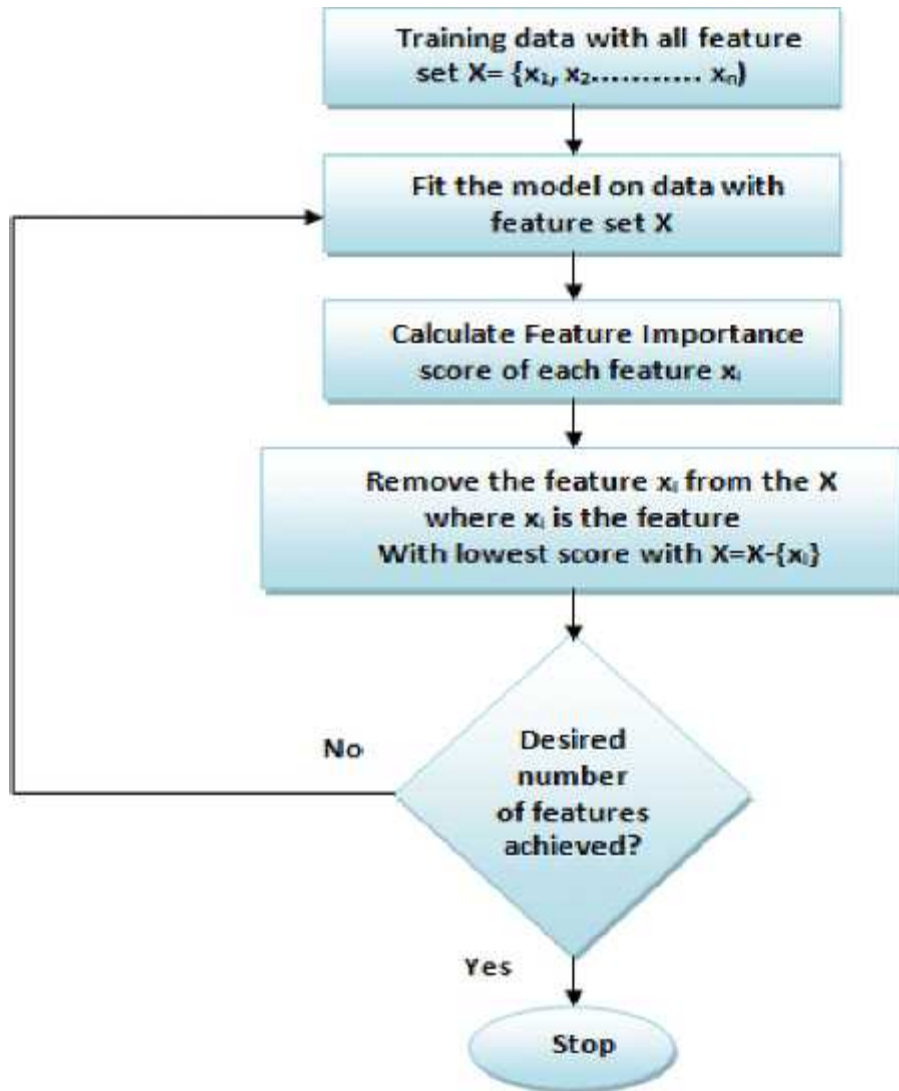


Fig. 4.3: Recursive Feature Elimination Flow

4.4 Light GBM

Light GBM, short for Light Gradient Boosting Machine, is a disseminated slope supporting framework for AI that was made by Microsoft. It is utilized for positioning, order, and other AI applications and depends on choice tree calculations. The accentuation during improvement is on execution and versatility.

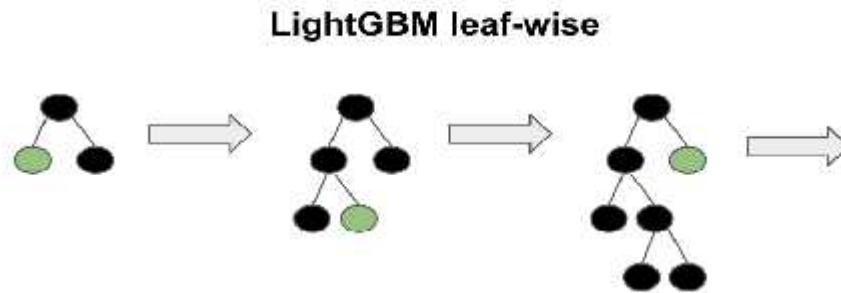


Fig. 4.4: Light GBM Architecture

4.5 LR Model

Feature selection is a feature engineering component that removes irrelevant features and chooses the best set of features for training a robust ML model. The curse of dimensionality is avoided by using feature selection methods to reduce the dimensionality of the data.

To remove redundant features from a dataset, Lasso Regression (LR with L1-regularization) can be used. L1-regularization adds sparsity to the dataset and reduces the value of redundant feature coefficients to 0. It is a highly handy technique or hacks for reducing the dataset's dimensionality by deleting extraneous features.

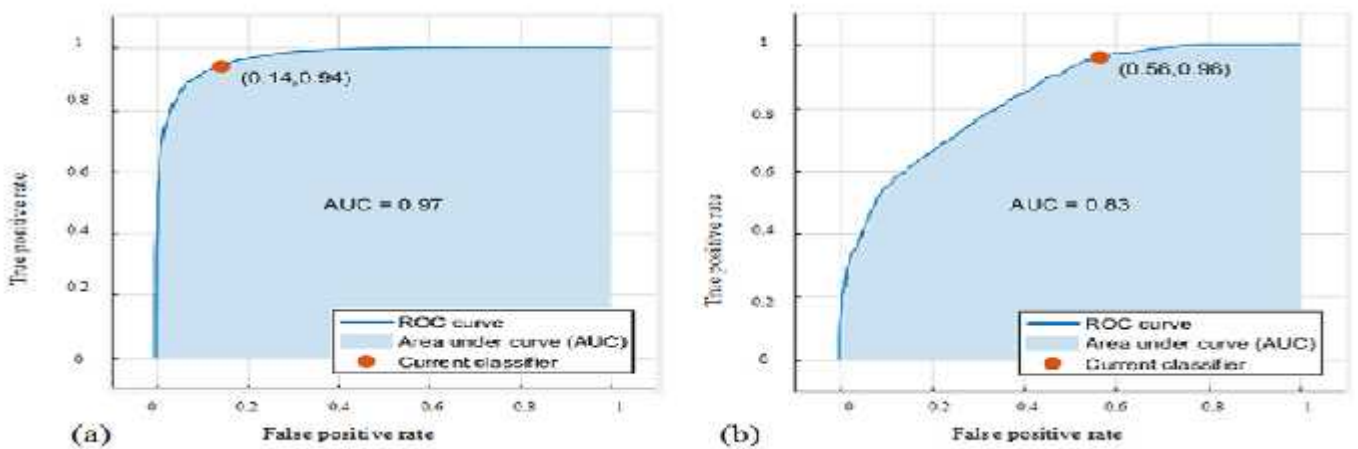


Fig. 4.5: Selection of the parameter () in the LASSO model by 10-fold cross-validation based on minimum criteria.

4.6 RF for Feature Selection

They additionally offer two basic element choice techniques: mean lessening pollutant and mean abatement exactness. An irregular woodland is comprised of a few choice trees. Each hub in the choice trees is a condition on a specific component, to part the dataset into two gatherings with comparative reaction values.

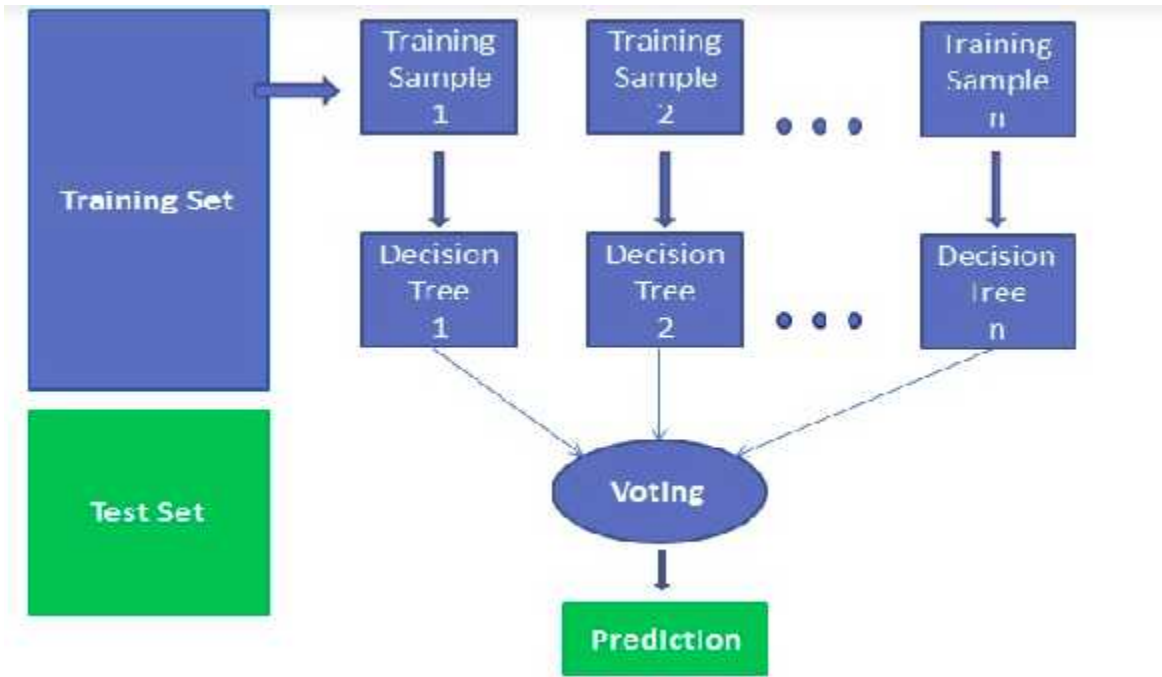


Fig. 4.6(a): RF for Feature Selection

Now it will be compared which feature has the majority number of supports in terms of feature selection algorithms. As it is clear from figure 3 the top 5 most contributing features are Cholesterol, Max heart Rate achieved, Age, Old peak, and slope upsloping.

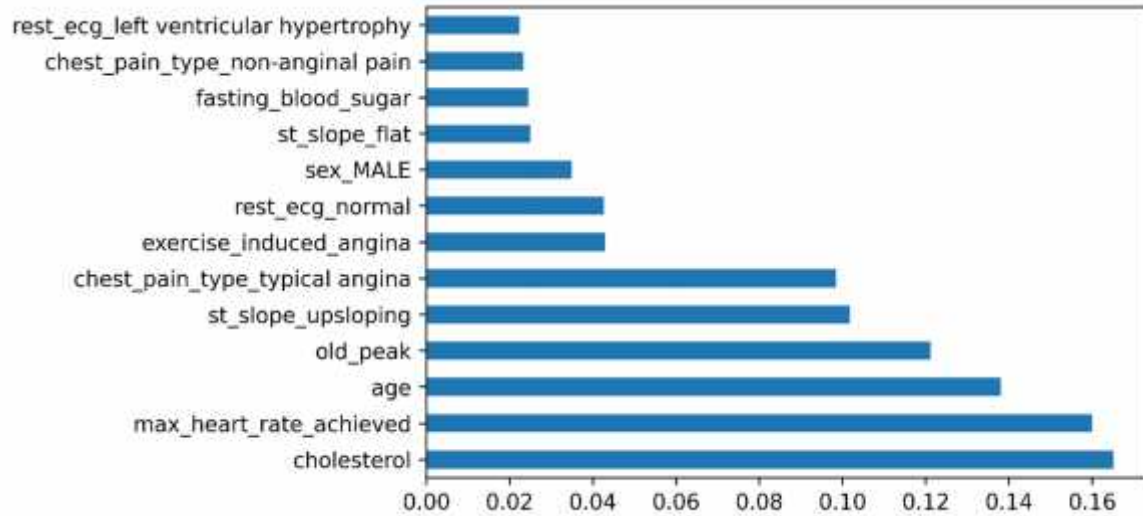


Fig. 4.6(b): Distribution of most contributing features

Now, 10-fold cross-validation is applied to filter out the top 5 ML classification techniques after isolating the most relevant features on the application of the feature selection process. The top 5 ML techniques which are: RF, DT, XG Boost, Extra Tree, and Gradient Boosting are now used to create a Soft voting ensemble model

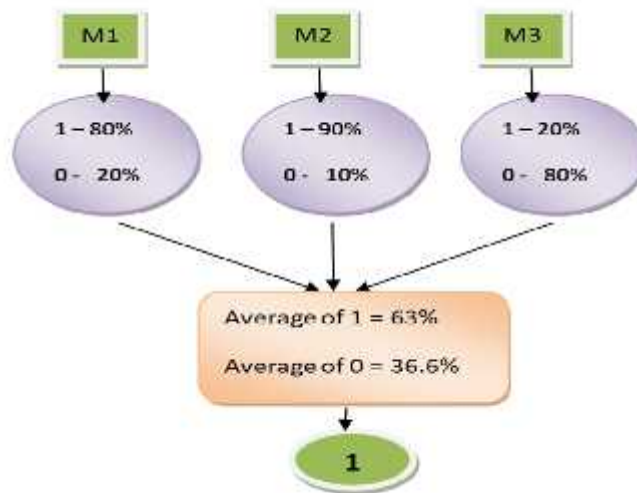


Fig. 4.6(c): Hard vs Soft voting Classifier

further which soft voting is compared with all the 10 ML techniques. Table 4 shows all the 10 ML techniques used along with the soft voting ensemble technique with their Accuracy, Precision, Sensitivity, Specificity, F1 Score, ROC, Log Loss, and Mathew correlation coefficient. ROC Curve and Precision-Recall curve is shown in Figure 39 & 40.

Table 4(b): Analysis of ML Techniques with Soft Voting

Sr. No.	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log Loss	Mathew correlation coefficient
0	Soft Voting	0.914062	0.901408	0.941176	0.883333	0.920863	0.912255	2.968220	0.827865
1	RF entropy	0.908854	0.893023	0.941176	0.872222	0.916468	0.906699	3.148113	0.817697
2	MLP2	0.820312	0.813953	0.857843	0.777778	0.835322	0.817810	6.206270	0.638980
3	K-NN2	0.833333	0.836538	0.852941	0.811111	0.844660	0.832026	5.756534	0.665067
4	Extra tree classifier	0.927083	0.935644	0.926471	0.927778	0.931034	0.927124	2.518480	0.853737
5	XGB2	0.906250	0.900000	0.926471	0.883333	0.913043	0.904902	3.238054	0.811796
6	SVC2	0.796875	0.781250	0.857843	0.727778	0.817757	0.792810	7.015791	0.592767
7	SGD2	0.778646	0.769231	0.833333	0.716667	0.800000	0.775000	7.645408	0.555295
8	AdaBoost	0.786458	0.779817	0.833333	0.733333	0.805687	0.783333	7.375568	0.570817
9	CART	0.898438	0.890995	0.921569	0.872222	0.906024	0.896895	3.507892	0.796147
10	GBM	0.820312	0.797357	0.887255	0.744444	0.839907	0.815850	6.206282	0.641208

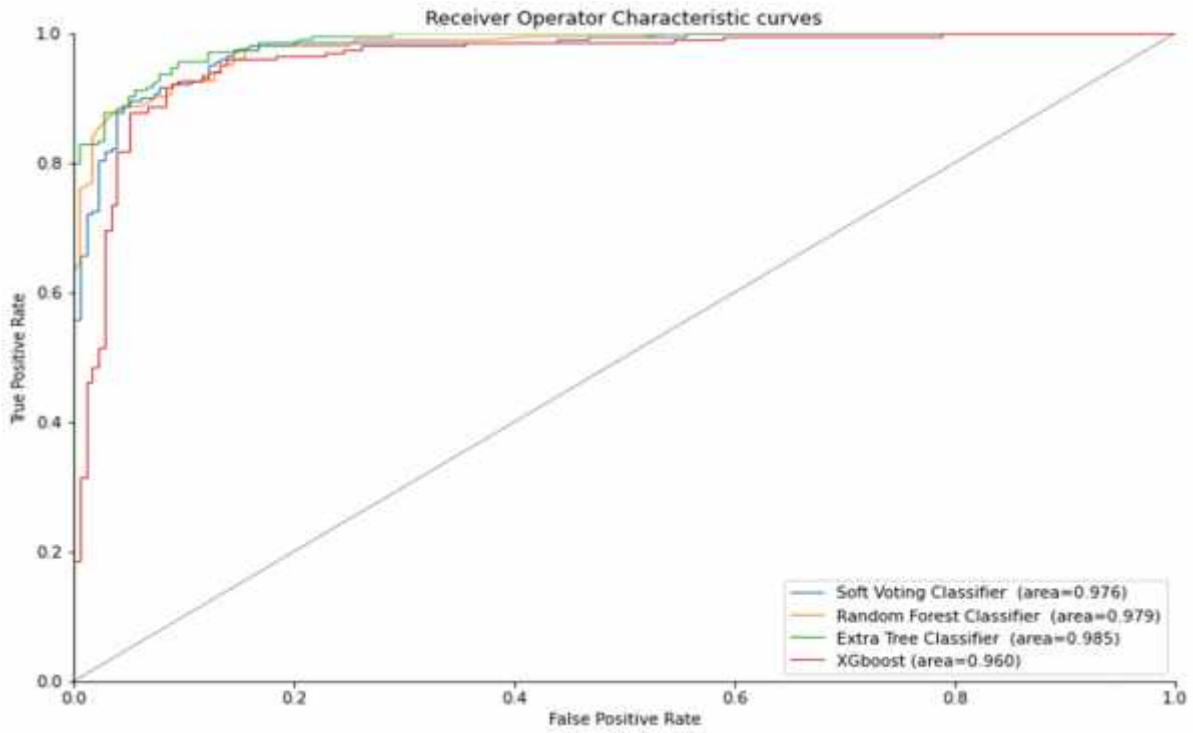


Fig. 4.6(d): ROC Curve

The most noteworthy typical region under the bend (AUC) of 0.985 is accomplished by Extra Tree Classifier.

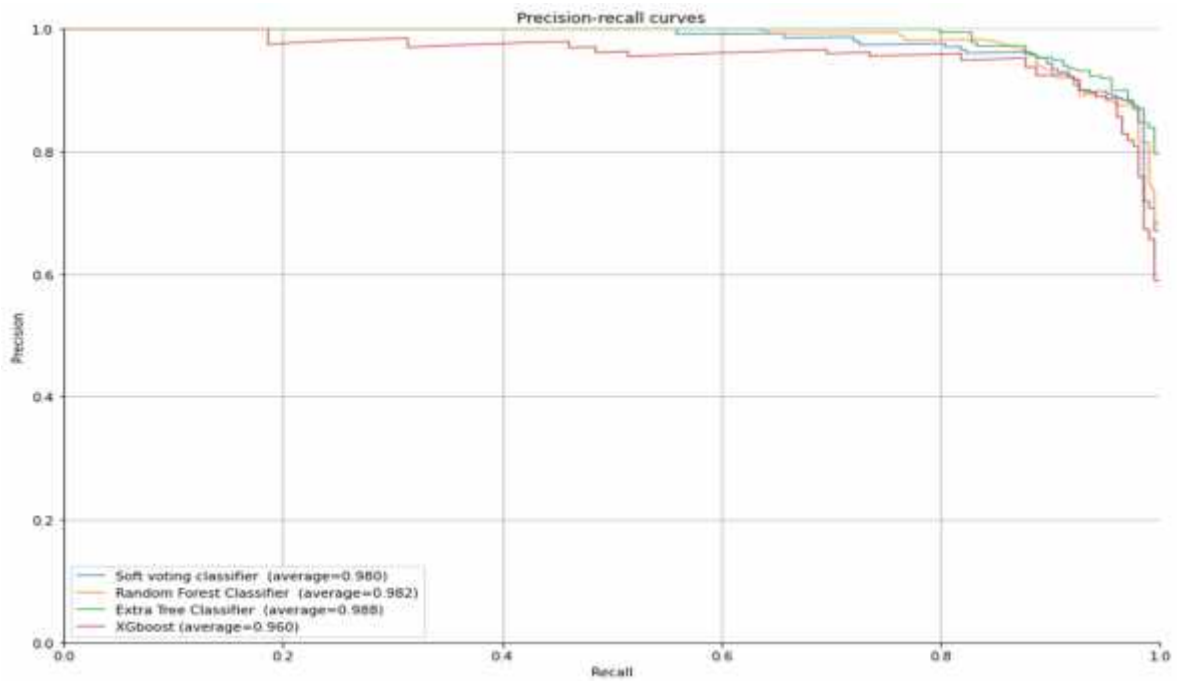


Fig. 4.6(e): Precision-Recall Curve

CHAPTER 4: CONCLUSION AND FUTURE SCOPE

With the rising figure of deaths because of coronary illness, it has become important to plan a framework that can practically and definitively expect coronary illness. The current work used multiple ML methods to create a heart disease prediction model. Information assortment is done utilizing various sources that are essential variables liable for any kind of coronary illness expectation. The results show that the Extra tree classifier performs admirably by yielding an accuracy of 0.927 and may thus be used as a predictive tool for fetal health monitoring. In Table. 2 Extra Tree Classifier provides an accuracy of 0.9270, sensitivity of 0.9313, specificity of 0.922, and highest f1-score of 0.9313 whereas, whereas Soft Voting yields an accuracy of 0.9088, sensitivity of 0.9362, and specificity of 0.8777. We have also discovered the top 5 most contributing features that are Cholesterol, Max heart Rate achieved, Age, Old peak, and st slope upsloping. The paper's future scope is to anticipate cardiac illnesses utilizing advanced approaches and algorithms that require less time.

REFERENCES

1. Anjan Nikhil Repaka, Ramya G Franklin, Design and Implementation Heart Disease Prediction Using Naive Bayesian, International Conference on Trends in Electronics and Information (ICOEI 2019).
2. Fahd Saleh Alotaibi, Implementation of ML Model to Predict Heart Failure Disease, International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
3. Nagaraj M Lutimath, Chethan C., Prediction of Heart Disease using ML, International Journal of Recent Technology and Engineering, (2S10), pp 474-477, 2019.
4. T. Nagamani, S. Logeswari, Heart Disease Prediction using Data Mining with MapReduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
5. Rishabh Magar, Rohan Memane, Heart disease prediction using ML, International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.7, I
6. Apurb Rajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using ML, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020).
7. P. Umasankar, V. Thiagarasu, “Decision Support System for Heart Disease Diagnosis Using Interval Vague Set and Fuzzy Association Rule Mining”, 2018 4th International Conference on Devices, Circuits and Systems (ICDCS).
8. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid ML Techniques”, IEEE Access, 2019, Volume: 7.
9. Aditi Gavhane, GouthamiKokkula, Isha Pandya, Prof. Kailas Devadkar, “Prediction of Heart Disease Using ML”, 2018 Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA).
10. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, 2018, 4th International Conference on Computational Intelligence & Communication Technology (CICT).
11. Harshit Jindal et al 2021, Heart disease prediction using ML algorithms, IOP Conf. Ser.:

Mater. Sci. Eng. 1022 012072.

12. Ravindhar NV, Anand, Hariharan Shanmugasundaram, Ragavendran, Godfrey Winstler. Intelligent Diagnosis of Cardiac Disease Prediction using ML. Volume-8 Issue-11, September 2019, ISSN: 2278-3075 (Online). Page No: 1417-1421. DOI: 10.35940/ijitee.J9765.0981119.
13. Akella, Aravind and Akella, Sudheer. ML algorithms for predicting coronary artery disease: efforts toward an open-source solution. Future Science OA Volume 7, Number 6, Pages FSO698, 2021.
14. N. Saranya, P. Kaviyarasu, A. Keerthana, C. Oveya. Heart Disease Prediction using ML International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020, Page No: 700-70.
15. Aadar Pandita, Siddharth Vashisht, Aryan Tyagi, Prof. Sarita Yadav. "Prediction of Heart Disease using ML Algorithms", Volume 9, Issue V, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 2422-2429, ISSN: 2321-9653.
16. Garg, Apurv & Sharma, Bhartendu & Khan, Rizwan. Heart disease prediction using ML techniques. 2021 IOP Conference Series: Materials Science and Engineering. 1022. 012046. 10.1088/1757-899X/1022/1/012046.
17. Chaitanya Suvarna, Abhishek Sali, Sakina Salmani, "Efficient heart disease prediction system using optimization technique", International Conference on Computing Methodologies and Communication (ICCMC) 2017.
18. Rashmi G Saboji, "A scalable solution for heart disease prediction using classification mining technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).
19. Sowmiya, C., and P. Sumitra. 2017. Analytical study of heart disease diagnosis using classification techniques. In the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS).
20. Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System" Vol:85, pp: 962 – 969, (2018)
21. S. Bhagavathy, V. Gomathy, S. Sheeba Rani, Sujatha. K, "Early Heart Disease Detection Using Data Mining Techniques with Hadoop MapReduce", International Journal of Pure

- and Applied Mathematics, 119(12), 1915-1920, 2018
22. Muthuvel Marimuthu & Sivaraju Deivarani & Ramamoorthy Gayathri (AISGSC 2019) "Analysis of heart disease prediction using various ML techniques"
 23. C. T. and A. Choudhary, "Heart Disease Diagnosis using a ML Algorithm," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019, pp. 1-4
 24. M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telemat. Informatics, vol. 36, no. November, pp. 82–93, 2019.
 25. L. Verma, S. Srivastava, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 178, 2016, DOI: 10.1007/s10916-016-0536-z.
 26. Frantisek Babi , Jaroslav Olejář, Zuzana Vantová, Jan Paralic, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, Prague, 2017, ISSN 2300-5963 ACSIS, Vol. 11, DOI: 10.15439/2017F219, pp. 155–163.
 27. R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, pp. 459–468, DOI: 10.1016/j.procs.2015.09.132.
 28. Ch. Yadav, S. Lade, and M. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining", International Journal of Computer Applications, vol. 87, no. 4, 2014, pp. 9-13.
 29. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K, "Smartphone-based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design." In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.
 30. E. Maini, B. Venkateswarlu, and A. Gupta, "Applying ML Algorithms to Develop a Universal Cardiovascular Disease Prediction System," in International Conference on Intelligent Data Communication Technologies and Internet of Things, 2018, pp. 627– 632.
 31. S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool," in Soft

- Computing and Signal Processing, Singapore: Springer, 2019, pp. 111–117.
32. K. Deepika and S. Seema, “Predictive analytics to prevent and control chronic diseases,” in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016.
 33. Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., “Disease prediction by ML over big data from healthcare communities,” *IEEE Access*, 5, pp.8869-8879
 34. Tikotikar, A., & Kodabagi, M., “A survey on technique for prediction of disease in medical data,” In 2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con) (pp. 550-555). IEEE.
 35. Raju, C., Philips, E., Chacko, S., Suresh, L.P. and Rajan, S.D., 2018, “A Survey on Predicting Heart Disease using Data Mining Techniques.” In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS) (pp. 253-255). IEEE
 36. Praveen Kumar Reddy, M., Sunil Kumar Reddy, T., Balakrishnan, S., Syed Muzamil Basha, & Ravi Kumar Poluru., “Heart Disease Prediction Using ML Algorithm,” *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, Volume-8 Issue-10.
 37. Huang, F., Wang, S. and Chan, C.C., 2012, “Predicting disease by using data mining based on healthcare information systems,” In 2012 IEEE International Conference on granular computing (pp. 191-194).
 38. Amiri, A.M. and Armano, G., 2013, “Early diagnosis of heart disease using classification and regression trees,” In The 2013 International Joint Conference on NNs (IJCNN) (pp. 1-4).
 39. Banu, M.N. and Gomathy, B., 2014, “Disease forecasting system using data mining methods,” In 2014 International conference on intelligent computing applications (pp. 130-133). IEEE.
 40. Ankita Dewan, Meghna Sharma, “Prediction of heart disease using a hybrid technique in data mining classification”, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).
 41. Jabbar, Shirina Samreen, “Heart disease prediction system based on hidden NB classifier”, 2016 International Conference on Circuits, Controls, Communications and Computing (I4C)

42. Monika Gandhi, Shailendra Narayan Singh, “Predictions in heart disease using techniques of data mining”, 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)
43. K. Prasanna Lakshmi, C. R. K. Reddy, “Fast rule-based heart disease prediction using associative classification mining”, 2015 International Conference on Computer, Communication, and Control (IC4)
44. Jagdeep Singh, Amit Kamra, Harbhag Singh, “Prediction of heart diseases using associative classification”, 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON)
45. N. Priyanka, Pushpa Ravikumar, “Usage of data mining techniques in predicting the heart diseases — NB & DT”, 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT).
46. J. Thomas, R Theresa Princy, “Human heart disease prediction system using data mining techniques”, 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT).
47. Purushottam, Kanak Saxena, Richa Sharma, “Efficient heart disease prediction system using DT”, 2015, International Conference on Computing, Communication & Automation.
48. R. Wu, W. Peters, M.W. Morgan, “The Next Generation Clinical Decision Support: Linking Evidence to Best Practice”, *Journal of Healthcare Information Management*. 16(4), pp. 50-55, 2002.
49. Mary K. Obenshain, “Application of Data Mining Techniques to Healthcare Data”, *Infection Control and Hospital Epidemiology*, vol. 25, no.8, pp. 690–695, Aug. 2004.
50. G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J.D. MartinGuerrero, E. Soria-Olivas, L. Alonso-Chorda, J. Moreno, “Robust support vector method for hyperspectral data classification and knowledge discovery.” *Trans. Geosci. Rem. Sens.* vol.42, no.7, pp.1530–1542, July.2004.
51. R.B. Burrows, J. Montpetit, M. Faucher, J. Walmsley, “CART-NEUROFUZZY statistical data modeling”, 14th Conference on Probability and Statistics in the Atmospheric Sciences, pp. 160–167. Amer. Meteorological Society, MA.,1998.
52. JSR Jang, CT Sun, E Mizutani, *Neuro-Fuzzy and Soft-Computing- A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall, Englewood Cliffs, NJ,

1997.

53. Taho Yang, Huan-Chang Lin and Meng-Lun Chen “Metamodeling approach in solving the machine parameters optimization problem using NN and genetic algorithms: a case study”. *Rob. Comp-Int. Manuf.* vol. 22, no.4, pp. 322– 331, Aug. 2006.
54. Stuart Russell, and Peter Norvig, “Artificial Intelligence: A Modern Approach”. Prentice-Hall, Englewood Cliffs, NJ,1995
55. Sellappan Palaniappan, Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques” International Conference on Computer Systems and Applications, April 2008. pp.108- 115, AICCSA 2008. IEEE/ACS.
56. Magnus Stensmo and Terrence J. Sejnowski “Automated Medical Diagnosis based on Decision Theory and Learning from Cases “World Congress on NNs 1996 International NNsociety pp. 1227-123
57. McGill H. C., McMahan C. A., Gidding S. S. Preventing heart disease in the 21st century. *Circulation.* 2008;117(9):1216–1227. DOI: 10.1161/circulationaha.107.717033.
58. Amin M. S., Chiam Y. K., Varathan K. D. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics.* 2019; 36:82–93. DOI: 10.1016/j.tele.2018.11.007.
59. Nalluri S., Saraswathi R. V., Ramasubbareddy S., Govinda K., Swetha E. *Data Engineering and Communication Technology.* Singapore: Springer; 2020. Chronic heart disease prediction using data mining techniques, advances in intelligent systems and computing; pp. 903–912.
60. Louridi N., Amar M., Ouahidi B. E. Identification of cardiovascular diseases using ML. Proceedings of the 2019 7th Mediterranean Congress of Telecommunications (CMT); October 2019; Fez, Morocco. IEEE; pp. 1–6.
61. Shah D., Patel S., Bharti S. K. Heart disease prediction using ML techniques. *SN Computer Science.* 2020;1(6):1–6. DOI: 10.1007/s42979-020-00365-y.
62. Kumar A., Kumar P., Srivastava A., Kumar V. D. A., Venkatesan K., Singhal A. Comparative analysis of data mining techniques to predict heart disease for diabetic patients. Proceedings of the International Conference on Advances in Computing and Data Sciences; April 2020; Valletta, Malta. Springer; pp. 507–518.
63. Pires I. M., Marques G., Garcia N. M., Ponciano V. ML for the evaluation of the presence

- of heart disease. *Procedia Computer Science*. 2020; 177:432–437. DOI: 10.1016/j.procs.2020.10.058.
64. Kavitha M., Gnaneswar G., Dinesh R., Sai Y. R., Suraj R. S. Heart disease prediction using hybrid ML model. Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT); January 2021; Coimbatore, India. IEEE; pp. 1329–1333.
 65. Spencer R., Thabtah F., Abdelhamid N., et al. Exploring feature selection and classification methods for predicting heart disease. *Digital health*. 2020;6 DOI: 10.1177/2055207620914777.
 66. Khan M. A. An IoT framework for heart disease prediction based on MDCNN classifier. *IEEE Access*. 2020;8 DOI: 10.1109/access.2020.2974687.34717.
 67. Mohan S., Thirumalai C., Srivastava G. Effective heart disease prediction using hybrid ML techniques. *IEEE Access*. 2019;7 DOI: 10.1109/access.2019.2923707.81542.
 68. Magesh G., Swarnalatha P. Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence*. 2021;14(2):583–593. DOI 10.1007/s12065-019-00336-0.
 69. Mehmood A., Iqbal M., Mehmood Z., et al. Prediction of heart disease using deep convolutional NNs. *Arabian Journal for Science and Engineering*. 2021;46(4):3409–3422. DOI: 10.1007/s13369-020-05105-1.
 70. Vinutha H. P., Poornima B., Sagar B. M. Detection of outliers using interquartile range technique from intrusion dataset. *Advances in Intelligent Systems and Computing*. 2018; 701:511–518. DOI: 10.1007/978-981-10-7563-6_53.
 71. Shanker M., Hu M. Y., Hung M. S. Effect of data standardization on NN training. *Omega*. 1996;24(4):385–397. DOI: 10.1016/0305-0483(96)00010-2.
 72. Chandrashekar G., Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014;40(1):16–28. DOI: 10.1016/j.compeleceng.2013.11.024.
 73. Kukreja S. L., Löfberg J., Brenner M. J. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC proceedings volumes*. 2006;39(1):814–819. DOI: 10.3182/20060329-3-au-2901.00128.
 74. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward NNs. Proceedings of the Thirteenth International Conference on Artificial Intelligence and

- Statistics, PMLR; May 2010; Sardinia, Italy. pp. 249–256.
75. He K., Zhang X., Ren S., Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Proceedings of the IEEE International Conference on Computer Vision (ICCV); December 2015; Santiago, Chile. pp. 1026–1034.
 76. Ramprakash P., Sarumathi R., Mowriya R., et al. Heart disease prediction using deep NN. Proceedings of the International Conference on Inventive Computation Technologies (ICICT); February 2020; Coimbatore, India. IEEE; pp. 666–670.
 77. Gao X. Y., Ali A. A., Hassan H. S., Anwar E. M. Improving the accuracy for analyzing heart disease prediction based on the ensemble method. *Complexity*. 2021; 2021:10. DOI10.1155/2021/6663455.6663455.
 78. Ali M. M., Kumar B. P., Ahmad K., Francis M. B., Julian M. W. Q., Moni M. A. Heart disease prediction using supervised ML algorithms: performance analysis and comparison. *Computers in Biology and Medicine*. 2021;136104672.
 79. Rahman M., Zhan M. M., Islam L. Effective prediction on heart disease: anticipating heart disease using data mining techniques. Proceedings of the 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT); November 2019; Tirunelveli, India. pp. 536–541.
 80. Ponikowski, P., Voors, A.A., Anker, S.D., Bueno, H., Cleland, J.G., Coats, A.J., Falk, V., González-Juanatey, J.R., Harjola, V.P., Jankowska, E.A., et al.: 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European society of cardiology (ESC) Developed with the special contribution of the heart failure association (HFA) of the ESC. *Eur. Heart J.* 37(27), 2129–2200 (2016).
 81. Aljaaf, A., Al-Jumeily, D., Hussain, A., Dawson, T., Fergus, P., Al-Jumaily, M.: Predicting the likelihood of heart failure with a multi-level risk assessment using a DT. In: 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECE), pp. 101–106. IEEE (2015)
 82. Son, C.S., Kim, Y.N., Kim, H.S., Park, H.S., Kim, M.S.: Decision-making model for early diagnosis of congestive heart failure using rough set and DT approaches. *J. Biomed. Inform.* 45(5), 999–1008 (2012)

83. Hartmann, C., Varshney, P., Mehrotra, K., Gerberich, C.: Application of information theory to the construction of efficient DTs. *IEEE Trans. Inf. Theory* 28(4), 565–577 (1982)
84. Haykin, S.: *NNs: a comprehensive foundation*, 2nd edn. Prentice-Hall PTR, Upper Saddle River, NJ, USA (1998)
85. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: *Classification and Regression Trees*. CRC Press (1984)
86. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11(1), 10–18 (2009)
87. Zadeh, L.A.: Fuzzy sets. In: Lotfi A.Z. (ed.) *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*, pp. 394–432. World Scientific (1996)
88. Hartigan, J.A., Wong, M.A.: Algorithm as 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* 28(1), 100–108 (1979)

ACCEPTANCE OF PAPER 1 PRESENTATION

ICSIS-2022 Acceptance for paper presentation 5886 Inbox x



ICSIS-2022 <icis2022@easychair.org>
to me ▾

Thu, 5 May, 11:02 ☆ ↶ ⋮

Dear Author - Abhinav Dubey

Greetings from Poomima Institute of Engineering and Technology, Jaipur!!

Please accept our gratitude towards your paper in our International Conference "ICSIS-2022" scheduled on 6-7 May 2022.

This has reference to your manuscript submitted for ICSIS-2022,

Paper id: 5886

Title: Machine Learning for Heart Disease Prediction: Recent Trends and Major Challenges

Based on reviewers consent, we are pleased to inform you that your paper is accepted for presentation in the conference. Please quote your paper-ID for future correspondence.

You are advised to Register by Below Google form link for the conference latest by today only before 3.00 pm . You would require the camera ready paper, copyright form & fee receipt at the time of registration.

Registration Link: <https://forms.gle/rvZRmLWFr7BxJmQA>

You are also requested to send the Camera ready paper in AIP format, signed copyright form & fee receipt on icis2022@poomima.org with subject line containing paper ID for your smooth participation in the conference.

For Camera Ready template, Copyright form and conference fee, kindly go through below link on website
<http://convergence.poomima.org/?controller=pages&action=registration>

You can pay the conference fee through online mode.
Bank Detail for Registration Fee in ICSIS2022:

SUBMISSION OF PAPER 2

SoCTA2022 submission 3539 Inbox X



SoCTA2022 <socta2022@easychair.org>
to me ▾

14:58 (1 hour)

Dear authors,

We received your submission to SoCTA2022 (Soft Computing: Theories and Applications):


Authors : Abhinav Dubey, Neeraj Baishwar and Hirdesh Varshney
Title : Heart Disease Prognosis Tool Using Machine Learning
Number : 3539

The submission was uploaded by Abhinav Dubey
<yashdubeyko0@gmail.com>. You can access it via the SoCTA2022
EasyChair Web page

<https://easychair.org/conferences/?conf=socta2022>

Thank you for submitting to SoCTA2022.

Best regards,
EasyChair for SoCTA2022.

 Reply

 Forward

ACCEPTANCE OF PAPER 2 PRESENTATION

Decision on Manuscript submitted at 7th International Conference of Soft Computing: Theories and Applications (SoCTA-2022) for paper ID-3539 



hirdesh varshney <hirdeshvarshney@gmail.com>
to me ▾

17:23 (22 minutes ago) ☆ ↵ ⋮

From: **SoCTA2022** <socta2022@easychair.org>
Date: Tue, 24 May, 2022, 5:10 pm
Subject: Decision on Manuscript submitted at 7th International Conference of Soft Computing: Theories and Applications (SoCTA-2022) for paper ID-3539
To: Hirdesh Varshney <hirdeshvarshney@gmail.com>

*****PLEASE DO NOT REPLY TO THIS MAIL*****

Dear Varshney

Congratulations! On behalf of the SoCTA-2022 Chairs, we are pleased to inform you that your paper:

Paper id: 3539

Title: Heart Disease Prognosis Tool Using Machine Learning

Abhinav Dubey, Neeraj Baishwar, Hirdesh Varshney

has been conditionally accepted* for the ORAL presentation (in Virtual Format) at SoCTA-2022 and for publication in the conference proceedings published by LNNS Series of Springer.

*Conditional acceptance: All the satisfactory modifications/revisions should be incorporated as suggested by the reviewer(s) comments. The Editorial Board decides the final decision based on the revision or revised version of the paper received.

Secondly, the best papers based on the review scores plus deep paper analysis will be awarded under different categories. The result of the same would be available on the conference website (Awards Sections). The award ceremony will be at the closing session on Dec 18, 2022. All winners of the best paper selections will get the certificate along with Mementoes* (sent through Courier

Service)



BABU BANARASI DAS UNIVERSITY, LUCKNOW
CERTIFICATE OF FINAL THESIS SUBMISSION

(To be submitted in duplicate)

1. Name: **Abhinav Dubey**
2. Enrollment No.: **12004490673**
3. Thesis Title: **“Heart Disease Prognosis Tool Using Machine Learning”**
4. Degree for which the thesis is submitted: **M.Tech. (SE)**
5. School (of the University to which the thesis is submitted):

School of Engineering

- | | |
|--|----------------|
| 6. Thesis Preparation Guide was referred to for preparing the thesis. | YES/NO |
| 7. Specifications regarding thesis format have been closely followed. | YES /NO |
| 8. The contents of the thesis have been organized based on guidelines. | YES /NO |
| 9. The thesis has been prepared without resorting to plagiarism. | YES /NO |
| 10. All sources used have been cited appropriately. | YES /NO |
| 11. The thesis has not been submitted elsewhere for a degree. | YES /NO |
| 12. All the corrections have been incorporated. | YES /NO |
| 13. Submitted 2 hard bound copies plus 2 CD. | YES /NO |

Signature:

Name: Ms. Hirdesh Varshney
Assistant Professor
Department of Computer Science & Engineering

.....

(Signature of Candidate)
Name: **Abhinav Dubey**
Enrollment No.: **12004490673**
ROLL NO: **1200449001**

Signature:

Name: Mr. Neeraj Baishwar
Associate Professor
Department of Computer Science & Engineering