A Project Report

On

HEART DISEASE PREDICTION USING MACHINE LEARNING

Submitted to

Amity University Uttar Pradesh

inpartial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

(*Computer Science & Engineering*)

By

HARSH SHARMA(A2305218648)

AYUSH BHARDWAJ(A2305218614)

under the guidance of

Dr. SUMIT KUMAR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
AMITY UNIVERSITY UTTAR PRADESH
NOIDA (U.P.)

May-June, 2022

DECLARATION

We Harsh Sharma and Ayush Bhardwaj, students of B.Tech Computer Science and Engineering hereby declare that the project titled "Heart Disease Prediction using Machine Learning" which is submitted by us to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of requirement for the award of the degree of Bachelors of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. We hereby declare that we have gone through project guidelines including policy on health and safety, policy on plagiarism etc.

Noida

Date: Harsh Sharma:

Ayush Bhardwaj:

CERTIFICATE

On the basis of declaration submitted by Harsh Sharma, Ayush Bhardwaj, students of B.Tech (Computer Science and Engineering), We hereby certify that the Project titled "Heart Disease Prediction using Machine Learning" which is submitted to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in "Computer Science and Engineering" is an original contribution with existing knowledge and faithful record of work carried out by them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Noida

Date:

(Dr. Sumit Kumar)
(Associate Professor)
Department of Computer Science and Engineering
Amity School of Engineering & Technology
Amity University Uttar Pradesh, Noida

CONSENT FORM

(To be submitted separately)

This is to certify that we, Harsh Sharma and Ayush Bhardwaj students of B. Tech (*Computer Science and Engineering*) of 2018-2022 batch presently in the VIII Semester at *Amity School of Engineering and Technology*, Domain of Engineering and Technology, Amity University Uttar Pradesh, give our consent to include all our personal details (i.e., Name, Enrollment ID, etc.) for all accreditation purposes.

Place: Delhi

Date:

Harsh Sharma
A2305218648
Ayush Bhardwaj
A2305218614

ACKNOWLEDGEMENT

It is high privelege for us to express our deep sense of gratitude to those entire faculty members who helped us in the completion of the project, specially our internal guide **Dr. Sumit Kumar** who was always there at hour of need.

Our special thanks to all other faculty members, batchmates & seniors of **Amity School of Engineering & Technology**, Amity University Uttar Pradesh for helping us in the completion of project work and its report submission.

Harsh Sharma
A2305218648
Ayush Bhardwaj
A2305218614

Abstract

History reveals that out all the diseases heart disease is the difficult to predict and extremely complicated to cure in medical sector. According to various surveys, each minute a person dies because of heart disease. Earlier it was difficult to understand and foresee whether a person is suffering from cardiovascular disease but with advancement of technology it is now possible to predict such diseases with favourable accuracies. Data Science can aid in clinical industry by accumulating information associated with patients as well as process it which can additionally be made use of in producing forecasts by utilizing numerous machine learning strategies. Anticipating the cardiovascular disease is a difficult task to handle, so it is crucial to make advancements of making prediction with acceptable accuracy in order to initiate the process of treatment as soon as possible. In such cases where a particular technique fails to understand the situation, there is a need to use various techniques together to understand and compare the results. In this project work, dataset made use of is UCI machine learning dataset. In this project work, numerous techniques are contrasted according to the accuracies and various performance metrics given by the techniques.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Diseases related to the heart are one of the major reasons for worldwide death every year. Cardiovascular diseases involve multiple risk factors and it takes time to calculate accuracy, diagnosis approach, to make a move on the early stages of the disease. Machine learning is being used for processing enormous data and developing solutions to achieve accuracy and solutions in limited time in the healthcare domain. Many researchers are working with machine learning classification techniques to analyze massive data and generating favorable outcomes to predict cardiovascular disease.

1.1. Background Information

According to the reports of WHO, cardiovascular disease is major cause of causalities over past couple of years. Around 32% fatalities throughout the globe are triggered by cardiovascular diseases [26]. Heart is an essential organ of human body which circulates the blood in each part of the body, any abnormal condition to heart affects the whole body because of its fundamental role in maintaining flow of blood all over the body. Heart and blood vessels together form cardiovascular system. Any abnormal symptom or uneasiness to the capillary and heart is called cardiovascular disease (CVD). Diseases like strokes, coronary heart disease (CHD), endocarditis, peripheral vascular disease, etc., comes under cardiovascular diseases [42].

According to the reports of WHO around 17 million patients die every year only because of heart disease [26]. Cardiac arrest and strokes are the significant reason because of which people die, more than four out of five patient suffering from CVD dies because of strokes and heart attacks. One third of the total deaths caused by CVD covers people below the age of 70 [43]. Numerous symptoms associated with CVD is the substantial cause to diagnose the problem in a short period of time with efficiency. The major challenge in handling these diseases is the quality service and fast as well as affective and accurate diagnosis of CVD.

There are multifarious causes why one suffers from such conditions, physical inactivity, unhealthy diet, improper sleep schedule, consumption of tobacco and alcohol, stress, psychological factors like obesity, high bp, hyper tension, cholesterol, the prime reasons why one suffers from CVD [44]. Identifying those who are at high risk of CVD at early stage and taking early medical measures can prevent premature deaths. This data can be set side by side to a real-life application. Weights of the characteristics are divided on the premise of their impact on making predictions.

Therefore, it is important to utilize the knowledge as well as experience of such professionals in preparing databases in order to assist diagnosis and classification methods should be used by the database to generate predictions and comparing accuracies [45]. With the assistance of such technique's prediction models can be designed, such models can be made into usage in various medical fields by professionals in prediction cardiovascular diseases based on the medical record of patients. Therefore, by carrying out a system for forecasting making use of machine learning approaches, one can expect much more exact as well as analytical outcomes whether an individual is dealing with cardiovascular disease or not.

1.2. Purpose of Plan

One of the major challenges faced by healthcare organizations is the availability of efficient services at affordable price. These services represents efficient diagnosis of patients and administering effective treatments. Heart disease prediction models plays a critical role in medical fields. Detecting such kind of diseases in early stages can prevent premature death.

This project work presents an analysis of multiple supervised classification techniques, deep learning techniques and ensemble learning techniques in order to predict the accuracy for diagnosing as well as predicting heart diseases. To boost accuracy and prediction, feature selection must be done, feature selection is the method of decreasing the frequency of input values to reduce cost of modelling to intensify the efficiency of the prediction model. Dimensionality reduction can also be used to improve overall attainment of machine learning classification model. All these techniques will be measured on different accuracy parameters like accuracy, F1 score, recall, precision.

Machine learning classification systems are used in numerous fields around the globe. It plays a necessary part in predicting malignant and benign classes of various diseases like locomotor disorder, breast cancer, skin cancer, heart diseases, etc. Making predictions using such techniques is important because if one is able to predict well in advance with efficiency then it can provide important insights to doctors which can help diagnosis and treatment.

The major purpose of this project work is to compose a projection model for cardiovascular disease on heart disease dataset provided by professionals in order to make prediction about a patient whether that person is diagnosed with cardiovascular disease or not, which is a binary outcome. The outcomes of numerous classification models are then going to be analyzed to recognize the classification models which are more efficient to predict heart diseases.

Positive result (patient is diagnosed with heart disease) = 1,

Negative result (patient is not diagnosed with heart disease) = 0.

We begin by reviewing and evaluating some of the most recent cardiovascular disease prediction research work done by various researchers. At the last of this report, we will endeavor to find the efficacious method for heart disease prediction and on the basis of that we build a web application and deployed that web application on Heroku platform.

1.3. Scope

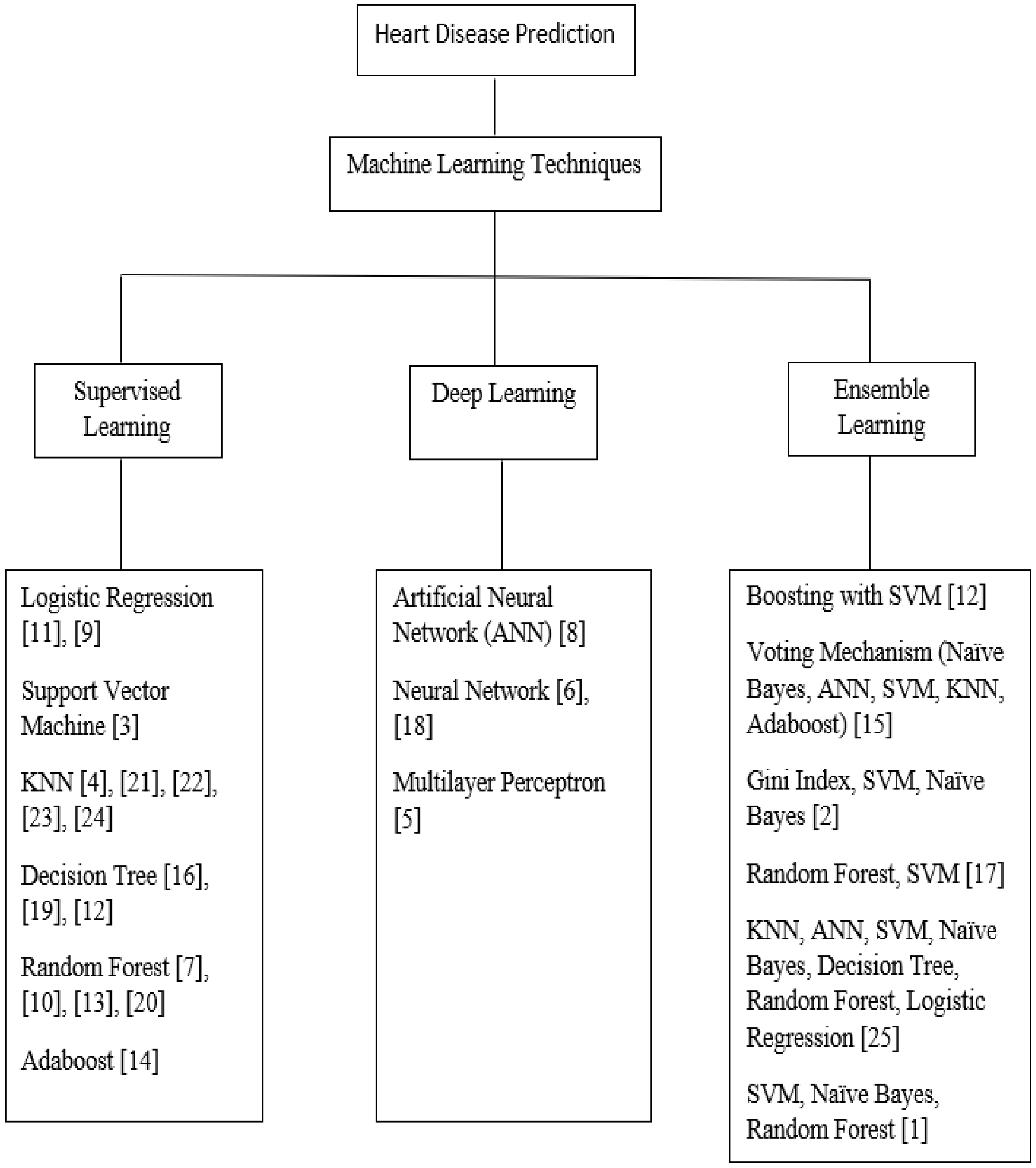
Project goal: To use numerous machine learning approaches on heart disease dataset to make prediction whether a person is suffering from cardiovascular diseases order, the various predictions generated by machine learning classification models are analyzed with intention to understand the outcomes generated by different classifier to make predictions using heart disease dataset.

In this project, classification methods such KNN, Decision tree, Naïve Bayes, etc., various ensemble techniques and deep learning techniques are collated with the assistance of heart disease dataset and favorable outcomes were attained. Various methods which are appraised for analysis of classification models are confusion matrix, sensitivity, F1 score and precision. Before using classification techniques, the dataset should be normalized in order to avoid overfitting of the training model which can hinder accuracies of prediction drastically.

Recent developments in machine learning can facilitate increase healthcare access in developing countries and innovate cancer diagnosis and treatment. In future, classification techniques will be used on larger training dataset to gain much more precise accuracy. Analyzing the scope of usage of machine learning in the medical sectors, more than a million various data points will be maintained in records of health system. Curing heart related diseases can get complicated sometimes due to lack of knowledge and late detection of any irregular activity in the body. This happens mostly when early detection is ignored, this can make the predicament of the patient more complicated in a very short period of time which might result in death.

Chapter 2: Literature Review

Various classification methods are used by researchers in order to predict cardiovascular disease prediction. Multiple algorithms are used by researchers in order to achieve utmost accuracy.



Heart Disease Prediction
Machine Learning Techniques
Supervised Learning
Deep Learning
Ensemble Learning
Logistic Regression [11], [9]
Support Vector Machine [3]
KNN [4], [21], [22], [23], [24]
Decision Tree [16], [19], [12]
Random Forest [7], [10], [13], [20]
Adaboost [14]
Artificial Neural Network (ANN) [8]
Neural Network [6], [18]
Multilayer Perceptron [5]
Boosting with SVM [12]
Voting Mechanism (Naïve Bayes, ANN, SVM, KNN, Adaboost) [15]
Gini Index, SVM, Naïve Bayes [2]
Random Forest, SVM [17]
KNN, ANN, SVM, Naïve Bayes, Decision Tree, Random Forest, Logistic Regression [25]
SVM, Naïve Bayes, Random Forest [1]

Fig. 2.1 Taxonomy of Heart Disease Prediction

Table 2.1 Literature Review of Heart Disease Prediction

Serial Number	**Title of the Research Paper**	**Authors of the Paper**	**Journal in which paper was published**	**Dataset Used**	**Methodology Used**	**Remarks**
[1]	Comprehensive review of Heart Disease Prediction using Machine Learning.	Dr. Dilbag Singh and Jagjit Singh Samagh	Journal Of Critical Reviews	Cleveland Heart Disease database	Naïve Bayes, SVM, Decision Tree, KNN, J48, Neural Networks, PCA, Random Forest.	Various techniques which can enhance the performance of ML algos for predicting heart disease like Ensemble Techniques and Optimization Algorithms.
[2]	Heart Disease Prediction using ML Techniques	V.V. Ramalingam, Ayantan Dandapath, M Karthik Raj	International Journal of Engineering & Technology	Cleveland Heart Disease database	Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models	Ensemble technique of SVM, KNN, ANN gives the accuracy of 94.12%; Naïve Bayes, SVM, Decision Tree gives the accuracy of 82%; Gini Index, SVM, Naïve Bayes gives the accuracy of 98%.

[3]	Machine Learning Techniques for Heart Disease Prediction	A. Lakshmanarao, Y. Swathi, P. Sri Sai Sundareswar	International Journal of Engineering & Technology	Framingham Heart Disease Dataset	Logistic Regression, KNN, Adaboost, Decision Tree, Naive Bayes, Random Forest, SVM, Extra Tree Classifier, Gradient Boosting. Authors applied sampling techniques on the dataset: Random over sampling, Synthetic minority oversampling, Adaptive synthetic sampling approach.	With Random Oversampling technique SVM gives an accuracy of 99.7%. This technique gives more accuracy in all sampling techniques. For Synthetic Minority Oversampling, Random Forest and Extratree Classifier given the best accuracy of 91%. For Adaptive synthetic sampling, Random Forest and Extratree Classifier given the best accuracy of 90%.
[4]	Heart Disease Prediction and Classification using Machine Learning algorithms optimized by Particle Swarm Optimization and Ant Colony Optimization	Youness Khourdifi, Mohamed Bahaj	International Journal of Intelligent Engineering and System	Heart Disease Dataset of UCI Machine Learning Repository	FCBF), KNN, SVM, Naïve Bayes, Random Forest, Multilayer Perceptron, Artificial Neural Network optimized by PSO combined with ACO.	Evaluate the effectiveness of all classifiers according to 3 steps: 1. Classifiers without optimization 2. Classifiers optimized by FCBF 3. Classifiers optimized by FCBF, PSO and ACO. And it shows that that the optimization hybrid approach increases the predictive accuracy of medical dataset.

[5]	Prediction of Heart Disease using Machine Learning	Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar	IEEE Xplore	Cleveland Dataset from UCI Library	Neural network algorithm multi-layer perceptron (MLP)	The output of the system will give a result if the person has a heart disease, in terms of Yes or No. It gives the average precision of 91%.
[6]	An Analysis of Heart Disease Prediction using Data Mining Techniques	Nidhi Bhatla, Kiran Jyoti	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease database	Neural Network, Decision Tree, Combination of Genetic algorithms and Decision Tree	Neural Network has shown the highest accuracy i.e., 100% so far. Decision Tree performed well with 99.62% accuracy with the aid of 15 attributes. Additionally, Genetic algorithm with unification of Decision tree gives the accuracy of 99.2%.
[7]	Effective Heart Disease Prediction using Hybrid Machine Learning Techniques	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava	IEEE Access	Cleveland Dataset from UCI Library	Hybrid Random Forest with Linear Model	Authors generated the boosted performance level with an accuracy degree of 88.7% through the prediction model for CVD with HRFLM.

[8]	Machine Learning Techniques for Heart Disease Prediction: A Review	MaryamI. Al-Janabi, Mahmoud H. Qutqut, Mohammad Hijjawi	International Journal of Engineering & Technology	Cleveland Dataset from UCI Library (Mostly Used)	Naïve Bayes, Artificial Neural Network, Radial Basis Function, Decision Tree, K-Nearest Neighbor, Support Vector Machine (SVM), Genetic Algorithm, Ensemble Learning	Authors conclude that the researchers who produced the highest accuracy were Dangare and Apte using Artificial Neural Network (ANN), WEKA tool and a combination of the Cleveland and Statlog heart disease datasets.
[9]	Comparative Analysis of Classification Function Techniques for Heart Disease Prediction	Dr. S. Vijayarani1, S. Sudha2	International Journal of Innovative Research in Computer and Communication Engineering	Cleveland cardiovascular disease dataset from UCI repository	Logistic Regression, Multi-Layer Perceptron, Sequential Minimal Optimization	The paper analyses the performance of various classification function techniques in data mining for predicting the heart disease from the heart disease data set. The classification function algorithms used and tested in this work are Logistics, Multi-Layer Perception and Sequential Minimal Optimization algorithms.

[10]	Heart Disease Prediction using Data Mining Techniques	H. Benjamin Fredrick David and S. Antony Belcy	ICTACT Journal on Soft Computing	StatLog dataset in UCI repository	Random Forest, Decision Tree, Naïve Bayes	The main objective of this significant study is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out. It is found that Random Forest algorithm performs best with 81% precision after comparing to other algorithms for heart disease prediction.
[11]	Heart Disease Prediction using Machine Learning	Rishabh Magar, Rishabh Magar, Rohan Memane, Suraj Raut, Prof. V. S. Rupnar	Journal of Emerging Technologies and Innovative Research (JETIR)	Cleveland heart diseases data set from the University of California Irvine (UCI)	Support Vector Machine (SVM), Decision Tree, Naïve Bayes Algorithm, Logistic Regression	Found out that Logistic Regression algorithm has the most efficient out of the four and resulted with an accuracy of 82.89%. Decision tree and Naïve Bayes had accuracy of 80.43% and 80.43% respectively, and SVM was having 81.57%

[12]	Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning	Hemanth Gadde	International Journal of Engineering Trends and Applications (IJETA)	Cleveland heart dataset from the UCI Machine Learning Repository	Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Ensemble Methods	It is evident that most of the machine learning algorithms are performing well in predicting and diagnosing of cardio vascular or heart diseases, may be some algorithms are poor in results in terms of performance and accuracy measures, Random Forest, Decision Tree algorithms generally work well on the data related to over fitting, whereas algorithms like SVM and Naive Bayes will work for real-world problems and on data sets.
[13]	Ensemble approach for developing a smart heart disease prediction system using classification algorithms	Mustafa Jan, Akber A Awan, Muhammad S Khalid, Salman Nisar	Research Reports in Clinical Cardiology 2018	Cleveland and Hungarian dataset from UCI Machine Learning Repository	Naïve Bayes, Neural Networks, Support Vector Machine (SVM), Random Forest, Ensemble Method	Lowest accuracy is 93.22% for regression analysis and the highest accuracy is 98.17% for the RF algorithm. The RF ensemble algorithm and SVM performed well and further hybridization through voting of each algorithm with more than 93% prediction probability has enhanced reliability of the system. More emphasis is given to select the algorithms having high true positive rate, as being the core measure for early diagnosis of cardiovascular disease.

[14]	Impact of ensemble learning algorithms towards accurate heart disease prediction	H. Benjamin Fredrick David	ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2020, VOLUME: 10, ISSUE: 03	StatLog Dataset	Bagging, Stacking and AdaBoost Support Vector Machine, Naive Bayes and K-Nearest Neighbour.	The main objective of the research work is to identify the best performing ensemble classification algorithm for heart disease prediction. For this purpose, the UCI data repository is used for performing the comparative analysis of three algorithms such as AdaBoost, Bagging and Stacking. From the research work, it has been experimentally proven that AdaBoost provides perfect results as compared to competitors.
[15]	Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations.	YING-TSANG LO, HAMIDO FUJITA and TUN-WEN PAI	Journal of Mechanics in Medicine and Biology Vol. 16, No. 1 (2016) 1640010 World Scientific Publishing Company	Cleveland Heart Disease Dataset	Naïve Bayes, ANN, SMO, KNN, AdaBoost, J48, and Random Forest,	In this study, seven machine learning methods were applied to make predictions based on CAD datasets. The results have shown that proposed ensemble learning/voting mechanism provided the best prediction performance. Analytical procedures directly imply that CAD exhibits positive correlation with age, heart rate, blood pressure, smoking habits, and cholesterol factors.

[16]	Heart Disease Prediction Using Classification with Different Decision Tree	K. Thenmozhi , P.Deepika	International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014	Cleveland Heart Disease database	Decision tree, Artificial Neural Network and Bayesian Classifier.	In this work Various techniques and data mining classifiers are defined for effective cardiovascular disease diagnosis. In this, Decision tree resulted with 99.62% accuracy with the help of using 15 attributes. Furthermore, in combination with genetic, Decision tree has shown 99.2% performance.
[17]	Predicting the presence of heart disease using machine learning	Akshay Jayraj Suvarna, Arvind Kumar M, Ajay Billav, Muthamma K M, Asst. Prof. Gadug Sudhamsu	International Journal of Computer Science and Mobile Computing	Cleveland Heart Disease Dataset	Random Forest and Support Vector Machine	When it is the initial stage of such disease, it is significant to identify in the initial phase only. So, when the disease is identified then it becomes very important to give the proper treatment to the patient. Hence, machine learning becomes very useful in such cases to predict such disease in prior.

[18]	A data mining approach for prediction of heart disease using neural network	Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte	International Journal of Computer Engineering and Technology (IJCET), ISSN 0976–6367(Print), ISSN0976-6375(Online) Volume 3, Issue 3,	Cleveland Heart Disease database	Neural network.	This work provides the prediction system for cardiovascular disease using data mining as well as ANN techniques. From ANN, a multilayer perceptron neural network is made use of to construct the system. Authors in their research study states that neural network led to 100% accuracy.
[19]	Heart Disease Prediction Using Machine learning and Data Mining Technique	Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel	IJCSC	Cleveland database of UCI repository	J48 algorithm, Logistic model tree algorithm and Random Forest algorithm.	In this research work it was concluded thatJ48 tree technique turned out to be best classifier for cardiovascular disease prediction because it contains more accuracy and least total time to build. J48 on UCI data resulted with supreme accuracy i.e., 56.76% and the total time to build model is 0.04 seconds
[20]	Heart Disease Prediction using Machine Learning	Apurb Rajdhan, Dundigalla Ravi, Milan Sai, Avi Agarwal	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease Dataset	Naive Bayes, Decision Tree, Logistic Regression and Random Forest.	The trial results verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

[21]	Heart Disease Prediction Using Machine Learning Algorithms	Archana Singh, Rakesh Kumar	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease Dataset	k-nearest neighbor, decision tree, linear regression and support vector machine (SVM)	To evaluate the different classification techniques KNN gives the best technique for heart disease prediction.
[22]	Heart disease prediction using machine learning algorithms	Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath	IOP Conf. Series: Materials Science and Engineering	UCI repository with patient's medical history and attributes	Logistic Regression and KNN	KNN to get an accuracy of an average of 87.5% on the prediction model which is better than the previous models having an accuracy of 85%. Accuracy of KNN is highest between the methods that were used i.e., 88.52%.
[23]	A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach	M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, V. Pavithra	International Journal of Computer Applications (0975 – 8887)	Cleveland database of UCI repository	Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbor (KNN), Naïve Bayes and Support Vector Machine (SVM).	The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN, SMO and Naïve Bayes. KNN gives the best technique for heart disease prediction.
[24]	Heart Disease Prediction using Machine Learning Techniques	Devansh Shah, Samir Patel, Santosh Kumar Bharti	SN Computer Science (2020)	dataset from the Cleveland database of UCI repository	Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm.	The highest accuracy score is achieved with KNN.

[25]	Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms	Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun	Hindawi Mobile Information Systems	Cleveland heart disease dataset	logistic regression, K-NN, ANN, SVM, NB, DT, and random forest	In this research study, a hybrid intelligent machine-learning based predictive system was contemplated.The system was tested on Cleveland heart disease dataset. Seven well-known classifiers such as logistic regression, K-NN, ANN, SVM, NB, DT, and random forest were used with feature selection algorithms Relief, mRMR, and LASSO used to select the important features. The K-fold cross-validation method was used in the system for validation

2.1. Heart Disease Prediction Using Supervised Learning: Supervised learning algorithms are algorithms which are trained with the help of labelled data for a specific output in order to analyze the hidden patterns and correlation between the input and output labels, various researchers are using supervised learning techniques so that favourable output can be generated, in the review process it was found out that different researchers used different methodologies to generate desired results, Prof. V. S. Rupnar, Rishabh Magar, Suraj Raut, Rohan Memane, in their research paper different supervised learning methods were used on heart disease dataset from UCI repository and they found out that logistic regression generated the highest accuracy of 82% in comparison to the other algorithms used which were SVM, decision tree, Logistic regression and Naïve Bayes [11], similarly A. Lakshmanarao, Y. Swathi, P. Sri Sai Sundareswar also used supervised learning approach to predict if a person has cardiovascular disease, in their research it was discovered that SVM along with random oversampling technique gave the highest accuracy of 99.7% while other used algorithms generated the accuracy of around 90% only [3].

For generating efficient outcomes, researchers also use various optimization techniques over supervised techniques which helps to increase the accuracy of output generated by the classifiers. Youness Khourdifi, Mohamed Bahaj did the same thing and optimized ANN technique by PSO which was combined with ACO, in their research work they found out that optimization hybrid techniques increase the predictivity of the models and helps achieving much more favorable results [4]. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava also considered optimization to achieve better results and used hybrid random forest and linear model and generated a boosted result with 88.7% accuracy [7].

Machine learning is developing various boosting methods regularly which are helping in achieving in better results along with mainstream techniques like supervised learning. Benjamin Fredrick David, used boosting technique called Ada boost with supervised learning algorithms and proved that boosting technique used provided much efficient outcomes in comparison to the other methods [14].

2.2. Heart Disease Prediction Using Deep Learning: Deep learning is a section of machine learning and is completely based on the usage of artificial neural network and this type of learning can be unsupervised, semi-supervised as well as supervised.

MaryamI. Al-Janabi, Mahmoud H. Qutqut, Mohammad Hijjawi used deep learning method in their research work and concluded that among all the algorithms used only ANN gave efficient outcome as compared to other techniques like decision tree, naïve bayes, etc. [8]. Nidhi Bhatla, Kiran Jyoti also used deep learning technique along with some other techniques like decision tree and in their research work it was stated that neural network generated an accuracy of 100%, highest so far, and decision tree resulted with an accuracy of 99.67% [6]. Multilayer perceptron (MLP) was used in the research work of Gouthami Kokkula, Isha Pandya, Aditi Gavhane, Kailas Devadkar and an accuracy of 91% was generated with the help of deep learning methods.

2.3. Heart Disease Prediction Using Ensemble Learning: When one wants to improve the results generated by a classification model and to reduce the absolute mean error, ensemble learning method can be used. In machine learning, ensemble learning is done by combining the results of multiple classification algorithm which will ultimately increase the productivity. Tun-Wen Pai, Ying-Tsang Lo, Hamido Fujita, they used ensemble methods on CAD dataset and used 7 machine learning algorithms in their research work. In their research paper it was concluded that ensemble learning provided much better results than the results provided by individual classification technique. Ensemble learning provided with the most efficient results and reduced the mean error in their research work [15].

In ensemble learning, multiple classification models are combined in a strategical manner in order to solve a particular problem which requires the intelligence of machine learning with efficient output. Such ensemble technique is used in the research work of V.V. Ramalingam, M Karthik Raj, Ayantan Dandapath, in their research paper, some of the algorithms used in this research work are; SVM, KNN, naïve bayes, decision tree, random forest, and ensemble models. The ensemble technique of KNN, SVM, and ANN gave an accuracy of 94% and ensemble method of SVM and gini index provided the highest accuracy of 98% in their research work [2].Jian Ping Li, Muhammad Hammad Memon, Amin Ul Haq, Shah Nazir, Ruinan Sun developed a projection model for cardiovascular disease which was using hybrid classification methods on heart disease dataset. The algorithms which were used in this work are; SVM, ANN, KNN, naïve bayes, decision tree. For the validation, researchers used validation method called k-fold cross in order to confirm the employed analytical procedure for specific testing [25].

Chapter 3: Approach to Design/Methodology

This chapter will discuss various steps taken which helped in solving the problem statement along with complete explanation and details of the approach. The steps and methodology which are used in order to achieve desired outcomes will be explained in this chapter, in what ways various classification methods are used to solve the problem statement is being discussed in this chapter.

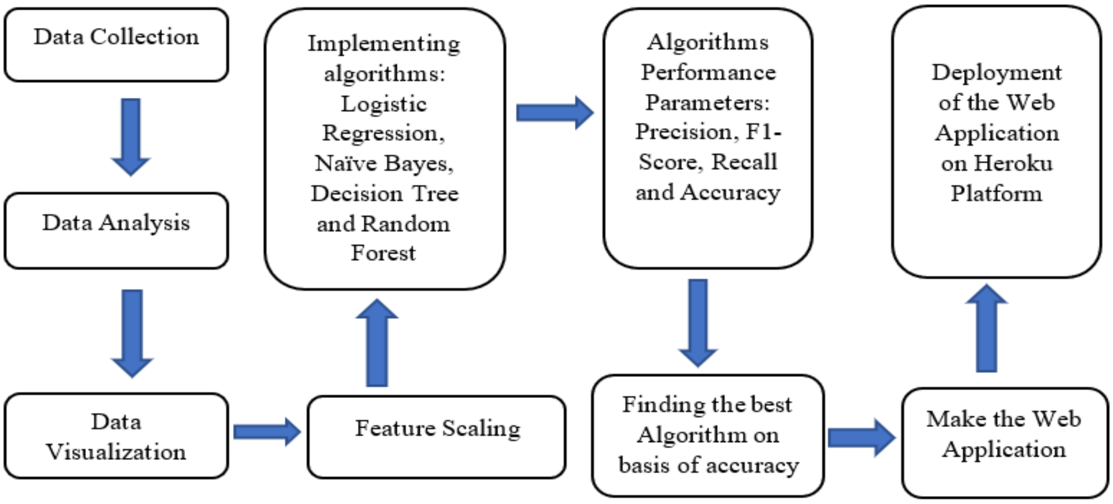

Data Collection
Data Analysis
Data Visualization
Feature Scaling
Implementing algorithms: Logistic Regression, Naïve Bayes, Decision Tree and Random Forest
Algorithms Performance Parameters: Precision, F1-Score, Recall and Accuracy
Finding the best Algorithm on basis of accuracy
Make the Web Application
Deployment of the Web Application on Heroku Platform

Fig. 3.1 Methodology Used for Heart Disease Prediction

3.1. Data Collection

In order to generate a classification model, initial step that should be taken into consideration is of collection and processing of raw data [32]. In this project heart disease dataset called Cleveland dataset from UCI repository is taken into consideration.

This dataset consists of 76 attributes, various researchers who have worked on cardiovascular disease researches using ML used this data due to its large variety of attributes and it’s easier to understand such a processed dataset, thus this dataset used in this project work to generate classification model on the basis of subsets of 14 of the attributes from the dataset.

For easier analysis and classifier generation, this dataset is already pre-processed and is available on the internet for public usage.

3.2. Data Analysis

Analysis refers to the thorough inspection and examination of the structure, design, working and approach used of a research project. One of the steps to be taken into account is the data analysis. Data analysis means inspecting, transforming, cleansing and modelling the data which then helps in achieving efficient outcomes and results [29].

Raw data or the real-world data consists of various issues and discrepancies like noise, incompleteness, inconsistency and outliers, thus it is important for the data to be analysed and pre-processed before using it to create classifiers so that one can achieve utmost accuracy.

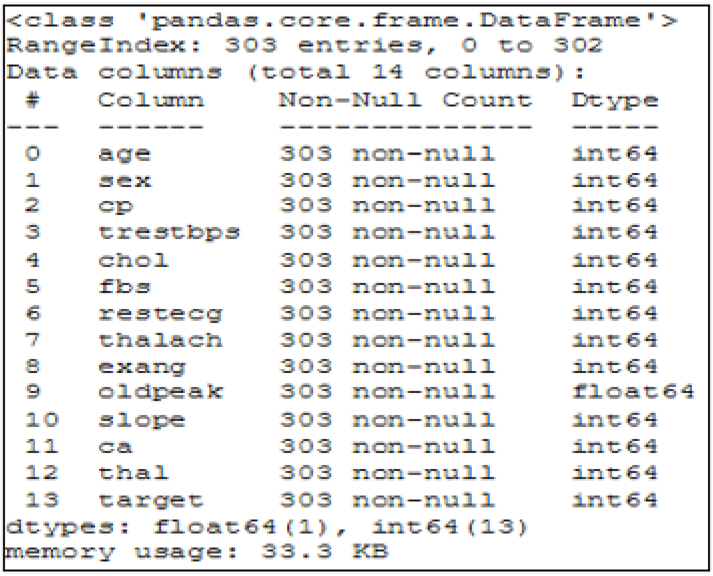
Accuracy, accessibility, interpretability, consistency and completeness are key features for a well-accepted multidimensional view of data quality. Various sectors like business, industry, medical, social science domains are using various approaches for the analysis of the data to generate efficient results because today everything depends on technology.

In order to be successful and at the top one should be using most refined ways of technology and that's why different sectors are using different methods for the analysis of the data. Visualization and integration are closely related to data analysis.

In order to generate accurate outcomes for the classifiers generated with the help of the dataset, data analysis follows few major steps which are as follows [29]:

- Data cleaning: this refers to the method of wadding data voids in the dataset, smooth the data which is noisy in nature, detach outliers and settle conflict issues.
- Data integration: It means integration as well as summing up of collective data cubes or files.
- Data transformation: data transformation consists of methodologies like standardization and normalization which are used to transform one format of data to another.
- Data reduction: Data reduction is the technique for optimization of capacity planning in which data is reduced to a simpler format in order to free up space on a storage device.

After fulfilling the steps mentioned above on the Cleveland heart disease dataset, it is evident that there is no null value in the dataset which is being used for this project work.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Fig. 3.2 Data Analysis of the Cleveland Dataset

3.3. Data Visualization

Data visualization is used to represent the data in the form of pie charts, graphs, histograms, scatterplots, etc. in order to understand the trends, patterns, issues, outliers in the data. It basically is graphical representation of data which helps to understand the correlation between attributes, the differences in the pattern, the approach to consider in the further research and to analyse which attributes effects the outcome the most. In today’s world, it is tough to work with massive portion of data so data visualization helps to understand and analyse data patterns in a much informative way to make data-driven decisions. Data visualization helps to visualize the data so that users can easily understand the complicated concepts and identify various patterns. Data can also be visualized in an interactive pattern to drill down deeply inside the graphs, charts, for much clear details [30].

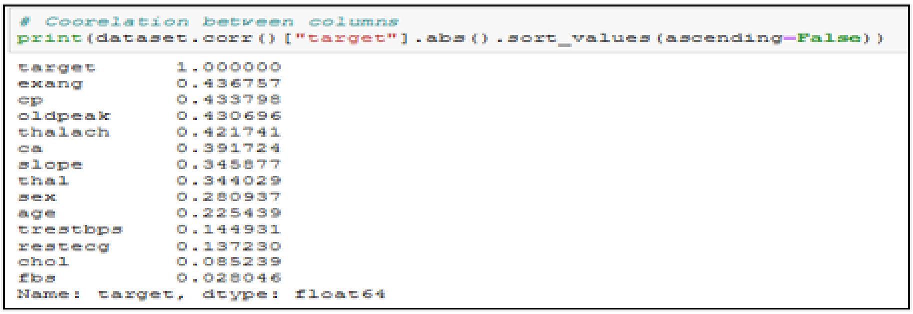
For data visualization in machine learning Matplotlib library is used in order to form graphs, pie charts of the data for identifying patterns. This library allows users to interpret data in the form of pie charts, graphs, histograms, scatterplots, etc. Basically, this library is used for data analysing and helps with numerical plotting library.

Studies states that humans grasp visualized information easily than flat table numbers. From business perspective it feels more presentable and informative to present the data in visualized manner, which allows to make decisions quickly.

The following are some of the advantages of data visualization [30][31]:

- It helps decision making easier and efficient with less consumption and helps to understand informative data in a much more interactive way which results in quick business decision making.
- It helps to target on the attributes which hinders the output most so that more attention can provided to the respective attribute.
- Helps to work and get comfortable with large amount of data in a pictorial format to understand unseen patterns, reveal complication and correlation.
- Data visualization helps to reveal previously unnoticed key points about the data sources in order to make data analysis reports with efficiency.

As it can be observed in the below figure that most of the columns are relatively corelated to the target variable but 'fbs' is not strongly corelated.



```
# Coorelation between columns
print(dataset.corr()["target"].abs().sort_values(ascending=False))
```

```
target      1.000000
exang       0.436757
cp          0.433798
oldpeak     0.430696
thalach     0.421741
ca          0.391724
slope       0.345877
thal        0.344029
sex         0.280937
age         0.225439
trestbps    0.144931
restecg     0.137230
chol        0.085239
fbs         0.028046
Name: target, dtype: float64
```

Fig. 3.3 Corelation between columns

3.4. Training and Testing

To analyse the execution of generated model, fresh test cases must be subjected to it. Testing and training is an important step for successful performance of the project as it shows how compatible or flawed the created model is, if the model is having multiple issues, then it will hinder the efficiency of the output [33].

Performance of the model can be tested in various steps; it can be accomplished either by running the model through production or by dividing the dataset into testing and training datasets. If the model is having issues, then the production method may fail. In the machine learning world, training and testing is considered to be an important step for the machine learning project to work properly. In this project the dataset is divided into 2 different sets named: training and testing [33].

The purpose of testing and training of the model is to make sure whether the model is working properly, if not then some changes can be done before the deployment and to make some improvements if possible. For this project work, train_test_split() function of sklearn with training size equals to 80 and testing size of 20 is used for testing and training purpose.

3.5. Machine Learning Algorithms Used

After the division of testing and training data, following algorithms are implemented on the Cleveland dataset of cardiovascular diseases. Implementation of methods are done using the direct functions of applied machine learning classification techniques provided by sklearn.

3.5.1. Logistic Regression

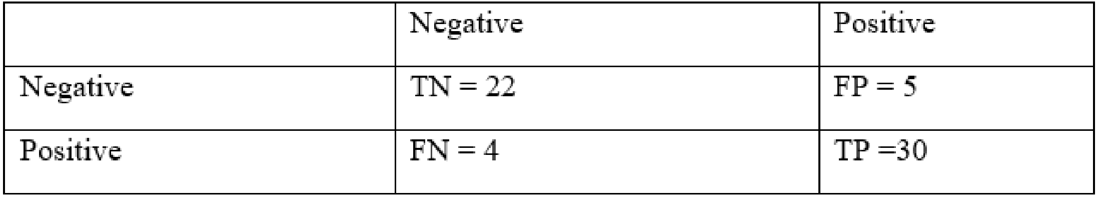
One of the majorly used and studied classification method in machine learning is logistic regression. It is a classic statistical model which is used to predict dependable variables which means categorical in nature, it predicts the dependable variables with the help of provided independent and dependent variables which are continuous [34].

When there is a choice between two options then logistic regression is used which means where binary classification is required, for example; Insurance company wants to verify whether the customer is suffering from cardiovascular disease or not (No as 0 or yes as 1).

Alogrithm-1: *Logistic Regression*

```
1. Import Logistic Regression from sklearn.linear_model
2. Use the LogisticRegression() of sklearn
3. Train the model on X_train and Y_train
4.After the training of the model, predict target variable for X_test and results are stored
in Y_predict_lr
```

Table 3.1 Confusion Matrix of Logistic Regression



	Negative	Positive
Negative	TN = 22	FP = 5
Positive	FN = 4	TP =30

3.5.2. Naïve Bayes

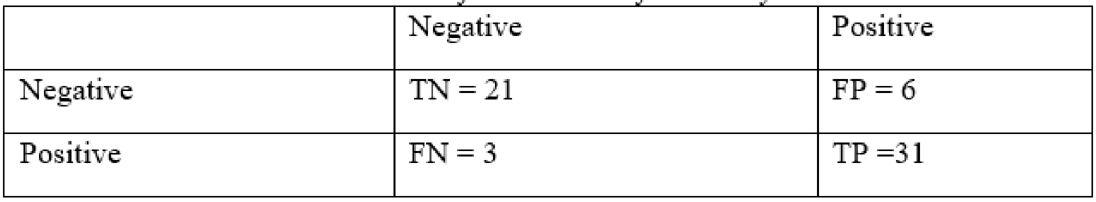
Naïve bayes theorem is a classification technique which is basically established on Bayes theorem that is based on the conditional probabilities of an event. Naïve bayes theorem simply states that no presence of features is corelated to the presence of any other feature. It is mainly used with massive datasets and is to build, it is used with large datasets due to its property of independent contribution of features. For example, a fruit is contemplated to be an apple only if it is red, 1.5 inches in radius and round, these features may depend on each other but are contributing individually and independently to the probability of fruit being the apple, this is the reason why this classification technique is named as naïve [34].

Naïve bayes is used in various research works and it selected by researchers due to its less complexity and higher efficiency in comparison to other classification techniques. Bayes' theorem is also called Bayes' Rule, and is used to examine the probability of a hypothesis generated by the model which depends on probability condition[34].

Alogrithm-2: *Naïve Bayes*

```
1. Import GaussianNB from sklearn.naive_bayes
2. Use the GaussianNB() of sklearn
3. Train the model on X_train and Y_train
4. After training the model, predict the target variable for X_test and results are stored in Y_predict_nb
```

Table 3.2 Confusion Matrix of Naïve Bayes



	Negative	Positive
Negative	TN = 21	FP = 6
Positive	FN = 3	TP =31

3.5.3. Decision Tree

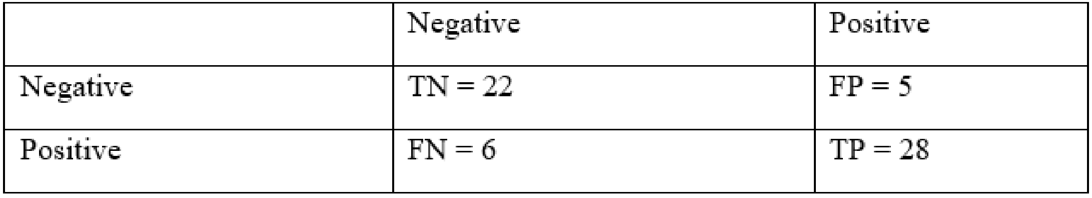
A decision tree as its name suggests is a representation of a flowchart like tree structure in which a feature is represented by an interior node, decision is represented by the branches and the outcome is given by leaf node. Decision tree begins with the root node which is also the topmost node of the tree. Decision tree basically works on various attribute values and according to these values the partition is done in recursive partitioning manner.

Decision tree is used in events mainly where decision analysis is required which helps in identifying a plan or strategy to reach the end goal, key features that’s why it’s used in research works are that it can easily be combined with other decision-making techniques and it helps to identify the worst, expected or best values for various scenarios. For a decision tree to be effective it must consider all the possibilities which must be mutually exclusive [34].

Alogrithm-3: *Decision Tree*

1. Import DecisionTreeClassifier from sklearn.tree
2. Use the DecisionTreeClassifier() of sklearn
3. Train the model on X_train and Y_train
4. After training the model, predict the target variable for X_test and results are stored in Y_predict_dt

Table 3.3 Confusion Matrix of Decision Tree



	Negative	Positive
Negative	TN = 22	FP = 5
Positive	FN = 6	TP = 28

3.5.4. Random Forest

For solving and understanding regression and classification problems various researchers use random forest machine learning technique due to its ability to work with ensemble learning techniques which are methods created to solve much more complex and complicated problems by combining various classification methods. This creates a forest which is then trained by bootstrap aggregation or bagging [34].

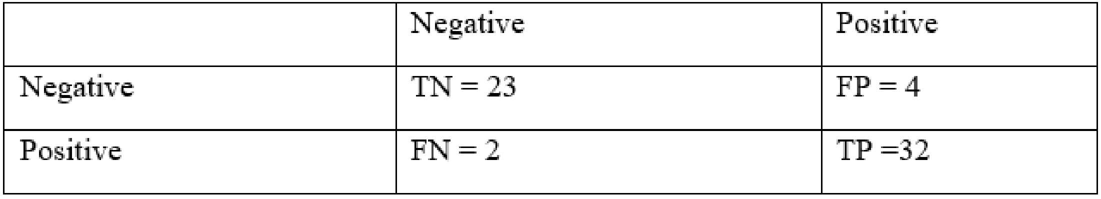
By combining multiple classification technique, the accuracy of outcome which will we get from the classification models will be much more efficient. This grouping of different algorithms together is called bagging. Random forest depends on the decision tree in order to make predictions with accuracy. It forecasts by averaging the output of various trees [34].

In this classification model, more the frequency of trees formed the accuracy will be much better. Many researchers use random forest over decision trees due to its ability to avoid overfitting and improved precision, which are the drawbacks of decision tree classification technique.

Alogrithm-4: *Random Forest*

1. Import RandomForestClassifier from sklearn.ensemble
2. Use the RandomForestClassifier() of sklearn
3. Train the model on X_train and Y_train
4. After training the model, predict the target variable for X_test and results are stored in Y_predict_rt

Table 3.4 Confusion Matrix of Random Forest



	Negative	Positive
Negative	TN = 23	FP = 4
Positive	FN = 2	TP =32

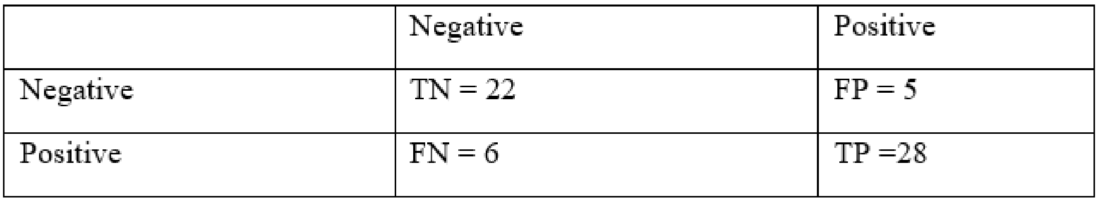
3.5.4. XGboost

For classification and regression problems, boosting technique i.e., XGBoost (extreme gradient boosting algorithm) is used in various research works. XGboost have separate methods to solve respective problems of classification and regression. In this project work, classification is being dealt with that’s why XGBClassfier() is used in order to achieve favourable results [35].

In order to work as a boosting technique, booster gtree parameter is used by XGBoost which is a tree grown after the others and tries to decrease the misclassification rate in iterations. Due to its uasge of parallel processing this classification technique is supposed to work faster than others and it can also handle missing values and outliers [35].

In this project work not only XGBoost function is used from Sklearn but Xgboost hyper parameters are also tuned with the help of grid search by setting up a grid that will be checked during the search. Parameters and options of each aparameter which we have used for Xgboost are: {'max_depth': [2,4,6], 'n_estimators': [50,100,200]}.

Table 3.5 Confusion Matrix of XGBoost



	Negative	Positive
Negative	TN = 22	FP = 5
Positive	FN = 6	TP =28

Alogrithm-5: *XGBoost Classifier*

1. Import xgboost.
2. Import GridSearchCV from sklearn.model_selection
3. Use the XGBClassifier() of sklearn
4. Set up the parameter grid.
5. Use the GridSearchCV() of sklearn.
6. Train the model on X_train and Y_train
7. After training the model, predict the target variable for X_test and results are stored in Y_predict_xgb

3.5.5. Catboost

Catboost algorithm is basically used for supervised machine learning problem statement and this boosting technique works on gradient boosting. Catboost is used on categorical data as this technique can transform the format of categorical data to numerical data on its own and can be used in both the cases of classification and regression [36].

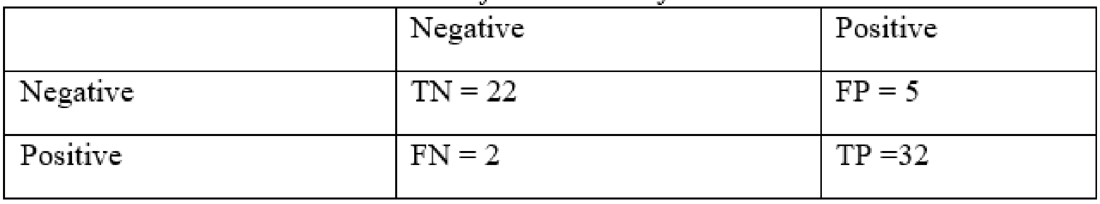
As mentioned above, CatBoost works on gradient boosting and during the training of the model, a collection of decision tree is built consecutively. As training moves ahead, each tree is generated with a downfall loss compared to the previous tree [36].

Catboost can easily be integrated with various fields like deep learning, data mining machine learning, key feature of CatBoost is that it can work with diverse data types which can help solving a variety of problems industries faces today.

Alogrithm-6: *CatBoost Algorithm*

1. Import CatBoostClassifier from catboost
2. Use the CatBoosttClassifier() of sklearn
3. Train the model on X_train and Y_train
4. After training the model, predict the target variable for X_test and results are stored in Y_predict_cat

Table 3.6 Confusion Matrix of Catboost



	Negative	Positive
Negative	TN = 22	FP = 5
Positive	FN = 2	TP =32

3.5.6. Ensemble Learning

Ensemble learning models are predictive models that combines various predictions from different classification models. In a predictive project it is important for the accuracy to be favourable according to the researcher, this is where ensemble learning techniques are used in the cases where accuracy of the prediction is the priority.

This method generates its prediction by combining one or more predictions from different prediction models, these prediction models can be of both same as well as different type and may not be trained on the same dataset like others. The predictions in ensemble learning are combined using different statistical methods such as mean, mode, etc. Ensemble technique uses multiple models at the same time and thus the time required to train such models and maintenance results in increase of complexity and computational cost [37].

There are various ensemble methods which are available in public to generate accurate predictions, in this project voting ensemble method is used with the help of VotingClassifier() with hard voting, Hard voting predicts the class with the largest sum of votes from models [37].

For ensemble learning the techniques which are used in this project are, random forest, logistic regression, naïve bayes and decision tree. Men score of 82.07% is achieved after combining the predictions of mentioned algorithms with the help of ensemble learning.

Alogrithm-7: *Ensemble Learning*

```
1. Import cross_val_score from sklearn.model_selection
2. Import RepeatedStratifiedKFold from sklearn.model_selection
3. Import VotingClassifier from sklear.ensemble
4. Import pipeline from sklear.pipeline
5. put all the algorithms in pipeline one by one.
6. Use the VotingClassifier() with voting=’hard’ as parameter.
7. Calculate the cross validation score using cross_val_score().
8. Calculate the mean score.
```

3.6. Web Application

After successfully making the machine learning model, we made the web application that will be based on the algorithm which gives the highest accuracy i.e., Random Forest Algorithm. In this, web application use will be required to input the values and as a result the web application will predict the heart disease in 0 and 1.

3.6.1. Programming Language and Platform Used

3.6.1.1. Programming Language Used

Python is a high-level programming language. With its popularity python has become a default choice for building applications involving machine learning, deep learning, data analytics as it supports numerous frameworks and third-party libraries. The specifications of the language and packages used to build and test the machine learning model is given below:

- **NumPy:** This library supports the massive multi- dimensional arrays and matrices. As its name suggests, it consists of built-in functions which can be used in order to perform mathematical operations with ease and efficiency. This library is not only used by programmers externally but is also used internally in other libraries like TensorFlow [38].

- **Pandas:** A library written in python for data manipulation and analysis. As its name suggests, it consists of built-in functions which can be used in order to perform mathematical operations with ease and efficiency. This library provides flexible, expressive, high-level data structures and a large variety of tools for analysis. This library helps with cleaning, manipulation and analysis of the data in an uncomplicated method [38].
- **Scikit-learn:** It's a free to use machine learning library, featuring various classification, regression methods designed to interoperate with NumPy [38].
- **Seaborn:** In analysis of the data, seaborn is used for plotting numerical data. This library allows users to interpret data in the form of pie charts, graphs, histograms, scatterplots, etc. Basically, this library is used for data analysing and helps with numerical plotting library [38].
- **Pickle:** It is used for serializing and de-serializing python object structures. The process to converts any kind of python objects (list, dict, etc.) into byte streams (0s and 1s) is called pickling or serialization or flattening or marshalling.

3.6.1.2. Platform Used

Jupyter Notebook: There are various platforms which can be used in order to work with data science, the platform which is taken into consideration for this project work is Jupyter Notebook. It was launched in 2014 and was born out of IPython Project.

Jupyter notebook is an interactive environment which can work with data science, the only purpose of jupyter notebook is not only to be used as an IDE i.e., Integrated development environment but can also be used as a platform for educational and presentation purposes. It is an open-source application which allow users to share, create, modify the files containing visualization, mathematical representations and live code [40].

Jupyter notebook is helpful as it includes various tools which can help users in data cleaning, numerical simulation, data visualization, machine learning, transformation, etc.

Spyder: Spyder is one of the IDE which is free and open-source environment. This platform is written in Python and for python, and designed via way of means of and for scientists, engineers and facts analysts. It functions a completely unique mixture of the superior editing, analysis, debugging, and profiling capability of a complete improvement device with the facts exploration, interactive execution, deep inspection, and exquisite visualization skills of a systematic package [40].

Fig. 3.4 Python, Spyder with Importing Libraries

3.6.2. Web Framework Used

For the development of the web application based on machine learning model, in this project work we used one of the most popular frameworks named, Streamlit.

Streamlit: Web apps are efficient tools for researchers to make their projects available to the users on their system through the means of internet or intranet. For developing such application for machine learning and data science, one of the most popular open-source frameworks which helps in developing web applications is Streamlit. With the help of Streamlit, web applications web application can be developed and deployed in a short span of time with better efficiency [40].

Below are the steps mentioned to install Streamlit on windows:

Before installing Streamlit, anaconda must be installed in the system as per the Streamlit documentations. Anaconda navigator is the official supported environment manager on windows for Streamlit.

Step 1: Download and install Anaconda (skip this step if you already have Anaconda installed). navigate to Anaconda Download web page and download Anaconda Individual Edition.

Step 2: Using anaconda navigator set up the environment

Step 3: In the new environment open the terminal.

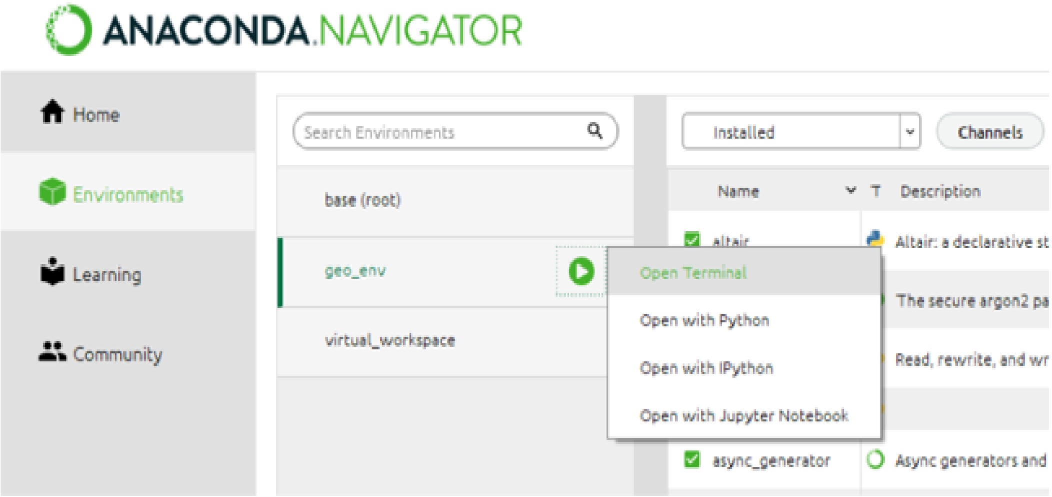
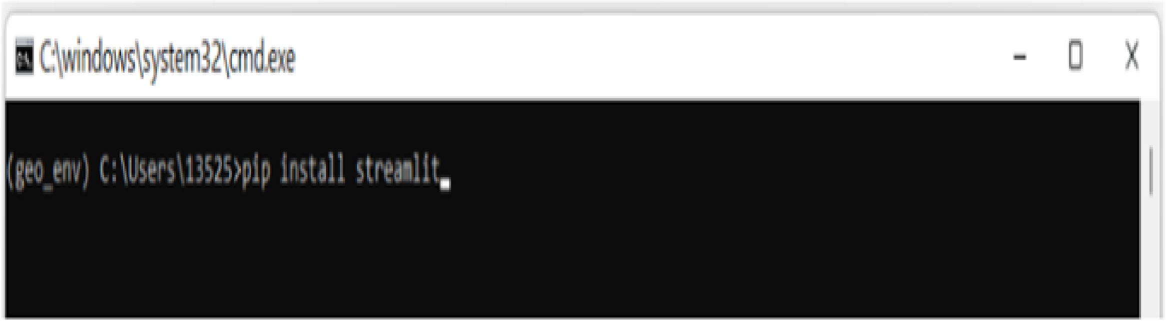


Fig. 3.5 Anaconda Dashboard

Step 4: Install Streamlit

After installing a terminal will be displayed, type the following command:

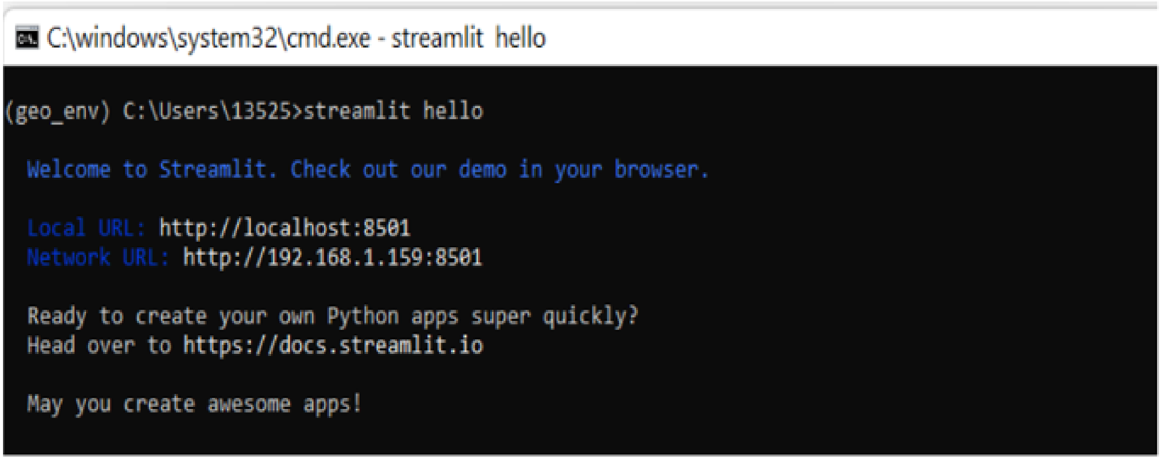


```
(geo_env) C:\Users\13525>pip install streamlit_
```

Fig. 3.6 Command Terminal

Step 5: Check whether Streamlit is install properly or not.

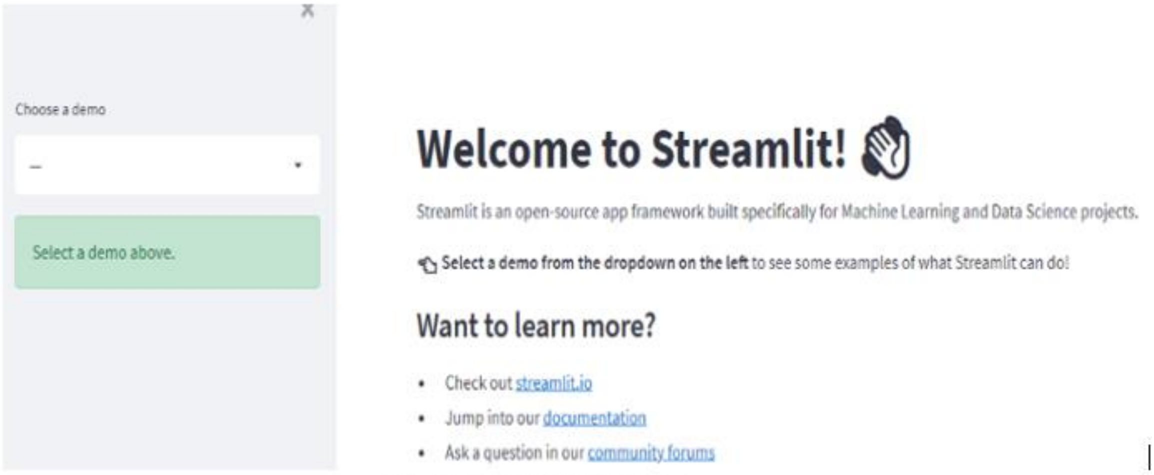
In the command prompt type, the command 'Streamlit hello'. After successful installation, a message will pop up in the terminal and Streamlit's Hello app appear in your web browser.



C:\windows\system32\cmd.exe - streamlit hello

```
(geo_env) C:\Users\13525>streamlit hello

  Welcome to Streamlit. Check out our demo in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.159:8501

  Ready to create your own Python apps super quickly?
  Head over to https://docs.streamlit.io

  May you create awesome apps!
```

Fig. 3.7 Command Terminal



Choose a demo

–

Select a demo above.

Welcome to Streamlit! 👋

Streamlit is an open-source app framework built specifically for Machine Learning and Data Science projects.

👈 **Select a demo from the dropdown on the left** to see some examples of what Streamlit can do!

Want to learn more?

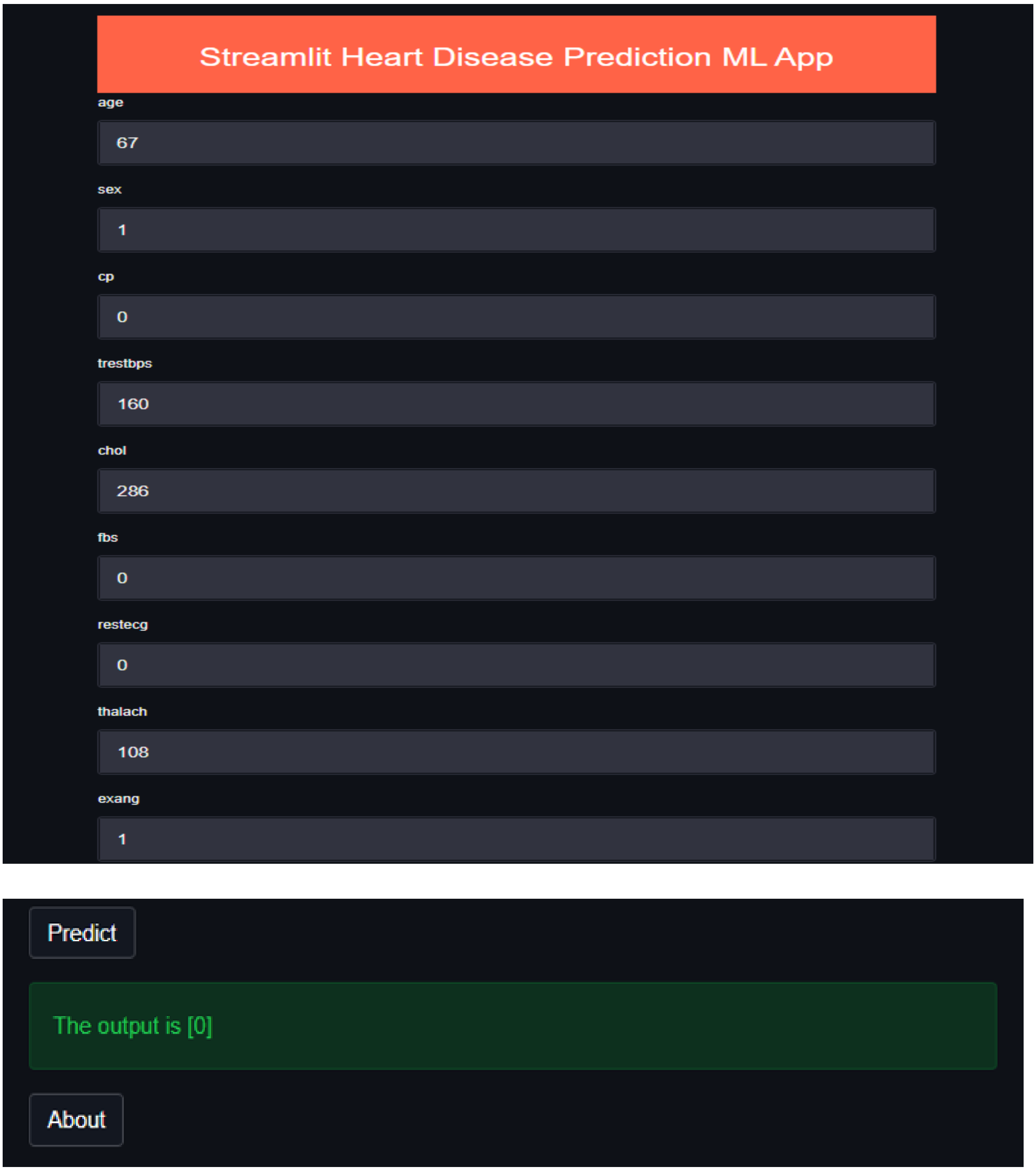
- Check out streamlit.io
- Jump into our documentation
- Ask a question in our community forums

Fig. 3.8 Streamlit Web App

3.6.3. Functionality of the Web Application

After the implementation of various classification techniques on Cleveland heart disease dataset a web application is created with the help of Streamlit framework in this project. After implementation of multiple algorithms which are decision tree, random forest, XGboost, Naïve bayes, logistic regression, catboost and ensemble learning. After the implementation of each and every technique it is analysed that classification technique with the most efficient output is random forest. So, in the development of the application using Streamlit framework the classification technique random forest is taken into consideration. In this project the web application is developed in such a way that users will input the values of text-fields and after filling out all the required fields with the desired data the user needs to submit the data. After submission of the data the application will make predictions whether the patients data given by the user signifies patient is suffering from cardiovascular disease in the form of 0 or 1.

This figure below shows the screenshot of the web application:



Streamlit Heart Disease Prediction ML App

age: 67

sex: 1

cp: 0

trestbps: 160

chol: 286

fbs: 0

restecg: 0

thalach: 108

exang: 1

The output is [0]

Fig. 3.9 Heart Disease Prediction Web Application

3.7. Deployment of the Web Application

After the web application is developed successfully, it is important for the web application to be deployed and a web link should be generated. There are different platforms with which one can deploy the developed web application, some of them are; Amazon AWS, Microsoft Azure, GitHub, Heroku, etc; this project uses Heroku platform for the deployment of the web application.

3.7.1 Platform Used for Deployment

Heroku: Heroku is widely used by the developers in order to manage, scale and deploy web applications, in order to present applications in the market developers highly considers Heroku as a means of deployment due to its elegant, flexible and easy to use features. In today's market Heroku is recommended as the top choice for many developers to use for the deployment of the application. Heroku allows developers to focus on enhancing their application experience for the user by managing the hardware and servers on its own and providing time to the developer to work with the application. Heroku ensures that users receive the highest quality experiences as possible [41].

Heroku can be with multiple programming languages like Java, Scala, Python, PHP, Node.js, etc., as this platform supports multiple languages it can be used by variety of developers who works with different programming languages [41].

Heroku uses unique domain names for applications which are then used to route HTTP requests to the desired container, these containers are developed in order to run and package services. Heroku manages and runs application that's why there is no need to manage OS or other system configurations [41].

3.7.2 Steps Performed for Deployment

Following are the steps followed to deploy the Heart Disease Prediction application on Heroku platform:

Step 1: Login to https://id.heroku.com/login

Step 2: Click on "New" in top-right on the dashboard.

Step 3: Then click on "Create New App".

Step 4: Then give the name to your app and choose the region and then click on "Create App".

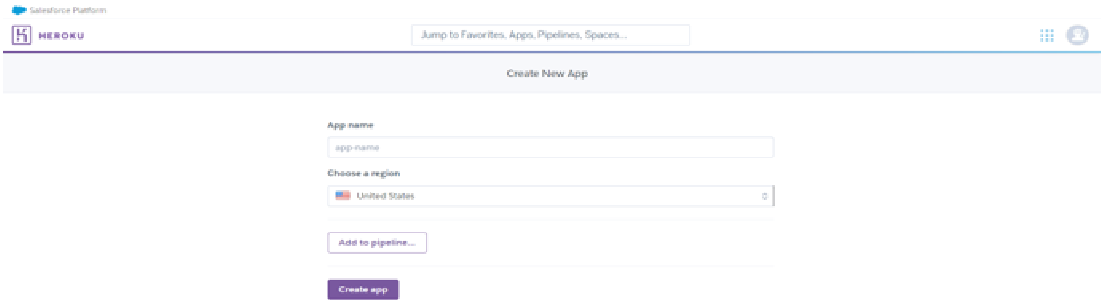
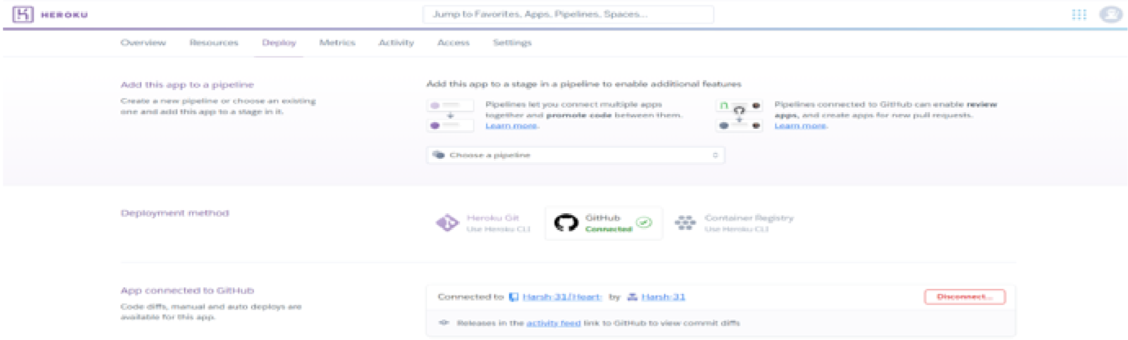


Fig. 3.10 Create New App on Heroku

Step 5: Then select the deployment method as "GitHub" from the newly opened dashboard.

Step 6: Search for the repository name with which you saved your project on your GitHub account.

Step 7: Then click on "Connect" to make the connection of Heroku with your GitHub repository.



Add this app to a pipeline

Create a new pipeline or choose an existing one and add this app to a stage in it.

Add this app to a stage in a pipeline to enable additional features

Pipelines let you connect multiple apps together and **promote code** between them. Learn more.

Pipelines connected to GitHub can enable **review apps**, and create apps for new pull requests. Learn more.

App connected to GitHub

Code diffs, manual and auto deploys are available for this app.

Connected to Harsh-31/Heart by Harsh-31

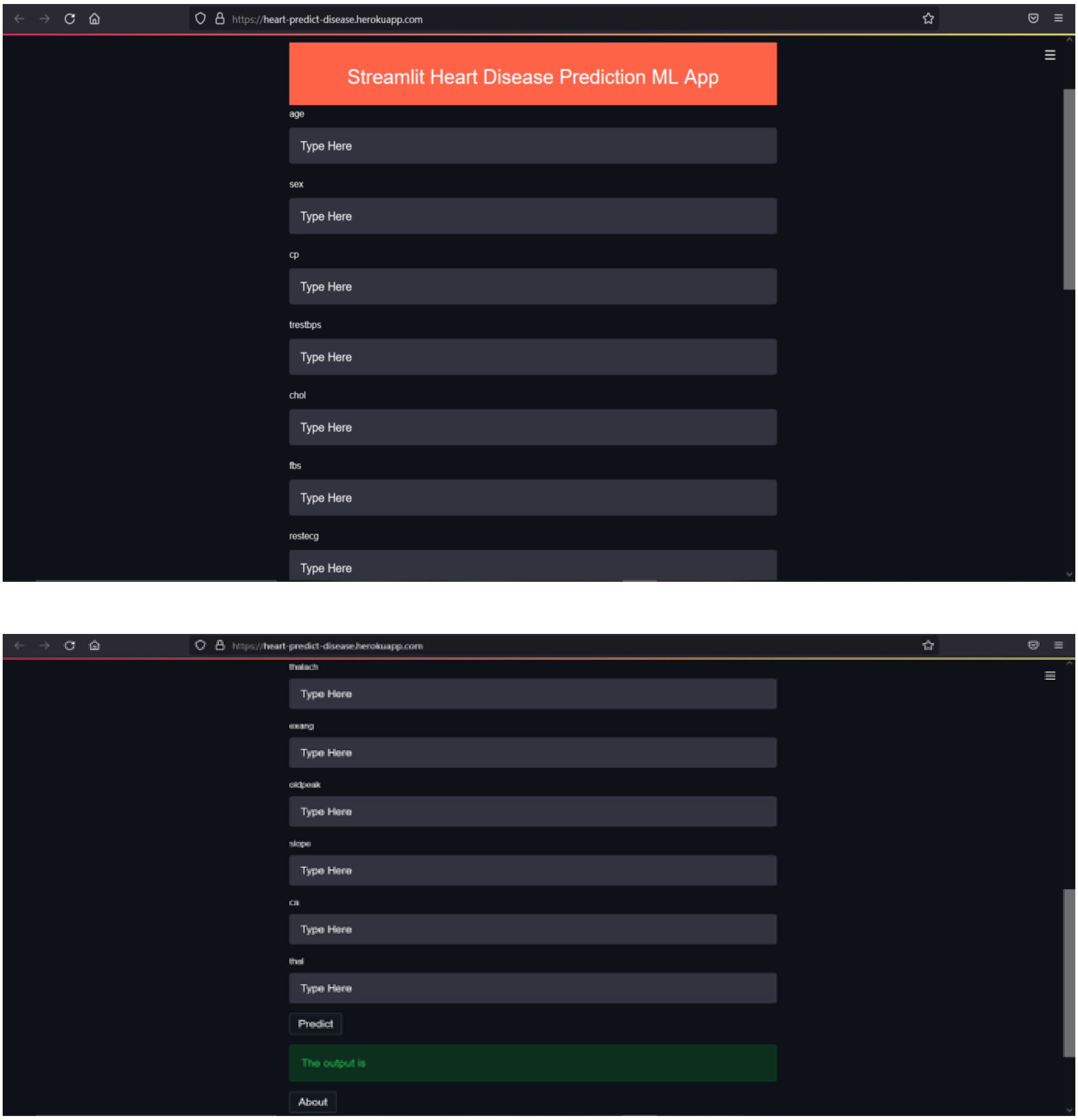
Releases in the activity feed link to GitHub to view commit diffs

Fig. 3.11 Application Deployment on Heroku

Step 8: Then click on "Deploy Branch" tab in Manual Deploy section.

Step 9: And then application deployment starts and a link will be generated for the web application. For our heart disease prediction, the generated working link after performing above all steps is*: https://heart-predict-disease.herokuapp.com/*

And finally, our application will be opened for use and below is the screenshot of the deployed web application:



https://heart-predict-disease.herokuapp.com

Streamlit Heart Disease Prediction ML App

age

Type Here

sex

Type Here

cp

Type Here

trestbps

Type Here

chol

Type Here

fbs

Type Here

restecg

Type Here

https://heart-predict-disease.herokuapp.com

thalach

Type Here

exang

Type Here

oldpeak

Type Here

slope

Type Here

ca

Type Here

thal

Type Here

The output is

Fig. 3.12 Heart Disease Prediction Web Application after deployment on Heroku

Chapter 4: Simulation, Experimentation and Analysis

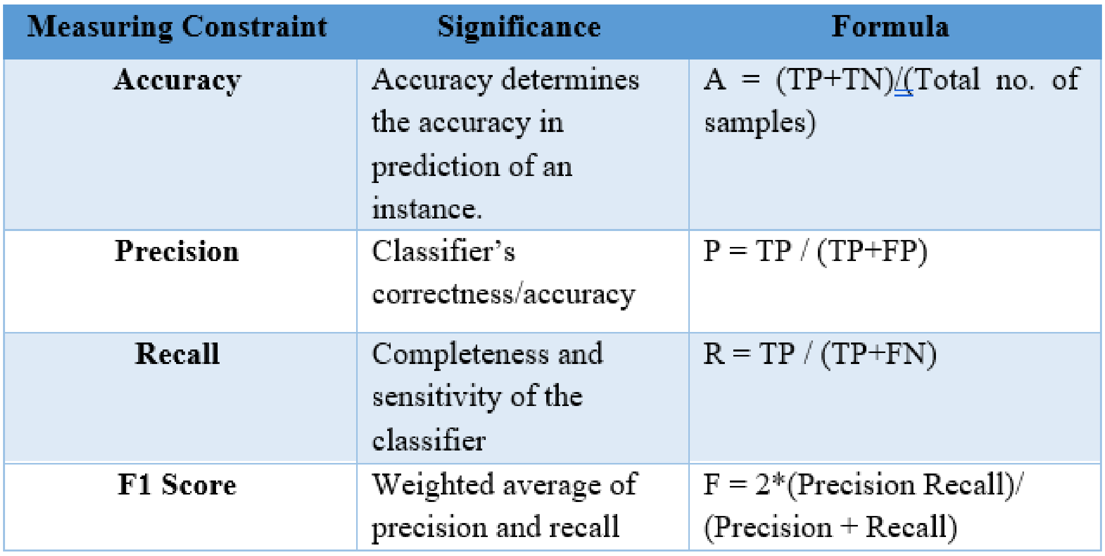
In the experiment, the aim is to find out the performance of proposed methodology on heart disease dataset. The effectiveness of classification methods is accessed using a set of indicators which indicates the performance of the model. A confusion matrix is generated in order to evaluate the parameters, TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) for actual data and predicted data.

Since confusion matrix helps to evaluate the score of the classifier which depicts how good our model is running on the training dataset. It is a squared matrix which provides a clear grasp of our classifier's prediction and how it relates to the actual testing label.

Apart from using confusion matrix, for this project work we have use classification report for every machine learning technique. This classification report depicts multiple performance measures named, precision, accuracy, recall, f1 score.

Below table shows the significance and formula used for the calculation of different measuring constraints of classification score.

Table 4.1 Different Performance Measures



Measuring Constraint	Significance	Formula
Accuracy	Accuracy determines the accuracy in prediction of an instance.	A = (TP+TN)/(Total no. of samples)
Precision	Classifier's correctness/accuracy	P = TP / (TP+FP)
Recall	Completeness and sensitivity of the classifier	R = TP / (TP+FN)
F1 Score	Weighted average of precision and recall	F = 2*(Precision Recall)/ (Precision + Recall)

Below are the results of F1 score, precision and recall of all the methods applied in this project which are Naïve Bayes, Logistic Regression, Decision Tree, Random Forest.

Table 4.2 Performance Measures of ML Algorithms

Algorithms	**Performance Parameters**					
	Precision		**Recall**		**F1 Score**	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Logistic Regression	0.85	0.86	0.81	0.88	0.83	0.87
Naïve Bayes	0.88	0.84	0.78	0.91	0.82	0.87
Decision Tree	0.79	0.85	0.81	0.82	0.80	0.84
Random Forest	0.92	0.89	0.85	0.94	0.88	0.91
XGboost	0.79	0.85	0.81	0.82	0.80	0.84
Catboost	0.92	0.86	0.81	0.94	0.86	0.90
Ensemble Learning	0.79	0.85	0.81	0.82	0.80	0.84

This table represents the accuracy percentage of all the methods applied. Those methods are, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest.

Table 4.3 Accuracy Percentage of ML Techniques Used

Algorithms	**Accuracy Percentage**
Logistic Regression	85.25%
Naïve Bayes	85.25%
Decision Tree	81.97%
Random Forest	90.16%
XGBoost	81.97%
CatBoost	88.52%
Ensemble Learning	82.07%

Chapter 5: Discussion of Results

In this project, apart from using the algorithms which have already been used in the base research paper which we referred for this project, we used some more techniques/algorithms of machine learning with an aim to get comparison of all the machine learning algorithms.

Since in the base paper the authors used some of the algorithms like Naïve Bayes, Logistic Regression, Decision Tree, Random Forest. As per the result of their research, Random Forest outperformed all the other algorithms which have been used in their research work which was giving the accuracy of 90.16%.

So, in this project we move forward with the study of the authors from the place where they left it. We started it by implementing all the algorithms which have been used by the authors so that we can use the results of these algorithms to carry forward the project work. After implementing all the algorithms, we also got the same accuracy percentage as the researchers mentioned in their study. For logistic regression it is 85.25%, for decision tree it was 81.97%, in case of naïve bayes it is 85.25%, for random forest it is 90.16%.

Hereafter, we started implementing some more machine learning techniques like, XGboost Algorithm, Catboost Algorithm and then we used Ensemble Learning (Naive Bayes, Logistic Regression, Decision Tree, Random Forest). For each and every technique we plotted confusion matrix, classification report. Consequently, we got to know that XGboost algorithm was giving the accuracy rate of 81.97% and Catboost Algorithm was showing the accuracy rate of 88.52% and ensemble learning technique was depicting the accuracy of 82.07%.

On analysing the accuracy rates of all the different machine learning algorithm, we go to know that Random Forest still outperformed XGboost, Catboost and ensemble learning. But as per the base paper researchers’ study, after random forest it was Naïve Bayes and Logistic Regression which were giving highest accuracy rate of 85.25% but their classification reports are different. So, if we took classification reports into consideration then in some cases Naïve Bayes outperformed Logistic Regression and, in some cases, Logistic Regression outperformed Naïve Bayes.

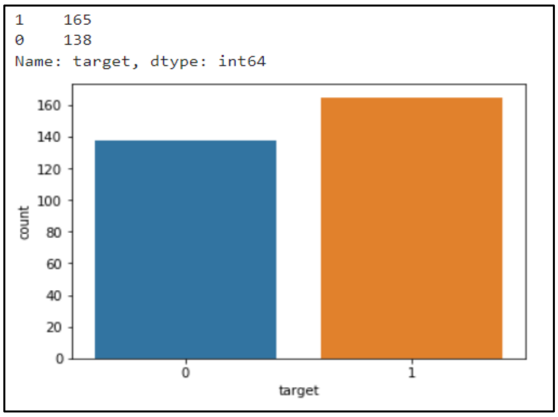
But when we implement three more techniques then we got to know that, catboost algorithm outperformed Naïve Bayes and Logistic Regression by giving the accuracy rate of 88.52% even its classification report parameters (precision, recall, F-Measure) outperformed both Naïve Bayes and Logistic Regression.

Moreover, the accuracy rate and the classification reports of Decision Tree and XGBoost are similar numerically.

After analysing all the results of algorithms, we decided to make a web application using streamlit framework which will give the result in 0 and 1 using the accuracy rate of Random Forest algorithm.

Chapter 6: Presentation of Results and their Analysis

There are various facts and figures which were clearly evident from the graphs which we plotted for every attribute of the Cleveland Heart Disease dataset. After analysing the corelation between the columns of the dataset, then firstly we analysed the target variable of the dataset by plotting the countplot.



```
1    165
0    138
Name: target, dtype: int64
```

Fig. 6.1 Countplot of target feature of dataset used

Percentage of patients without any heart issue: 45.54%

Percentage of patients having a heart problem: 54.46%

Subsequently, we analysed features of dataset named, ‘sex’, ‘cp’, ‘restecg’, ‘exang’, ‘slope’, ‘ca’ and ‘thal’.

In the below figure, we can notice that the ‘sex’ feature has two unique features and it was analysed that females have a higher tendency to be suffering from heart disease than males.

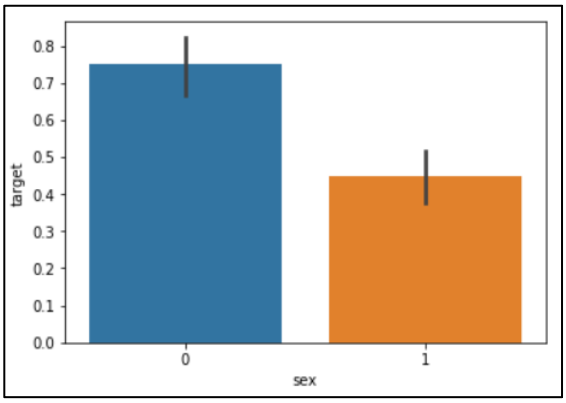


Fig. 6.2 Countplot of sex feature of dataset used

Then we analysed the ‘chest paintype’ feature and it is clearly evident that it has values from 0 to 3 and the chest pain of type ‘0’.

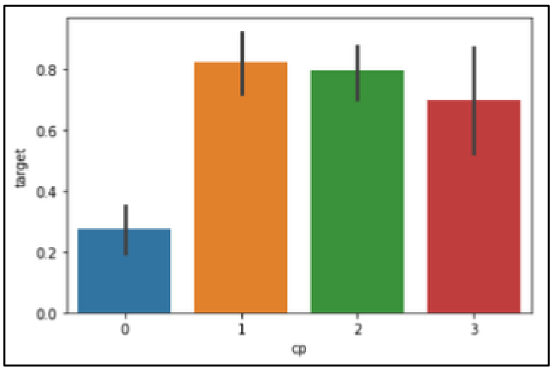


Fig. 6.3 Countplot of cp feature of dataset used

‘fbs’ feature has also been analysed and patients record with the value ‘1’ and ‘0’ have almost equal chances of having heart problems.

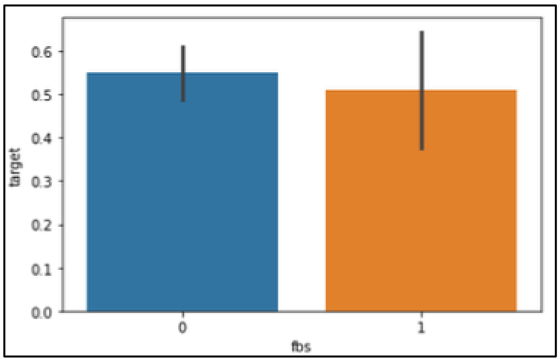


Fig. 6.4 Countplot of fbs feature of dataset used

From the below plotted graph of ‘exang’ feature, it is clearly evident that the people with exang=1 i.e. exercise including anigma are less likely to have cardiovascular issues.

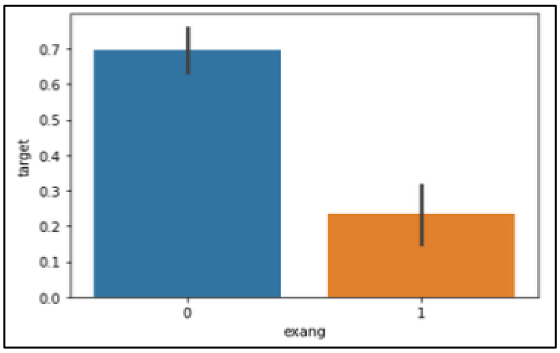


Fig. 6.5 Countplot of exang feature of dataset used

From the below plotted graph of ‘restecg’ feature, it is evident that the people with resecg ‘1’ and ‘0’ are much more likely to have a heart disease than with restecg ‘2’.

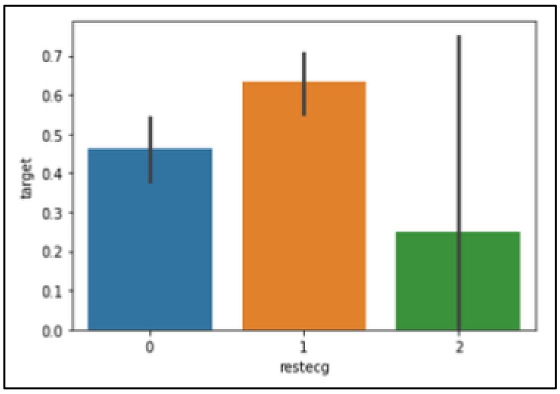


Fig. 6.6 Countplot of restecg feature of dataset used

We observe that the slope ‘2’ causes heart pain much more than slope ‘0’ and ‘1’.

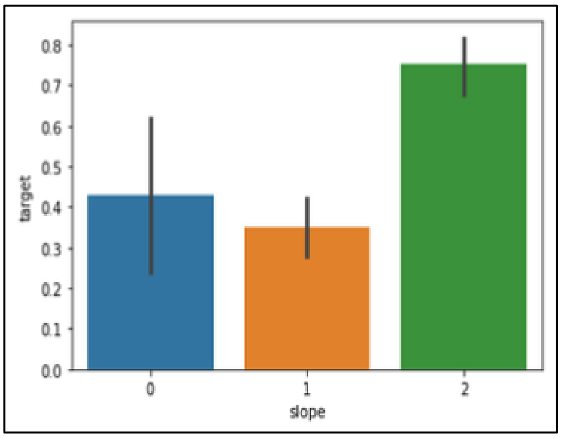


Fig. 6.7 Countplot of slope feature of dataset used

It is clearly observable from the below plotted graph that the ca=4 has astonishingly large number of heart patients.

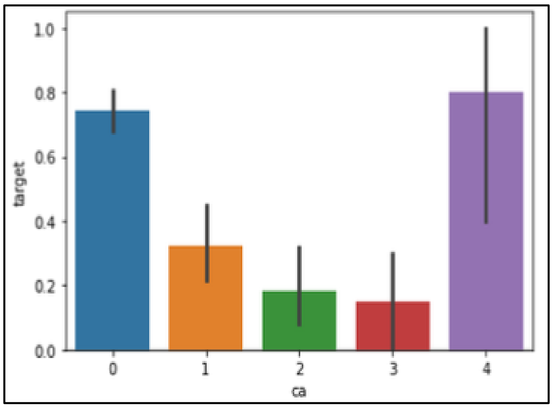


Fig. 6.8 Countplot of ca feature of dataset used

After implementing all the algorithms using sklearn functions, we plotted the ROC-AUC curve where, ROC is used to plot the graph and AUC is used to compute the area under the curve. Moreover, higher the AUC of the algorithm, better is the performance of that particular algorithm.

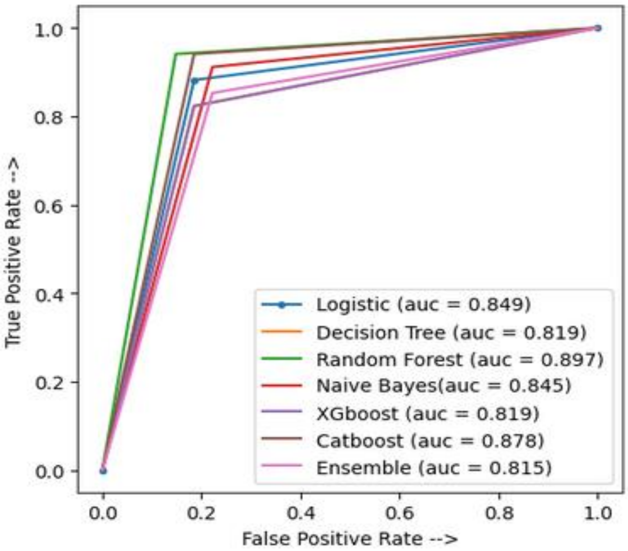


Fig. 6.9 ROC-AUC Curve of ML Algorithms Used

As it is evident from the above plotted ROC-AUC curve that, the area under the curve of Random Forest is greater than all the other used machine learning algorithms. Moreover, its True Positive Rate is greater and False Positive Rate is smaller as compared to others.

Chapter 7: Conclusion

Heart Disease forecast at a beginning phase can help in going to consider precautionary measures. We carried out literature survey to compare and contrast the different methods and approaches of carrying out heart disease prediction which were used by various researchers and also analyzed some of the optimization techniques utilized by various authors in their work to predict heart disease. In this project work, different characterization techniques were tried and an efficient end was drawn based on the exhibition of the algorithms on execution estimates like accuracy, recall, F1 Score and precision. It has been seen that for the dataset that has been utilized for this work, Random Forest Classifier played out the best in contrast to other classifier calculations. It accomplished an accuracy of 90.16% and the subsequent best is catboost with an accuracy of 88.52%. Also, the determined different exhibition estimates portrays that the Random Forest classifier outflanks the other classification algorithms though the catboost accomplished the subsequent position in performance measure. The model created was able to predict the label in binary values in 0 and 1 and it is clearly evident from the web application which we created in this project work, since on the submission of the input values by the user on the web application, the output will be in 0 and 1.

Chapter 8: Future Prospects

Machine Learning methodologies are certainly advantageous in anticipating the cardiovascular disease by considering various parameters. After analysing the work of various researchers, it is evident that there is lot of scope by which the accuracy of cardiovascular disease prediction can be further improved. Some of the researchers directly used the algorithms on dataset and evaluated the efficiency of algorithms on various measures like accuracy, precision, recall, F1 Score etc. While some of the researchers in their work used some optimization techniques like Particle Swarm Optimization, Ant Colony Optimization to further enhance the accuracy rate. Moreover, Ensemble learning technique was also utilised in predicting the cardiovascular disease and as a result it also proves to be very effective.

Since, there are some other algorithms and techniques which can also be taken into consideration to predict heart disease, so with the help of those algorithms and techniques the accuracy rate can also be estimated and compared with the existing work.

References

[1] Singh, D., & Samagh, J. S. (2020). A comprehensive review of heart disease prediction using machine learning. *Journal of Critical Reviews*, *7*(12), 281-285.

[2] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[3] Lakshmanarao, A., Swathi, Y., & Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

[4] Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, *12*(1), 242-252.

[5] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1275-1278). IEEE.

[6] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[8] Aljanabi, M., Qutqut, H. M., & Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: A review. *International Journal of Engineering & Technology*, *7*(4), 5373-5379.

[9] Vijayarani, S., & Sudha, S. (2013). Comparative analysis of classification function techniques for heart disease prediction. *International Journal of Innovative Research in Computer and Communication Engineering*, *1*(3), 735-741.

[10] David, H., & Belcy, S. A. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES. *ICTACT Journal on Soft Computing*, *9*(1).

[11] Rishabh Magar ; Rohan Memane ; Suraj Raut; Prof. V. S. Rupnar ;HEART DISEASE PREDICTION USING MACHINE LEARNING

[12] Gadde, H. Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning.

[13] Mustafa, J., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology*, *9*, 33.

[14] David, H. B. F. Impact of Ensemble Learning Algorithms Towards Accurate Heart Disease Prediction.

[15] Lo, Y. T., Fujita, H., & Pai, T. W. (2016). Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations. *Journal of Mechanics in Medicine and Biology*, *16*(01), 1640010.

[16] Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

[17] Suvarna, A. J., Kumar, A., Billav, A., KM, M., & Sudhamsu, G. (2019). Predicting The Presence of Heart Disease Using Machine Learning. *International Journal of Computer Science and Mobile Computing, IJCSMC*, *8*(5), 119-125.

[18] Dangare, C., & Apte, S. (2012). A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering and Technology (IJCET)*, *3*(3).

[19] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

[20] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

[21] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

[22] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[23] Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, *181*(18), 20-25.

[24] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 1-6.

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

[26] (2021, June 21). Cardiovascular Disease https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[27] Streamlit Framework https://www.analyticsvidhya.com/blog/2021/06/build-web-app-instantly-for-machine-learning-using-streamlit/ (Accessed on 2022, March, 24)

[28] Heroku Plaform https://mentormate.com/blog/what-is-heroku-used-for-cloud-development/ (Accessed on 2022, March, 26)

[29] Data Analysis https://www.simplilearn.com/data-analysis-methods-process-types-article (Accessed on 2022, March 22)

[30] Data Visualization https://blog.datumize.com/top-five-advantages-of-data-visualization (Accessed on 2022, March 22)

[31] Data Visualization https://wisdomplexus.com/blogs/pros-cons-data-visualization/ (Accessed on 2022, March 2022)

[32] Data Collection https://waverleysoftware.com/blog/data-collection-for-machine-learning-guide/ (Accessed on 2022, March 2022)

[33] Training and Testing https://machinelearningmastery.com/train-to-the-test-set-in-machine-learning/ (Accessed on 2022, March 2022)

[34] Machine Learning Algorithms https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501 (Accessed on 2022, March 2022)

[35] Xgboost Algorithm https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (Accessed on 2022, March 2022)

[36] Catboost Algorithm https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/ (Accessed on 2022, March 2022)

[37] Ensemble Learning https://www.analyticsvidhya.com/blog/2021/12/a-detailed-guide-to-ensemble-learning/ (Accessed on 2022, March 2022)

[38] Python libraries and frameworks https://vivekjaglan.medium.com/frequently-used-functions-of-numpy-pandas-matplotlib-seaborn-scikit-learn-tensorflow-and-keras-a9cb04e2339f (Accessed on 2022, March 2022)

[39] Jupyter Notebook and Spyder https://ssiddique.info/pycharm-vs-spyder-vs-jupyter.html (Accessed on 2022, March 2022)

[40] Streamlit Framework https://www.analyticsvidhya.com/blog/2021/06/build-web-app-instantly-for-machine-learning-using-streamlit/ (Accessed on 2022, April 2022)

[41] Heroku Platform https://mentormate.com/blog/what-is-heroku-used-for-cloud-development/ (Accessed on 2022, April 2022)

[42] Symptoms of CVD https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 (Accessed on 2022, May 2022)

[43] CVD Death Ratio https://www.who.int/india/health-topics/cardiovascular-diseases#:~:text=In%202016%20India%20reported%2063,well%20as%20overweight%20and%20obesity (Accessed on 2022, May 2022)

[44] cardiovascular disease causes https://www.nhs.uk/conditions/cardiovascular-disease/ (Accessed on 2022, May 2022)

[45] Machine learning in cardiovascular disease https://www.nyu.edu/about/news-publications/news/2021/july/machine-learning-for-cardiovascular-disease-improves-when-social.html (Accessed on 2022, May 2022)

Appendix

AMITY UNIVERSITY

UTTAR PRADESH

Minor Project Report

On

Heart Disease Prediction using Machine Learning

Submitted to

Amity University Uttar Pradesh

AMITY UNIVERSITY

In partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

In

Computer Science & Engineering

By

HARSH SHARMA (A2305218648)
AYUSH BHARDWAJ (A2305218614)
(Group Number: 195)

Under the guidance of

Dr. Sumit Kumar

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
AMITY UNIVERSITY UTTAR PRADESH
NOIDA (U.P.)

Session: 2021-2022

DECLARATION

We Harsh Sharma and Ayush Bhardwaj, students of B.Tech Computer Science and Engineering hereby declare that the project titled "Heart Disease Prediction using Machine Learning" which is submitted by us to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of requirement for the award of the degree of Bachelors of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Noida

HarshSharma

Ayush Bhardwaj

Date: 13/12/21 Name and signature of Student(s)

CERTIFICATE

On the basis of declaration submitted by Harsh Sharma, students of B.Tech (Computer Science and Engineering), I hereby certify that the Project titled "Heart Disease Prediction using Machine Learning" which is submitted to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in "Computer Science and Engineering" is an original contribution with existing knowledge and faithful record of work carried out by him/her under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Noida

Date

(Signature of the Guide)
Department of Computer Science and Engineering
Amity School of Engineering & Technology
Amity University Uttar Pradesh, Noida

ACKNOWLEDGEMENT

We take this opportunity to extend our sincere thankfulness to our Vice Chancellor Dr. Balvinder Shukla and Head of the Department Dr. Sanjeev Thakur for bestowing us with the opportunity to work on a project that has immensely helped us to understand the concepts of Machine Learning. We would like to extend our gratitude to our mentor and faculty guide Dr. Sumit Kumar for his commendable guidance, supervision and encouragement thorough the course of this major project. We would also like to thank our parents and friends for their constant encouragement and guidance throughout the course of this training.

Harsh Sharma (A2305218648)
Ayush Bhardwaj (A2305218614)
B. Tech CSE
Amity School of Engineering and Technology, Noida

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

History reveals that out all the diseases heart disease is the difficult to predict and extremely complicated to cure in medical sector. According to various surveys, each minute a person dies because of heart disease. Earlier it was difficult to predict whether a person is suffering from cardiovascular disease or not but with advancement of technology it is now possible to predict such diseases with favourable accuracies. Data Science can aid in clinical industry by accumulating information associated with patients as well as process it which can additionally be made use of in producing forecasts by utilizing numerous machine learning strategies. Anticipating the cardiovascular disease is a difficult task to handle, so it is crucial to make advancements of making prediction with acceptable accuracy in order to initiate the process of treatment as soon as possible. In such cases where a particular technique fails to understand the situation, there is a need to use various techniques together to understand and compare the results. In this research work dataset used is UCI machine learning heart disease dataset. In this research study, dataset made use of is UCI machine learning dataset. In this study, numerous techniques are contrasted according to the accuracies given by the techniques.

CHAPTER 1
INTRODUCTION

1.1 Background Information

According to the reports of WHO (World Health organization) cardiovascular disease is the vital cause of causalities over the past couple of years. Around 32% fatalities throughout the globe are triggered by cardiovascular diseases [26]. Human heart is an essential part of human body which regulates blood in each part of the body, any abnormal condition to heart affects the whole body because of its fundamental role in maintaining flow of blood all over the body. Heart and blood vessels together form cardiovascular system. Any abnormal symptom or uneasiness to the capillary and heart is called cardiovascular disease (CVD). Diseases like strokes, coronary heart disease (CHD), endocarditis, peripheral vascular disease, etc., comes under cardiovascular diseases.

According to the reports of WHO around 17 million patients die every year only because of heart disease [26]. Cardiac arrest and strokes are the significant reason because of which people die, more than four out of five patient suffering from CVD dies because of strokes and heart attacks. One third of the total deaths caused by CVD covers people below the age of 70. Numerous symptoms associated with CVD is the substantial cause to diagnose the problem in a short period of time with efficiency. The major challenge in handling these diseases is the quality service and fast as well as affective and accurate diagnosis of CVD.

There are multifarious causes why one suffers from such conditions, physical inactivity, unhealthy diet, improper sleep schedule, consumption of tobacco and alcohol, stress, psychological factors such as obesity, high blood pressure, hyper tension, high cholesterol, the prime reasons why one suffers from CVD. Identifying those who are at high risk of CVD at early stage and taking early medical measures can prevent premature deaths. These databases can be compared to real life application. Weights of the characteristics are divided on the premise of their impact on making predictions. Therefore, it is important to utilize the knowledge as well as experience of such professionals in preparing databases in order to assist diagnosis and machine learning methods can be used by the database to make predictions and comparing accuracies. With the assistance of such techniques prediction models can be designed, such models can be made into usage in various medical fields by professionals in prediction cardiovascular diseases based on the medical record of patients. Therefore, by carrying out a system for forecasting making use of machine learning approaches, one can expect much more exact as well as analytical outcomes whether an individual is dealing with cardiovascular disease or not.

1.2 Purpose of Plan

Heart disease prediction models plays a critical role in medical fields. Detecting such kind of diseases in early stages can prevent premature death. This research presents a comparative examination of various supervised classification techniques in order to select which method is most affectively suitable for diagnosing as well as predicting heart diseases. To boost accuracy and prediction, feature selection must be done, feature selection is the process of reducing the number of input values to reduce cost of modelling to enhance the performance of the prediction model. Dimensionality reduction can also be used to improve overall attainment of machine learning classification model.

Machine learning classification systems are used in numerous fields around the globe. It plays a necessary part in predicting malignant and benign classes of various diseases like locomotor disorder, breast cancer, skin cancer, heart diseases, etc. Making predictions using such techniques is important because if one is able to predict well in advance with efficiency then it can provide important insights to doctors which can help diagnosis and treatment.

The major objective of the research is to compose a prediction system for cardiovascular disease on heart disease dataset provided by professionals in order to make prediction about a patient whether that person is diagnosed with cardiovascular disease or not, which is a binary outcome. The outcomes of numerous classification models are then going to be compared to recognize the much suitable one to predict heart diseases.

Positive result (patient is diagnosed with heart disease) = 1,

Negative result (patient is not diagnosed with heart disease) = 0. We begin by reviewing and evaluating some of the most recent cardiovascular disease prediction research work done by various researchers. At the last of this report, we will endeavor to find the efficacious method for heart disease prediction.

1.3. Scope

Project Goal: To use numerous machine learning approaches on heart disease dataset to make prediction whether a person is suffering from cardiovascular diseases order, the various predictions generated by machine learning classification models are compared and analysed with intention to pick one of the most ideal category techniques to make predictions using heart disease dataset.

In this project, classification techniques like Decision tree, KNN, Naïve Bayes, etc., is collated with the assistance of heart disease dataset and favourable outcomes were attained. Various methods which are appraised for comparison analysis are confusion matrix, sensitivity, F1 score and precision. Before using classification techniques, the dataset should be normalised in order to avoid overfitting of the training model which can hinder accuracies of prediction drastically.

Recent developments in machine learning can facilitate increase healthcare access in developing countries and innovate cancer diagnosis and treatment. In future, classification techniques will be used on larger training dataset to gain much more precise accuracy. Analysing the scope of usage of machine learning in the medical sectors, more than a million various data points will be maintained in records of health system. Curing heart related diseases can get complicated sometimes due to lack of knowledge and late detection of any irregular activity in the body. This happens mostly when early detection is ignored, this can make the predicament of the patient more complicated in a very short period of time which might result in death.

Due to day-to-day stress, irregular physical activities, unhealthy diet and non-stop tension, the situation has worsened from before. In order to keep the irregularities of heart in control regular check-ups must be practised in order to detect the symptoms in early stages to avoid any causalities. In order to avoid such issues, one should maintain a healthy lifestyle, doing regular exercise, consuming healthy stuff, controlling mind and soul, avoiding consumption of alcohol and tobacco, etc., can decrease the chances of a person to get diagnosed with CVD.

The increasing utilization of machine learning in medical sectors indicates that in upcoming future various intelligent systems may get developed that will help to analyse and predict the particular method of treatment a patient should be given according to their symptoms of the disease they carry. Currently there are many systems already under development which excels in predicting cardiovascular diseases, to attain predictions almost perfectly whether a patient is diagnosed with heart issue or not.

Techniques like machine learning will come handy in future in deciding the method of treatment to be given to a patient by excerpting and scrutinizing the data given to the system by medical professionals.

Tasks: Classification and result analysis.

Costs: Zero Hardware cost.

1.4 Limitations

Following can be the limitations of the project:

1. Considering that it is extremely vital to understand the serviceability of system. So, if there is lack of awareness of any system then it can be the most significant constraints of any kind of project.

2. The mathematical implementation of machine learning methods is quite complex hence carrying out this project using machine learning algorithms without understanding the mathematical aspect can be difficult.

3. For any software project, security is an important aspect of any big software. Since this project is dealing with the heart disease forecasting only, so the security issues with the web application may be created.

4. An important characteristic of any good software is response time.

CHAPTER 2

LITERATURE REVIEW

Various researchers used classification techniques in order to predict cardiovascular disease prediction by using various algorithms. These researchers have applied different algorithms and get high accuracies.

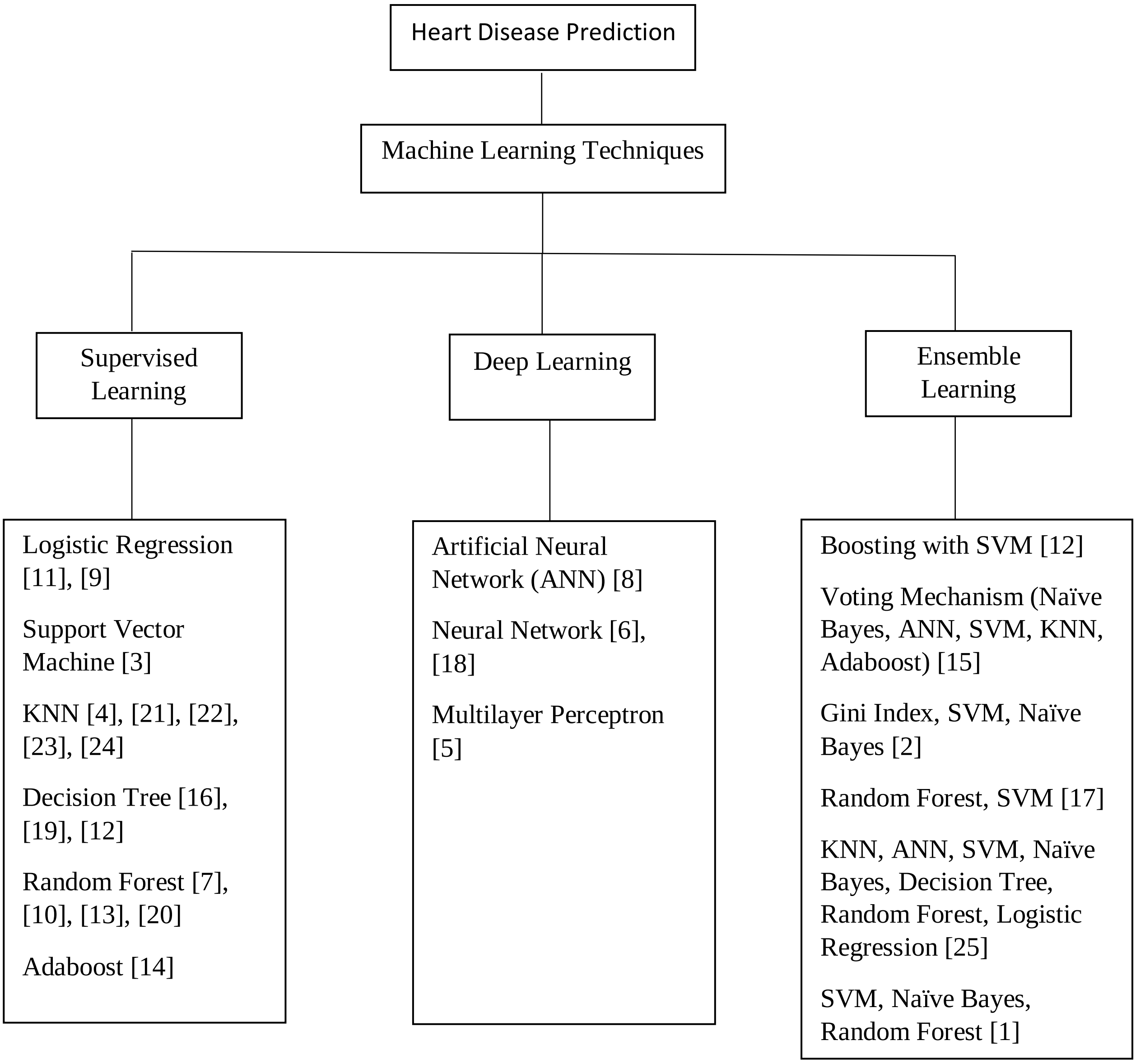

Heart Disease Prediction
Machine Learning Techniques
Supervised Learning
Deep Learning
Ensemble Learning
Logistic Regression [11], [9]
Support Vector Machine [3]
KNN [4], [21], [22], [23], [24]
Decision Tree [16], [19], [12]
Random Forest [7], [10], [13], [20]
Adaboost [14]
Artificial Neural Network (ANN) [8]
Neural Network [6], [18]
Multilayer Perceptron [5]
Boosting with SVM [12]
Voting Mechanism (Naïve Bayes, ANN, SVM, KNN, Adaboost) [15]
Gini Index, SVM, Naïve Bayes [2]
Random Forest, SVM [17]
KNN, ANN, SVM, Naïve Bayes, Decision Tree, Random Forest, Logistic Regression [25]
SVM, Naïve Bayes, Random Forest [1]

Fig. 2.1 Taxonomy of Heart Disease Prediction

Table 2.1 Literature Review of Heart Disease Prediction

Serial Number	**Title of the Research Paper**	**Authors of the Paper**	**Journal in which paper was published**	**Dataset Used**	**Methodology Used**	**Remarks**
[1]	Comprehensive review of Heart Disease Prediction using Machine Learning.	Dr. Dilbag Singh and Jagjit Singh Samagh	Journal Of Critical Reviews	Cleveland Heart Disease database	Naïve Bayes, SVM, Decision Tree, KNN, J48, Neural Networks, PCA, Random Forest.	Various techniques which can enhance the performance of ML algos for predicting heart disease like Ensemble Techniques and Optimization Algorithms.
[2]	Heart Disease Prediction using ML Techniques	V.V. Ramalingam, Ayantan Dandapath, M Karthik Raj	International Journal of Engineering & Technology	Cleveland Heart Disease database	Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models	Ensemble technique of SVM, KNN, ANN gives the accuracy of 94.12%; Naïve Bayes, SVM, Decision Tree gives the accuracy of 82%; Gini Index, SVM, Naïve Bayes gives the accuracy of 98%.

[3]	Machine Learning Techniques for Heart Disease Prediction	A. Lakshmanarao, Y. Swathi, P. Sri Sai Sundareswar	International Journal of Engineering & Technology	Framingham Heart Disease Dataset	Logistic Regression, KNN, Adaboost, Decision Tree, Naive Bayes, Random Forest, SVM, Extra Tree Classifier, Gradient Boosting. Authors applied sampling techniques on the dataset: Random over sampling, Synthetic minority oversampling, Adaptive synthetic sampling approach.	With Random Oversampling technique SVM gives an accuracy of 99.7%. This technique gives more accuracy in all sampling techniques. For Synthetic Minority Oversampling, Random Forest and Extratree Classifier given the best accuracy of 91%. For Adaptive synthetic sampling, Random Forest and Extratree Classifier given the best accuracy of 90%.
[4]	Heart Disease Prediction and Classification using Machine Learning algorithms optimized by Particle Swarm Optimization and Ant Colony Optimization	Youness Khourdifi, Mohamed Bahaj	International Journal of Intelligent Engineering and System	Heart Disease Dataset of UCI Machine Learning Repository	FCBF), KNN, SVM, Naïve Bayes, Random Forest, Multilayer Perceptron, Artificial Neural Network optimized by PSO combined with ACO.	Evaluate the effectiveness of all classifiers according to 3 steps: 1. Classifiers without optimization 2. Classifiers optimized by FCBF 3. Classifiers optimized by FCBF, PSO and ACO. And it shows that that the optimization hybrid approach increases the predictive accuracy of medical dataset.

[5]	Prediction of Heart Disease using Machine Learning	Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar	IEEE Xplore	Cleveland Dataset from UCI Library	Neural network algorithm multi-layer perceptron (MLP)	The output of the system will give a result if the person has a heart disease, in terms of Yes or No. It gives the average precision of 91%.
[6]	An Analysis of Heart Disease Prediction using Data Mining Techniques	Nidhi Bhatla, Kiran Jyoti	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease database	Neural Network, Decision Tree, Combination of Genetic algorithms and Decision Tree	Neural Network has shown the highest accuracy i.e., 100% so far. Decision Tree performed well with 99.62% accuracy with the aid of 15 attributes. Additionally, Genetic algorithm with unification of Decision tree gives the accuracy of 99.2%.
[7]	Effective Heart Disease Prediction using Hybrid Machine Learning Techniques	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava	IEEE Access	Cleveland Dataset from UCI Library	Hybrid Random Forest with Linear Model	Authors generated the boosted performance level with an accuracy degree of 88.7% through the prediction model for CVD with HRFLM.

[8]	Machine Learning Techniques for Heart Disease Prediction: A Review	MaryamI. Al-Janabi, Mahmoud H. Qutqut, Mohammad Hijjawi	International Journal of Engineering & Technology	Cleveland Dataset from UCI Library (Mostly Used)	Naïve Bayes, Artificial Neural Network, Radial Basis Function, Decision Tree, K-Nearest Neighbor, Support Vector Machine (SVM), Genetic Algorithm, Ensemble Learning	Authors conclude that the researchers who produced the highest accuracy were Dangare and Apte using Artificial Neural Network (ANN), WEKA tool and a combination of the Cleveland and Statlog heart disease datasets.
[9]	Comparative Analysis of Classification Function Techniques for Heart Disease Prediction	Dr. S. Vijayarani1, S. Sudha2	International Journal of Innovative Research in Computer and Communication Engineering	Cleveland cardiovascular disease dataset from UCI repository	Logistic Regression, Multi-Layer Perceptron, Sequential Minimal Optimization	The paper analyses the performance of various classification function techniques in data mining for predicting the heart disease from the heart disease data set. The classification function algorithms used and tested in this work are Logistics, Multi-Layer Perception and Sequential Minimal Optimization algorithms.

[10]	Heart Disease Prediction using Data Mining Techniques	H. Benjamin Fredrick David and S. Antony Belcy	ICTACT Journal on Soft Computing	StatLog dataset in UCI repository	Random Forest, Decision Tree, Naïve Bayes	The main objective of this significant study is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out. It is found that Random Forest algorithm performs best with 81% precision after comparing to other algorithms for heart disease prediction.
[11]	Heart Disease Prediction using Machine Learning	Rishabh Magar, Rishabh Magar, Rohan Memane, Suraj Raut, Prof. V. S. Rupnar	Journal of Emerging Technologies and Innovative Research (JETIR)	Cleveland heart diseases data set from the University of California Irvine (UCI)	Support Vector Machine (SVM), Decision Tree, Naïve Bayes Algorithm, Logistic Regression	Found out that Logistic Regression algorithm has the most efficient out of the four and resulted with an accuracy of 82.89%. Decision tree and Naïve Bayes had accuracy of 80.43% and 80.43% respectively, and SVM was having 81.57%

[12]	Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning	Hemanth Gadde	International Journal of Engineering Trends and Applications (IJETA)	Cleveland heart dataset from the UCI Machine Learning Repository	Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Ensemble Methods	It is evident that most of the machine learning algorithms are performing well in predicting and diagnosing of cardio vascular or heart diseases, may be some algorithms are poor in results in terms of performance and accuracy measures, Random Forest, Decision Tree algorithms generally work well on the data related to over fitting, whereas algorithms like SVM and Naive Bayes will work for real-world problems and on data sets.
[13]	Ensemble approach for developing a smart heart disease prediction system using classification algorithms	Mustafa Jan, Akber A Awan, Muhammad S Khalid, Salman Nisar	Research Reports in Clinical Cardiology 2018	Cleveland and Hungarian dataset from UCI Machine Learning Repository	Naïve Bayes, Neural Networks, Support Vector Machine (SVM), Random Forest, Ensemble Method	Lowest accuracy is 93.22% for regression analysis and the highest accuracy is 98.17% for the RF algorithm. The RF ensemble algorithm and SVM performed well and further hybridization through voting of each algorithm with more than 93% prediction probability has enhanced reliability of the system. More emphasis is given to select the algorithms having high true positive rate, as being the core measure for early diagnosis of cardiovascular disease.

[14]	Impact of ensemble learning algorithms towards accurate heart disease prediction	H. Benjamin Fredrick David	ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2020, VOLUME: 10, ISSUE: 03	StatLog Dataset	Bagging, Stacking and AdaBoost Support Vector Machine, Naive Bayes and K-Nearest Neighbour.	The main objective of the research work is to identify the best performing ensemble classification algorithm for heart disease prediction. For this purpose, the UCI data repository is used for performing the comparative analysis of three algorithms such as AdaBoost, Bagging and Stacking. From the research work, it has been experimentally proven that AdaBoost provides perfect results as compared to competitors.
[15]	Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations.	YING-TSANG LO, HAMIDO FUJITA and TUN-WEN PAI	Journal of Mechanics in Medicine and Biology Vol. 16, No. 1 (2016) 1640010 World Scientific Publishing Company	Cleveland Heart Disease Dataset	Naïve Bayes, ANN, SMO, KNN, AdaBoost, J48, and Random Forest,	In this study, seven machine learning methods were applied to make predictions based on CAD datasets. The results have shown that proposed ensemble learning/voting mechanism provided the best prediction performance. Analytical procedures directly imply that CAD exhibits positive correlation with age, heart rate, blood pressure, smoking habits, and cholesterol factors.

[16]	Heart Disease Prediction Using Classification with Different Decision Tree	K. Thenmozhi , P.Deepika	International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014	Cleveland Heart Disease database	Decision tree, Artificial Neural Network and Bayesian Classifier.	In this work Various techniques and data mining classifiers are defined for effective cardiovascular disease diagnosis. In this, Decision tree resulted with 99.62% accuracy with the help of using 15 attributes. Furthermore, in combination with genetic, Decision tree has shown 99.2% performance.
[17]	Predicting the presence of heart disease using machine learning	Akshay Jayraj Suvarna, Arvind Kumar M, Ajay Billav, Muthamma K M, Asst. Prof. Gadug Sudhamsu	International Journal of Computer Science and Mobile Computing	Cleveland Heart Disease Dataset	Random Forest and Support Vector Machine	When it is the initial stage of such disease, it is significant to identify in the initial phase only. So, when the disease is identified then it becomes very important to give the proper treatment to the patient. Hence, machine learning becomes very useful in such cases to predict such disease in prior.

[18]	A data mining approach for prediction of heart disease using neural network	Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte	International Journal of Computer Engineering and Technology (IJCET), ISSN 0976–6367(Print), ISSN0976-6375(Online) Volume 3, Issue 3,	Cleveland Heart Disease database	Neural network.	This work provides the prediction system for cardiovascular disease using data mining as well as ANN techniques. From ANN, a multilayer perceptron neural network is made use of to construct the system. Authors in their research study states that neural network led to 100% accuracy.
[19]	Heart Disease Prediction Using Machine learning and Data Mining Technique	Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel	IJCSC	Cleveland database of UCI repository	J48 algorithm, Logistic model tree algorithm and Random Forest algorithm.	In this research work it was concluded thatJ48 tree technique turned out to be best classifier for cardiovascular disease prediction because it contains more accuracy and least total time to build. J48 on UCI data resulted with supreme accuracy i.e., 56.76% and the total time to build model is 0.04 seconds
[20]	Heart Disease Prediction using Machine Learning	Apurb Rajdhan, Dundigalla Ravi, Milan Sai, Avi Agarwal	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease Dataset	Naive Bayes, Decision Tree, Logistic Regression and Random Forest.	The trial results verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

[21]	Heart Disease Prediction Using Machine Learning Algorithms	Archana Singh, Rakesh Kumar	International Journal of Engineering Research & Technology (IJERT)	Cleveland Heart Disease Dataset	k-nearest neighbor, decision tree, linear regression and support vector machine (SVM)	To evaluate the different classification techniques KNN gives the best technique for heart disease prediction.
[22]	Heart disease prediction using machine learning algorithms	Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath	IOP Conf. Series: Materials Science and Engineering	UCI repository with patient's medical history and attributes	Logistic Regression and KNN	KNN to get an accuracy of an average of 87.5% on the prediction model which is better than the previous models having an accuracy of 85%. Accuracy of KNN is highest between the methods that were used i.e., 88.52%.
[23]	A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach	M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, V. Pavithra	International Journal of Computer Applications (0975 – 8887)	Cleveland database of UCI repository	Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbor (KNN), Naïve Bayes and Support Vector Machine (SVM).	The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN, SMO and Naïve Bayes. KNN gives the best technique for heart disease prediction.
[24]	Heart Disease Prediction using Machine Learning Techniques	Devansh Shah, Samir Patel, Santosh Kumar Bharti	SN Computer Science (2020)	dataset from the Cleveland database of UCI repository	Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm.	The highest accuracy score is achieved with KNN.

[25]	Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms	Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun	Hindawi Mobile Information Systems	Cleveland heart disease dataset	logistic regression, K-NN, ANN, SVM, NB, DT, and random forest	In this research study, a hybrid intelligent machine-learning based predictive system was contemplated.The system was tested on Cleveland heart disease dataset. Seven well-known classifiers such as logistic regression, K-NN, ANN, SVM, NB, DT, and random forest were used with feature selection algorithms Relief, mRMR, and LASSO used to select the important features. The K-fold cross-validation method was used in the system for validation

CHAPTER 3
METHODOLOGY

With the assistance of machine learning, prediction models can be designed, prediction model can be utilized in various medical fields by professionals in prediction cardiovascular diseases based upon medical record of patients. Hence, by implementing a prediction system using classification techniques one can anticipate more accurate and statistical results whether a person is struggling with cardiovascular disease or otherwise. In this scrutinized work the main essence was on numerous methods that are used in different classification models in order to predict such diseases at early stages so that adequate and cogent treatment can be provided to the respective patient.

By using the data provided by patients, large datasets can be created by processing the data which can further be used to make predictions. By using classification techniques numerous patterns and covered information can be extracted from such large datasets and by analysing these patterns predictions can be made with utmost accuracy.

The principal motive of this study is to create a classification model by utilising classification techniques like naïve bayes, logistic regression, decision tree, KNN, random forest, etc. and comparing the results of each classification technique after applying them on cardiovascular disease dataset and comparing their results and choosing out the algorithm with most favourable outcome and check if the results can be enhanced by using various boosting techniques.

3.1 Data Collection

The most favourable steps to create a classification model are, first collecting data and pre-processing it. Cleveland heart disease dataset from UCI repository is taken into account. The dataset is made up of 76 attributes, this dataset is considered in this research because many different authors of various research papers selected this dataset and used it to make classification models. This dataset is available for public use and is already processed so that it can be easy to use for analysis.

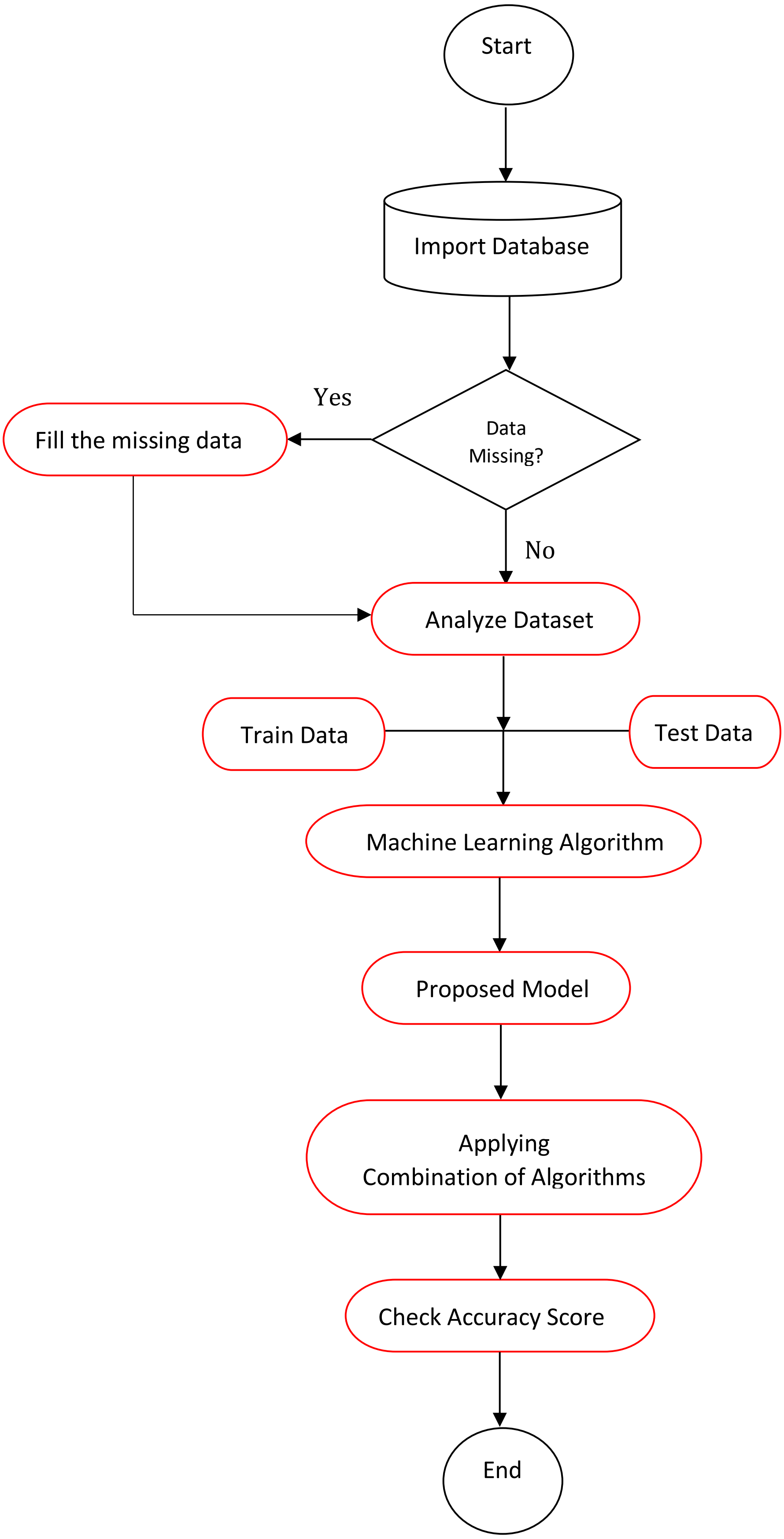


Fig. 3.1 Methodology

3.2 Data Pre-processing

There are numerous approaches like random forest which does not sustains void values in the dataset which can impede the results in a adverse side that’s why it is important to pre-process the information as well as transform the raw information from the dataset right into usable data. Essentially, this pre-processing is done to inspect if there is any missing value existing in the dataset or not. Moreover, Normalization and Standardization is also carried out to bring the values in the ranges.

3.3 Data Analysis

Data Evaluation is essential to determine the patterns and eventually to understand the new end results that will certainly aid to conveniently comprehend and also eventually to refine the data to make predictions, in this technique the data from the dataset is explored in such a form which can be easily analysed.

3.4 Training and Testing on Proposed Model

Next step is to divide the dataset right into testing and also training components in order to utilize a part of dataset to train the model and the other component to examine it whether the predictions created by the model are sustainable or not. After that we will examine the recommended model in which the combination of algorithms will certainly be made use of. Highest possible three algorithms which will reveal the optimum accuracy will be taken into account for combination. Then the proposed model will be used to make the predictions.

3.5 Output

Since the predictions are done to inspect which model is ideal fit for the medical area to predict an illness. In view of the fact that, this proposed model can be proved extremely significant in anticipating heart related problem so it ends up being very crucial to evaluate the accuracy of the outcomes by taking various specifications to make sure that the proposed model can be used by physicians to predict the heart disease of patients at early stage and accordingly ideal actions can be thought about.

Hence, the efficiency of the proposed model will certainly be assessed on numerous specifications. like accuracy, precision, Recall as well as F1 Score by making use of Confusion Matrix.

CHAPTER 4
CONCLUSION AND FUTURE SCOPE

Machine Learning methodologies are certainly advantageous in anticipating the cardiovascular disease by considering various parameters. After analysing the work of various researchers, it is evident that there is lot of scope by which the accuracy of cardiovascular disease prediction can be further improved. Some of the researchers directly used the algorithms on dataset and evaluated the efficiency of algorithms on various measures like accuracy, precision, recall, F1 Score etc. While some of the researchers in their work used some optimization techniques like Particle Swarm Optimization, Ant Colony Optimization to further enhance the accuracy rate. Moreover, Ensemble learning technique was also utilised in predicting the cardiovascular disease and as a result it also proves to be very effective.

Since, there are some other algorithms and techniques which can also be taken into consideration to predict heart disease, so with the help of those algorithms and techniques the accuracy rate can also be estimated and compared with the existing work.

References

[1] Singh, D., & Samagh, J. S. (2020). A comprehensive review of heart disease prediction using machine learning. *Journal of Critical Reviews*, *7*(12), 281-285.

[2] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[3] Lakshmanarao, A., Swathi, Y., & Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

[4] Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, *12*(1), 242-252.

[5] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1275-1278). IEEE.

[6] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[8] Aljanabi, M., Qutqut, H. M., & Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: A review. *International Journal of Engineering & Technology*, *7*(4), 5373-5379.

[9] Vijayarani, S., & Sudha, S. (2013). Comparative analysis of classification function techniques for heart disease prediction. *International Journal of Innovative Research in Computer and Communication Engineering*, *1*(3), 735-741.

[10] David, H., & Belcy, S. A. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES. *ICTACT Journal on Soft Computing*, *9*(1).

[11] Rishabh Magar ; Rohan Memane ; Suraj Raut; Prof. V. S. Rupnar ;HEART DISEASE PREDICTION USING MACHINE LEARNING

[12] Gadde, H. Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning.

[13] Mustafa, J., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology*, *9*, 33.

[14] David, H. B. F. Impact of Ensemble Learning Algorithms Towards Accurate Heart Disease Prediction.

[15] Lo, Y. T., Fujita, H., & Pai, T. W. (2016). Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations. *Journal of Mechanics in Medicine and Biology*, *16*(01), 1640010.

[16] Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

[17] Suvarna, A. J., Kumar, A., Billav, A., KM, M., & Sudhamsu, G. (2019). Predicting The Presence of Heart Disease Using Machine Learning. *International Journal of Computer Science and Mobile Computing, IJCSMC*, *8*(5), 119-125.

[18] Dangare, C., & Apte, S. (2012). A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering and Technology (IJCET)*, *3*(3).

[19] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

[20] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

[21] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

[22] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[23] Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, *181*(18), 20-25.

[24] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 1-6.

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

[26] (2021, June 21). Cardiovascular Disease https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

DrillBit

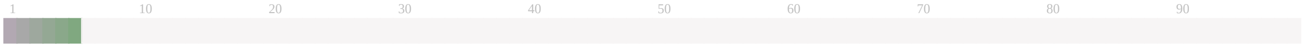
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

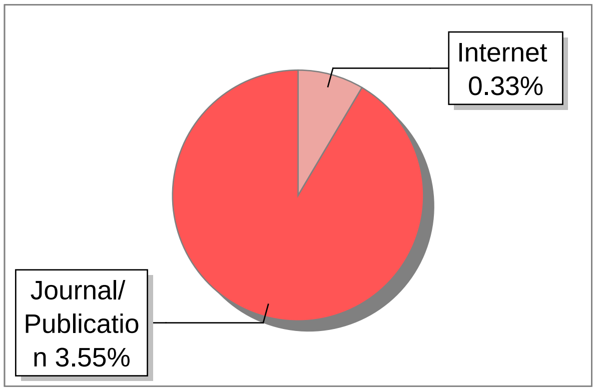
Author Name	Group-195
Title	Harsh
Paper/Submission ID	510870
Submission Date	2022-05-05 09:25:34
Total Pages	48
Document type	Thesis

Result Information

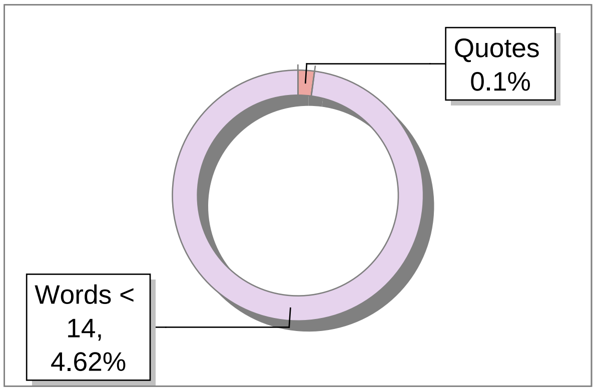
Similarity **6 %**



Sources Type



Report Content



Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words Similarity	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

6	51	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	IEEE 2014 5th International Conference- Confluence The Next Generati by	<1	Publication
2	IEEE 2020 IEEE International Conference on Electronics, Computing and Communic	<1	Publication
3	www.researchgate.net	<1	Internet Data
4	MIFH A Machine Intelligence Framework for Heart Disease Diagnosis by Gupta-2019	<1	Publication
5	A Study on Coronary Disease Prediction Using Boosting-based Ensemble Machine Lea by Sulta-2019	<1	Publication
6	Data Mining Classification Algorithms for Kidney Disease Prediction by S-2015	<1	Publication
7	Prediction of heart disease and classifiers sensitivity analysis by Almustafa-2020	<1	Publication
8	www.dx.doi.org	<1	Publication
9	researchspace.ukzn.ac.za	<1	Publication
10	C Gmez-Campo,Editors, ,Biology of Brassica Coenospecies (1999) Elsevier Scienc by Rdige-2000	<1	Publication
11	arxiv.org	<1	Publication

12	www.researchgate.net	<1	Internet Data
13	www.dx.doi.org	<1	Publication
14	www.aensiweb.net	<1	Publication
15	Removal of Methyl Orange from Water Using Sulfur-Modified nZVI Supported on Bioc by Yang-2018	<1	Publication
16	Multilabel Classification of Tweets- www.ijcaonline.org	<1	Publication
17	Fault Isolability Analysis and Optimal Sensor Placement for Fault Diagnosis in S by Trothe-2019	<1	Publication
18	Article Published at Procedia - Social and Behavioral Sciences - SCIDIR.	<1	Internet Data
19	IEEE 2014 23rd Australian Software Engineering Conference (ASWEC) - by	<1	Publication
20	Ensemble Learning With Attention-Integrated Convolutional Recurrent Neural Netwo by Ai-2020	<1	Publication
21	www.researchtrend.net	<1	Publication
22	Computational Intelligence based QoS-aware Web Service Composition A by Jatoth-2015	<1	Publication
23	Article Published by Journal of Computations & Modelling - www.scienpress.com	<1	Publication
24	IEEE 2019 IEEE Jordan International Joint Conference on Electrical E	<1	Publication
25	www.researchgate.net	<1	Internet Data
26	www.jestr.org	<1	Publication
27	www.hindawi.com	<1	Internet Data
28	www.dx.doi.org	<1	Publication

29	www.dx.doi.org	<1	Publication
30	www.dx.doi.org	<1	Publication
31	www.dx.doi.org	<1	Publication
32	Type-2 fuzzy logic aggregation of multiple fuzzy controllers for airplane flight by Cervantes-2015	<1	Publication
33	Thesis submitted to shodhganga - shodhganga.inflibnet.ac.in	<1	Publication
34	Thesis submitted to shodhganga - shodhganga.inflibnet.ac.in	<1	Publication
35	Student Article Published in www.conferenceworld.in	<1	Publication
36	Physicochemical properties of Iranian ziziphus honey and emerging appr by Faal-2019	<1	Publication
37	Performance and economic viability assessment of a 50 KWp rooftop solar photovol by Mukherji-2020	<1	Publication
38	Paper published at International Journal of Computer Science & Mobile Computing - www.ijcsmc.com	<1	Publication
39	Mapping plant functional types in Northwest Himalayan foothills of India using r by Srinet-2020	<1	Publication
40	IEEE 2020 11th International Conference on Computing, Communication and Networ	<1	Publication
41	Empirical Evidence Supporting the Use of Multiple Choice Models in Analyzing a P by Gensch-1987	<1	Publication
42	Classification of Handwritten Devanagari Number An analysis of Pattern Recog by Prashanth-2020	<1	Publication
43	Biomarkers of severity and threshold of allergic reactions during oral peanut ch by Santos-2020	<1	Publication

44	Bayesian network learning algorithm based on unconstrained optimizatio by Wang-2012	<1	Publication
45	Application of fuzzy logic and genetic algorithm in heart disease risk level pre by Sharma-2017	<1	Publication
46	An overview of cardiovascular disease infection A comparative analysis of boost by Adeboye-2020	<1	Publication
47	An ensemble machine learning approach through effective feature extraction to cl by Hakak-2021	<1	Publication
48	An artificial neural network approach for predicting hypertension using NHANES d by Lpez-Martnez-2020	<1	Publication
49	IEEE 2019 1st International Conference on Innovations in Information	<1	Publication
50	Role of machine learning algorithms over heart diseases prediction by Jonnavithula-2020	<1	Publication
51	ijarcce.com	<1	Publication

COMMENT BY EXTERNAL EXAMINER (Page to be added at the end of the Report)

Name & Signature of the External Examiner: