A Project Report on

CardioVascular Disease Prediction Using Machine Learning Algorithms

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

B.PRAJNAYA
(20H51A0507)

V.SRI VIDYA
(20H51A0524)

B.SHARANYA
(20H51A05K2)

Under the esteemed guidance of

Mr. A.VIVEKANAND
(Associate Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2023- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project report entitled " **CardioVascular Disease Prediction Using Machine Learning Algorithms**" was being submitted by B.Prajnaya(20H51A0507),V.Sri Vidya (20H51A0524), B. Sharanya (20H51A05K2) in partial fulfillment for the award of Bachelor of Technology in Computer Science and Engineering is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree.

A.Vivekanand
Associate Professor
Dept. Of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. Of CSE

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express our heartfelt gratitude to all the people who helped in making this project a grand success.

We are grateful to **Mr.A.Vivekanand, Associate Professor**, Department of CSE for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of CSE, CMR College of Engineering and Technology, who is the major driving force to complete our project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of CSE for their cooperation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary& Correspondent, CMR Group of Institutions, and Shri Ch Abhinav Reddy, CEO, CMR Group of Institutions for their continuous care and support.

Finally, we extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly or indirectly in completion of this project work.

B.PRAJNAYA (20H51A0507)

V.SRIVIDYA (20H51A0524)

B.SHARANYA (20H51A05K2)

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Detecting cardiovascular problems early is crucial for timely treatment. In our study, we employed machine learning to analyze a diverse set of information about individuals' lives and health, to predict cardiovascular disease. Ensuring data accuracy and addressing missing information were prioritized in our approach. Experimenting with different solo ML & ensemble ML methods, composed of Random Forest and XGBoost with tuning, we achieved a notable 92% accuracy in identifying potential heart issues. Remarkably, combining multiple machine learning methods through ensemble learning proved even more effective than individual methods. Expanding our methodology to include Light GBM, Extra Tree, Decision Tree, SVM, Naive Bayes, QDA, & Adaboost enhanced the comprehensiveness of our analysis. Additionally, delving into ensemble learning methods such as bagging, boosting, tuning, & stacking further pushed the boundaries of predictive accuracy. In essence, our research stands out the potency of diverse ensemble machine-learning techniques and algorithms in early cardiovascular prediction. Ensemble methods, which combine different algorithms, emerged as powerful tools without relying on complex terminology.

Keywords— Cardiovascular Disease (CVD), Artificial Intelligence(AI), Machine Learning (ML), Deep Learning (DL), Ensemble Learning, Heart disease.

CHAPTER 1

INTRODUCTION

CHAPTER 1
INTRODUCTION

As per WHO CVD is the deadliest disease almost a third of all deaths are caused by cardiovascular disease it takes over 17 million lives over a year. In the year 2023, it took over 20.5 million lives [1]. Advanced knowledge can help us to prevent deaths caused by cardiac arrest. To identify cardiac diseases there are symptoms like chest pain, chest pressure, shortness of breathing, and fainting. Heart diseases are identified by clinical examinations, medical history, and diagnostic tests.Cardiovascular disease (CVD) is a disease that affects the heart and blood vessels. There are distinct types of cardiovascular diseases. They are Coronary heart disease, Peripheral heart disease, stroke, and Heart failure. Cardiovascular disease is caused due to raised blood pressure, high cholesterol levels, lack of exercise, diabetes, smoking, tobacco, and improper Diet. Healthcare professionals or providers can use Artificial Intelligence to detect heart diseases. Leads to diagnosis at the advanced stage of the disease. Artificial Intelligence is predominantly used in identifying cardiovascular disease. In the presented study we hypothesize that ensemble, ML functionalities are superior to solo and deep learning algorithms due to the scarcity of available clinical data. Artificial intelligence, with its capacity for data analysis and pattern recognition, holds immense promise in revolutionizing the detection and management of cardiovascular disease. Leveraging machine learning (ML) algorithms, healthcare professionals can sift through vast troves of clinical data to identify subtle indicators of cardiac pathology, facilitating earlier interventions and mitigating the risk of adverse outcomes. Moreover, the advent of ensemble learning—an approach that combines multiple ML models to yield more robust predictions—offers a particularly tantalizing prospect for enhancing diagnostic accuracy in the context of limited clinical data availability. In light of these considerations, the present study endeavors to explore the comparative efficacy of ensemble ML methodologies vis-à-vis solo and deep learning algorithms in the realm of cardiovascular disease diagnosis. By elucidating the potential advantages conferred by ensemble approaches in the face of data scarcity, this research aims to chart a course towards more effective and efficient strategies for combating the scourge of CVD.

1.1.Problem Statement

Early detection and effective management of CVD are critical for reducing the burden of this disease on individuals and healthcare systems. However, traditional approaches to CVD prediction and diagnosis often have limitations, such as reliance on a limited set of risk factors and the complexity of CVD etiology.Furthermore, the availability of comprehensive and high-quality data for developing accurate predictive models for CVD is often limited, posing a significant challenge for researchers and healthcare providers. Machine learning (ML) algorithms offer a promising solution to these challenges by leveraging large datasets to identify complex patterns and relationships that may not be apparent through conventional methods.

Therefore, the problem statement revolves around developing robust ML-based predictive models for CVD that can address the limitations of traditional approaches, including data scarcity and predictive accuracy. These models aim to enhance early detection and intervention strategies for CVD, ultimately improving patient outcomes and reducing the burden on healthcare systems.

1.2.Research Objective

To develop and validate predictive models using machine learning (ML) algorithms for the early identification of individuals at risk of developing cardiovascular diseases (CVDs) before symptoms manifest. These models aim to leverage the comprehensive integration of diverse data sources to detect complex patterns, enhancing the accuracy of CVD risk assessment compared to traditional risk scoring systems. Furthermore, the research seeks to evaluate the potential impact of these ML-based predictive models in reducing the mortality rates associated with CVDs and alleviating the economic burden on healthcare systems globally.

1.3.Project Scope

- Develop a web-based application focused on predicting cardiovascular disease (CVD) using machine learning algorithms.
- The application will serve healthcare providers, including physicians, nurses, and other professionals, by providing them with accurate predictions to make more informed decisions about patient care.
- Additionally, the application will be useful to individual users, allowing them to assess their risk of CVD and receive personalized recommendations for preventive measures and dietary adjustments.
- Incorporate machine learning models capable of analyzing diverse data sources to predict the likelihood of CVD development in both clinical and non-clinical settings.
- Provide an intuitive user interface that allows healthcare providers to input patient data and receive instantaneous predictions regarding CVD risk.
- For individual users, offer a user-friendly platform where they can input personal health data and lifestyle factors to receive personalized risk assessments and dietary recommendations.
- Ensure data privacy and security measures are implemented to protect sensitive health information provided by both healthcare providers and individual users.
- Conduct rigorous testing and validation of the machine learning models to ensure accuracy and reliability in predicting CVD risk for diverse populations.
- Collaborate with healthcare professionals, nutritionists, and other experts to develop evidence-based dietary recommendations tailored to individual risk profiles.
- Provide ongoing support and updates to the web application to incorporate new research findings and improve prediction accuracy over time.
- Aim to improve overall health outcomes by facilitating early detection and prevention of CVD through personalized risk assessment and dietary intervention strategies.

CHAPTER 2

BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

Cardiovascular diseases (CVDs) are a major global health concern, contributing significantly to morbidity and mortality rates worldwide. Understanding the epidemiology of CVDs is crucial for identifying trends, prevalence rates, and disparities across regions. Through epidemiological studies, researchers gain insights into demographic factors like age, gender, and socioeconomic status that influence CVD risk. This understanding helps in developing targeted interventions and healthcare policies to address the diverse needs of populations.

Various risk factors contribute to the development and progression of CVDs. Established factors include hypertension, hyperlipidemia, diabetes, smoking, obesity, and physical inactivity. Additionally, emerging factors like genetic predisposition, psychosocial stress, and environmental exposures add complexity to CVD etiology. Recognizing the interplay between these factors is essential for crafting effective prevention and management strategies tailored to individual risk profiles.

Current diagnostic methods for CVDs typically involve clinical assessments, medical history evaluations, and laboratory tests. Traditional risk scoring systems, such as the Framingham Risk Score, have been widely used to estimate an individual's CVD risk. However, these methods may have limitations, especially in diverse populations or when considering complex interactions among multiple risk factors.

Early identification of individuals at risk of CVDs before symptoms appear is a key challenge. Delayed diagnosis can result in missed opportunities for preventive interventions and poorer patient outcomes. Machine learning presents promising solutions by leveraging comprehensive datasets to develop predictive models that identify at-risk individuals. By analyzing diverse data sources, including electronic health records, genomic data, and wearable device data, machine learning algorithms can detect complex patterns and interactions among risk factors, leading to more accurate risk assessments.

The clinical implications of machine learning-based predictive models for CVDs are substantial. Early identification of at-risk individuals enables timely interventions, which can improve patient outcomes and reduce healthcare costs. Moreover, developing predictive models tailored to individual risk profiles holds promise for personalized medicine approaches in CVD prevention and management. However, ethical considerations, such as data privacy and algorithm transparency, must be addressed to ensure responsible and equitable use of predictive modeling tools in clinical practice.

2.1.Literature Review

In this literature review, we explore the latest research on machine learning-based approaches for CVD prediction. We examine the different machine learning algorithms employed in CVD risk prediction, the datasets used for model development and validation, evaluation metrics for assessing model performance, and recent advancements in the field. By synthesizing findings from the literature, we attempt to contribute insights into the potential of machine learning to revolutionize CVD risk stratification and inform personalized preventive strategies in clinical practice.

Muktevi Sri Venkatesh et al[2] In their discussion of the early methods for forecasting cardiovascular illness, suggested a prediction using an SVM, Naïve Bayes classifier, Random Forest RF, and logistic regression. In comparison to other machine learning algorithms, logistic regression has a higher accuracy rate (77.06%), according to his research.

Ankur Gupta et al [3] suggested a framework for machine intelligence MIFH is used to diagnose heart disease. They suggested a framework called MIFH, which can be utilized to forecast the occurrences of either heart patients or normal persons. Their sensitivity was 91%, compared to 89.28% for MIFH.

Abu Yazid et al[4] researched the ANN and used the Cleveland dataset to achieve 90.9% of accuracy and also worked with the statlog dataset and achieved an accuracy of 90%.

Talha Javed et al [5] suggested deep learning and machine learning methods based on ensembles to forecast cardiovascular illness. The models' performance was evaluated based on how accurate they were. Their accuracy rate was 88.70%.

Pradhan et al [6] considered the UCI repository dataset and five methods (support vector machine, logistic regression, main component analysis, multi-layer perceptron classifier, and achieved approximately 90% accuracy.

Vicky Singh et al [7] used machine learning algorithms in this examination, and a recommendation system based on variables like age, blood pressure, and so on was presented. They concluded that SVM and decision tree classifiers provide 85% accuracy.

R. Karthikeyan et al [8] suggested utilizing a convolution neural network and deep learning to predict cardiovascular illness.

Uma Maheshwari et al[9] engaged a unique method for predicting cardiac disease by combining neural networks with logistic regression analysis. Initially, the foremost risk indicators for illness prediction are chosen using logistic regression. The statistical p-value is produced. With an accuracy of 84%, the combination of logistic regression and neural network is used to predict cardiac disease.

Senthil Kumar et al[10] By using machine learning approaches, created an excellent approach that improves the accuracy of cardiovascular disease detection by identifying important aspects. He achieved an improved performance level of 88.7% using HRFLM.

2.2.Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

2.2.1 Introduction

Senthilkumar Mohan et al.[10] Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN),Decision Trees (DT), Genetic algorithm (GA), and NaiveBayes (NB).In this work, numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques.This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR),

Atrial flutter (AFL), Premature Ventricular Contraction (PVC)), andSecond degree block (BII) to find out the exact condition of the patient in relation to heart disease. The dataset with a radial basis function network (RBFN) is used for classification, where 70% of the data is used for training and the remaining 30% is used for classification and also introduced Computer Aided Decision Support System (CADSS) in the field of medicine and research.

2.2.2 Merits and Demerits

Merits:

- Designed a model called Hybrid Random Forest with Linear Model (HRFLM) which gives accuracy of 89%.
- Unlike many other studies that impose limitations on feature selection for algorithmic use, HRFLM utilizes all features without any constraints on feature selection

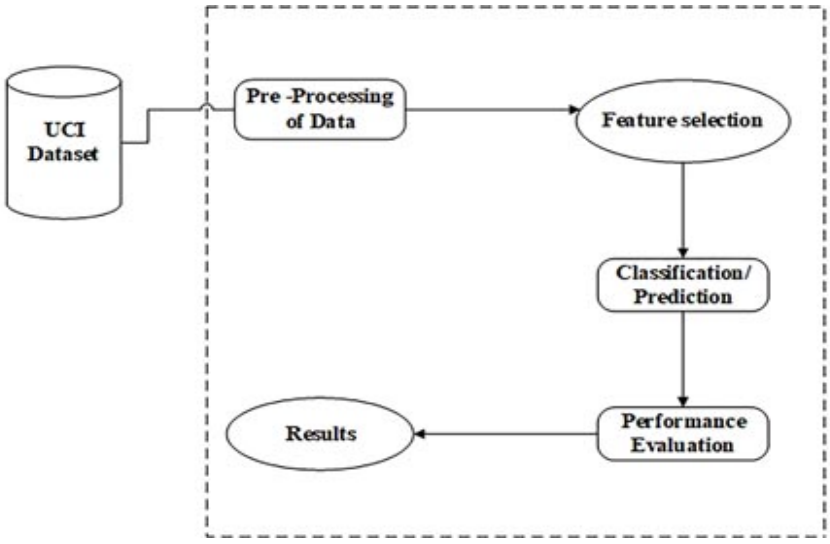
Demerits :

- Combining multiple models may increase the risk of overfitting the data, especially if not carefully controlled, which can lead to reduced generalization and reliability of the predictions on new data .
- Integrating both the random forest and linear model techniques may lead to increased complexity in the implementation and understanding of the methodology, which could pose challenges for users with limited expertise.

2.2.3 Implementation

In HRFLM, a computational approach is used with the three association rules of mining namely, apriori, predictive and Tertius to find the factors of heart disease on the UCI Cleveland dataset. The available information points to the deduction that females have less of a chance for heart disease compared to males. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis. HRFLM makes use of ANN with back propagation along with 13 clinical features as the input. The obtained results are comparatively analyzed against traditional methods. The risk levels

become very high and a number of attributes are used for accuracy in the diagnosis of the disease.The Probabilistic Principal ComponentAnalysis (PPCA) method is proposed for evaluation, based on three datasets of Cleveland, Switzerland, and Hungarian in UCI respectively. The method extracts the vectors with high covariance and vector projection used for minimizing the feature dimension.

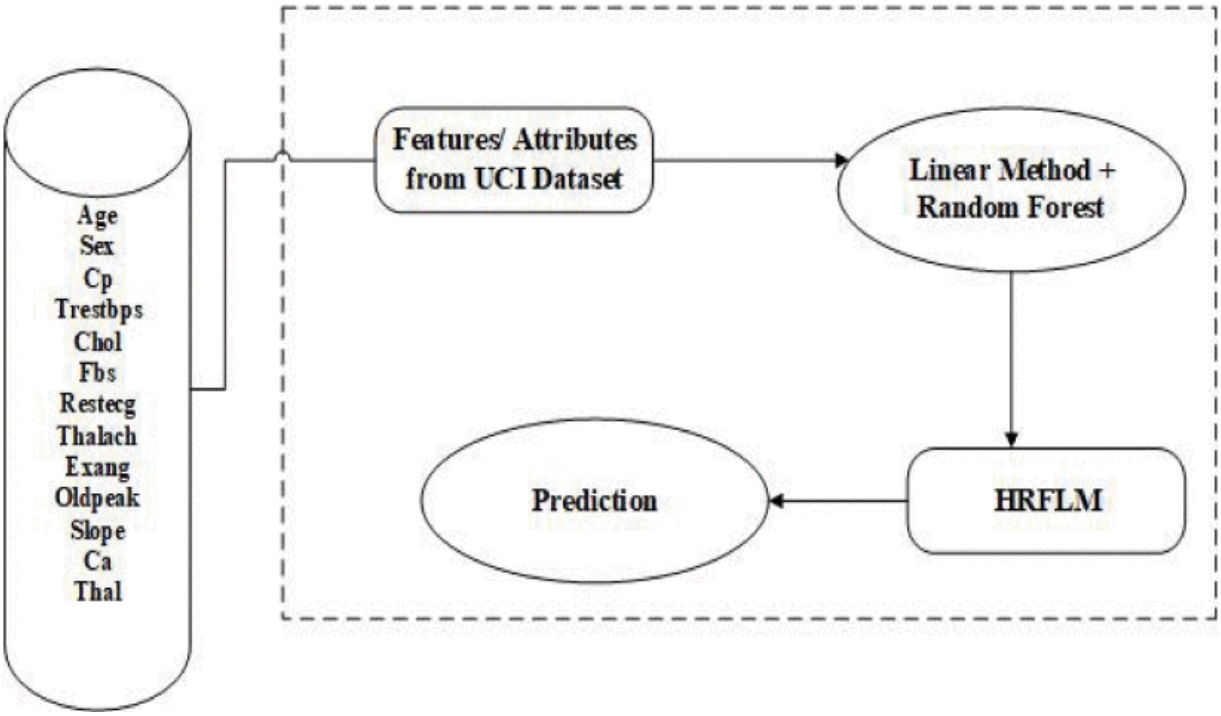


src : Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques paper by SenthilKumar

Figure 2.2.1: Experiment Workflow with UCI dataset

The proposed method effectively reduced the set of critical attributes. The remaining attributes are input for ANN subsequently. The heart disease prediction with multilayer perception of NN is proposed.

In this study, used an R studio rattle to perform heart disease classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The performance of each model generated based on 13 features and ML techniques used for each iteration and performance are recorded.



src : Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques paper by SenthilKumar

Figure 2.2.2: Prediction of Heart Disease using HRFLM

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records.Clinical records are vital to diagnosis and learning the severity of heart disease. In this experiment, several (ML) techniques are used namely, NB, GLM, LR,DL, DT, RF, GBT and SVM.

Table 2.2:Performance Metrics

Parameter	Performance Metrics
Accuracy	89.01
Precision	90.1
Recall	92.8
F-measure	0.90

2.3.Cardiovascular disease incidence prediction by machine learning and statistical techniques:

2.3.1 Introduction

Kamran Mehrabani et al.[11] The study adopts the most popular ML algorithms used in CVD prediction studies, including k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Artificial NeuralNetwork (ANN), and Gradient Boosting Machine (GBM)to develop suitable and efficient prediction models for predicting the future occurrence of CVD events based on the comprehensive set of risk factors in the framework of the long-term Isfahan Cohort Study (ICS), a population based cohort in the eastern Mediterranean region,Iran. This study also aimed to identify the most efficient predictors of future CVD incidence in participants who were healthy at the entrance to the ICS in order to find a high risk group for early CVD events. This study also attempted to compare the predictive abilities of the machine learning modeling approach with traditional statistical methods.This study revealed that age, systolic blood pressure, fasting blood sugar, two hour postprandial glucose, diabetes mellitus, history of heart disease, history of high blood pressure, and history of diabetes are the most contributing factors for predicting CVD incidence in the future. The main differences between the results of classification algorithms are due to the trade-off between sensitivity and specificity.

2.3.2 Merits and Demerits

Merits:

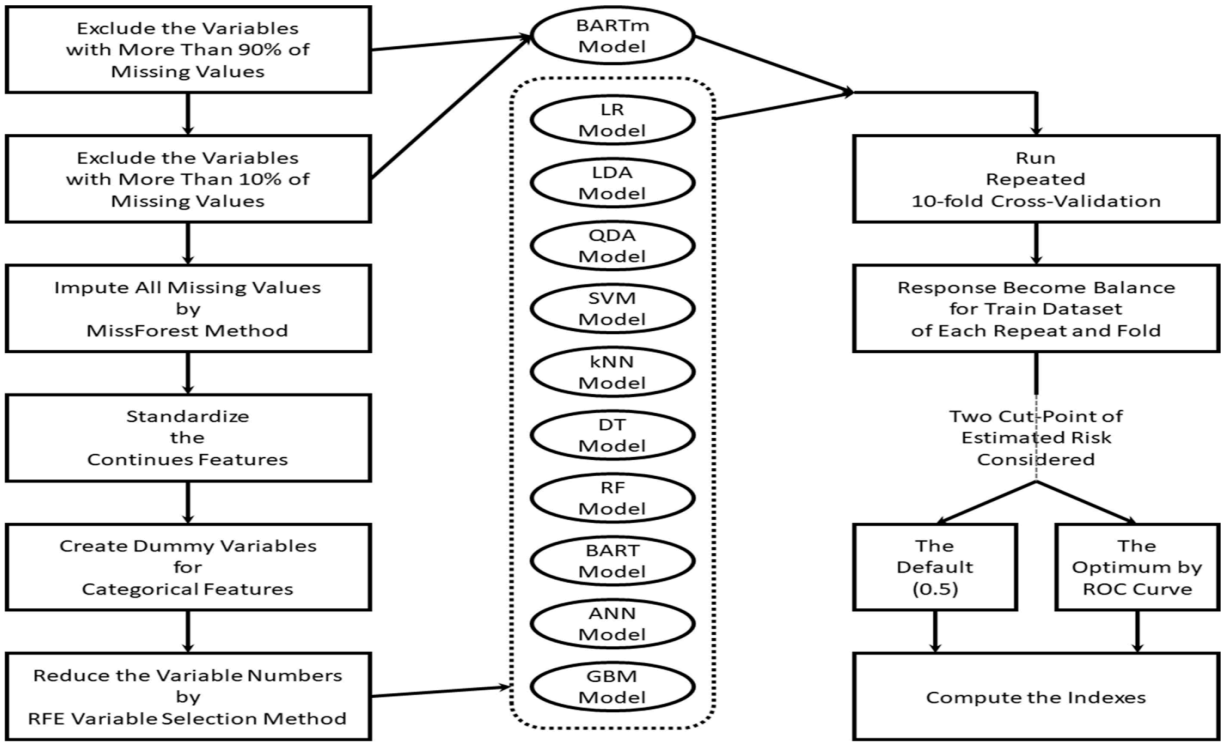
- Robust Handling of Missing Data: The utilization of techniques like "missingness incorporated in attributes" and MissForest enables effective handling of missing values, allowing for a more comprehensive analysis of the dataset and improved model performance.
- Feature Selection for Enhanced Insights: The Recursive Feature Elimination (RFE) method aids in selecting the most influential variables, thereby improving the model's interpretability and potentially enhancing the overall predictive performance.

Demerits:

- Complex Data Preprocessing: The process of managing missing values and employing imputation techniques, along with recursive feature elimination, may result in complex data preprocessing steps that require a comprehensive understanding and careful execution.
- Increased Computational Demands: Utilizing advanced techniques like BARTm and RFE may require significant computational resources, leading to higher processing times and potentially increased memory usage.
- Sensitivity to Hyperparameters: The performance of techniques like BARTm and MissForest can be sensitive to the selection of hyperparameters, and improper tuning may lead to suboptimal results and potentially biased analyses.

2.3.3 Implementation

Several ML algorithms have been utilized for CVD incidence prediction but there is no unique model with the highest predictive ability in all situations. A meta-analysis on 344 studies showed that the SVM and GBM have the highest predictive ability. A review article in 2022 indicated that RF and ANN have the best predictive performance .So, in this study, the various supervised classical statistical and machine learning classification models were used by considering their predictive power and popularity, including Logistic Regression (LR) , Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), SVM, kNN, DT, RF, Bayesian Adaptive Regression Trees (BART), missing incorporated to attributes within BART (BARTm), ANN and GBM. All models run according to the same procedure except BARTm. The BARTm model has a combined statistical and ML algorithm that makes it capable of accurately classifying data even with 90% of missing values, without any imputation.So, the BARTm model was implemented on the dataset with two missing value scenarios: (I) all variables with up to 90% of missing values were considered (515 variables); (II) only those variables with up to 10% missing values were considered (385 variables).These two model verifications of the BARTm model were denoted by BARTm.90% and BARTm.10%, respectively.The grid search cross-validation techniques were applied to tune the



src : Cardiovascular disease incidence prediction by machine learning and statistical techniques paper by Kamran Mehrabani

Figure 2.3: The flow of data analysis process

hyper-parameters of ML algorithms that determine the optimal values to achieve higher accuracy. The response variable in the current study was considered as any diagnosis of CVD events until 2017, which includes: fatal and non-fatal myocardial infarction, fatal and non-fatal stroke, sudden cardiac death, and unstable angina. Among all 5432 participants, CVD events occurred for 819 participants(15.08%) in the follow-up period; Hence, the response variable is imbalanced relevant techniques and evaluation metrics should be used during modeling .

Table 2.3:Performance Metrics

Parameter	Performance Metrics
Accuracy	75.5
Precision	73.37
Recall	85.45
F-measure	69.84

2.4 Neural Network based Heart Disease Prediction

2.4.1 Introduction

J. Jasmine et al .[9] Predictive analytics include various statistical techniques from predictive modeling, machine learning (ML)and data mining to make predictions based on the current or historical data. The use of predictive analytics are in the customer relationship management, healthcare industry and many other fields. Deep learning has a significant impact on predictive analytics. There are many models in predictive modeling [10] such as Naive Bayes, Logisticregression, Neural networks , Support VectorMachine , Classification and Regression trees etc.Artificial neural network (ANN) is one of the mathematical algorithmic approaches. The artificial neural network has connections,propagation direction and discrete layer.Each layer is made up of nodes with the arrows that represent the interconnections between them.In the neural network, there are many nodes in the input layer. These input layer nodes are connected to the hidden layer nodes. Each input is assigned with the weights. The input nodes in the network pass the data to the nodes in the hidden layer which performs some tasks or computations and send the processed data to the output node. The output layer has the node which yields the final result. This is an overview of the process of neural networks.

2.4.2 Merits and Demerits

Merits:

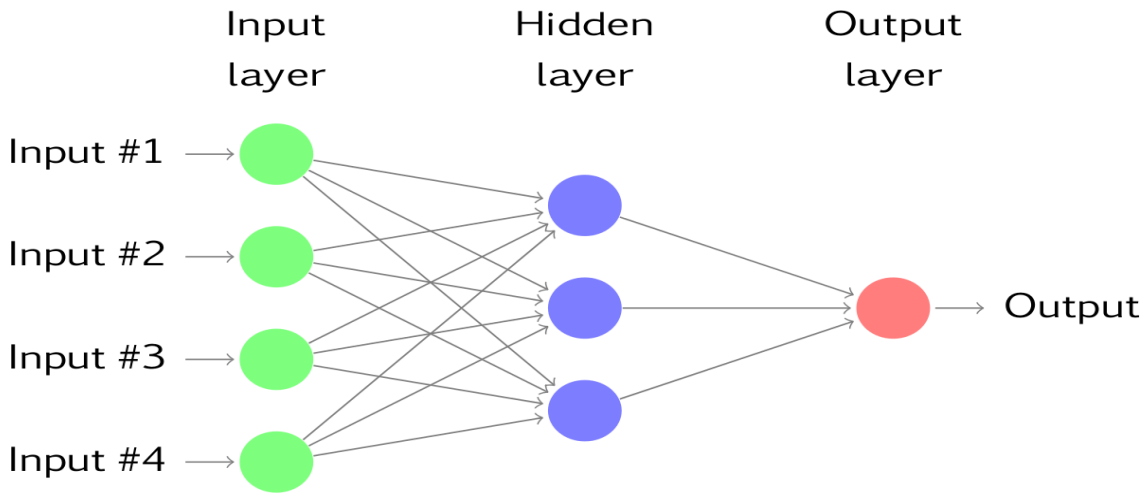
- Effective Data Processing: The neural network's ability to process complex data sets and perform intricate computations makes it well-suited for various predictive analytics tasks, ensuring comprehensive analysis and accurate predictions.
- Incorporation of Complex Relationships: The interconnected nodes and layers of the neural network enable the model to capture intricate relationships and patterns within the data, allowing for the identification of nuanced insights and predictive capabilities.

Demerits:

- Complex Model Interpretability: Neural networks often present challenges in terms of interpretability, as the complex interconnections and intricate computations within the network can make it difficult to understand the specific factors influencing the model's predictions.
- Requirement of Large Datasets: Training effective neural network models typically necessitates large and diverse datasets, which may pose challenges in data collection and preprocessing, especially in domains with limited data availability.
- Computational Complexity: The implementation and training of neural networks can be computationally intensive, requiring substantial computational resources and potentially leading to longer processing times and increased energy consumption.

2.4.3 Implementation

The ultimate goal is to combine the logistic regression model and neural network based approach in the prediction of heart disease. The heart disease dataset has 303 observations of individuals out of which 297 observations are taken for consideration. The proposed system mainly consists of two parts. The first part is to find the important risk factors in predicting the heart disease from the available risk factors in the dataset based on the p-value. This p-value yields the significant codes for each attribute. And the second part is to divide the dataset into training and testing dataset. The Neural network is built for the training dataset and the learned neural network is able to predict the testing dataset.Logistic regression model is one of the statistical regression models and it has the capacity to measure the relationship between the categorical dependent variable and one or more independent variables. Here the independent variables are age, sex, chest pain type, resting blood pressure,serum cholesterol, fasting blood sugar, resting electrographic results, maximum heart rate, exercise induced angina,old peak-slope of the peak, slope of the peak exercise, blood vessels affected, thal defect. The dependent variable is the class which is to be predicted as healthy or having heart disease.



src : Neural Network based Heart Disease Prediction research paper by J.Jasmine

Figure 2.4: Sample Artificial Neural Networks

The neural network is a computational model based on biological neural networks. Artificial neural networks(ANN) is based on observation of a human brain. Humanbrain is a very complicated web of neurons . AnalogicallyANN is an interconnected set of three units such as input,hidden and output units.The effectiveness of artificial neural networks was proven in medicine. ANN areused in predicting coronary heart disease. Here the input layer consisting of 8 neurons corresponds to 8 significant attributes. There is one output class variable which takes the value either 0 or 1. The value 0 represents that the individual is not suffering from heart disease and the value 1 represents that the individual suffers from heart disease. The number of nodes used in the hidden layer are 3.

Table 2.4:Performance Metrics

Parameter	Performance Metrics
Accuracy	84.0
Precision	77.5
Recall	91.5
F-measure	83.9

CHAPTER 3

PROPOSED SYSTEM

CHAPTER 3

PROPOSED SYSTEM

For our predictive modeling approach, we have utilized XGBoost, a powerful machine learning algorithm known for its effectiveness in handling complex datasets and achieving high predictive accuracy. To optimize the performance of the XGBoost model, we employed hyperparameter tuning using a technique called Cross-Validation Grid Search. This approach systematically explores a range of hyperparameter values and selects the combination that maximizes the model's performance metrics.

After identifying the optimal hyperparameters through Cross-Validation Grid Search, we trained the XGBoost model on our dataset. This training process involved feeding the model with labeled data containing various features relevant to cardiovascular disease risk prediction. The model learned patterns and relationships within the data to make accurate predictions about individuals' risk of developing CVDs.

Once the XGBoost model was trained and validated, we proceeded to deploy it in a web application designed for CVD risk prediction. The web application provides an intuitive interface where users can input their health information and receive personalized risk assessments for cardiovascular diseases. Leveraging the trained XGBoost model, the web application generates predictions in real-time, allowing users to access valuable insights about their cardiovascular health status.

The deployment of the XGBoost model within the web application enables easy accessibility and usability for both healthcare professionals and individual users. By integrating advanced machine learning techniques with user-friendly web interfaces, our proposed system facilitates early identification of individuals at risk of developing CVDs and empowers users to make informed decisions about their health.Overall, the proposed system leverages the predictive capabilities of XGBoost combined with hyperparameter tuning to build a robust model for CVD risk prediction. The integration of this model into a web application enhances accessibility and usability, ultimately contributing to improved health outcomes and reduced morbidity associated with cardiovascular diseases.

3.1 Materials and Methods

In machine learning, data is paramount for accuracy. This collected dataset contains 19 variables of which 12 are arithmetical and 7 are categorical. The number of instances is 308854 and the dataset does not contain missing values.

Table 3.1: CVD Dataset

Serial Number	Attribute	Description
1	General Health	Well-being, fitness
2	Check-up	Examination to ensure health or wellness
3	Exercise	Activity for fitness and health.
4	Heart Disease	Cardiac condition
5	Skin_cancer	Describe different parts of your skin or conditions that can affect it.
6	Other_Cancer	Those who indicated they have experienced any other forms of cancer
7	Depression	Feeling sad, hopeless, or down for a long time.
8	Diabetics	Having too much sugar in your blood for a long time.
9	Arthritis	Pain and swelling in your joints makes it hard to move.
10	Gender	0-Female,1-Male
11	Age	In days

12	Height	In Cent Meter
13	Weight	Kilograms
14	BMI	It is a number that shows if a person is a healthy weight for their height.
15	Smoking History	Whether Patient Smokes or Not
16	Alcohol consumption	Whether patient smokes or not
17	Fruit consumption	Recording patients' fruit intake
18	Green Vegetables consumption	Recording patients' green vegetable intake
19	Friedpotato vegetable consumption	Recording patients' potato intake

3.2 Algorithms Used for Proposed System

3.2.1 Solo Machine Learning Algorithms

Algorithms for machine learning allow computers to recognize patterns and connections in data without the need for explicit programming. Based on input data, these algorithms employ statistical approaches to find patterns and provide predictions or choices. Three main categories can be used to group them[12]

Supervised Learning: In this method, an algorithm is trained using a labeled dataset in which every input has a matching output. To forecast or categorize newly discovered data, the algorithm gains knowledge from this labeled dataset.

Unsupervised Learning: When there are no labels on the input data, unsupervised learning algorithms are applied. Without any indication, the algorithm looks for facts in the statistics. This learning comprises approaches, clustering algorithms.

Reinforcement learning: Teaching an agent through reinforcement learning involves guiding it to interact with its environment to maximize rewards. The agent learns through trial and error receiving feedback in the form of rewards or penalties based on its performance.

Logistic Regression

In the domain of learning algorithms logistic regression stands out as a choice in machine learning. It is utilized for predicting outcomes based on a set of variables. The core concept behind regression is fitting a S shaped function instead of a straight line to predict binary outcomes that indicate the likelihood of an event occurring.The sigmoid function is referred to as an activation function for logistic regression and is defined as:

$$f(x) = 1 / (1+e^{-x})$$

where,

e = base of natural logarithms

value = numerical value one wishes to transform

Ridge Classifier

To combat overfitting issues in machine learning models methods like Ridge Regression and Ridge Classifier are employed. Overfitting occurs when a model performs well on training data but poorly on data.In Ridge Regression, an additional term known as L2 regularization is included in the linear regression equation to avoid overfitting. This term penalizes coefficients helping to manage the complexity of the model. Likewise a Ridge Classifier employs L2 regularization to prevent overfitting, in tasks involving class classification.

Support Vector Machine

SVM can be used as a regression tool, though classification is its main area.In reality, these support vectors determine where the line goes and in what direction it would slant; thus giving birth to SVM as its name suggests.The algorithm exhibits great efficiency in classifying data.

Let the training samples have a dataset Data=$\{y_i, x_i\}$; i=1,2,…,n where $x_i \in Rn$ represent the ith vector and yi$\in$Rn represent the target item. The linear SVM finds the optimal hyperplane of the form $f(x)=w^Tx+b$ where w is a dimensional coefficient vector and b is an offset.

$$Min_{w,b,\xi i} \ \frac{1}{2} \ w^2 + C\sum_{i=1}^{n} \xi_i$$

$$s.\ t.\ y_i\ (w^T x_i + b) \geq 1-\xi_i,\ \xi_i \geq 0, \ \forall_i \in \{1,2,...,m\}$$

K-Nearest Neighbour

Regression problems and classification tasks are addressed by utilizing the K-Nearest Neighbors (KNN) algorithm within machine learning. KNN is an easy to use and adaptable technique making it applicable in different areas like pattern recognition, data mining and intrusion detection.This means that it does not presume anything about how data is distributed.This flexibility makes it handy for real-life situations where data can be messy or irregular. To use KNN, you start with some known data (called training data), which has points already labeled. Then, when you get a new point, KNN compares it to the known points and makes predictions based on their proximity.It extracts the knowledge based on the samples Euclidean distance function d(xi,xj) and the majority of k-nearest neighbors.

$$d(x_{i,}\ x_{i,}) = (x_{i,1}-x_{j,1}) + \cdots + (x_{i,m} - x_{j,m})^2$$

Decision Tree

Although it is frequently used for classification, a DT is a useful tool in supervised learning; it would handle both regression & classification tasks. It resembles a tree-shaped flowchart in which decisions are made at each stage depending on a particular data attribute.The selection of features to employ and the timing of decisions are made easier by this algorithm.

As a result, a decision tree is a visual tool for determining possible results or solutions for an issue by making decisions based on variables in the data. Because it begins with a root node and grows into stems to form a structure corresponding to a tree, it is known as a DT.

$$Entropy = -\sum_{j=1}^{m} P_{ij}\ log_2 P_{ij}$$

Naive Bayes

One sort of supervised learning utilized for classification tasks—particularly for text classification with huge datasets—is the Naïve Bayes functionality.

Probabilistic Classifier: Naïve Bayes makes predictions based on how likely it is that an object will fall into a specific class. For example, it might determine whether an email is spam or not by looking at the likelihood of specific terms showing up in spam emails. "Naïve" The assumption: Because it presumes that features are independent of one another, it is referred to as "naïve".This learning model applies Bayes rules through independent features. Every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability

$$P(X_f) = priority \in (0:1)$$

$$P(X_{f1}, X_{f2}, \ldots, X_{fn} \mid c)$$

$$= \prod_{i=1}^{n} P(X_{fi} \mid c)$$

$$P(X_f \mid c_i)$$

$$= P(c_i \mid X_f)\, P(X_f)\, P(c_i) \quad c \in \{benign , malignant\}$$

3.2.2 Ensemble Learning Techniques

Ensemble learning is a powerful technique in machine learning where many models are joined to enhance predictive accuracy and robustness. The underlying principle is that while individual models may have limitations or biases, combining their predictions can mitigate these weaknesses and yield better overall performance.

Gradient Boosting

Gradient Boosting is a potent machine learning technique that combines weak models, typically decision trees, to create a strong predictive model. It works by minimizing errors iteratively, using gradient descent to guide the training of each new model. Unlike AdaBoost, it uses residuals of the previous model as labels for training the next one. Gradient Boosted Trees is a popular variant where each weak learner is a decision tree. Gradient Boosting's iterative approach allows it to incrementally improve model performance, making it particularly effective for complex datasets.

LightGBM

LightGBM introduces two key innovations, GOSS and EFB, which greatly enhance efficiency and reduce memory usage compared to traditional gradient boosting frameworks. GOSS prioritizes instances with significant gradient contributions while subsampling others, accelerating training without sacrificing model performance. EFB packs exclusive features together, reducing dimensionality and conserving memory. These techniques distinguish LightGBM from traditional GBDT frameworks, leading to superior machine learning performance by enabling efficient and effective gradient boosting beyond histogram-based algorithms.

XGBoost

XGBoost, short for Extreme Gradient Boosting, is a leading optimized distributed GBoosting library renowned for its efficiency and scalability in training ML models. It combines knowledge from weak models to create robust predictions, excelling with massive datasets across diverse ML tasks. Notably, XGBoost efficiently handles missing values, reducing preprocessing complexities and speeding up the modeling pipeline. With built-in support for parallel processing, it accelerates model training on large-scale datasets. Overall, XGBoost's boosting techniques, regularization methods, efficient split finding algorithms, and handling of missing values make it a versatile and powerful tool widely trusted by data scientists and ML practitioners.

Random Forest

Random Forest is highly scalable and effective for handling large, high-dimensional datasets with complex relationships between features.Despite its computational complexity, Random Forest offers high predictive accuracy and is widely used across various domains due to its ease of implementation and excellent performance.This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, $X=\{x_1,x_2,x_3,\ldots,x_n\}$ with responses $Y=\{x_1,x_2,x_3,\ldots,x_n\}$ which repeats the bagging from b=1 to B . The unseen samples x' is made by averaging the predictions $\sum Bb=1fb(x')$ from every individual trees on x' :

$$j = 1/B \sum_{b=1}^{B} fb(x')$$

The uncertainty of prediction on these tree is made through its standard deviation,

$$\sigma = \sum_{b=1}^{B} (fb(x') - f)^2 / (B-1)$$

3.2.3 Ensemble Methods

After our first model did not make great predictions we applied various ensemble methods on it. Level 2 Models are created by combining Level 1 Model's predictions through weighted averages or simple pooling techniques. These techniques were employed to boost the overall predictive power of the model.

Bagging

Firstly, bagging involves the creation of numerous subsets of the original dataset through a process known as bootstrap sampling. This entails randomly selecting instances from the dataset with replacements, thereby generating subsets of the same shape as the original result set. Secondly, a base model, typically a decision tree although other models can also be used, is trained on each of these bootstrap samples. As a result of the variations in the training data introduced by bootstrap sampling, each model trained on a different subset will inherently be slightly different from the others. Finally, once all the models have been trained, predictions are made for unseen data using each model.

Stacking

Other words used interchangeably with stacking include stacked generalizations or stacked.A good example would be ensemble as another effective method of group instruction. Instead of just averaging the predictions of several systems, stacking involves training a meta-model or meta-learner to figure out how to optimally combine the predictions of the base models. By using the underlying models' predictions as input features, this metamodel learns to produce predictions based on these inputs. In essence, it figures out how best to combine or weigh the underlying models' predictions to get the final one. Stacking is an ensemble technique that is more advanced than majority voting or simple averaging, which are employed in bagging methods.

Boosting

Boosting is also well-known as an ensemble learning method in machine learning, which emphasizes training several weak learners in turn giving a powerful ensemble system. In contrast to independent model training, boosting trains models iteratively by having each new model place greater emphasis on cases that the preceding model misclassified. Essentially, boosting is the process of combining several weak models—usually shallow decision trees, or "weak learners"—to produce a strong and precise predictive model.

Tuning

In machine learning parlance, "tuning" is modifying a few variables, or "hyperparameters," to maximize a machine learning model's performance. In contrast to a model's parameters, which are determined during training, hyperparameters are predetermined and have an impact on the learning process. Hyperparameter tuning is essential for getting the most out of a machine-learning model, regardless of the method employed.It often involves a trade-off between computational resources, such as time and hardware, and the quality of the resulting model. Effective hyperparameter tuning can lead to improved model accuracy, generalization, and robustness across different datasets and applications.

Voting

Voting is a technique commonly used in ensemble learning, where multiple machine learning models are combined to make predictions. It operates on the principle of aggregating the individual predictions of multiple models to produce a final prediction that is more robust and accurate than any single model.

Two main types of voting in ensemble learning:

Hard Voting: Each model in the ensemble makes its own prediction, and the final prediction is determined by a majority vote. It's commonly used for classification tasks with discrete class labels.

Soft Voting: Models' predicted probabilities for each class are averaged, and the class with the highest average probability is chosen as the final prediction. This method is suitable for classification tasks where the output is probability scores rather than discrete class labels.

3.3 Stepwise Implementation and Code :

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
df = pd.read_csv(r'C:\Users\shara\OneDrive\Desktop\project\CVD.csv')
df.head()

from sklearn.preprocessing import LabelEncoder
#  'df' is a DataFrame with categorical variables
categorical_columns = df.select_dtypes(include=['object']).columns
label_encoder = LabelEncoder()
# Iterate through each categorical column and transform the values
for column in categorical_columns:
df[column] = label_encoder.fit_transform(df[column])
df.head()

import pandas as pd
from sklearn.preprocessing import MinMaxScaler
original_df = pd.DataFrame(df)
# Create a copy to keep the original data
original_df_copy = original_df.copy()
# Define min-max scaler with range 0 to
scaler = MinMaxScaler(feature_range=(0, 1))
# Transform the data
df_scaled = scaler.fit_transform(original_df_copy)
# Convert the scaled data back to a DataFrame
df_scaled = pd.DataFrame(df_scaled, columns=original_df_copy.columns)
# Replace the original DataFrame 'df' with the scaled data
df = df_scaled.copy()
```

```
# Now, 'df' contains the scaled data
df.head()

# Counting NaN values in all columns
nan_count = df.isna().sum()
print(nan_count)
# Create a bar plot for the number and percentage of fraudulent vs non-fraudulent transcations
plt.bar(['No','Yes'], df['Heart_Disease'].value_counts(), color=['r','b'])
plt.xlabel('Heart_Disease')
plt.ylabel('Number of samples')
plt.annotate('{}\n({:.4}%)'.format(df['Heart_Disease'].value_counts()[0], df['Heart_Disease'].
value_counts()[0]/df['Heart_Disease'].count()*100),(0.20, 0.45), xycoords='axes fraction')
plt.annotate('{}\n({:.4}%)'.format(df['Heart_Disease'].value_counts()[1], df['Heart_Disease'].
value_counts()[1]/df['Heart_Disease'].count()*100), (0.70, 0.45), xycoords='axes fraction')
plt.tight_layout()
plt.savefig("output-0.jpg")
plt.show()

y = df['Heart_Disease']
X = df.drop('Heart_Disease', axis=1, inplace=False)

#Applying Imbalanced technique SMOTE
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
X_train,X_test,y_train,y_test = train_test_split(X, y, test_size=0.1,stratify=y, random_state=0)
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)
from sklearn.linear_model import LogisticRegression

# Creating and training the logistic regression model
logistic_regression_model = LogisticRegression()
logistic_regression_model.fit(X_resampled, y_resampled)
```

```
# Making predictions
y_pred = logistic_regression_model.predict(X_test)
# Evaluating the model
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
print(classification_report(y_test, y_pred))
print("\nAccuracy Score:", accuracy_score(y_test, y_pred))

# Creating and training the Naive Bayes model
naive_bayes_model = GaussianNB()
naive_bayes_model.fit(X_resampled, y_resampled)
# Making predictions
y_pred = naive_bayes_model.predict(X_test)
print(classification_report(y_test, y_pred))
print("\nAccuracy Score:", accuracy_score(y_test, y_pred))
# Make predictions on the test set using Lightgbm Algorithm
y_pred_lgbm = lgbm_model.predict(X_test)
accuracy_lgbm = accuracy_score(y_test, y_pred_lgbm)
classification_rep_lgbm = classification_report(y_test, y_pred_lgbm)
print("Accuracy:", accuracy_lgbm)
print("Classification Report:\n", classification_rep_lgbm)
rf_model = RandomForestClassifier(random_state=0)
rf_model.fit(X_resampled, y_resampled)
# Make predictions on the test set
y_pred_rf = rf_model.predict(X_test)
# Evaluate the model
accuracy_rf = accuracy_score(y_test, y_pred_rf)
classification_rep_rf = classification_report(y_test, y_pred_rf)
# Print the results for RandomForestClassifier
print("Accuracy:", accuracy_rf)
print("Classification Report:\n", classification_rep_rf)
```

```
#using XGBoost
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, classification_report
xgb = XGBClassifier(random_state=42)
xgb.fit(X_resampled, y_resampled)
# Make predictions on the test set
y_pred_xgb = xgb.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred_xgb)
classification_rep = classification_report(y_test, y_pred_xgb)
print("Accuracy:", accuracy)
print("Classification Report:\n", classification_rep)

#Hyperparameter tuning
# Initialize XGBoost model
xgb_model = XGBClassifier()
# Define the hyperparameters grid
param_grid = {
    'max_depth': [7],
    'learning_rate': [0.1],
    'n_estimators': [100],
    'subsample': [1.0],
    'colsample_bytree': [0.8]
}
# Initialize GridSearchCV
grid_search = GridSearchCV(xgb_model, param_grid, scoring='accuracy', cv=5, n_jobs=-1)
# Fit the model with resampled data
grid_search.fit(X_resampled, y_resampled)
# Get the best parameters
best_params = grid_search.best_params_
```

```
# Train the model with the best parameters on the original training set
best_model = XGBClassifier(**best_params)
best_model.fit(X_train, y_train)
# Make predictions on the test set
y_predxgb = best_model.predict(X_test)
# Evaluate the model
accuracyxgb = accuracy_score(y_test, y_predxgb)
classification_repxgb = classification_report(y_test, y_predxgb)
print("Best Hyperparameters:", best_params)
print("Accuracy:", accuracyxgb)
print("Classification Report:\n", classification_repxgb)
#importing and loading final model
import pickle
filename='trained_model2.sav'
pickle.dump(best_model,open(filename,'wb'))
```

3.4 Designing

UML Diagram

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML comprises two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with;UML .

A) Use Case Diagram :

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

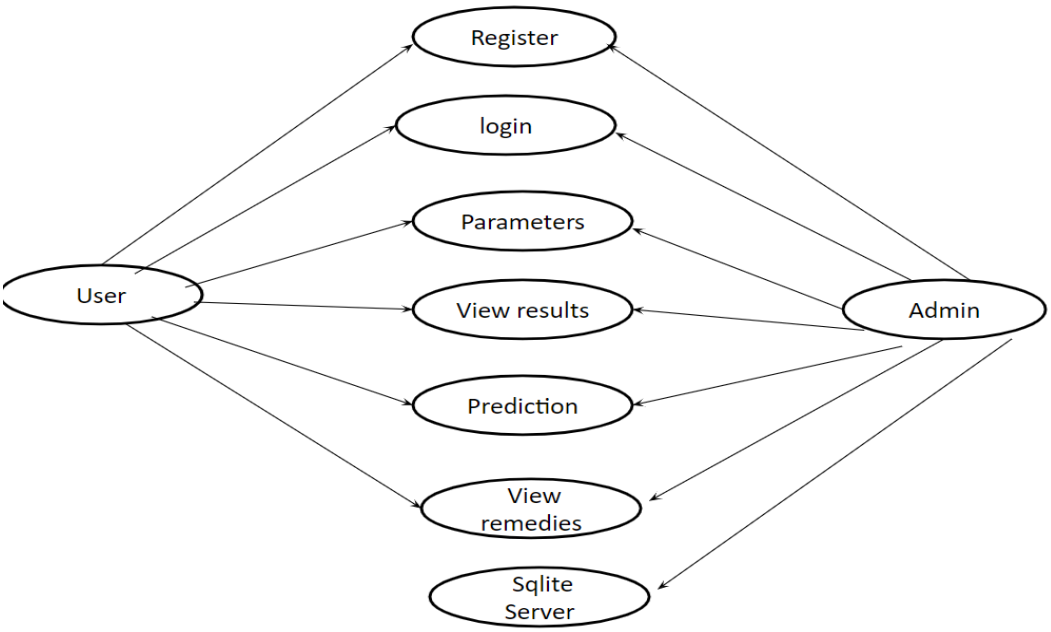


Figure 3.4.1: UseCase Diagram

B) Activity Diagram:

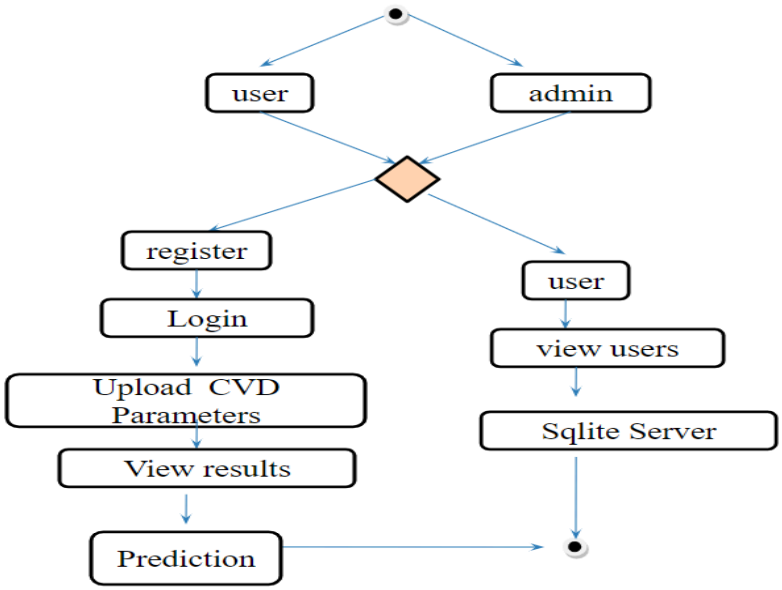


Figure 3.4.2: Activity Diagram

C) Class Diagram

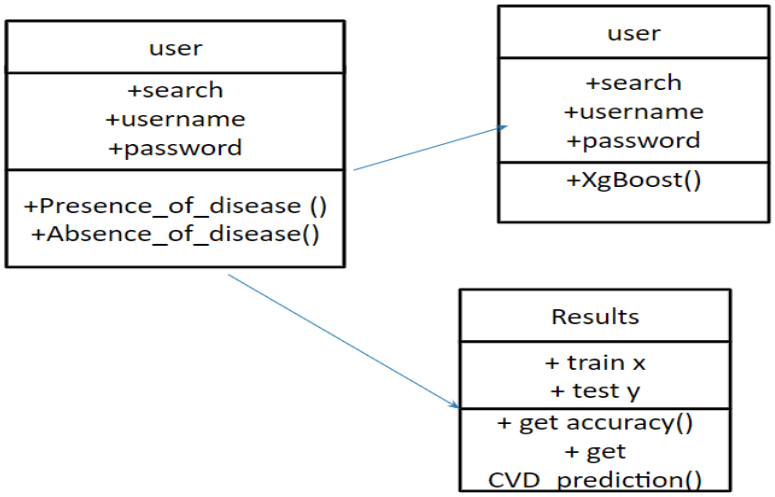


Figure 3.4.3: Class Diagram

CHAPTER 4

RESULTS AND DISCUSSIONS

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data Preprocessing

Dataset Contains 12 Categorical values so we owned a Label Encoder to transfigure Categorical to Numerical.

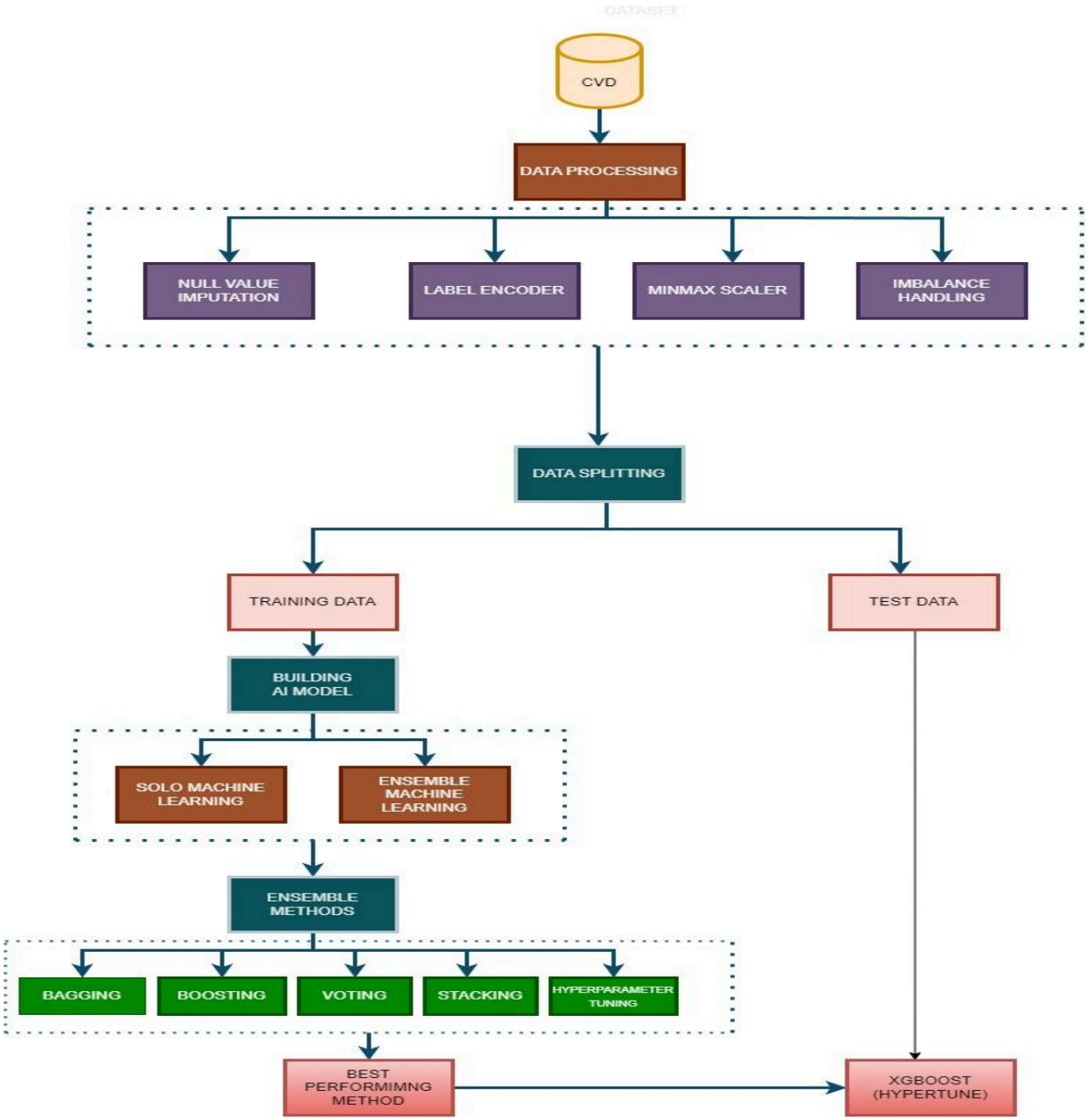


Fig 4.1: Architecture Diagram

4.1.1 Label Encoder

In the realm of machine learning, the label encoder plays a pivotal role in transforming categorical or textual data into a format that can be readily processed by various algorithms. Its primary function is to convert qualitative information into a quantitative representation, which is essential for many machine learning models designed to operate on numerical inputs. When initializing a label encoder, an instance is created, and during the training phase, the encoder is 'fitted' with the categorical labels from the dataset.

4.1.2 Min Max Scaler

MIN MAX Scaler is widely used to proportionate the nominal value from 0 to 1. To scale numerical characteristics inside a given range, usually between 0 and 1, machine learning uses the preprocessing approach known as min-max scaling. Our objective is to standardize the data and assign a uniform scale to every characteristic so that variations in their magnitudes do not lead to one feature predominating over others. This normalization might be especially crucial for algorithms, such gradient based optimization methods, that depend on gradients or distance measurements.

4.1.3 Imbalance Data Handling

In handling imbalanced data, we experimented with two techniques, SMOTE and ADASYN.

SMOTE

Machine learning uses the SMOTE data augmentation technique to solve the issue of class imbalance in classification jobs. When one class in the target variable has much fewer instances than another, it is said to be imbalanced. As a result, the minority lesson will see inefficient performance from one-sided models .

ADASYN

ADASYN (Adaptive Synthetic Sampling) was created to solve some of its shortcomings. Similar to SMOTE, ADASYN is pre owned to manage class imbalance in datasets used for ML, especially in cases of classifying the input data into two mutually exclusive categories where one class is greatly underrepresented. We tried both approaches and discovered that the best accuracy was obtained when combining SMOTE with the XGBoost algorithm.

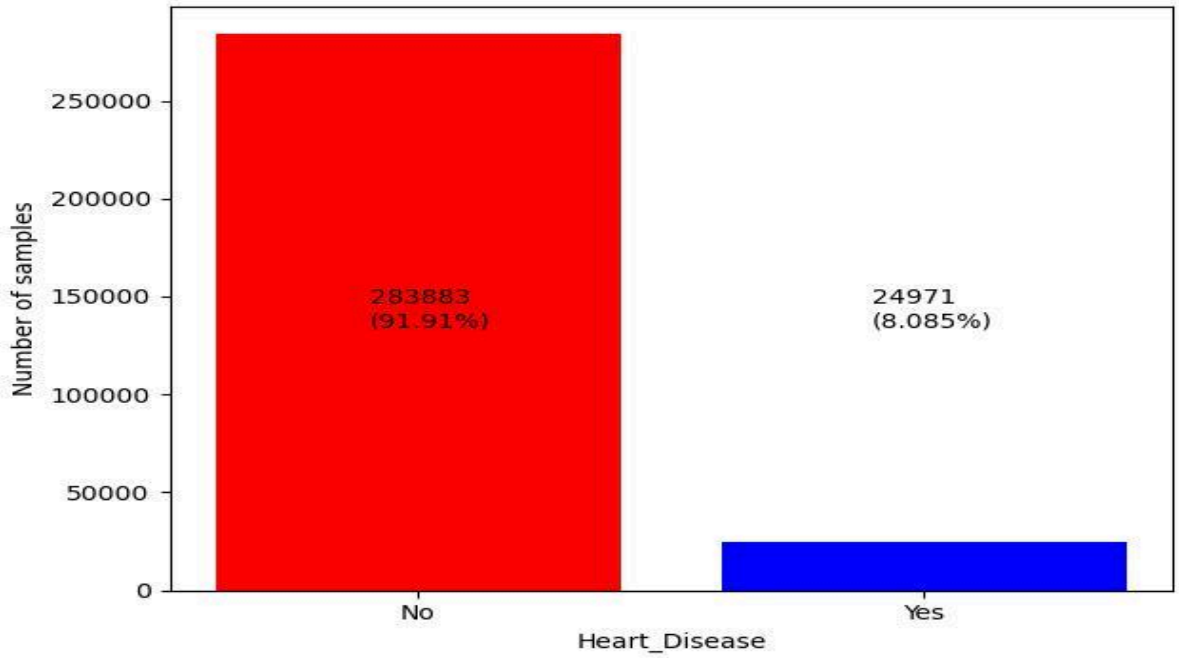


Figure 4.1.1:TargetVariable distribution before using SMOTE

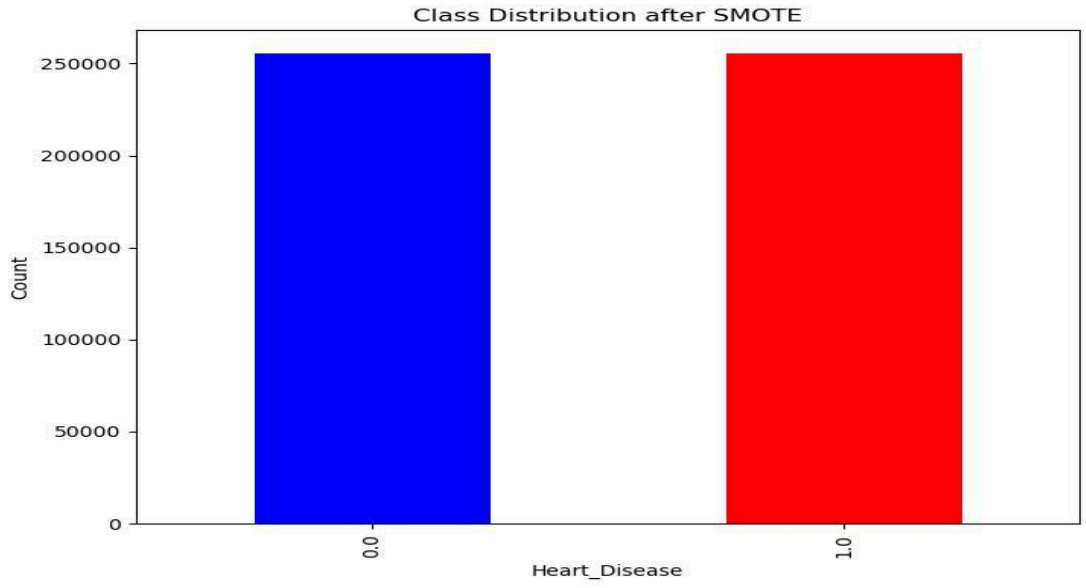


Figure 4.1.2 : Target Variable Distribution after using SMOTE

4.2 Data splitting

We used the K fold Cross Validation technique and tested our set of results with 10 folds,4 folds, 2folds, 5 folds and took the model with the validation that gives more performance metrics and k=10 with Extreme Gradient Boosting gives superlative rightness.

Table 4.2 : Performance Metrics of different algorithms(k=10 folds)

Model	Accuracy	AUC	Recall	Precision	F1 Score
xgboost	0.917	0.8288	0.9013	0.8858	0.883
lightgbm	0.9009	0.8182	0.836	0.8884	0.8863
random forest	0.8832	0.8064	0.9171	0.8805	0.8865
ada	0.8502	0.8065	0.9	0.8801	0.8835
et	0.8256	0.7929	0.854	0.8781	0.8883
gbc	0.8256	0.8304	0.8456	0.8894	0.8883
dt	0.8613	0.5739	0.8613	0.8719	0.897
svm	0.7508	0	0.7592	0.9074	0.8804
lr	0.7476	0.8358	0.7508	0.9125	0.7951
ridge	0.7413	0	0.7476	0.9127	0.803
lda	0.7412	0.8348	0.7413	0.9127	0.7984
knn	0.6839	0.5606	0.6839	0.862	0.7532
nb	0.6091	0.801	0.6091	0.911	0.696
qda	0.5701	0.787	0.5701	0.9116	0.6615

4.3 Model selection

After separating our result set into training and testing sets and preprocessing it, choosing XGBoost as our model. The next stage is to train the model on the training set and assess its effectiveness. Since you suggested utilizing K Fold Cross Validation by having K=10, we used cross validation to obtain a reliable performance estimate for our model.To determine which machine learning algorithm would work best for our dataset, we investigated a variety of models in this study. LightGBM, XGBoost, GB Classifier, Random Forest (RF), Extra Trees, AdaBoost (ADA), Decision Tree (DT), SVM, Logistic Regression (LR), Ridge Classifier (Ridge), Linear Discriminant Analysis (LDA), K-nearest neighbors (KNN), Naive Bayes (NB), and Quadratic Discriminant Analysis (QDA) are among the models we have developed. Among them, XGBoost appeared to be the most appropriate choice for our purposes based on such factors as accuracy, robustness, and interpretability. It has been shown to possess some of the best characteristics regarding accuracy and reliability.

Therefore, we will now refine this model by modifying hyperparameters examining feature importance and preparing it for production. Henceforth it is crucial that we validate carefully comprehending these results in order to ascertain if the XGboost that we selected aligns with our customized machine learning task. Because of this, in addition to hyperparameter tuning, Bagging, Boosting, Stacking and Voting will increase the performance of our chosen model XGBoost.We intend to further enhance the system's robustness as well as its predictability using bagging , boosting , stacking , voting which build on top of ensemble approach.Bagging will involve training top 5 models on different subsets from dataset; thus reducing overfitting while improving stability.Boosting concatenates a sequential training process, enabling the model to compare and progressively correct its errors, holding intricate relationships in the data. Stacking embraces diverse model predictions as inputs into a new meta-learner, refining a final output. Voting involves consolidating the predictions of multiple models, contributing to a more comprehensive decision map. Hyperparameter tuning involves an iterative adjustment of configuration settings of the model, attempting to discern the order that yields the best results.We aim to extract the maximum potential from XGBoost, through numerous adjustments.

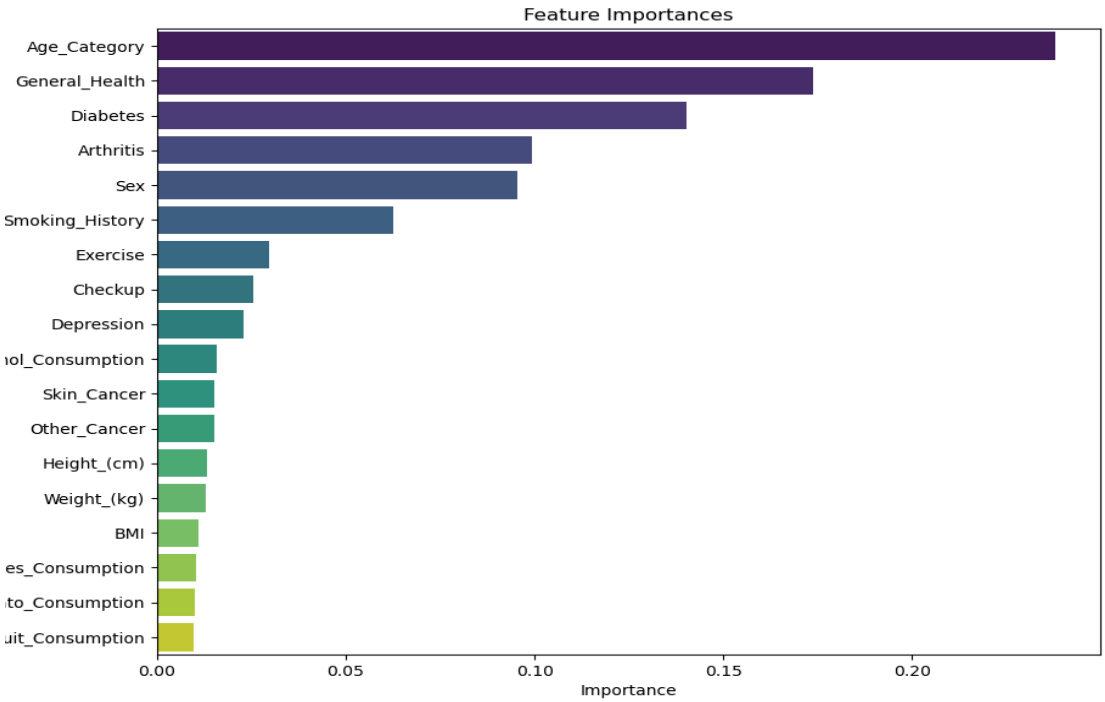


Fig 4.3.1: Feature importance

Table 4.3.1: Performance Metrics obtained using Bagging

Model	Accuracy	Precision	Recall	F1 Score
lightgbm	0.9190	0.8838	0.923	0.8900
xgboost	0.9189	0.8827	0.9156	0.8888
Random Forest	0.9065	0.8829	0.9073	0.8934
Adaboost	0.82192	0.8847	0.8009	0.8541
gbc	0.8256	0.8900	0.8456	0.8883

Table 4.3.2: Performance Metrics obtained using stacking

Model	Accuracy	Precision	Recall	F1 Score
stacking_model	0.90704	0.8835	0.9132	0.8918

Table 4.3.3: Performance Metrics obtained using voting

Model	Accuracy	Precision	Recall	F1 Score
voting_model	0.91031	0.8818	0.9090	0.8868

Table 4.3.4: Performance Metrics obtained using Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1 Score
xgboost	0.9200	0.8900	0.9200	0.8900
lightgbm	0.9114	0.8835	0.9110	0.8866
gbc	0.9065	0.8629	0.8973	0.8834
Adaboost	0.8956	0.8847	0.8909	0.8541
Random Forest	0.8256	0.8900	0.8256	0.8183

4.4 Model Evaluation

Following our XGBoost hyperparameter tuning model, we obtained outstanding model performance outcomes. The accuracy was a remarkable 92% demonstrating the general accuracy of our predictions.Additionally, the % of real positive predictions among all positive predictions is indicated by the precision of 0.89. This measure is very useful when trying to reduce false positives.The model's recall of 0.92 shows that it can accurately identify the majority of true positive cases.It highlights how well the model detects positive situations by showing a low percentage of false negatives.The F1-Score in our case was 0.89, indicating a balanced performance in recall and precision. Together, these assessment indicators show how reliable and efficient our adjusted XGBoost model is. These outcomes, in our opinion, demonstrate the model's capacity to fulfill the goals set out in our machine learning job.

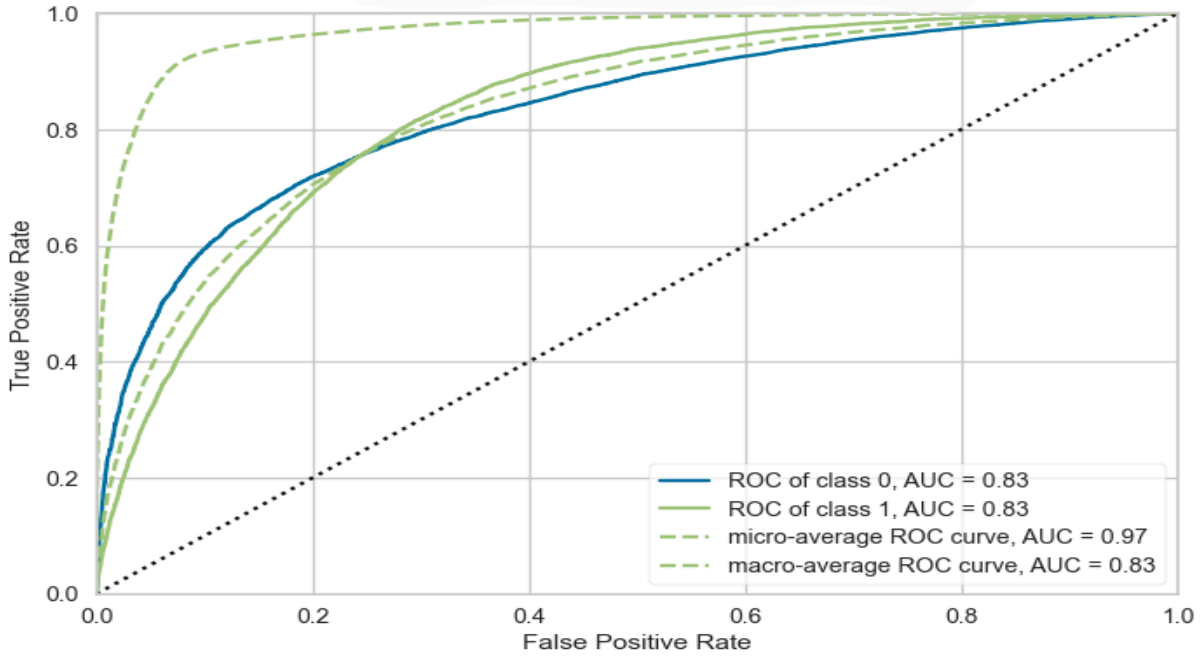


Figure 4.3.2: AUROC performance curve of XGBoost Hypertuning.

$$\text{Accuracy} = \frac{TP+TN+FP+FN}{TP+TN}$$

In this formula:

True Positives (TP) are instances correctly predicted as positive.

True Negatives (TN) are instances correctly predicted as negative.

False Positives (FP) are instances incorrectly predicted as positive.

False Negatives (FN) are instances incorrectly predicted as negative.

Precision: Precision measures the proportion of true positive predictions out of all positive predictions made.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall (Sensitivity): Recall measures the proportion of true positive predictions out of all actual positive instances in the dataset.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives +False Negatives}}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

$$\text{F1 Score} = \frac{2*\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.5. Proposed Method Output

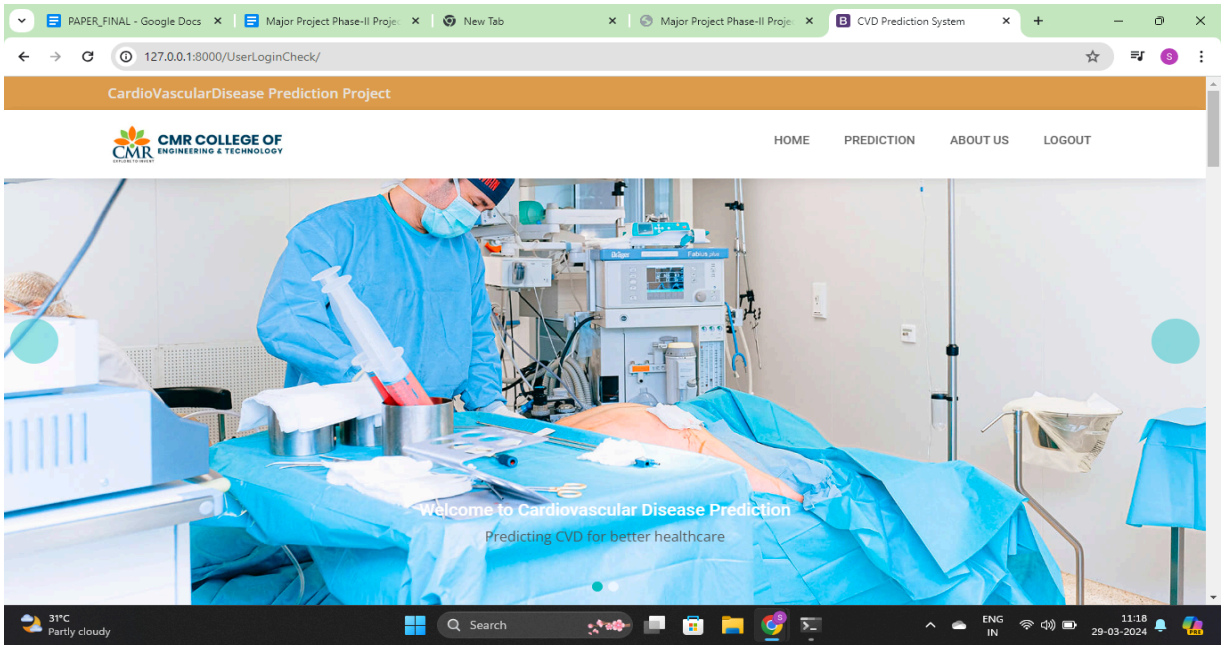


Figure 4.5.1: Home page of a proposed system

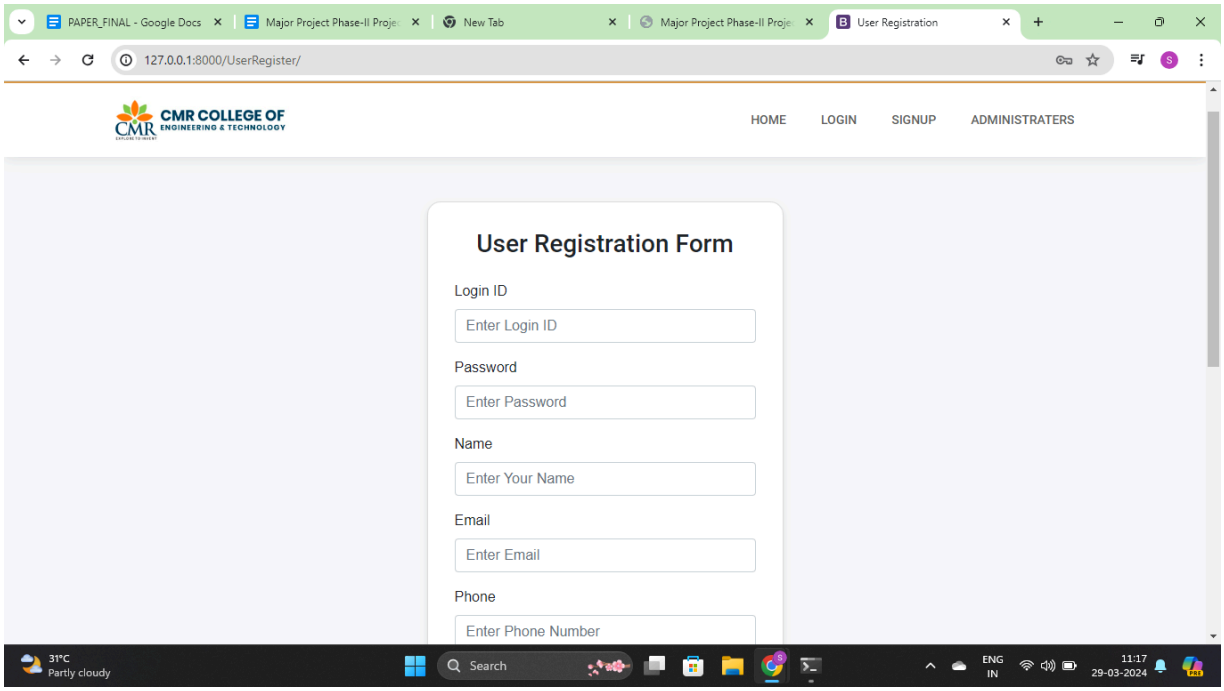


Figure 4.5.2: Registration Form of a User

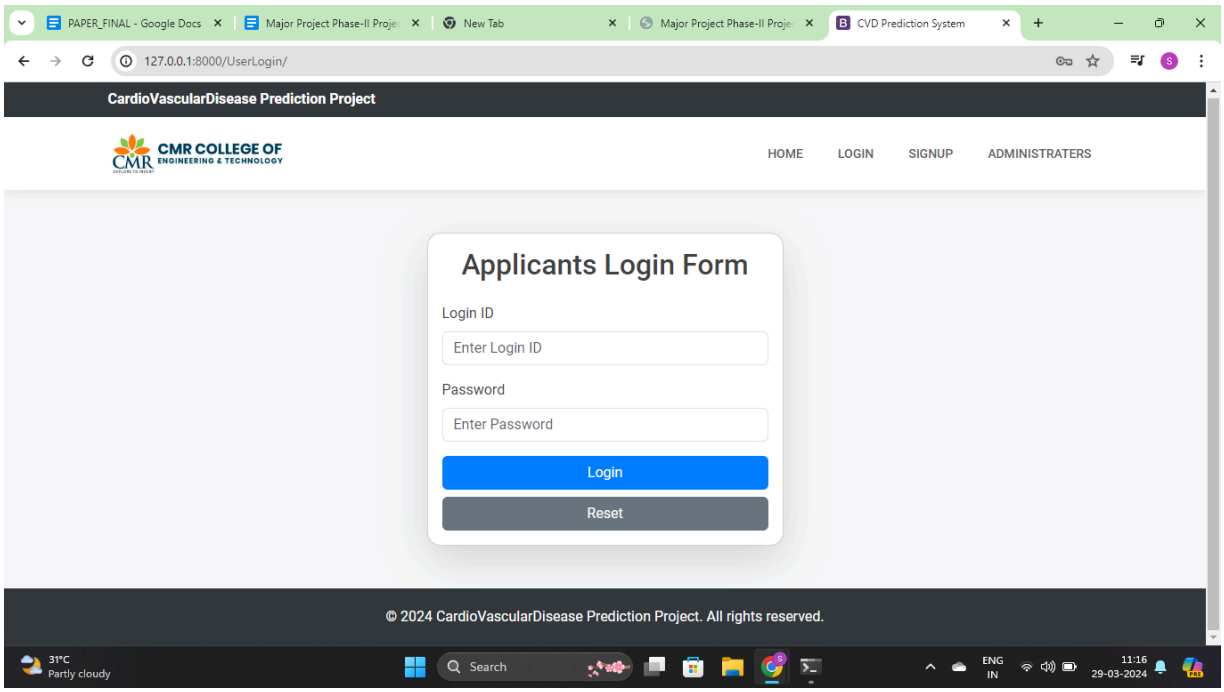


Figure 4.5.3: Login Page

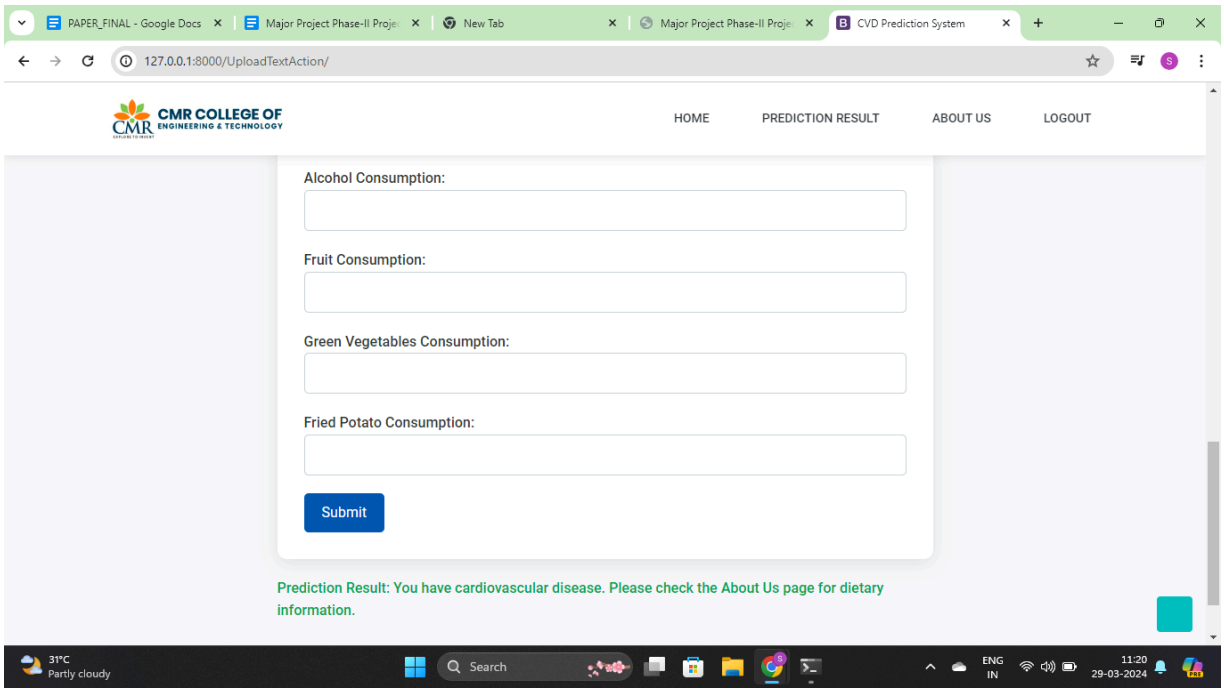


Figure 4.5.4: Disease Prediction Form

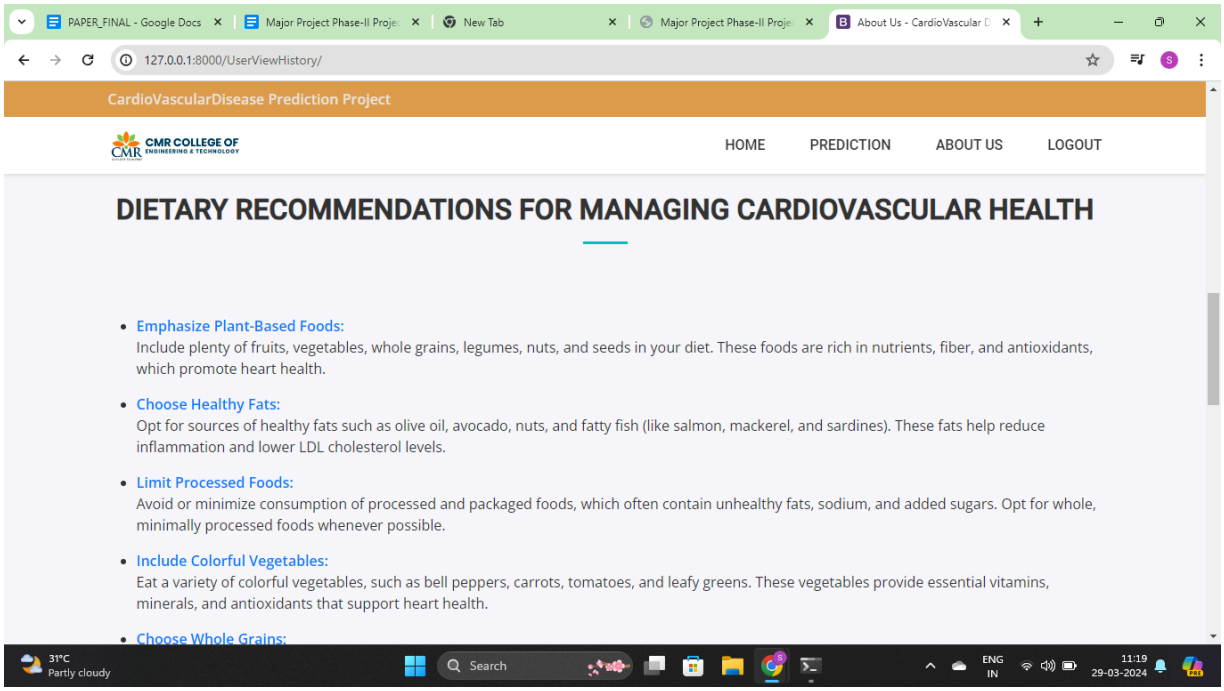


Figure 4.5.4: Dietary Recommendation Page

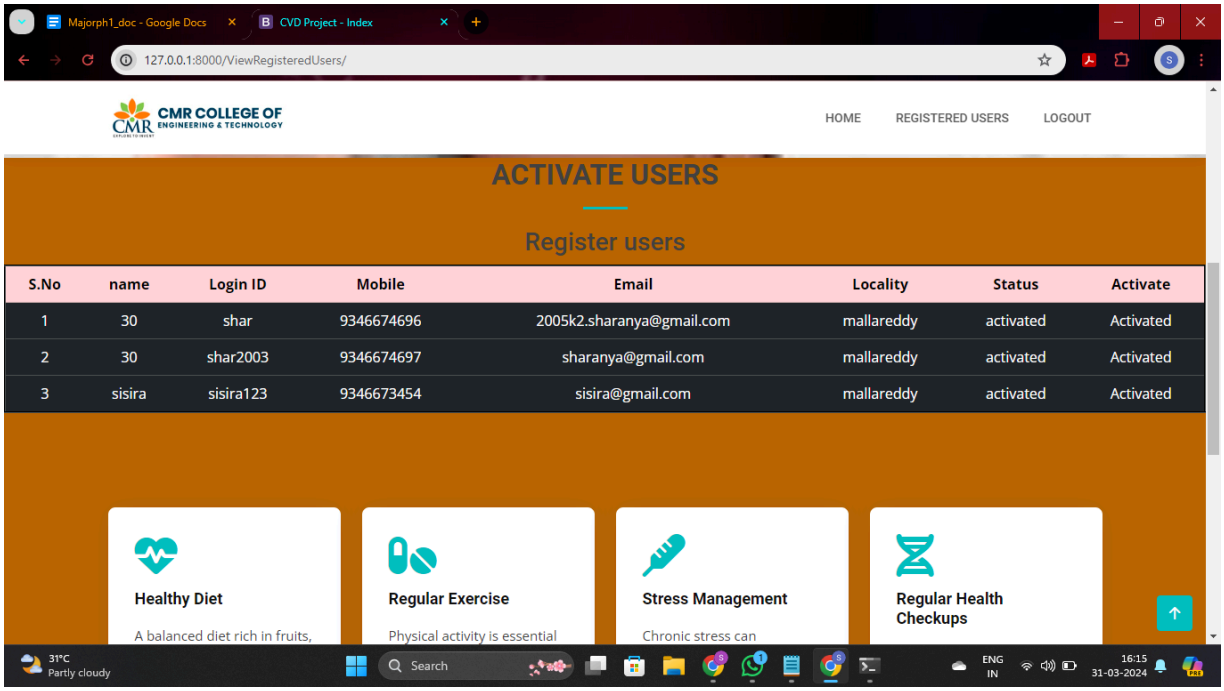


Figure 4.5.6: Admin Activate Users

CHAPTER 5

CONCLUSION

CHAPTER 5
CONCLUSION

5.1 CONCLUSION AND FUTURE ENHANCEMENT:

In conclusion, our aim of predicting Cardiovascular Disease (CVD) using advanced ensemble techniques leads us to the hypothesis that combining multiple machine learning models results in higher accuracy. Developed a reliable tool for early detection of CVD and intervention for those at risk. Through our exploration of various ensemble methods, including Random Forests, AdaBoost, Gradient Boosting, and XGBoost, we successfully created an ensemble model with an accuracy of 92%. This finding suggests that the hypothesis holds – ensembling techniques contribute to improved predictive accuracy in the context of CVD. The ensemble approach, blending the strengths of different models, plays a crucial role in achieving this high accuracy. By avoiding overfitting and demonstrating effectiveness in real-world healthcare scenarios, our ensemble model stands out as a promising tool for identifying individuals at risk of cardiovascular complications. As we move forward, it becomes evident that our hypothesis aligns with the outcomes of this study. Ensembling machine learning models, as demonstrated in our research, offers a practical avenue for healthcare professionals and policymakers to enhance the accuracy of predictive tools. This, in turn, contributes to proactive healthcare interventions and the prevention of cardiovascular diseases in at-risk populations .

In the future, continuous monitoring and updating of the model with new data, along with advancements in machine learning methodologies, to ensure its relevance and effectiveness in real-world healthcare settings. Lastly, collaboration with healthcare professionals and stakeholders to validate the model's predictions and integrate it into clinical practice for early detection and intervention of cardiovascular disease.

REFERENCES

REFERENCES

[1]WHO, Geneva. "WHO methods and data sources for country-level causes of death." (2014)

[2] Singirikonda, Bhagyalaxmi, and Muktevi Srivenkatesh. "An Approach to Prediction of Cardiovascular Diseases using Machine and Deep Learning Models." International Journal of Intelligent Systems and Applications in Engineering 10

[3] Gupta, Ankur, Rahul Kumar, Harkirat Singh Arora, and Balasubramanian Raman. "MIFH: A machine intelligence framework for heart disease diagnosis." IEEE access 8 (2019)

[4] Yazid, M. Haider Abu, Muhammad Haikal Satria, Shukor Talib, and Novi Azman. "Artificial neural network parameter tuning framework for heart disease classification."

[5] Alqahtani, Abdullah, Shtwai Alsubai, Mohammed Sha, Lucia Vilcekova, and Talha Javed. "Cardiovascular disease detection using ensemble learning." Computational Intelligence and Neuroscience 2022 (2022).

[6] Pradhan, M. R. "Cardiovascular disease prediction using various machine learning algorithms." Journal of Computer Science 18, no. 10 (2022): 993-1004.

[7] Singh, Vicky, and Brijesh Pandey. "Prediction of Cardiac Arrest and Recommending Lifestyle Changes to Prevent It Using Machine Learning." In International Conference on Intelligent Technologies & Science, pp. 1-6. 2021.

[8] Karthikeyan, R., D. Vijendra Babu, R. Suresh, M. Nalathambi, and S. Dinakaran. "Cardiac Arrest Prediction using Machine Learning Algorithms." In Journal of Physics: Conference Series, vol. 1964, no. 6, p. 062076. IOP Publishing, 2021

[9] Maheswari, K. Uma, and J. Jasmine. "Neural network based heart disease prediction."

[10] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554..

[11] Mehrabani-Zeinabad, Kamran, Awat Feizi, Masoumeh Sadeghi, Hamidreza Roohafza, Mohammad Talaei, and Nizal Sarrafzadegan. "Cardiovascular disease incidence prediction by machine learning and statistical techniques: a 16-year cohort study from eastern Mediterranean region." BMC Medical Informatics and Decision Making 23, no. 1 (2023): 72.

[12] El Naqa, Issam, and Martin J. Murphy. What is machine learning?. Springer International Publishing, 2015.

GITHUB LINK: https://github.com/sisira1485/major_project/tree/main

iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.58888

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

An Efficient Ensemble Machine Learning Model for Cardiovascular Disease Prediction Using Digital Health Records

B. Sharanya[1], V. Srividya[2], B. Prajnaya[3], Bandari Gayathri[4]
[1, 2, 3]*UG Student Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana*
[4]*Assistant Professor, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana*

Abstract: ***Detecting cardiovascular problems early is crucial for timely treatment. In our study, we employed machine learning to analyze a diverse set of information about individuals' lives and health, to predict cardiovascular disease. Ensuring data accuracy and addressing missing information were prioritized in our approach. Experimenting with different solo ML & ensemble ML methods, comprised of Random Forest and XGBoost with tuning, we achieved a notable 92% accuracy in identifying potential heart issues. Remarkably, combining multiple machine learning methods through ensemble learning proved even more effective than individual methods. Expanding our methodology to include Light GBM, Extra Tree, Decision Tree, SVM, Naive Bayes, QDA, & Adaboost enhanced the comprehensiveness of our analysis. Additionally, delving into ensemble learning methods such as bagging, boosting, tuning, & stacking further pushed the boundaries of predictive accuracy. In essence, our research outstands the potency of diverse ensemble machine-learning techniques and algorithms in early cardiovascular prediction. Ensemble methods, which combine different algorithms, emerged as powerful tools without relying on complex terminology.***

Keywords: Cardiovascular Disease (CVD), Artificial Intelligence(AI), Machine Learning (ML), Deep Learning (DL), Ensemble Learning, Heart disease.

I. INTRODUCTION

As per WHO CVD is the deadliest disease almost a third of all deaths are caused by cardiovascular disease it takes over 17 million lives over a year. In the year 2023, it took over 20.5 million lives [1]. Advanced knowledge can help us to prevent deaths caused by cardiac arrest. To identify cardiac diseases there are symptoms like chest pain, chest pressure, shortness of breathing, and fainting. Heart diseases are identified by clinical examinations, medical history, and diagnostic tests. Cardiovascular disease (CVD) is a disease that affects on heart and blood vessels. There are distinct types of cardiovascular diseases. They are Coronary heart disease, Peripheral heart disease, stroke, and Heart failure. Cardiovascular disease is caused due raised blood pressure, high cholesterol levels, lack of exercise, diabetes, smoking, tobacco, and improper Diet. Healthcare professionals or providers can use Artificial Intelligence to detect heart diseases. Leads to diagnosis at the advanced stage of the disease. Artificial Intelligence is predominantly used in identifying cardiovascular disease. In the presented study we hypothesize that ensemble, ML functionalities are superior to solo and deep learning algorithms due to the scarcity of available clinical data. This paper is structured across 5 different modules. module 2 concisely scrutinizes literature related to CVD prediction using ML and DL algorithms. module 3 describes the materials and methods used in the prediction of CVD using ML and DL algorithms. module 4 embraces the results. In module 5 we present a comparative study of implemented model with the state-of-art techniques for predicting CVD.

II. LITERATURE REVIEW

In this literature review, we explore the latest research on machine learning-based approaches for CVD prediction. We examine the different machine learning algorithms employed in CVD risk prediction, the datasets used for model development and validation, evaluation metrics for assessing model performance, and recent advancements in the field. By synthesizing findings from the literature, we attempt to contribute insights into the potential of machine learning to revolutionize CVD risk stratification and inform personalized preventive strategies in clinical practice.

Muktevi Sri Venkatesh et al[2] In their discussion of the early methods for forecasting cardiovascular illness, suggested a prediction using an SVM, Naïve Bayes classifier, Random Forest RF, and logistic regression. In comparison to other machine learning algorithms, logistic regression has a higher accuracy rate (77.06%), according to his research.

Ramdas Kapila et al[3] introduced the Mcclusky Binary Classifier (QMBC) model, which makes use of an ensemble of seven models, an adopted set of machine learning
techniques. They all used ANOVA and chi-square methods to determine the top ten traits. The Cleveland dataset and the CV dataset were used in this investigation. It provided a 98.36% accuracy rate.

Liaqat Ali et al [4] focus on issues related to overfitting and underfitting. The chi-square statistical model, which they devised, is used to remove unnecessary features when searching with an exhaustive search method for the best-configured deep neural network. The presented mixed diagnostic system's temporal complexity was not examined in this study. The expected accuracy of the suggested model is 93.3%.

Hemantha Kumar Kalluri et al [5] implemented a model for disease forecasting that leverages Convolutional Neural Networks (CNNs) to achieve predictive accuracy. Comprising two convolution layers, 2 dropout layers, and a single resulted layer, this model demonstrates a reported accuracy of 94.78%. Through the strategic integration of convolutional layers, the model captures hierarchical patterns in input data, crucial for disease prediction. Dropout layers enhance model generalization by reducing overfitting, while the output layer synthesizes learned features into actionable predictions. With its high accuracy, this CNN-based model presents a promising avenue for disease forecasting, offering valuable insights into potential diagnoses and treatment strategies.

Mayank Sharma et al [6] proposed a model that gains performance in detecting cardiac disease. They used a Tree classifier, hybrid CNN, and Bi-LSTM to predict heart illness. In this study, comparative studies are also examined. This approach is capable of producing an accuracy of 96.66%.

James Meng et al [7] presented the first clinical knowledge-enhanced ML model for predicting IHD. They included key steps such as statistical analysis, preprocessing, feature selection, and model learning evaluation. This model based on SVM achieved an accuracy of 94.4%.

Abu Yazid et al[8] researched the ANN and used the Cleveland dataset to achieve 90.9% of accuracy and also worked with the statlog dataset and achieved an accuracy of 90%

B.B.Gupta et al [9] created a model for accurately predicting cardiovascular illness. IoT models and machine learning were employed. With an accuracy of 87.72%, it seems that more complicated classifiers, such as SVM and Random Forest, produced superior results.

Ankur Gupta et al [10] suggested a framework for machine intelligence MIFH is used to diagnose heart disease. They suggested a framework called MIFH, which can be utilized to forecast the occurrences of either heart patients or normal persons. Their sensitivity was 92.8%, compared to 89.28% for MIFH.

Talha Javed et al [11] suggested deep learning and machine learning methods based on ensembles to forecast cardiovascular illness. The models' performance was evaluated based on how accurate they were. Their accuracy rate was 88.70%.

Pradhan et al [12] considered the UCI repository dataset and five methods (support vector machine, logistic regression,
main component analysis, multi-layer perceptron classifier, and achieved approximately 90% accuracy.

Vicky Singh et al [13] used machine learning algorithms in this examination, and a recommendation system based on variables like age, blood pressure, and so on was presented. They concluded that SVM and decision tree classifiers provide 85% accuracy.

R. Karthikeyan et al [14] suggested utilizing a convolution neural network and deep learning to predict cardiovascular illness.

Uma Maheshwari et al[15] engaged a unique method for predicting cardiac disease by combining neural networks with logistic regression analysis. Initially, the foremost risk indicators for illness prediction are chosen using logistic regression. The statistical p-value is produced. With an accuracy of 84%, the combination of logistic regression and neural network is used to predict cardiac disease.

Senthil Kumar et al[16] By using machine learning approaches, created an excellent approach that improves the accuracy of cardiovascular disease detection by identifying important aspects. He achieved an improved performance level of 88.7% using HRFLM.

III. MATERIALS & METHODS

A. Description of the Data Set

In machine learning, data is paramount for accuracy. This collected dataset contains 19 variables of which 12 are arithmetical and 7 are categorical. The number of instances is 308854 and the dataset does not contain missing values.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Table 1: CVD Dataset

Serial Number	Attribute	Description
1	General Health	Well-being, fitness
2	Check-up	Examination to ensure health or wellness
3	Exercise	Activity for fitness and health.
4	Heart Disease	Cardiac condition
5	Skin_cancer	Describe different parts of your skin or conditions that can affect it.
6	Other_Cancer	Those who indicated they have experienced any other forms of cancer
7	Depression	Feeling sad, hopeless, or down for a long time.
8	Diabetics	Having too much sugar in your blood for a long time.
9	Arthritis	Pain and swelling in your joints makes it hard to move.
10	Gender	0-Female,1-Male
11	Age	In days
12	Height	In Cent Meter
13	Weight	Kilograms
14	BMI	It is a number that shows if a person is a healthy weight for their height.
15	Smoking History	Whether Patient Smokes or Not
16	Alcohol consumption	Whether patient smokes or not
17	Fruit consumption	Recording patients' fruit intake
18	Green Vegetables consumption	Recording patients' green vegetable intake
19	Friedpotato vegetable consumption	Recording patients' potato intake

B. Methods

1) Solo Machine Learning Algorithms

[17]Algorithms for machine learning allow computers to recognize patterns and connections in data without the need for explicit programming. Based on input data, these algorithms employ statistical approaches to find patterns and provide predictions or choices. Three main categories can be used to group them

Supervised Learning: In this method, an algorithm is trained using a labeled dataset in which every input has a matching output. To forecast or categorize newly discovered data, the algorithm gains knowledge from this labeled dataset. Neural networks, decision trees, SVmachine, logistic regression, and linear regression are common techniques in supervised learning.

Unsupervised Learning: When there are no labels on the input data, unsupervised learning algorithms are applied. Without any indication, the algorithm looks for facts in the statistics. This learning comprises approaches, clustering algorithms like k-means, and hierarchical clustering. In unsupervised learning, dimensionality reduction methods like (SVD) and (PCA) are frequently employed.

Reinforcement learning: Teaching an agent through reinforcement learning involves guiding it to interact with its environment to maximize rewards. The agent learns through trial and error receiving feedback in the form of rewards or penalties based on its performance. Algorithms, like Q learning and deep Q networks (DQN) are commonly used in robotics, gaming and autonomous systems for reinforcement learning tasks.

a) Logistic Regression

In the domain of learning algorithms logistic regression stands out as a choice in machine learning. It is utilized for predicting outcomes based on a set of variables. Unlike regression, which is used for regression tasks logistic regression tackles classification challenges by providing values between 0 and 1 instead of definite values of 0 or 1. The core concept behind regression is fitting a S" shaped function instead of a straight line to predict binary outcomes that indicate the likelihood of an event occurring.

b) Ridge Classifier

To combat overfitting issues in machine learning models methods like Ridge Regression and Ridge Classifier are employed. Overfitting occurs when a model performs well on training data but poorly, on data.In Ridge Regression, an additional term known as L2 regularization is included in the linear regression equation to avoid overfitting. This term penalizes coefficients helping to manage the complexity of the model. Likewise a Ridge Classifier employs L2 regularization to prevent overfitting, in tasks involving class classification.

c) Support Vector Machine

Machine learning is principally built upon SVM, which is an abbreviation of Support Vector Machine. It can also be used as a regression tool, though classification is its main area. At its very essence, SVM looks for a decision horizon that splits n-dimensional space into well-defined groups; this decision horizon is usually referred to as a hyperplane. Thereafter, when new data points are encountered in the future, it will be easy to place them into different classes if there was proper interpretation of the hyperplane initially made. The selection of critical elements called support vectors that define the hyperplane is central to how SVM functions. Consequently, suport vectors are those which have great impact on where the line passes through and what direction it takes on the graph.In reality, these support vectors determine where the line goes and in what direction it would slant; thus giving birth to SV machine as its name suggests.The algorithm exhibits great efficiency in classifying data points when using these support vectors via optimization of the hyperplane itself.

d) K-Nearest Neighbour

Regression problems and classification tasks are addressed by utilizing K-Nearest Neighbors (KNN) algorithm within machine learning.The following sentence provides an explanation for KNN:Think about some points that you have plotted on a graph with each one being assigned either category or value.Whereas KNN examines the closest neighbors.Then a vote is taken (for classification) or averages (for regression) obtain their labels or values in order to find the label or value of the new point. KNN is an easy to use and adaptable technique making it applicable in different areas like pattern recognition, data mining and intrusion detection. Another thing why KNN is considered good because it is “non-parametric.” This means that it does not presume anything about how data are distributed.This flexibility makes it handy for real-life situations where data can be messy or irregular. To use KNN, you start with some known data (called training data), which has points already labeled. Then, when you get a new point, KNN compares it to the known points and makes predictions based on their proximity.

e) Decision Tree

Although it is frequently used for classification, a DT is a useful tool in supervised learning; it would handle both regression & classification tasks. It resembles a tree-shaped flowchart in which decisions are made at each stage depending on a particular data attribute.

Nodes: Decision nodes and leaf nodes are the 2 foremost types. Decisions are taken at decision nodes, which then lead to branches. The ultimate results, or leaf nodes, have no offshoots. Making Decisions: The characteristics of the data are used to inform decision-making. For instance, we might choose a fruit based on its size or color to determine if it is an orange or an apple. Visual Representation: Consider it as a tree that branches out from a root node. Every branch symbolizes a choice made in response to a characteristic, with results appearing on the leaves. Building the Tree: To build the tree, we employ a technique known as CART (Classification and Regression Tree). The selection of features to employ and the timing of decisions are made easier by this algorithm. As a result, a decision tree is a visual tool for determining possible results or solutions for an issue by making decisions based on variables in the data. Because it begins with a root node and grows into stems to form a structure corresponding to a tree, it is known as a DT.

f) Naive Bayes

One sort of supervised learning utilized for classification tasks—particularly for text classification with huge datasets—is the Naïve Bayes functionality. WKT being easy to use but efficient in creating prediction models quickly. Probabilistic Classifier: Naïve Bayes makes predictions based on how likely it is that an object will fall into a specific class. For example, it might determine whether an email is spam or not by looking at the likelihood of specific terms showing up in spam emails. "Naïve" The assumption: Because it presumes that features are independent of one another, it is referred to as "naïve". For example, if we're trying to identify fruits based on color, shape, and taste, Naïve Bayes treats each feature (color, shape, taste) separately. So, even though these features might be related (like red apples being more likely to be sweet), Naïve Bayes assumes they're independent.

2) Ensemble Learning Techniques

Ensemble learning is a powerful technique in machine learning where many models are joined to enhance predictive accuracy and robustness. The underlying principle is that while individual models may have limitations or biases, combining their predictions can mitigate these weaknesses and yield better overall performance. One common approach in ensemble learning is through voting methods, where predictions from multiple models are aggregated, such as through major voting for classification tasks. Boosting is another widely used technique, wherein models are trained orderly, with each successive model concentrating on eradicating the errors made by its predecessors. This iterative process often results in highly accurate predictions. Stacking, on the other hand, involves combining predictions from a variety of models through a meta-learner. a meta-learner learns how to effectively integrate the expectations of base models to produce the final output. In this study, we employed various ensemble learning algorithms for our predictive modeling.[18]

3) Gradient Boosting

Gradient Boosting is a powerful method in machine learning that creates strong predictive models by combining multiple weaker models. Here's how it works: Combining Weak Models: It starts with simple models, often decision trees, and gradually improves them. Minimizing Loss: Each new model is trained to reduce the errors of the previous model. It does this by minimizing a loss function, like mean squared error or cross-entropy. Gradient Descent: The algorithm calculates the gradient (slope) of the loss function concerning the predictions of the current ensemble. This gradient guides the training of the new model. Training New Models: A new weak model is then trained to minimize this gradient. It focuses on correcting the errors made by the current ensemble. Adding Predictions: The predictions of the new model are added to the ensemble, improving the overall predictions. Iterative Process: This process repeats until a stopping point is reached, like when no further improvements are seen. Unlike AdaBoost, which adjusts the weights of training instances, Gradient Boosting uses residuals of the previous model as labels for training the next one. One popular form of Gradient Boosting is Gradient Boosted Trees, where each weak learner is DT (specifically CART - Classification and Regression Trees).

4) LightGBM

LightGBM offers a sizeable advancement in terms of efficiency and memory footprint due to its innovative techniques: GOSS and EFB.

GOSS redefines sample selection by assigning priority to instances with substantial gradient contributions while subsampling those with high gradients, thus hastening training without compromising model performance. On the other hand, EFB proposes another way of feature representation whereby exclusive features are packed into one bundle reducing dimensionality and conserving memory consumption. In summary, these techniques form the characteristic traits of LightGBM which differentiate it from traditional GBDT frameworks. By going beyond histogram-based algorithms, LightGBM paves way for efficient and effective gradient boosting which has led to superior machine learning performance

5) XGBoost

XGBoost, which is an abbreviation for Extreme Gradient Boosting, represents the state of the art in optimized distributed GBoosting libraries designed for efficient and scalable training of ML models. This ensemble learning approach merges knowledge from many weak models to produce a robust and powerful prediction machine.It is known for being able to handle massive datasets effectively and perform well across a wide range of machine learning tasks, making XGBoost one of the most widely used and useful tools in the field of machine learning. One key quality that makes it stand out is its ability to handle missing values in real world data sets. Consequently, this tool can work with such data directly reducing unnecessary preprocessing complexities; thus improving modeling pipeline speediness.In addition, XGBoost comes with built-in support for parallel processing which speeding up model training on large-scale data sets as opposed to traditional methods that are time-consuming

6) Ensemble Methods

After our first model did not make great predictions we applied various ensemble methods on it.Level 2 Models are created by combining Level 1 Model's predictions through weighted averages or simple pooling techniques. These techniques were employed to boost the overall predictive power of the model.

a) Bagging

Firstly, bagging involves the creation of numerous subsets of the original dataset through a process known as bootstrap sampling. This entails randomly selecting instances from the dataset with replacements, thereby generating subsets of the same shape as the original result set. Due to the nature of sampling with replacement, these subsets may contain duplicate instances, and each subset represents a slightly different perspective of the overall dataset.Secondly, a base model, typically a decision tree although other models can also be used, is trained on each of these bootstrap samples. As a result of the variations in the training data introduced by bootstrap sampling, each model trained on a different subset will inherently be slightly different from the others. Finally, once all the models have been trained, predictions are made for unseen data using each model. In regression tasks, the final forecast may entail averaging the predictions of all models, whereas, in classification tasks, the final values may be decided by a majority vote among the predictions of all models. The ensemble model's generalization performance is eventually enhanced by this combination of predictions from other models, which helps to minimize overfitting and lower variance. Bagging is particularly effective at reducing overfitting because it leverages the diversity introduced by training models on different subsets of result sets. By averaging the predictions of multiple models, the ensemble model tends to exhibit lower error rates compared to individual models, thus enhancing predictive accuracy. Popular algorithms that utilize bagging include Random Forests, XGBoost, LightGBM, and AdaBoost which employ bagging with decision trees as base learners, and Bagged Trees, which can utilize any base learning algorithm. Bagging's versatility in ensemble learning extends beyond decision trees, making it a widely applicable technique across various domains of machine learning.

b) Stacking

Other words used interchangeably with stacking include stacked generalizations or stacked.A good example would be ensemble as another effective method of group instruction. Instead of just averaging the predictions of several systems, stacking involves training a meta-model or meta-learner to figure out how to optimally combine the predictions of the base models. By using the underlying models' predictions as input features, this metamodel learns to produce predictions based on these inputs. In essence, it figures out how best to combine or weigh the underlying models' predictions to get the final one. Stacking is an ensemble technique that is more advanced than majority voting or simple averaging, which are employed in bagging methods. Enhancing predictive performance has been demonstrated to be highly beneficial, particularly when the underlying models are complementary and diversified. Nevertheless, stacking might need more precise hyperparameter adjustment and involve more computational work. Despite these difficulties, stacking is nevertheless a well-liked and effective method in the ensemble learning toolkit. e learning toolbox.

c) Boosting

Boosting is Additionally well-known as an ensemble learning method in machine learning, which emphasizes training several weak learners in turn giving a powerful ensemble system. In contrast to independent model training, boosting trains models iteratively by having each new model place greater emphasis on cases that the preceding model misclassified. Essentially, boosting is the process of combining several weak models—usually shallow decision trees, or "weak learners"—to produce a strong and precise predictive model. Boosting is well renowned for its capacity to generate extremely accurate models, frequently surpassing the performance of single models and even other ensemble techniques like bagging. Boosting algorithms, however, may be susceptible to noisy data and outliers, and to avoid overfitting, hyperparameters may need to be carefully adjusted.

d) Tuning

In machine learning parlance, "tuning" is modifying a few variables, or "hyperparameters," to maximize a machine learning model's performance. In contrast to a model's parameters, which are determined during training, hyperparameters are predetermined and have an impact on the learning process. Hyperparameter tuning is essential for getting the most out of a machine-learning model, regardless of the method employed.It often involves a trade-off between computational resources, such as time and hardware, and the quality of the resulting model. Effective hyperparameter tuning can lead to improved model accuracy, generalization, and robustness across different datasets and applications.

IV. RESULTS & DISCUSSIONS

A. Data Preprocessing

Dataset Contains 12 Categorical values so we owned a Label Encoder to transfigure Categorical to Numerical.

1) Label Encoder

In machine learning, a label encoder is a preprocessing tool that transforms textual or category input into an indexed representation. In particular, this is crucial when collaborating with distinct machine learning algorithms that need quantitative input because a large number of algorithms are built to work with numerical data. An instance of the label encoder is created at initialization. The training dataset's categorical labels are "fitted" to the encoder. The encoder learns the mapping between each distinct label and a corresponding numerical value in this step. The dataset's categorical labels can be converted into numerical representations using the encoder once it has been fitted. Using the mapping that was discovered during the fitting process, each distinct label is substituted with its assigned numerical value.

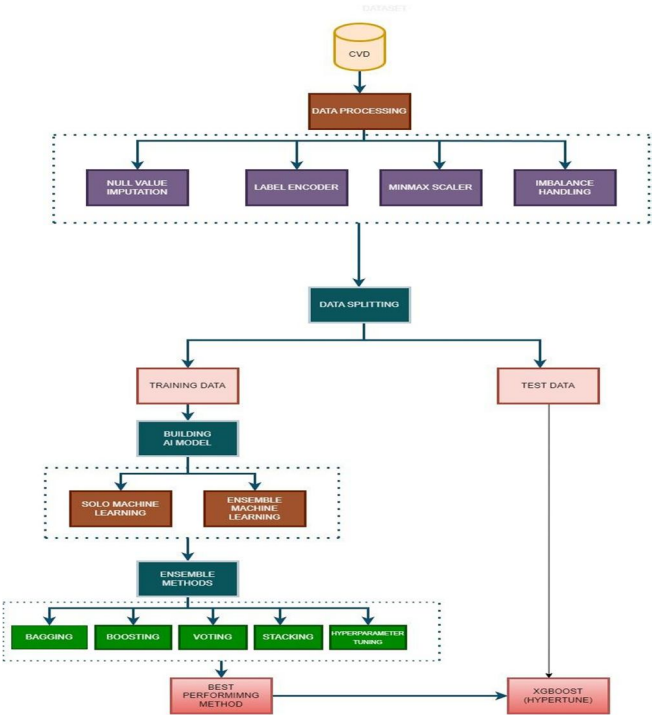


Fig1: Architecture Diagram

2) Min Max Scaler

MIN MAX Scaler is widely used to proportionate the nominal value from 0 to 1. To scale numerical characteristics inside a given range, usually between 0 and 1, machine learning uses the preprocessing approach known as min-max scaling. Our objective is to standardize the data and assign a uniform scale to every characteristic so that variations in their magnitudes do not lead to one feature predominating over others. This normalization might be especially crucial for algorithms, such gradient based optimization methods, that depend on gradients or distance measurements.

3) Imbalance Data Handling

In handling imbalanced data, we experimented with two techniques, SMOTE and ADASYN.

a) SMOTE

Machine learning uses the SMOTE data augmentation technique to solve the issue of class imbalance in classification jobs. When one class in the target variable has much fewer instances than another, it is said to be imbalanced. As a result, the minority lesson will see inefficient performance from one-sided models. To stabilize the distribution of classes, SMOTE generates synthetic examples with a concentration on the minority class.

b) ADASYN

ADASYN (Adaptive Synthetic Sampling) was created to solve some of its shortcomings. Similar to SMOTE, ADASYN is pre owned to manage class imbalance in datasets used for ML, especially in cases of classifying the input data into two mutually exclusive categories where one class is greatly underrepresented. We tried both approaches and discovered that the best accuracy was obtained when combining SMOTE with the XGBoost algorithm. This indicates that the best overall performance in predicting the target variable was obtained by using XGBoost to train our model and SMOTE to generate synthetic instances for the minority class.

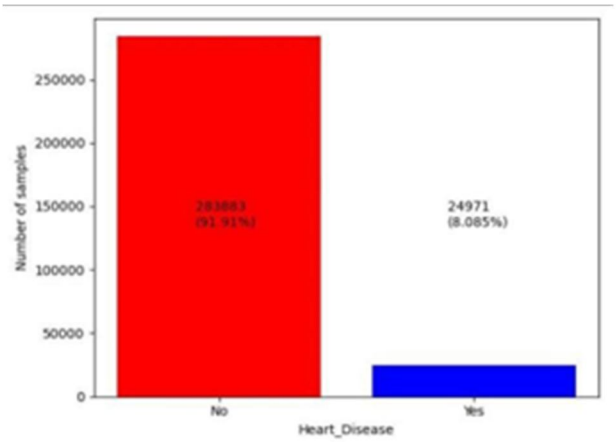


Fig2: Target Variable distribution before using SMOTE

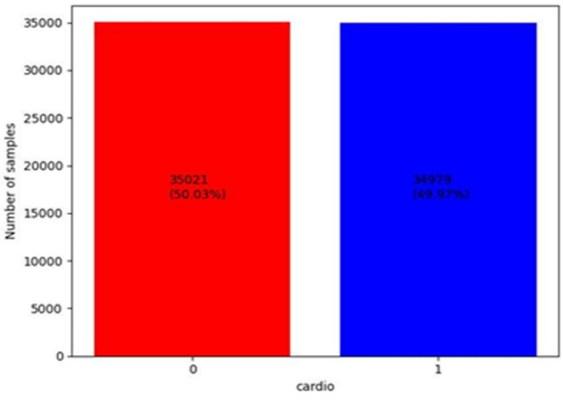


Fig3: Target Variable Distribution after using SMOTE

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

B. Data Splitting

We used the K fold Cross Validation technique and tested our set of results with 10 folds,4 folds,2folds,5 folds and took the model with the validation that gives more performance metrics and k=10 with Extreme Gradient Boosting gives superlative rightness.

Model	Accuracy	AUC	Recall	Precision	F1 Score
xgboost	0.917	0.8288	0.9013	0.8858	0.883
lightgbm	0.9009	0.8182	0.836	0.8884	0.8863
random forest	0.8832	0.8064	0.9171	0.8805	0.8865
ada	0.8502	0.8065	0.9	0.8801	0.8835
et	0.8256	0.7929	0.854	0.8781	0.8883
gbc	0.8256	0.8304	0.8456	0.8894	0.8883
dt	0.8613	0.5739	0.8613	0.8719	0.897
dummy	0.7892	0.6595	0.7892	0.8448	0.8664
svm	0.7508	0	0.7592	0.9074	0.8804
Lr	0.7476	0.8358	0.7508	0.9125	0.7951
ridge	0.7413	0	0.7476	0.9127	0.803
lda	0.7412	0.8348	0.7413	0.9127	0.7984
knn	0.6839	0.5606	0.6839	0.862	0.7532
nb	0.6091	0.801	0.6091	0.911	0.696
qda	0.5701	0.787	0.5701	0.9116	0.6615

Table2: Performance Metrics of different algorithms(k=10 folds)

C. Model Selection

After separating our result set into training and testing sets and preprocessing it, choosing XGBoost as our model. The next stage is to train the model on the training set and assess its effectiveness. Since you suggested utilizing K Fold Cross Validation by having K=10, we used cross validation to obtain a reliable performance estimate for our model. To determine which machine learning algorithm would work best for our dataset, we investigated a variety of models in this study. LightGBM, XGBoost, GB Classifier, Random Forest (RF), Extra Trees, AdaBoost (ADA), Decision Tree (DT), SVM, Logistic Regression (LR), Ridge Classifier (Ridge), Linear Discriminant Analysis (LDA), K-nearest neighbors (KNN), Naive Bayes (NB), and Quadratic Discriminant Analysis (QDA) are among the models we have developed. Among them, XGBoost appeared to be the most appropriate choice for our purposes based on such factors as accuracy, robustness, and interpretability. It has been shown to possess some of the best characteristics regarding accuracy and reliability. Therefore, we will now refine this model by modifying hyperparameters examining feature importance and preparing it for production. Henceforth it is crucial that we validate carefully comprehending these results in order to ascertain if the XGboost that we selected aligns with our customized machine learning task. Because of this, in addition to hyperparameter tuning, Bagging, Boosting, Stacking and Voting will increase the performance of our chosen model XGBoost. We intend to further enhance the system's robustness as well as its predictability using bagging , boosting , stacking , voting which build on top of ensemble approach.

Bagging will involve training top 5 models on different subsets from dataset; thus reducing overfitting while improving stability. Boosting concatenates a sequential training process, enabling the model to compare and progressively correct its errors, holding intricate relationships in the data. Stacking embraces diverse model predictions as inputs into a new meta-learner, refining a final output. Voting involves consolidating the predictions of multiple models, contributing to a more comprehensive decision map. Hyperparameter tuning involves an iterative adjustment of configuration setting, of the model, attempting to discern the order that yields the best results. By systematically tuning the model, we can reduce the system's performance on our specific dataset. We aim to extract the maximum potential from XGBoost, through numerous adjustments.

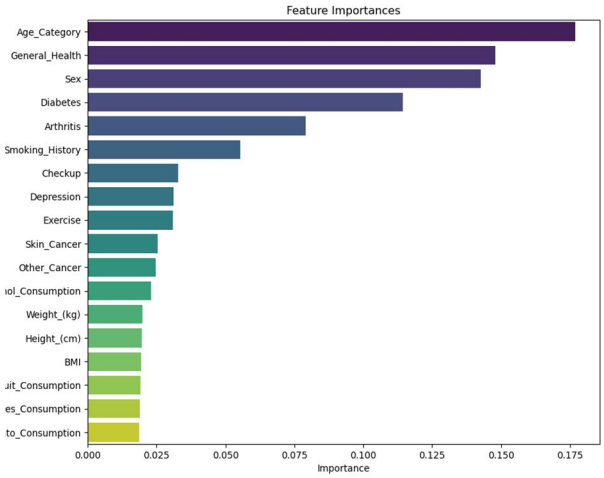


Fig 4: Feature importance

Model	Accuracy	Precision	Recall	F1 Score
lightgbm	0.9190	0.8838	0.923	0.8900
xgboost	0.9189	0.8827	0.9156	0.8888
Random Forest	0.9065	0.8829	0.9073	0.8934
Adaboost	0.82192	0.8847	0.8009	0.8541
gbc	0.8256	0.8900	0.8456	0.8883

Table 3: Performance Metrics obtained using Bagging

Model	Accuracy	Precision	Recall	F1 Score
stacking_model	0.90704	0.8835	0.9132	0.8918

Table 4: Performance Metrics obtained using stacking

Model	Accuracy	Precision	Recall	F1 Score
voting_model	0.91031	0.8818	0.9090	0.8868

Table 5: Performance Metrics obtained using voting

Model	Accuracy	Precision	Recall	F1 Score
xgboost	0.9200	0.8900	0.9200	0.8900
lightgbm	0.9114	0.8835	0.9110	0.8866
gbc	0.9065	0.8629	0.8973	0.8834
Adaboost	0.8956	0.8847	0.8909	0.8541
Random Forest	0.8256	0.8900	0.8256	0.8183

Table 6: Performance Metrics obtained using Hyperparameter Tuning

D. Model Evaluation

Following our XGBoost hyperparameter tuning model, we obtained outstanding model performance outcomes. The accuracy was a remarkable 92% demonstrating the general accuracy of our predictions.Additionally, the % of real positive predictions among all positive predictions is indicated by the precision of 0.89. This measure is very useful when trying to reduce false positives.The model's recall of 0.92 shows that it can accurately identify the majority of true positive cases.It highlights how well the model detects positive situations by showing a low percentage of false negatives.The F1-Score in our case was 0.89, indicating a balanced performance in recall and precision. Together, these assessment indicators show how reliable and efficient our adjusted XGBoost model is. These outcomes, in our opinion, demonstrate the model's capacity to fulfill the goals set out in our machine learning job.

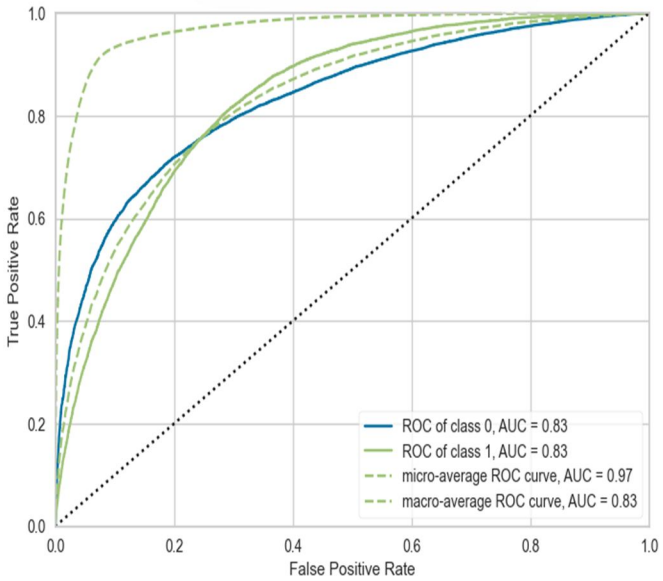


Fig 5: AUROC performance curve of XGBoost Hypertuning.

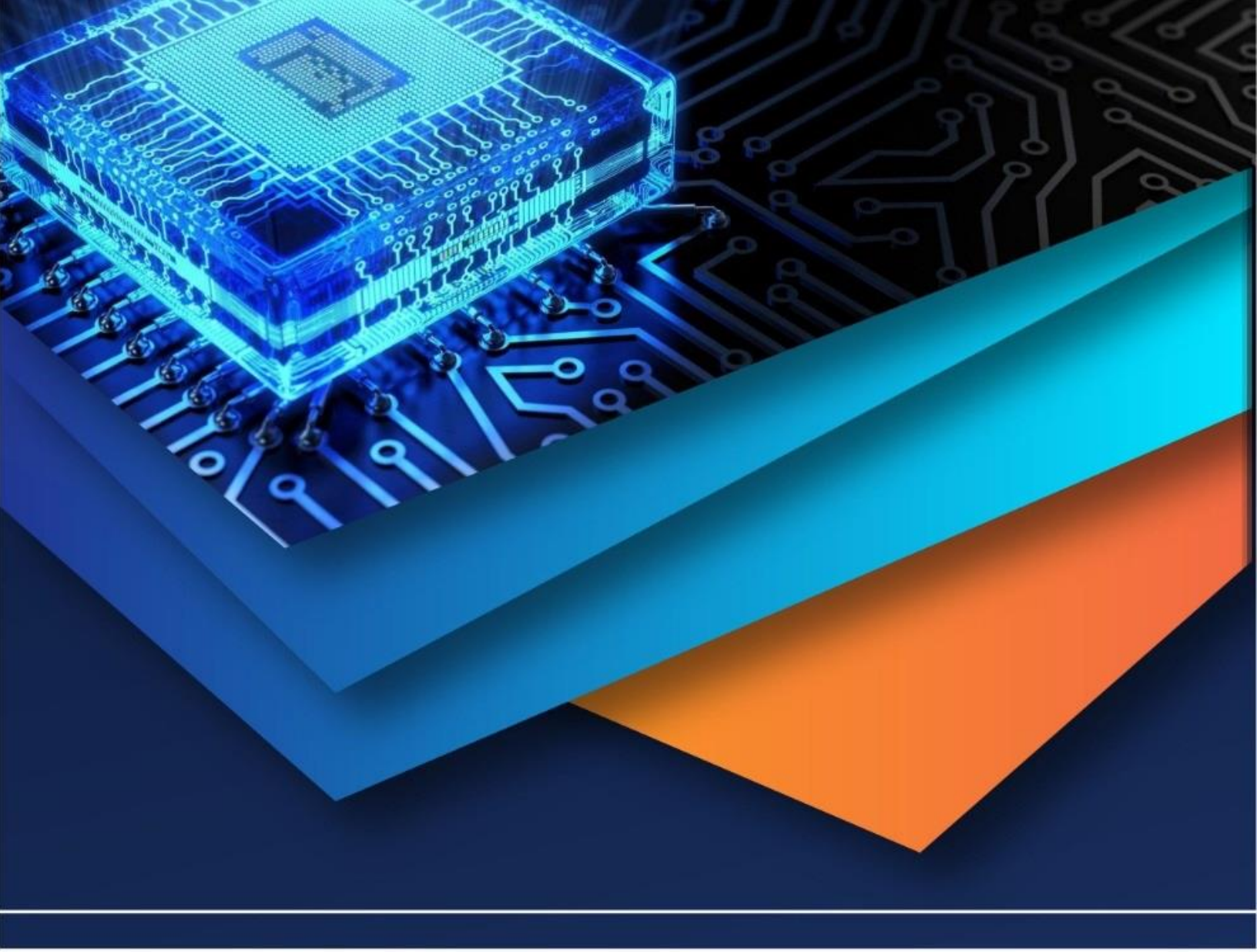
V. CONCLUSION

In conclusion, our study on predicting Cardiovascular Disease (CVD) using advanced ensemble techniques leads us to the hypothesis that combining multiple machine learning models results in higher accuracy. Our main aim was to develop a reliable tool for early detection of CVD and intervention for those at risk. Through our exploration of various ensemble methods, including Random Forests, AdaBoost, Gradient Boosting, and XGBoost, we successfully created an ensemble model with an accuracy of 92%. This finding suggests that the hypothesis holds – ensembling techniques contribute to improved predictive accuracy in the context of CVD. The ensemble approach, blending the strengths of different models, plays a crucial role in achieving this high accuracy. By avoiding overfitting and demonstrating effectiveness in real-world healthcare scenarios, our ensemble model stands out as a promising tool for identifying individuals at risk of cardiovascular complications.

As we move forward, it becomes evident that our hypothesis aligns with the outcomes of this study. Ensembling machine learning models, as demonstrated in our research, offers a practical avenue for healthcare professionals and policymakers to enhance the accuracy of predictive tools. This, in turn, contributes to proactive healthcare interventions and the prevention of cardiovascular diseases in at-risk populations.

REFERENCES

[1] WHO, Geneva. "WHO methods and data sources for country-level causes of death." (2014)
[2] Singirikonda, Bhagyalaxmi, and Muktevi Srivenkatesh. "An Approach to Prediction of Cardiovascular Diseases using Machine and Deep Learning Models." International Journal of Intelligent Systems and Applications in Engineering 10
[3] Kapila, Ramdas, T. Ragunathan, Sumalatha Saleti, T. Jaya Lakshmi, and Mohd Wazih Ahmad. "Heart Disease Prediction using Novel Quine McCluskey Binary Classifier (QMBC)."
[4] Ali, Liaqat, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, and Javed Ali Khan. "An automated diagnostic system for heart disease prediction based on ${\chi^{2}} $ statistical model and optimally configured deep neural network."
[5] Sajja, Tulasi Krishna, and Hemantha Kumar Kalluri. "A Deep Learning Method for Prediction of Cardiovascular Disease Using a Convolutional Neural Network."
[6] Shrivastava, Prashant Kumar, Mayank Sharma, and Avenash Kumar. "HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithms." Measurement: Sensors 25 (2023): 100657.
[7] Meng, James, and Ruiming Xing. "Inside the "black box": Embedding clinical knowledge in data-driven machine learning for heart disease diagnosis."
[8] Yazid, M. Haider Abu, Muhammad Haikal Satria, Shukor Talib, and Novi Azman. "Artificial neural network parameter tuning framework for heart disease classification.
[9] Ahamed, Jameel, Abdul Manan Koli, Khaleel Ahmad, Alam Jamal, and B. B. Gupta. "CDPS-IoT: cardiovascular disease prediction system based on IoT using machine learning." (2022).
[10] Gupta, Ankur, Rahul Kumar, Harkirat Singh Arora, and Balasubramanian Raman. "MIFH: A machine intelligence framework for heart disease diagnosis." IEEE access 8 (2019)
[11] Alqahtani, Abdullah, Shtwai Alsubai, Mohammed Sha, Lucia Vilcekova, and Talha Javed. "Cardiovascular disease detection using ensemble learning." Computational Intelligence and Neuroscience 2022 (2022).
[12] Pradhan, M. R. "Cardiovascular disease prediction using various machine learning algorithms." Journal of Computer Science 18, no. 10 (2022): 993-1004.
[13] Singh, Vicky, and Brijesh Pandey. "Prediction of Cardiac Arrest and Recommending Lifestyle Changes to Prevent It Using Machine Learning." In International Conference on Intelligent Technologies & Science, pp. 1-6. 2021.
[14] Karthikeyan, R., D. Vijendra Babu, R. Suresh, M. Nalathambi, and S. Dinakaran. "Cardiac Arrest Prediction using Machine Learning Algorithms." In Journal of Physics: Conference Series, vol. 1964, no. 6, p. 062076. IOP Publishing, 2021
[15] Arun Kumar, N., and P. Uma Maheshwari. "Neural Network Based Approach in Identifying Cardio Vascular Disease-A Survey."
[16] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554..
[17] El Naqa, Issam, and Martin J. Murphy. What is machine learning?. Springer International Publishing, 2015.
[18] Zhang, Cha, and Yunqian Ma, eds. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

ISSN No. : 2321-9653

IJRASET

International Journal for Research in Applied Science & Engineering Technology

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com

ISRA Journal Impact Factor: 7.429

45.98 INDEX COPERNICUS

THOMSON REUTERS Researcher ID: N-9681-2016

10.22214/IJRASET

TOGETHER WE REACH THE GOAL

SJIF 7.429

Certificate

It is here by certified that the paper ID : IJRASET58888, entitled

An Efficient Ensemble Machine Learning Model for Cardiovascular Disease Prediction Using Digital Health Records

by

B Sharanya

after review is found suitable and has been published in Volume 12, Issue III, March 2024

in

International Journal for Research in Applied Science & Engineering Technology

(International Peer Reviewed and Refereed Journal)

Good luck for your future endeavors

Editor in Chief, iJRASET

ISSN No. : 2321-9653

IJRASET

International Journal for Research in Applied Science & Engineering Technology

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com

ISRA Journal Impact Factor: 7.429

45.98 INDEX COPERNICUS

THOMSON REUTERS Researcher ID: N-9681-2016

10.22214/IJRASET

TOGETHER WE REACH THE GOAL

SJIF 7.429

Certificate

It is here by certified that the paper ID : IJRASET58888, entitled

An Efficient Ensemble Machine Learning Model for Cardiovascular Disease Prediction Using Digital Health Records

by

V. Srividya

after review is found suitable and has been published in Volume 12, Issue III, March 2024

in

International Journal for Research in Applied Science & Engineering Technology

(International Peer Reviewed and Refereed Journal)

Good luck for your future endeavors

Editor in Chief, iJRASET

ISSN No. : 2321-9653

IJRASET

International Journal for Research in Applied Science & Engineering Technology

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com

ISRA Journal Impact Factor: 7.429

45.98 INDEX COPERNICUS

THOMSON REUTERS Researcher ID: N-9681-2016

10.22214/IJRASET doi cross ref

TOGETHER WE REACH THE GOAL SJIF 7.429

Certificate

It is here by certified that the paper ID : IJRASET58888, entitled

An Efficient Ensemble Machine Learning Model for Cardiovascular Disease Prediction Using Digital Health Records

by

B. Prajnaya

after review is found suitable and has been published in Volume 12, Issue III, March 2024

in

International Journal for Research in Applied Science & Engineering Technology

(International Peer Reviewed and Refereed Journal)

Good luck for your future endeavors

Editor in Chief, iJRASET

ISSN No. : 2321-9653

iJRASET

International Journal for Research in Applied Science & Engineering Technology

IJRASET is indexed with Crossref for DOI-DOI : 10.22214

Website : www.ijraset.com, E-mail : ijraset@gmail.com

ISRA Journal Impact Factor: 7.429

45.98 INDEX COPERNICUS

THOMSON REUTERS Researcher ID: N-9681-2016

10.22214/IJRASET

TOGETHER WE REACH THE GOAL

SJIF 7.429

Certificate

It is here by certified that the paper ID : IJRASET58888, entitled

An Efficient Ensemble Machine Learning Model for Cardiovascular Disease Prediction Using Digital Health Records

by

Bandari Gayathri

after review is found suitable and has been published in Volume 12, Issue III, March 2024

in

International Journal for Research in Applied Science & Engineering Technology

(International Peer Reviewed and Refereed Journal)

Good luck for your future endeavors

Editor in Chief, iJRASET