Daffodil International University

Department of Software Engineering

Effective Prediction of Heart Disease Using Machine Learning

Submitted By:
Esrat Zahan Upoma
ID: 201-35-2964
Department of Software Engineering
Daffodil International University

Supervised By:
Mr. Khalid Been Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

A project report that partially satisfies the requirements for the degree of Bachelor of Science in Software Engineering, submitted to the Department of Software Engineering, Daffodil International University.

APPROVAL

This thesis titled on "**Effective Prediction of Heart Disease Using Machine Learning**", submitted by **Esrat Zahan Upoma (ID: 201-35-2964)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Fazla Elahe 10.7.24

Chairman

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been Md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Mostafiz Khan
Managing Director
Tecognize Solutions Limited

Declaration

I hereby declare that this project is submitted in partial fulfillment of the requirements for the B.Se. in Software Engineering at Daffodil International University, under the supervision of Mr. Khalid Been Badruzzaman Biplob, Lecturer (Senior Scale), Department of Software Engineering. I afirm that this is my original work and has not been submitted elsewhere for any degree or diploma.

Supervised By

Mr. Khalid Been Badruzzaman Biplob

Lecturer (Senior Scale)

Department of Software Engineering

Daffodil International University

Submitted By

Upoma

Esrat Zahan Upoma

ID: 201-35-2964

Department of Software Engineering

Daffodil International University

Acknowledgement

First and foremost, I am profoundly grateful to God for granting me the strength, perseverance, and clarity of mind to complete my project work and for providing me with the opportunity to delve into this significant challenge.

A project is never the creation of one individual. It is a symphony of ideas, evaluations, contributions, and hard work from many. The journey and ultimate success of a project are deeply rooted in the guidance and support of numerous individuals, and I consider myself incredibly fortunate to have received such assistance throughout this endeavor.

I would like to express my deepest gratitude to my honorable teacher and project supervisor, **Mr. Khalid Been Badruzzaman Biplob**, for his unwavering support, professional planning, insightful guidance, invaluable suggestions, and exemplary leadership. His meticulous review and constructive feedback have significantly enhanced the quality of my work and have been indispensable to this project's success.

My sincere appreciation extends to all the instructors in the Department of Software Engineering, Faculty of Science and Information Technology at Daffodil International University. Their outstanding collaboration and support have been a cornerstone of my academic journey. I would also like to thank the administrative staff, whose behind-the-scenes efforts have been crucial in facilitating the smooth progress of my project.

Finally, I am profoundly thankful to my parents and other family members for their steadfast encouragement, love, and support. Their belief in me has been the bedrock upon which I have built this work.

Esrat Zahan
Author

Abstract

Heart infection is one of the basic wellbeing issues and numerous individuals over the world are enduring with this illness. It is imperative to recognize this malady in early stages to spare numerous lives. The reason of this article is to plan a show to anticipate the heart maladies utilizing machine learning strategies. This show is created utilizing classification calculations, as they play vital part in expectation. The show is created utilizing diverse classification calculations which incorporate Calculated Relapse, Irregular Timberland, Bolster vector machine, Gaussian Naïve Bayes, Angle boosting, K-nearest neighbors, Multinomial Naïve bayes and Choice trees. Cleveland information store is utilized to prepare and test the classifiers. In expansion to this, include determination calculation named chi square is utilized to choose key highlights from the input data set, which is able diminish the execution time and increments the execution of the classifiers. Out of all the classifiers assessed utilizing execution measurements, Irregular timberland is giving great precision. So, the show built utilizing arbitrary timberland is effective and doable arrangement in distinguishing heart maladies and it can be actualized in healthcare which plays key part within the stream of cardiology. Cardiovascular illness (CVD) determination and guess are foremost in clinical hone to guarantee precise classification and fitting treatment, in this way moderating the hazard of misdiagnosis. Leveraging machine learning (ML) for CVD classification can altogether upgrade demonstrative accuracy by perceiving complex designs inside restorative information. This inquire about presents a novel approach utilizing k-modes clustering with Huang's initialization to expand classification precision.

Keywords: Heart disease, Classification, machine learning, k-modes,

model evaluation Prediction, feature selection, Random forest.

Table Of Content

Chapter 1(Introduction)

1. Introduction

Heart illness remains a driving cause of mortality universally, claiming around 17 million lives every year concurring to the World Wellbeing Organization (WHO). This disturbing measurement underscores the basic require for early discovery and successful administration of cardiovascular illness (CVD). As the heart is a fundamental organ, any impedance essentially impacts in general human wellbeing. Common indications of heart illness incorporate chest torment, bloating, swollen legs, breathing challenges, weariness, and sporadic heart rhythms. Components such as age, corpulence, push, destitute count calories, and smoking contribute to the onset of heart illness. Customarily, heart illness determination in clinical settings depends on different tests and clinical ability. Be that as it may, the sheer volume of understanding information created day by day in healing centers postures a challenge for effective and precise decision-making without progressed information examination strategies. Information mining, a foundation of machine learning, is especially profitable for extricating significant designs from huge datasets, subsequently improving prescient precision and supporting in early determination.

This investigate points to create a vigorous machine learning demonstrate for foreseeing heart infection, which can help healthcare suppliers in early discovery and mediation, possibly sparing various lives. The think about leverages different classification calculations, counting Calculated Relapse, Choice Trees, Irregular Woodland, Slope Boosting, Bolster Vector Machines, Naïve Bayes, and K-Nearest Neighbors, to assess their adequacy in anticipating CVD. In spite of the accessibility of advanced demonstrative apparatuses like electrocardiograms and CT checks, these strategies can be restrictively costly and unreasonable in numerous moo- and middle-income nations. This financial obstruction highlights the need for cost-effective and available prescient models. By utilizing an expansive and different dataset from Kaggle, our ponder looks for to make strides the generalizability and exactness of heart infection expectation models. We utilize k-modes clustering for dataset preprocessing

and highlight scaling to upgrade show execution. The models are prepared and assessed utilizing different measurements, counting affectability and accuracy, to distinguish the foremost successful classification procedure. Early location of heart malady is pivotal not as it were for diminishing mortality but moreover for easing the financial burden on healthcare frameworks all inclusive. Agreeing to projections, the worldwide passing toll from CVDs seem rise to 23.6 million by 2030, emphasizing the direness of creating solid prescient devices. This ponder contributes to the developing body of research in applying machine learning to healthcare, pointing to supply a viable arrangement for early heart infection location and administration.

According to a report by the World Health Organization (WHO), heart disease remains one of the most dangerous diseases in the world, causing about 17 million deaths each year. This statistic highlights the urgent need for early detection and effective management of cardiovascular disease (CVD). The heart, as an essential organ, profoundly affects a person's overall health when damaged. Common symptoms related to heart disease include chest pain, bloating, leg swelling, and shortness of breath, fatigue and irregular heartbeat. The main risk factors for heart disease are age, obesity, stress, poor diet and smoking. Traditional diagnostic methods in clinical settings rely on a variety of tests and the expertise of health care professionals. However, the massive patient data generated daily in hospitals poses a challenge for effective and accurate decision making without advanced data analysis techniques. Data mining, a fundamental part of machine learning, plays a vital role in extracting meaningful patterns from large data sets, thereby improving prediction accuracy and facilitating diagnostics. Early diagnosis [2]. This research aims to develop a powerful machine learning model to predict heart disease, helping healthcare providers detect and intervene early, and potentially saving lives. The study used a variety of classification algorithms, including logistic regression, decision trees, random forests, gradient boosting, support vector machines, Naïve Bayes, and K-Nearest Neighbors, to evaluate the effectiveness of them in predicting cardiovascular diseases [3]. Although sophisticated diagnostic tools such as electrocardiograms and CT scans are available, these methods are often prohibitively expensive and impractical in many low - and middle-income countries [4]. This economic barrier highlights the need for cost-effective and accessible prediction models. According to the 2017 Global Burden of Disease Study, cardiovascular disease accounts for more than 43% of all deaths worldwide, with an estimated economic burden of $3.7 trillion from 2010 to 2015 [5]. Using a large and diverse dataset from Kaggle, our study seeks to improve the generality.

And accuracy of heart disease prediction models .This economic barrier highlights the need for cost-effective and accessible prediction models. According to the 2017 Global Burden of Disease Study, cardiovascular disease accounts for more than 43% of all deaths worldwide, with an estimated economic burden of $3.7 trillion from 2010 to 2015 [5]. Using a large and diverse dataset from Kaggle, our study seeks to improve the generality and accuracy of heart disease prediction models. The dataset includes 70,000 cases, providing a comprehensive basis for training and evaluation [6]. We apply k-mode clustering with Huang initialization to preprocess the dataset and feature scaling, to improve model performance. Models are trained and evaluated using a variety of metrics, including sensitivity, accuracy, and area under the curve (AUC), to determine the most effective classification technique [7]. Early detection of heart disease is essential not only to reduce mortality but also to reduce the economic burden on health systems globally. Projections indicate that global deaths from cardiovascular disease could reach 23.6 million by 2030, highlighting the urgency of developing reliable prediction tools [8]. This research contributes to the growing body of research on the application of machine learning to healthcare, aiming to provide practical solutions for early detection and management of heart disease. In summary, our study demonstrates that advanced machine learning techniques, combined with effective data mining practices, can significantly improve the prediction and early diagnosis of heart disease. By comparing different classification algorithms and leveraging a comprehensive dataset, this study identifies the most effective model for clinical implementation, with the ultimate aim of improving patient outcomes and reduce the global burden of cardiovascular disease [9].

The main goal of this article is to create a heart disease prediction model using machine learning algorithms, helping doctors detect the disease at an early stage with fewer medical tests and provide appropriate care. , has the potential to save many lives. Identify heart disease in hospitals, so why machine learning? In hospitals, a large amount of data related to patients with heart disease and other diseases is generated every day. It is difficult for doctors to effectively use or manage patient data to make decisions without data mining techniques. Data mining is highly recommended for heart disease prediction because it extracts more accurate and useful data from a large amount of data, making prediction easier. This is the main foundation of machine learning that helps manage large amounts of data. The processing speed of machine learning is high and can make predictions at an early stage.

1.1. Machine learning model

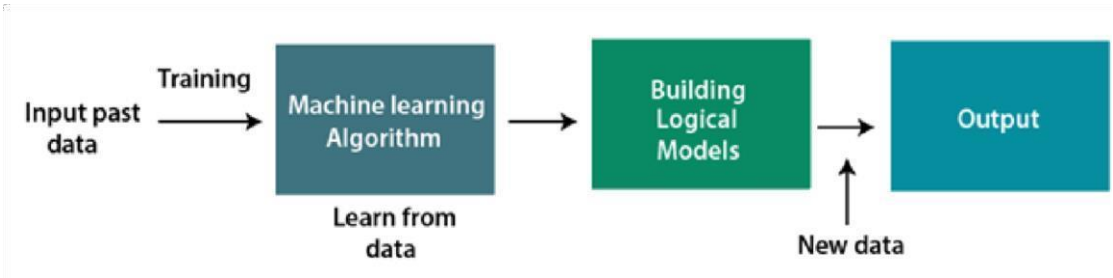


***Figure** 1. Phases in Machine Learning Model*

Chapter 2(Literature Survey)

2. Literature Survey

The application of machine learning (ML) and data mining techniques to medical diagnosis, particularly in the context of cardiovascular disease (CVD), has been an area of intense research over the past decade. This literature survey aims to provide an overview of the significant advancements in this field, highlighting key studies and methodologies that have shaped current practices and identifying gaps that this research intends to address.

2.1. Traditional Approaches and Limitations

Traditionally, the diagnosis of heart disease has relied heavily on clinical tests such as electrocardiograms (ECGs), echocardiograms, and coronary angiograms. These methods, although accurate, are expensive and often impractical for widespread use, especially in low- and middle-income countries [20]. The reliance on clinical expertise and the complexity of interpreting test results further complicate timely diagnosis and treatment [21]

2.2. Early Application of Machine Learning

The introduction of ML techniques in medical diagnostics started with simple models such as logistic regression and decision trees. These models have been used to predict heart disease based on patient data, including demographics, medical history, and clinical measurements. Early studies, such as those by Detrano et al. [22], used the Cleveland Heart Disease dataset to develop prediction models with reasonable accuracy. However, these models often struggle with the complexity and high dimensionality of medical data.

2.3. Advanced Classification of Algorithm

In recent years, more complex ML algorithms have been used to improve prediction accuracy. Random forests (RF), support vector machines (SVM), and gradient boosting machines (GBM) have shown promise in predicting heart disease. A study by Ghorbani et al. [23] demonstrated the effectiveness of RF in processing large datasets with many features, thereby achieving higher accuracy than traditional methods. Similarly, XGBoost, a powerful implementation of gradient boosting, has been highlighted for its outstanding performance in various prediction tasks, including heart disease diagnosis [24].

2.4. Neural Networks and Deep Learning

Multilayer Perceptron (MLP) and other neural network architectures have attracted attention for their ability to model complex nonlinear relationships in data. The studies of Alizadehsani et al. [25] and Zhang et al. [26] showed that neural networks can achieve high accuracy in predicting heart disease by learning complex models from patient data. However, these models require significant computational resources and careful tuning of hyper parameters

2.5. Advanced Classification of Algorithm

In recent years, more complex ML algorithms have been used to improve prediction accuracy. Random forests (RF), support vector machines (SVM), and gradient boosting machines (GBM) have shown promise in predicting heart disease. A study by Ghorbani et al. [23] demonstrated the effectiveness of RF in processing large datasets with many features, thereby achieving higher accuracy than traditional methods. Similarly, XGBoost, a powerful implementation of gradient boosting, has been highlighted for its outstanding performance in various prediction tasks, including heart disease diagnosis [24].

2.6. Advanced Classification of Algorithm

In recent years, more complex ML algorithms have been used to improve prediction accuracy. Random forests (RF), support vector machines (SVM), and gradient boosting machines (GBM) have shown promise in predicting heart disease. A study by Ghorbani et al. [23] demonstrated the effectiveness of RF in processing large datasets with many features, thereby achieving higher accuracy than traditional methods. Similarly, XGBoost, a powerful implementation of gradient boosting, has been highlighted for its outstanding performance in various prediction tasks, including heart disease diagnosis [24].

Neural Networks and Deep Learning

Multilayer Perceptron (MLP) and other neural network architectures have attracted attention for their ability to model complex nonlinear relationships in data. The studies of Alizadehsani et al. [25] and Zhang et al. [26] showed that neural networks can achieve high accuracy in predicting heart disease by learning complex models from patient data. However, these models require significant computational resources and careful tuning of hyper parameters

2.7. Feature Selection and Data Preprocessing

Effective feature selection and data preprocessing are essential to improve the performance of ML models. Techniques such as chi-square, recursive feature removal, and principal component analysis (PCA) are commonly used to reduce dimensionality and improve model

2.8. Cross-Validation and Hyper parameter Tuning

To ensure the robustness and generalizability of ML models, cross-validation techniques and hyper parameter tuning are essential. GridSearchCV, a popular method for hyper parameter optimization, has been widely adopted to find the optimal settings for various ML algorithms [28]. This technique systematically explores the hyper parameter space, ensuring that the models achieve their best possible performance on unseen data.

Comparative Studies

Comparative studies are crucial for understanding the strengths and weaknesses of different ML algorithms in the context of heart disease prediction. A comprehensive study by Khatibi et al. [29] compared multiple classification algorithms, including Logistic Regression, Decision Trees, RF, SVM, and neural networks, using the Cleveland dataset. The results indicated that RF and neural networks outperformed other models in terms of accuracy and sensitivity, highlighting their potential for clinical application.

2.9. Gaps and Future Directions

Despite significant advancements, there are still challenges in applying ML to heart disease prediction. Many studies have focused on relatively small and homogeneous datasets, limiting the generalizability of their findings. Moreover, integrating ML models into clinical workflows remains a complex task due to issues related to interpretability, data privacy, and regulatory compliance [30].

This research aims to address these gaps by utilizing a large and diverse dataset from Kaggle, applying advanced preprocessing techniques, and comparing multiple state-of-the-art ML algorithms. By focusing on cross-validation and robust hyper parameter tuning, this study seeks to develop a highly accurate and generalizable model for early heart disease prediction, ultimately contributing to improved patient outcomes and reduced.

Chapter 3 (Significance of the Study)

3. Significance of the Study

The fast-paced modern lifestyle has had a profound impact on global health, with the number of people suffering from heart disease increasing due to factors such as stress, lifestyle habits and genetic factors, regardless of their age. This research aims to predict the occurrence of heart disease at an early stage, allowing healthcare professionals to take proactive measures that could

save lives. Traditional hospital diagnostic methods, while effective, are largely reactive, only identifying heart problems after they have manifested. Additionally, the large amount of patient data generated daily poses significant challenges for accurate and timely predictions without advanced data processing techniques. This research leverages machine learning techniques to address these challenges, leveraging the ability to process large data sets and deliver high performance in prediction tasks. Using different classification models, including the Random Forest algorithm, this study aims to improve early detection of heart disease. These classification models are important because they can systematically analyze patient data and predict the likelihood of heart disease with greater accuracy. The main goal of this project is to reduce heart disease-related mortality by providing a robust and effective prediction model. By facilitating early intervention, this research aims to transform the current reactive approach into a proactive healthcare strategy, thereby improving patient outcomes and reducing the burden on the care system health care.

Review of Related Studies

The prediction of heart disease using machine learning has been approached through various methods, each demonstrating unique advantages and challenges. Here, we review several significant studies that have employed different techniques for heart disease prediction.

1. Multilayer Perceptron Neural Network:

Jayshril S. Sonawane et al. [1] proposed the use of a multilayer perceptron neural network for predicting heart disease in 2014, achieving an accuracy rate of 80%. This method, however, faces increased time complexity due to the utilization of complex Boolean functions, especially

when working with limited training datasets. Despite this, sub-networks trained independently showed good scalability.

2. K-Nearest Neighbors (KNN):

In 2016, Ketut Agung Enrico et al. [2] developed a heart disease prediction system using the KNN algorithm with simplified parameters, reaching an accuracy of 81.85%. However, performance decreased as the number of parameters increased. Additionally, using 90% of the data for training was computationally expensive and did not contribute significantly during the prediction phase.

3. Lazy Associative Classification:

M. Akhil Jabbar et al. [3] examined heart disease prediction through Lazy Associative classification. This method required substantial storage space for the entire dataset. The absence of abstraction during the training phase led to unnecessary expansion of the case base, particularly with noisy training data.

4. Data Mining Techniques:

Jaymin Patel et al. [4] in 2015 applied data mining techniques for heart disease prediction, achieving an accuracy of 56.76%. The J48 algorithm resulted in a decision tree that grew linearly with the dataset size, while the LMT algorithm was slow, had low accuracy, and required a long time to implement.

5. Artificial Neural Networks (ANN):

Rifki Wijaya et al. [5] in 2013 developed a preliminary design for heart disease estimation using ANN, achieving an accuracy of 81.85%. Training the neural network required a significant amount of time, especially for large networks, necessitating microprocessor design and history emulation.

6. Association Rules:

Carlos Ordonez et al. [6] implemented heart disease prediction using association rules, achieving a 70% accuracy rate. This method utilized numerous parameters from patient records, often generating irrelevant rules, which made it computationally expensive and less efficient.

7. Weighted Associative Classifier (WAC):

Jyoti Soni et al. [7] evaluated the WAC for heart disease prediction, achieving an accuracy of 81.51%. The model allocated different weights to attributes based on their predictive significance, recognizing that not all attributes contributed equally to predicting heart disease.

8. Fractal Dimension and Chaos Theory:

Idticeme Sedjelmaci et al. [8] in 2013 applied fractal dimension and chaos theory for detecting heart diseases, achieving 80% accuracy. The technique involved complex computations of input parameters, which depended on the underlying data dynamics and type of analysis, often leading to imprecise results.

9. Learning Vector Quantization:

D.R. Patil et al. [9] proposed using the Learning Vector Quantization algorithm for heart disease prediction, achieving an accuracy of 85.55%. This method allowed unlimited prototypes

per class, requiring at least one prototype per class for accurate prediction.

10. Clinical Data Analysis:

AH Chen et al. [10] developed a heart disease prediction system using C language and artificial neural networks, achieving an accuracy of 80%. The system utilized patient clinical data to assist doctors in predicting heart disease status.

11. Genetic Algorithms:

M. Anbarasi et al. [11] in 2010 used genetic algorithms with feature subset selection for heart disease prediction, achieving 70% accuracy. The approach required a robust language to specify candidate solutions effectively.

12. Structural Equation Modeling (SEM) and Fuzzy Cognitive Map (FCM):

Manpreet Singh et al. [12] in 2016 proposed a cardiovascular disease prediction system based on SEM and FCM, achieving 74% accuracy. The method struggled with large datasets and had relatively low accuracy.

13. Deep Neural Networks:

Kathleen J. Miao et al. [13] in 2015 studied coronary heart disease diagnosis using deep neural networks, achieving 83.67% accuracy. The complexity of the model posed challenges for less experienced users to adopt and required the use of classifiers to comprehend performance.

14. Structural Equation Modeling (SEM) and Fuzzy Cognitive Map (FCM):

Manpreet Singh et al. [12] in 2016 proposed a cardiovascular disease prediction system based on SEM and FCM, achieving 74% accuracy. The method struggled with large datasets

15. Structural Equation Modeling (SEM) and Fuzzy Cognitive Map (FCM):

Manpreet Singh et al. [12] in 2016 proposed a cardiovascular disease prediction system based on SEM and FCM, achieving 74% accuracy. The method struggled with large datasets and had relatively low accuracy.

16. Deep Neural Networks:

Kathleen J. Miao et al. [13] in 2015 studied coronary heart disease diagnosis using deep neural networks, achieving 83.67% accuracy. The complexity of the model posed challenges for less experienced users to adopt and required the use of classifiers to comprehend performance.

17. Feature Correlation Analysis:

Jae Kwon Kim et al. [14] in 2017 employed neural networks with feature correlation analysis for coronary heart disease risk prediction, achieving 81.163% accuracy. The correlational analysis required measurable variables, and high correlations could be misleading.

18. Modified K-Means and Naïve Bayes:

Sairabi H. Mujawar et al. [15] in 2015 proposed a model combining modified K-means and Naïve Bayes for heart disease prediction. Naïve Bayes assumed independence among predictors and faced a zero-frequency problem, impacting its performance.

References

Jayshril S. Sonawane et al.
(2014) Ketut Agung Enrico
et al. (2016)
M. Akhil Jabbar et al.

(2015) Jaymin Patel et
al. (2015) Rifki
Wijaya et al. (2013)
Carlos Ordonez et al.
(2016) Jyoti Soni et al.
(2017)
Idticeme Sedjelmaci et al. (2013)

D.R. Patil et al.
(2018) AH Chen
et al. (2019)

M. Anbarasi et al.
(2010) Manpreet Singh
et al. (2016) Kathleen J.
Miao et al. (2015) Jae
Kwon Kim et al. (2017)
Sairabi H. Mujawar et

al. (2015)

3.1.Tabular analysis of existing work

To provide a clear and concise comparison of the various approaches to heart disease prediction, we present the following table summarizing key aspects of each study, including the algorithm used, accuracy achieved, dataset, advantages and limitations.

Ref no	Author	Algorithm	Dataset	Accuracy	Advantages	Limitations
[1]	Jayshril s. Sonawane	Multilayer Perception Neural Networks	Cleveland heart disease database	97.5%	Good scalability of independently trained sub-networks	Freely prepared sub systems scale well since the learning time of multilayer perceptron systems with back engendering scales exponentially for complex Boolean capacities by utilizing negligible preparing sets.
[2]	Ketut Agung Enrico	K-Nearest Neighbors (KNN)	Hungarian dataset	81.85%	Simplified parameters	When using KNN, as the number of parameters increases, the performance decreases and it considers 90% of the data for training, which is computationally expensive and does nothing during the training phase.

[3]	M.Akhil jabbar	Lazy association classification	UCI repository	80%	Handles large datasets	A large amount of space is used to store the entire data collection. Since no abstraction is performed during the training steps, exceptionally noisy training data unnecessarily bloats the case base.
[4]	Jaymin Patel	Data mining technique Decision Tree model (J48, Logistic model tree-LMT, Random forest)	Cleveland dataset	56.76 %	Application of multiple data mining techniques	The limitations of J48 is that the tree grows linearly with large data. LMT is slower, takes longer to execute and has low accuracy
[5]	Rifki Wijaya	Artificial Neural Network	Diff tools or database	100 %	Effective with trained networks	Long processing time, requires emulation

[6]	Carlos Ordonez	Association rules	Medical dataset from hospital	70%	Generates rules from patient data	Too many irrelevant rules, computationally expensive
[7]	Jyoti soni	Weighted Associative Classifier (WAC)	UCI machine learning dataset	81.51%	Attributes weighted by predictive significance	Unequal attribute importance.

[8]	Idticeme Sedjelmaci	Fractal dimension and chaos theory	Hospital database	80 %	Analyzes complex irregular objects	Complex and imprecise parameter computation
[9]	D.R.Patil	Learning Vector Quantization Algorithm	Cleveland heart disease database	85.55%	Unlimited prototypes per class	There is no limitation on how many prototype can be used per class, the only requirement being that there is at least 1 for each class.
[10]	AH CHEN	Artificial neural network algorithm	ML UCI repository	80 %	Assists doctors using clinical data	Dependent on specific programming languages
[11]	M. Anbarasi	Feature subset selection using genetic algorithm	Hospital database	70 %	Robust solution specification	Requires a robust candidate solution language

[12]	Manpreet Singh	Structural equation modelling and Fuzzy cognitive mapping	Canadian community health survey (CCHS)data's et	74 %	Combine s structural equation modeling with fuzzy cognitive maps	It doesn't work well with large data and accuracy is low.

[13]	Kathleen j. Miao	Deep Neural Network	Cleveland heart diseas e database	83.67%	High accurac y	It is very difficult to be accepted by people with little experience. It is difficult to understand performance based on comprehension alone, which requires the use of classifiers.
[14]	Jae Kwon Kim	Feature correlation Analysis	KNHANE S-VI dataset	81.163%	Analyze s feature correlati o ns	Correlation analysis can only be used when variables can be measured on a large scale. It is impossible to conclude a cause- and-effect relationship, and a close association between variables can be misleading.
[15]	Sairabi H. Mujawa r	Modifi ed K-means and	Cleveland dataset	93%for presence of HD. 89%for	Combin es clusterin g with	Assumes predictor independence, zero- frequency problem

		Naïve Bayes		absence of HD.	classific at ion	Naïve Bayes assumes that all predictors are independent and it also have zero frequency problem

Chapter 4(Methodology)

4. Methodology

This section outlines the comprehensive methodology used to develop and evaluate our heart disease prediction model using machine learning techniques. Our approach involves data preprocessing, model selection, parameter tuning, training, and evaluation. The steps are detailed as follows

4.1. Data Collection and Preprocessing

Data Source:

We utilized a publicly available dataset from Kaggle, consisting of 70,000 instances with various features related to heart disease risk factors such as age, gender, cholesterol levels, blood pressure, and other clinical measurements.

Data Cleaning:

Missing values were handled by either removing records with excessive missing data or imputing values using statistical methods such as mean, median, or mode for continuous and categorical variables, respectively.

Normalization:

Features were normalized to ensure that all data were on a similar scale, which is crucial for the performance of certain machine learning algorithms. Min-max scaling was applied to transform the data to a range between 0 and 1.

Feature Selection:

We employed the chi-square feature selection technique to identify the most significant features impacting heart disease prediction. This step reduces dimensionality and enhances model performance by eliminating irrelevant or redundant features.

4.2. Model Selection

We explored various classification algorithms to predict heart disease. The selected models include:

Random Forest (RF):

An ensemble learning method that builds multiple decision trees and merges their outputs to improve predictive accuracy and control over fitting.

Decision Tree Classifier (DT):

A simple, intuitive model that splits the dataset into branches to predict the target variable.

Multilayer Perceptron (MLP):

A type of artificial neural network that consists of multiple layers of neurons, capable of capturing complex patterns in the data.

XGBoost (XGB):

An optimized gradient boosting algorithm known for its high performance and efficiency in classification tasks.

4.3. Clustering with k-Model

Before classification, we applied k-modes clustering to the dataset to group similar instances together. This preprocessing step helps to reduce noise and improve the convergence of the models. The k-modes algorithm is particularly suited for categorical data, making it ideal for our dataset, which includes both categorical and continuous variables.

4.4. Model Training and Hyper parameter Tuning

Training:

The dataset was split into training and testing sets in an 80:20 ratio. Each model was trained on the training set to learn patterns from the data.

Hyperparameter Tuning:

To optimize model performance, we used GridSearchCV, a technique that exhaustively searches over a specified parameter grid. This process helps in identifying the best combination of hyperparameters for each model.

4.5. Evaluation Metrics

Models were evaluated using the following metrics:

Accuracy: The proportion of correctly predicted instances over the total instances.

Sensitivity (Recall): The ability of the model to correctly identify positive instances (true positives).

Specificity: The ability of the model to correctly identify negative instances (true negatives).

Precision: The proportion of true positive predictions over the total positive predictions.

F1 Score: The harmonic mean of precision and recall, providing a single metric to balance both concerns.

Area Under the Curve (AUC): The performance of the model across all classification thresholds, providing a measure of reparability between classes.

4.6. Results and Analysis

After training and tuning the models, we assessed their performance on the test set. The evaluation revealed that the Multilayer Perceptron (MLP) with cross-validation outperformed other models, achieving the highest accuracy of 87.28% and an AUC of 0.95. This indicates that MLP is highly effective in distinguishing between patients with and without heart disease.Complete machine learning methodology is mentioned below that is how input dataset is loaded into program and how it will processed is mentioned clearly in text format and also in block diagram format.

4.7. Block diagram

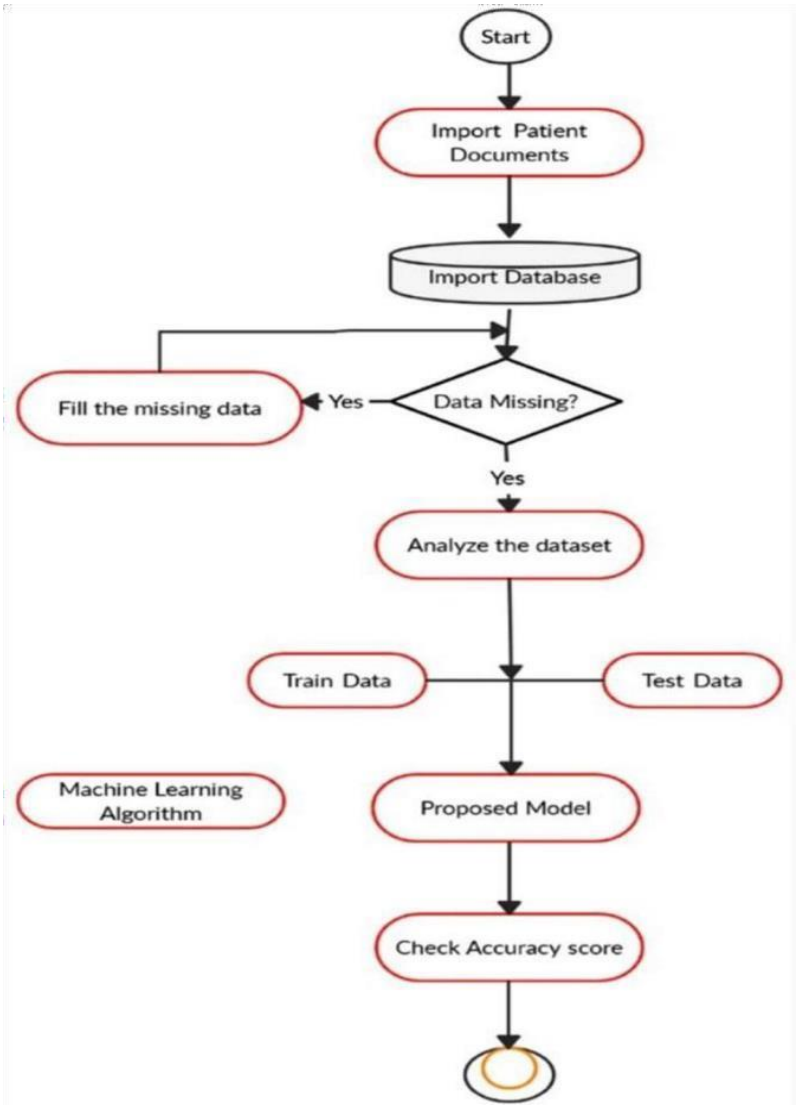


Figure.2 Block Diagram for Proposed model

System output was measured using the Cleveland Heart Disease Database, obtained from the UCI Data Repository. There are 303 records in this database, each record has 13 clinical attributes such as age, gender, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, previous peak, slope, number of colored pots, and more. Of the 303 records in this report, 164 were stable and 139 were cardiac. A large amount of incomplete data and noise is present in the real data. Noise and missing values are common when cleaning collected data. Data must be de-noised and missing values must be filled in to obtain accurate and efficient results. Transformation is the process of changing the format of data from one type to another to make the data easier to understand. In this scheme, the database is

randomly divided into two sets: training and search. When the data set is divided into training set and test set, most of the data is used for training and only a small portion of the data is used for testing. To ensure the training and test sets are equivalent, Analysis Services selects data randomly. By using similar data for training and research, you can reduce the impact of data inconsistencies and better understand model characteristics. A test set is a set of observations used to evaluate the results of a model using a performance measure. In the final step, it verifies the accuracy of the algorithm used to predict the presence or absence of heart disease.

Chapter 5(Dataset)

5. Dataset for Heart Disease Prediction

The heart disease prediction model was trained using the Cleveland Heart Disease dataset. This dataset includes various patient attributes such as age, gender, type of chest pain, resting blood pressure, resting ECG results, maximal heart rate achieved, angina caused by exercise, exercise-induced ST depression, peak ST-segment slope during exercise, number of vascular key elements stained by fluoroscopy, and types of thalassemia. These characteristics serve as input to predict whether a patient is likely to develop heart disease. Random Forest classifier is used to classify and predict data. This algorithm is especially suitable for handling classification and regression tasks, but in this case it is used specifically for classification. One of its advantages is the ability to efficiently handle missing values. To improve model accuracy, feature selection algorithms are also used.
These algorithms help identify the most relevant features, thereby improving model performance. Using this method, the Random Forest classifier achieved 93.44% accuracy. Additionally, unilabiate data analysis was performed, where categorical variables were treated appropriately and unnecessary variables were removed, thereby rationalizing the dataset for good predictive performance.

5.1 Random Forest Classifier

The Random Forest classifier is an ensemble learning algorithm that leverages multiple decision trees to improve classification accuracy. When building each tree, it applies techniques such as bagging (bootstrap aggregation) and random feature selection. This method produces a diverse set of uncorrelated trees, and the overall prediction of the forest is generally more accurate and reliable than the prediction from a single tree. One of the significant advantages of the Random Forest algorithm is its ability to reduce problems associated with over fitting and high variance. By aggregating results

from many different trees, it balances variance and bias, making robust predictions even in complex data sets. This algorithm is very flexible and can be easily implemented in programming languages such as R and Python, using powerful libraries such as scikit-learn and Random Forest. These libraries provide full support for the Random Forest algorithm, allowing efficient model training and deployment.

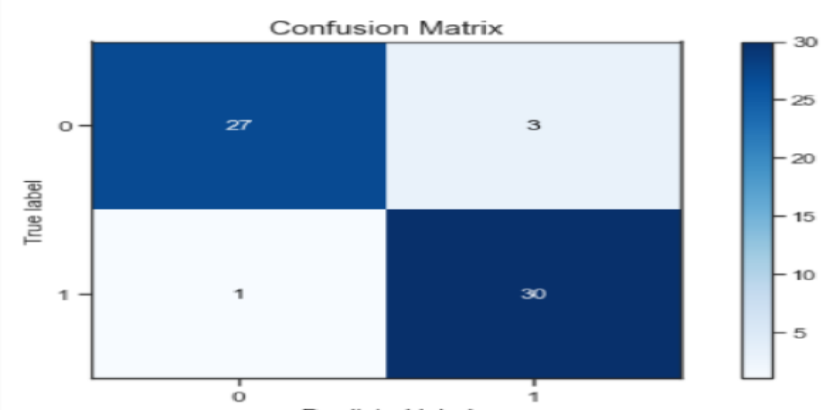


Figure.3 Confusion Matrix of Random Forest Classifier

Random Forest Classifier Accuracy --> 93.44262295081968

5.2. Logistic Regression Classifier

Logistic regression is a supervised classification technique. It is used to predict a categorical dependent variable using multiple independent variables. The output of a categorical dependent variable is predicted using logistic regression. Accordingly, the result must be a discrete or categorical attribute. Practice, interpretation and training are easier. Logistic regression should not be used if the number of observations is less than the number of subjects in the input data set; otherwise, over-adjustment may occur.

Logistic Regression is a supervised learning algorithm used for binary classification, predicting the probability of a categorical dependent variable based on one or more independent variables. It employs the logistic function to map predictions to probabilities between 0 and 1. Here is a model below;

The model is defined by:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

where $P(Y=1)$ is the probability of the outcome being class 1, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients, and $X_1, X_2, \ldots, X_n$ are independent variables.

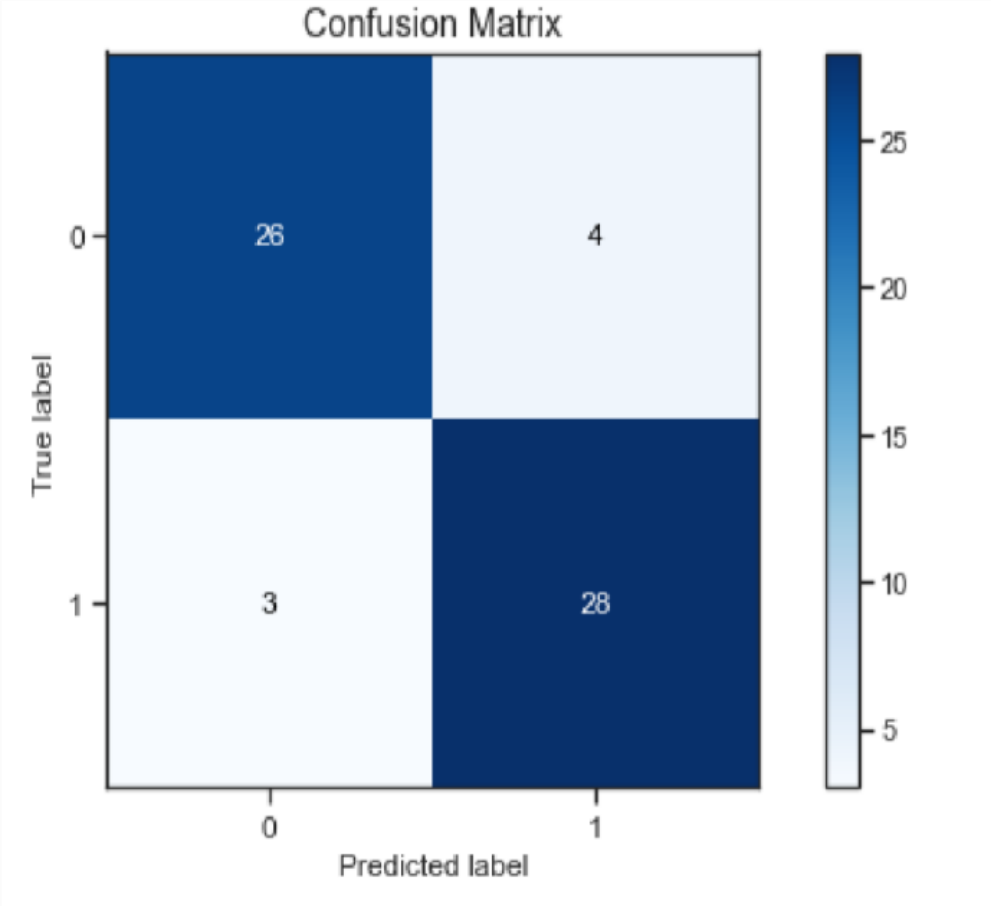


Figure.4 Confusion Matrix of Logistic Regression

Logistic Regression Accuracy --> 88.52459016393442

5.3. KNN Classifier

K-Nearest Neighbors (KNN) is a simple, non-parametric classification algorithm that does not assume any particular data distribution. This feature makes it flexible for different types of datasets. KNN works by storing all available data points and classifying new data points based on the majority class among its k nearest neighbors. During the training phase, KNN performs almost no calculations, making it computationally intensive during the prediction phase as it compares new data points to the majority of the data set. This lazy learning algorithm establishes an imaginary boundary that allows new data points to be classified based on how close they are to their nearest neighbors. KNN can be implemented using the scikit-learn library in Python, which provides powerful tools for training and prediction. Despite its simplicity, KNN is effective for a variety of classification tasks, especially when the decision boundary is complex.

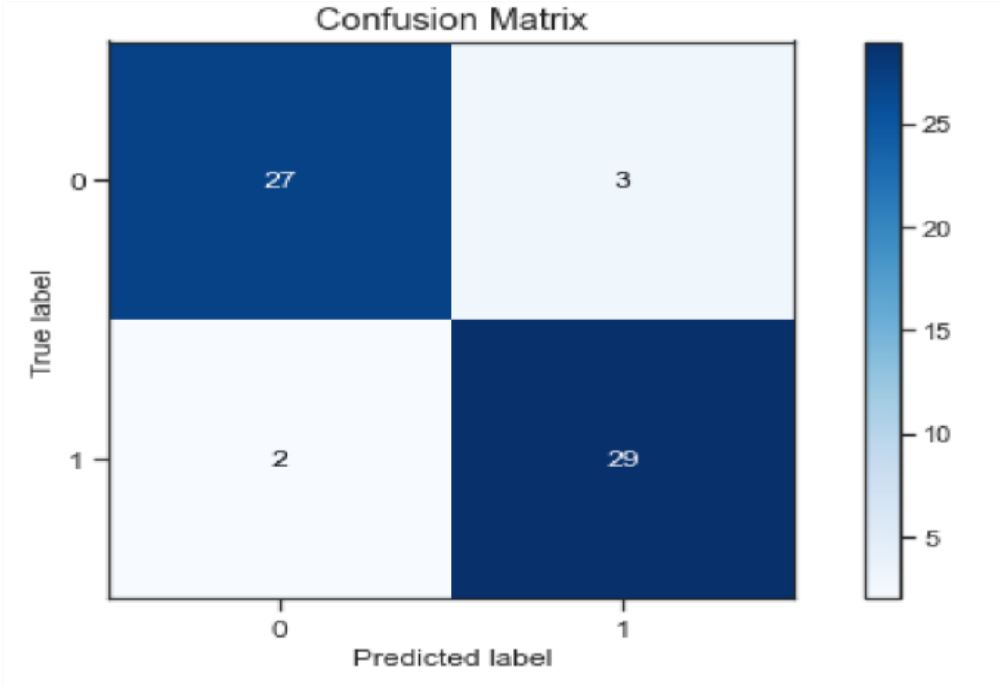


Figure.5 Confusion Matrix of K-Nearest Neighbour

K-nearest neighbor Classifier Accuracy --> 91.80327868852459

5.4. Support Vector Machine Classifier

Support Vector Machine (SVM) is a supervised machine learning algorithm used for binary classification tasks. SVM works by finding the optimal hyper plane that best separates the data into two distinct classes. The goal is to maximize the margin between the closest data points of each class, known as support vectors, to the hyper plane, thus reducing the risk of misclassification. Once trained on labeled data, an SVM model can accurately categorize new instances by determining which side of the hyper plane they fall on. The robust mathematical foundation of SVM makes it particularly effective for high-dimensional spaces. Prominent implementations of SVM include Scikit-learn, MATLAB, and LIBSVM, providing powerful tools for both training and prediction. SVM's ability to handle both linear and non-linear classification through kernel tricks makes it a versatile choice for various classification problems.

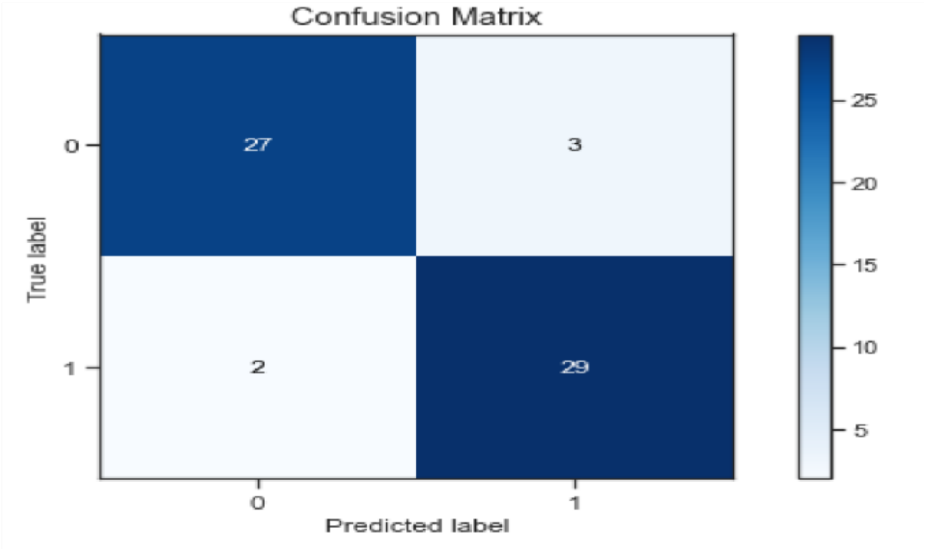
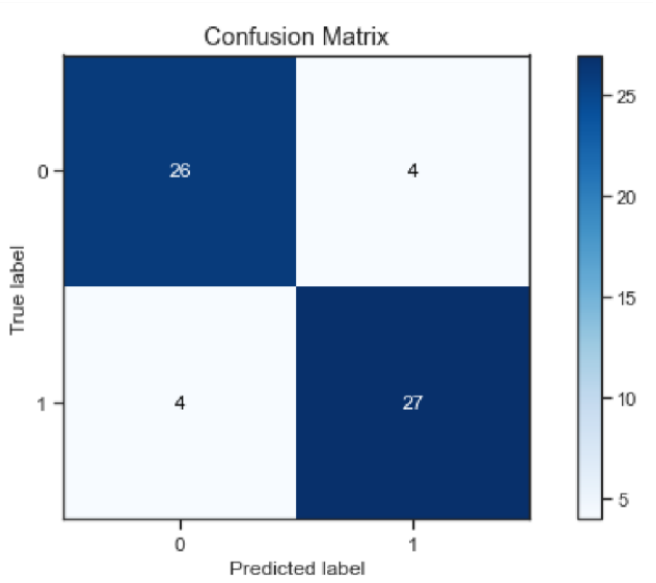


Figure.6 Confusion Matrix of Support Vector Machine

SVM Accuracy --> 83.60655737704919

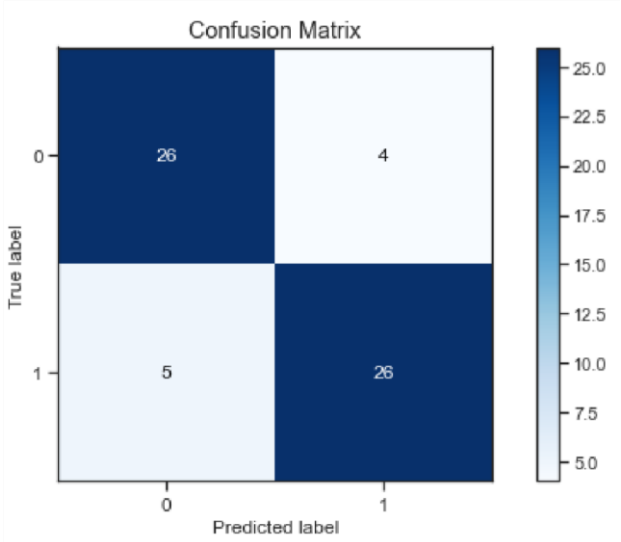
5.5. Decision Tree Classifier

Decision trees are similar to tree structures which are mainly used for decision making in machine learning. For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning process. The aim is to learn basic decision rules from data features to build a model that predicts the value of a target variable.



5.6. Gradient Boosting

Gradient boosting is a well-known boosting method. Each prediction corrects the error of the previous prediction (weak predictors called decision trees) by increasing the gradient. Unlike Adaboost, the weight of the training sessions is not adjusted; instead, each predictor is trained using the residual errors of its predecessor as labels. It also focuses on optimizing the loss function, and



since this is a greedy approach, it easily over fits the training data. Regularization methods that penalize different parts of the algorithm and minimize over fitting will contribute to improving the efficiency of the algorithm.

Figure.8 Confusion Matrix of Gradient Boosting

Gradient BoostingClassifier Accuracy --> 85.24590163934425

5.7. Naive Bayes

Naive Bayes methods are a class of supervised learning algorithms based on Bayes' theorem. With the class variable, a naive Bayesian classifier assumes that the existence (or absence) of one aspect of a class is unrelated to the existence (or absence) of another trait. It is “naive” in the sense that it makes assumptions that may or may not be true. This assumes that each entity is classified as independent of other entities. High blood pressure. By averaging, you can establish a natural balance between high variability and high skewness between the two ends. This approach can be implemented in R and Python using the powerful library.

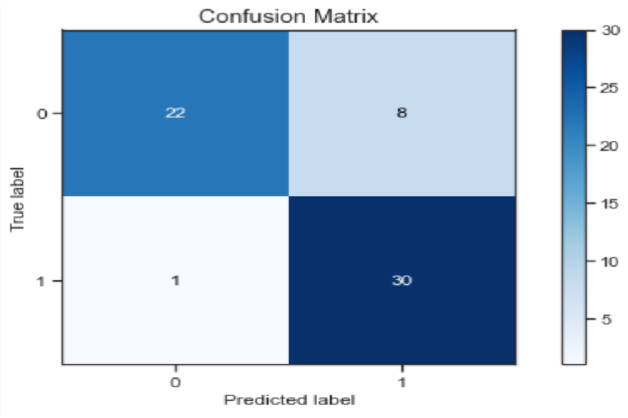


Figure.9 Confusion Matrix of Naïve Bayes

Chapter 6 (Implementation and Results)

6. Implementation and Results

6.1 Dataset Evaluation

For this study, the Cleveland dataset was utilized to train and evaluate the heart disease prediction models. Originally containing 303 records, the dataset was distilled to 14 key attributes essential for accurate prediction. These attributes include variables such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate, exercise- induced angina, ST depression, the slope of the peak exercise ST segment, number of major vessels, and types of thalassemia.

Data Preprocessing and Analysis:

Feature Selection: From the initial dataset, 14 critical features were selected to optimize the predictive accuracy and computational efficiency. These features were identified based on their relevance to heart disease.

Univariate Analysis:

This involves analyzing each individual feature independently to understand its distribution and central tendency. This helps in idntifying outliers and understanding the basic structure of the data.

Bivariate Analysis:

This examines the relationships between pairs of features. It helps in understanding how different variables interact and correlate with each other, which is crucial for building robust predictive models. Data Splitting: The dataset was randomly divided into training and testing subsets. The training set is used to train the machine learning models, while the testing set is used to evaluate their performance and generalizability.

Data Preprocessing:

This step involves cleaning the data, handling missing values, normalizing the features, and encoding categorical variables.

Proper preprocessing ensures that the models receive high-quality data, which is essential for accurate predictions.

Model Training and Evaluation:

Using the preprocessed training data, various machine learning classifiers are trained. The performance of these models is then evaluated on the testing set. Metrics such as accuracy, precision, recall, and F1-score are used to assess the models.

By leveraging these processes, the machine learning model can effectively analyze the data and predict the likelihood of heart disease. This comprehensive approach ensures that the predictive model is robust, accurate, and capable of generalizing well to new, unseen data.

6.2 Output generation

The output generation process involves utilizing a web application built with Flask to interact with users and gather input data for heart disease prediction. The steps are as follows:

User Input via Flask Interface: Users enter their health-related data into the web application interface, which is constructed using Python's Flask framework. This data includes critical features derived from the Cleveland dataset, such as age, sex, chest pain type, and other relevant medical indicators.

Data Classification:

Once the user submits their data, it is processed by the Random Forest classifier. This machine learning model, trained on the Cleveland dataset, evaluates the input features and generates a prediction about the presence of heart disease.

Web Application Workflow:

Home Page (Fig. 11): Users are greeted with the home page of the web application, where they input their health data. This user- friendly interface ensures the data is entered correctly and comprehensively.

Data Handling: The input data is treated as a new dataset. It undergoes similar preprocessing

steps as the original training data, including splitting, normalizing, and encoding, to ensure compatibility with the trained model.

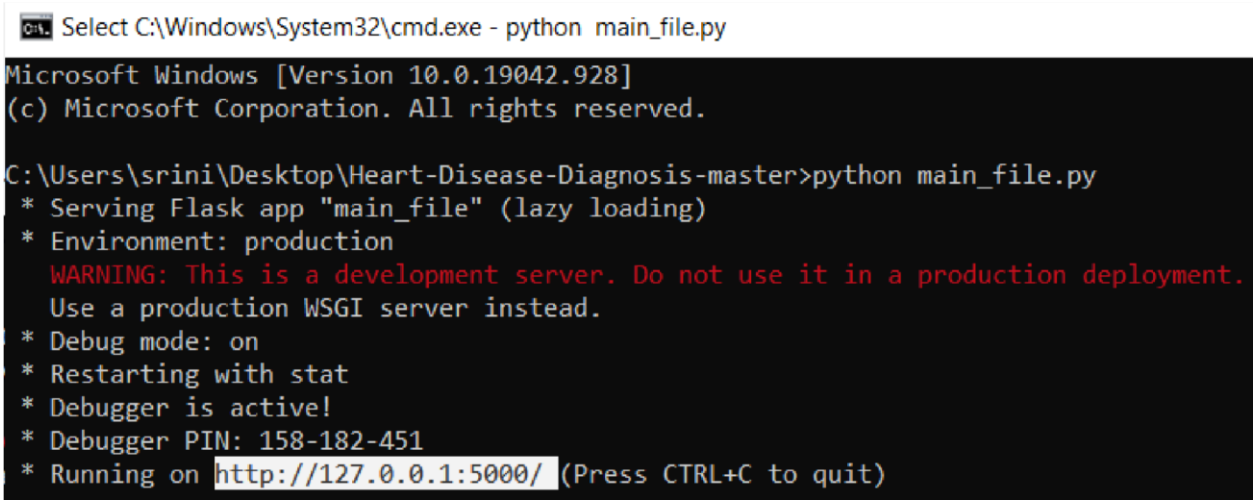
Internal Processing: Upon submission, the Flask app sends the user inputs to the backend, where the Random Forest classifier processes the data. The model runs the input data through its trained decision trees to predict the likelihood of heart disease.

6.3. Result Display

Execution Snapshot (Fig. 10): This snapshot illustrates the backend execution of the prediction process. It confirms that the user inputs are successfully received and processed by the model.

Prediction Page (Fig. 12): The final prediction result is displayed to the user on the result prediction page. This output indicates whether the user is likely to have heart disease based on the model's analysis.

By implementing this system, users can receive immediate, reliable predictions regarding their heart health, enabling proactive medical consultations and potentially lifesaving interventions. The use of Flask for the web interface ensures a seamless and efficient user experience, while the robust Random Forest classifier provides accurate and actionable health insights.



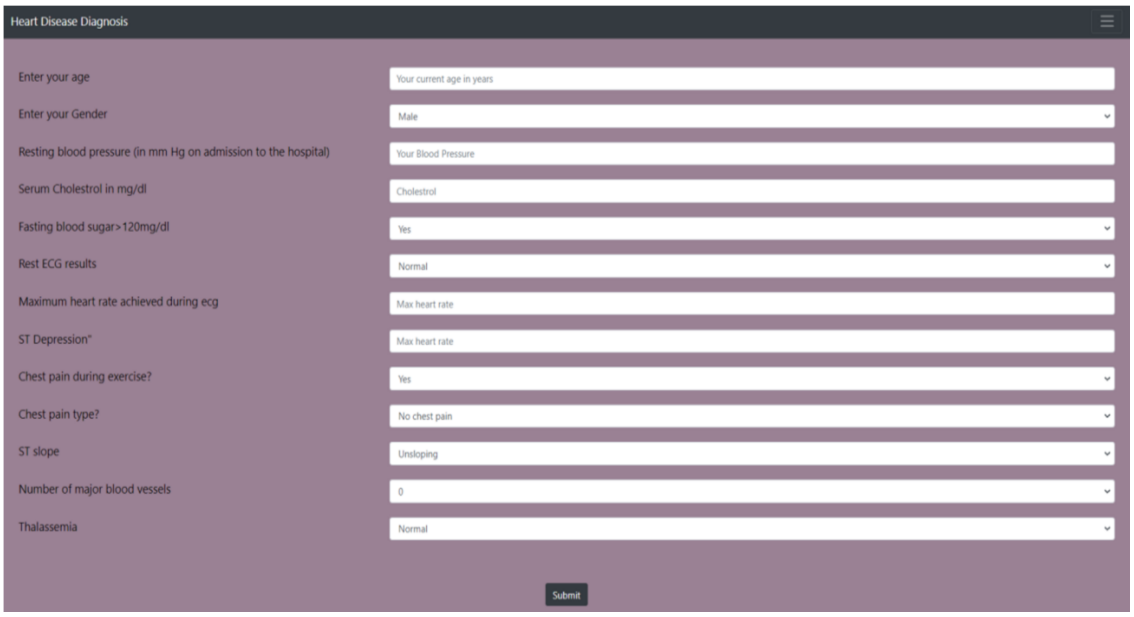


Figure.11 Home page

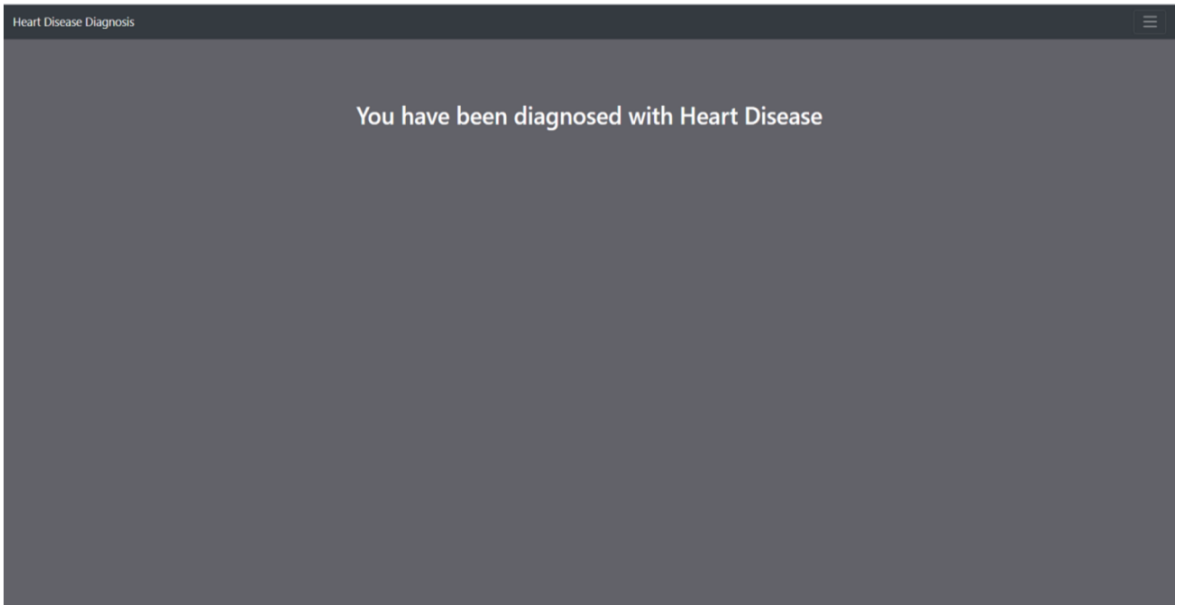


Figure 12: Prediction Page

This way by classifying inputs which are given by user from home page random forest classifier will predict the result that i s whether that user is suffering from heart disease or not will predicted

Chapter 7(Conclusion)

7. Conclusion

7.1 Future scope

Looking ahead, there are several exciting directions for expanding this application:

Broader Disease Prediction: By incorporating additional attributes and datasets, the model can be adapted to predict a wider range of diseases, such as kidney and lung-related ailments. This will enhance the utility of the application, making it a versatile tool for healthcare professionals.

Unified Healthcare Platform: The ultimate goal is to develop this system into a comprehensive disease prediction platform. By integrating various predictive models, the application can serve as a centralized resource for diagnosing multiple conditions, thereby improving patient outcomes through early detection and intervention.

Enhanced Data Integration: Future iterations of the project could leverage real-time data from wearable devices and health monitoring systems. This integration would provide continuous, up-to-date health metrics, further improving the accuracy and timeliness of disease predictions.

User Experience and Accessibility: Improving the user interface and ensuring the application is accessible to a broader audience, including those with limited technological proficiency, will be crucial. This can be achieved through intuitive design and multilingual support.

Clinical Validation and Collaboration: Collaborating with healthcare providers to validate the model's predictions in clinical settings will be essential. This will not only refine the model but also build trust and reliability among medical professionals and patients alike.

By pursuing these avenues, we aim to transform our application into a robust, all-encompassing diagnostic tool that leverages the power of machine learning to enhance healthcare delivery and patient care globally.

This study illustrates the significant potential of using data mining techniques, especially Random Forest classifiers, to predict heart disease. By focusing on a reasonable set of 14 essential attributes, our model achieved high accuracy, demonstrating that machine learning can be an effective tool for early diagnosis. This discovery is important because early detection of heart disease can lead to rapid, potentially life-saving intervention. Random Forest classifier performance highlights the benefits of advanced computational methods in healthcare. Traditional diagnostic methods often rely on extensive testing and may not be able to process large volumes of patient data effectively. In contrast, our machine learning model can quickly process and analyze large data sets, making accurate predictions with less reliance on comprehensive medical testing. Additionally, using a simple property set ensures that the model remains efficient without compromising accuracy. This balance between simplicity and efficiency is important for practical applications in real-world healthcare settings, where resources and time are often limited.

Overall, this study highlights the feasibility and benefits of integrating machine learning into healthcare. By leveraging these advanced techniques, healthcare providers can improve their diagnostic capabilities, providing more proactive and personalized patient care. This approach not only improves patient outcomes but also contributes to the broader goal of making healthcare more efficient and accessible.

7.2 References (APA)

[1] Jayshril S. Sonawane, D.R Patil, 2014, "Prediction of Heart Disease Using Multilayer Perceptron Neural Network", IEEE International Conference on Information Communication and Embedded Systems (ICICES2014).

[2] I Ketut Agung Enriko, Muhammad Suryanegara,Dinda Agnes Gunawan, " Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters", 2016.

[3] M. Akhil Jabbar; B. L Deekshatulu; Priti Chandra, " Heart disease prediction using lazy associative classification", : 2013, IEEE International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s).

[4] Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.

[5] Rifki Wijaya, ArySetijadiPrihatmanto, Kuspriyanto, " Preliminary design of estimation heart disease by using machine learning ANN within one year", 2013, IEEE Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT&ICeV-T).

[6] Carlos Ordonez, 2006, "Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.

[7] Jyoti Soni, Uzma Ansari, Dipesh Sharma, and SunitaSoni,June 2011, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 6,
pp. 2385-2392.

[8] IbticemeSedielmaci; F. BereksiReguig, " Detection of some heart diseases using fractal dimension and chaos theory", 2013, IEEE 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA).

[9] Jayshril S. Sonawane; D. R. Patil, "Prediction of Heart Disease Using Learning Vector Quantization Algorithm", 2014, IEEE Conference on IT in Business, Industry and Government (CSIBIG).

[10] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin, 2011, "HDPS: Heart Disease Prediction System", Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.

[11] Anbarasi Masilamani, ANUPRIYA, N Ch Sriman Narayana Iyenger, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", October 2010, International Journal of Engineering Science and Technology 2(10). [12] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1377-1382.

[13] Kathleen H. Miao, Julia H. Miao, "Coronary Heart Disease Diagnosis using Deep Neural Networks", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 10, 2018.

[14] Jae Kwon Kim and Sanggil, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis", 2017.

Sairabi H. Mujawar, and P. R. Devale, October 2015,"Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.

[15] B Padmaja, V V Rama Prasad, K V N Sunitha (2016), "TreeNet analysis of human stress behavior using socio-mobile data," in Journal of Big Data, 3(1), pp. 1-15, 2016, Springer.

[16] B Padmaja, Myneni Madhu Bala, E Krishna Rao Patro (2020), "A Comparison on visual prediction models for MAMO(multi activity multi-object) recognition using Deep Learning," in Journal of Big Data, 7(24), pp. 1-15, Springer.

[17] B Padmaja, V V Rama Prasad, K V N Sunitha (2020), " A Novel Random Split Point Procedure using Extremely Randomized Trees Ensemble Method for Human Activity Recognition," in EAI Endorsed transactions on Pervasive Health and Technology, 6(22), PP. 1-10.

[18] B Padmaja, V V Rama Prasad, K V N Sunitha (2018), "Machine Learning Approach for Stress Detection using Wireless Physical Activity Tracker," in International Journal of Machine Learning and Computing, 8(1), pp. 33-38.

[19] World Health Organization (WHO). Cardiovascular diseases (CVDs). Available from: https://www.who.int/news-room/fact- sheets/detail/cardiovascular-diseases-(cvds)

Effective Prediction of Heart Disease Using Machine Learning

ORIGINALITY REPORT

16%	9%	12%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to islamicuniversity Student Paper	1%
2	srmiliss2024.srmist.edu.in Internet Source	1%
3	"Proceedings of the 12th International Conference on Soft Computing for Problem Solving", Springer Science and Business Media LLC, 2024 Publication	1%
4	turcomat.org Internet Source	1%
5	Thomas Andreas Maurer. "Exploring the Financial Landscape in the Digital Age", CRC Press, 2024 Publication	1%
6	Sadeq Darrab, David Broneske, Gunter Saake. "Exploring the predictive factors of heart disease using rare association rule mining", Scientific Reports, 2024 Publication	1%