Enhancing the Performance of Heart Disease Prediction Models with Ensemble Learning

A project report

submitted in partial fulfillment of the requirements

for the award of the degree of

Master of Computer Applications

by

Anshul Kumar

52122108

by

Pushkar Joshi

52121205

by

Richa Singh

52112207

Under the supervision of

DR. Ashutosh Kumar Singh

Professor, Department of Computer Applications

DEPARTMENT OF COMPUTER APPLICATIONS

NATIONAL INSTITUTE OF TECHNOLOGY

KURUKSHETRA – 136119, HARYANA (INDIA)

CERTIFICATE

I hereby certify that the work which is being presented in the MCA Project report entitled **“”,**Enhancing the Performance of Heart Disease Prediction Models with Ensemble Learning in partial fulfillment of the requirements for the award of the **Master of Computer Applications** is an authentic record of my own work carried out during a period from June,2022 to May,2023 under the supervision of **Dr. Ashutosh Kumar Singh,Professor, Department of Computer Applications**, Computer Applications Department.

The matter presented in this thesis has not been submitted for the award of any other degree elsewhere.

Signature of Candidate
Anshul Kumar
Roll No. 52122108
Signature of Candidate
Pushkar Joshi
Roll No. 52121205
Signature of Candidate
Richa Singh
Roll No. 52112207

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of Supervisor
Date: **Dr. Ashutosh Kumar Singh,**
Professor, Department of Computer Applications

ACKNOWLEDGEMENT

We extend our sincerest gratitude to all individuals who have played a significant role in the successful completion of our research on heart disease prediction using an ensemble model.

First and foremost, we would like to extend our sincere gratitude to Dr. Ashutosh Kumar, our respected supervisor and adviser, for their exceptional guidance, direction, knowledge, and unwavering support throughout the whole study process.

Our deep appreciation also goes to the National Institute of Technology Kurukshetra, which provided us with the necessary resources, facilities, and infrastructure. Their support has been invaluable in facilitating the smooth execution of our research, including access to relevant datasets and computing resources.

Furthermore, we would like to acknowledge and express our gratitude to the research participants and the organizations involved in collecting and sharing the heart disease dataset used in our study. Their cooperation and willingness to contribute their data have been essential in enabling us to explore and develop accurate prediction models.

We would like to express our gratitude to our friends and colleagues who have given us priceless encouragement and support during our research journey. Their discussions, suggestions, and constructive criticisms have greatly contributed to the refinement of our ideas and methodology.

Last but certainly not least, we express our profound appreciation to our families for their unwavering support and understanding during the course of this research. Their encouragement, love, and patience have been our constant source of strength, enabling us to pursue our academic and research endeavors with dedication and enthusiasm.

Although it is not feasible to individually mention every person who has contributed, we genuinely appreciate the direct and indirect contributions of all individuals involved in this research project. Your support and involvement have played a vital role in its successful completion.

Sincerely,

(Anshul Kumar) (Pushkar Joshi) (Richa Singh)

ABSTRACT

Abstract:

Heart disease is a prevalent and potentially life-threatening condition that demands accurate and timely prediction for effective intervention. In this study, we propose an ensemble model-based approach for heart disease prediction. Ensemble learning, which combines multiple machine learning models, has proven effective in enhancing prediction accuracy and robustness.

Our study begins by gathering a comprehensive dataset comprising relevant clinical features associated with heart disease. The data is preprocessed to handle missing values and normalize the features. Subsequently, we select suitable machine learning algorithms known for their efficacy in heart disease prediction, such as decision trees, random forests, support vector machines, gradient boosting algorithms, and others.

The core of our approach involves constructing an ensemble model that combines the predictions of multiple individual models. Various ensemble methods, including bagging, boosting, and stacking, are explored to create a diverse and powerful ensemble. The individual models in the ensemble are trained on the dataset, and the performance of each model is assessed using measures including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve.

To optimize the ensemble model's performance, hyperparameter tuning techniques are employed to fine-tune the individual models. Validation of the ensemble model is conducted using an independent dataset or through cross-validation techniques to assess its generalization capabilities and robustness.

The proposed ensemble model offers several advantages, including improved prediction accuracy and enhanced reliability in heart disease prediction. Furthermore, the interpretability of the ensemble model provides insights into the contributing factors influencing heart disease predictions.

Our research aims to increase the accuracy of heart disease prediction, resulting in prompt interventions and better patient outcomes. The findings of this research contribute to the development of reliable and effective predictive models for heart disease, with potential applications in healthcare decision-making and risk assessment.

Contents

List of Figures

List of Algorithms

List of Abbreviations

VC Visual Cryptography

1 INTRODUCTION

Heart disease is a broad term that refers to a number of illnesses that affect the heart and blood arteries, such as coronary artery disease, heart failure, arrhythmias, and valve issues. High blood pressure, high cholesterol, diabetes, smoking, obesity, family history, and physical inactivity are risk factors for heart disease. Timely and accurate identification of individuals at risk of developing heart disease is crucial for implementing preventive measures, personalized interventions, and improving patient outcomes. With advancements in machine learning and data analysis techniques, there is an increasing focus on leveraging these tools to aid in the early prediction and diagnosis of heart disease.With the rapid advancements in machine learning and predictive analytics, researchers and healthcare professionals have turned to these technologies to develop robust and reliable models for heart disease prediction.

According to the World Health Organisation, heart disease is the leading cause of mortality worldwide, accounting for 17.9 million fatalities per year. Heart disease is the main cause of death in India, where it accounts for 28% of all fatalities. It ought to be noted that classic risk factors such as obesity, diabetes, and hypertension are becoming more and more common in India, especially in metropolitan areas. Preventing serious issues including heart attacks, strokes, and heart failure requires early detection and treatment of heart disease. The risk of heart disease can be reduced by maintaining routine check-ups, adopting a healthy lifestyle, and seeking immediate medical assistance for symptoms.

In this project, we describe the outcomes of our effort to create a model for predicting heart disease using ensemble learning methods. Heart illness has been detected using a variety of traditional machine learning techniques, such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN). These algorithms allow for the reliable identification of people with or without cardiac disease by classifying patients based on a number of parameters.

In the field of heart disease prediction, machine learning models have shown great potential for identifying individuals at high risk, aiding in prevention and targeted interventions. However, despite the advancements in machine learning algorithms, developing accurate and reliable heart disease prediction models remains a challenge. Individual models may have limitations in terms of bias, variance, or overfitting, which can impact their predictive performance and generalization to unseen data. To overcome these limitations, ensemble learning has gained significant attention.

ENSEMBLE LEARNING : Ensemble learning involves combining multiple individual machine learning models to create a more robust and accurate predictive model. The combination of diverse models allows for the exploitation of different learning algorithms and data representations, leading to improved prediction performance. By leveraging the strengths and diversity of different machine learning algorithms, ensemble learning can provide superior predictive performance compared to individual models.

The primary objective of our project was to design an ensemble learning system capable of accurately predicting the presence or absence of heart disease in patients based on a comprehensive set of clinical and demographic features.

For our project on heart disease prediction using ensemble learning, we collected a dataset from Kaggle. The dataset comprises patient information and symptoms related to heart disease.

The collected data includes the following features:

1. Age: A numerical representation of the patient's age.
2. Sex: The gender of the patient, represented as a categorical variable (0 for female, 1 for male).
3. Blood Pressure: The resting blood pressure of the patient.
4. Chest Pain Type: The type of chest pain experienced by the patient, categorized into different types (e.g., typical angina, atypical angina, non-anginal pain, or asymptomatic).
5. Cholesterol Levels: The cholesterol levels of the patient.
6. Fasting Blood Sugar: The patient's fasting blood sugar level is shown as a binary variable (1 for blood sugar levels greater than 120 mg/dL, 0 for blood sugar levels

lower than 120 mg/dL).

7. Resting Electrocardiographic Results: The electrocardiogram (ECG) test findings for a resting patient are categorised as normal, aberrant ST-T wave, and hypertrophic results.

8. Maximum Heart Rate Attained: The highest heart rate that could be attained while exercising, expressed in beats per minute.

The project also includes experimental findings and an assessment of the ensemble system's performance in comparison to separate base models. The predictive performance is evaluated using a variety of evaluation criteria, including accuracy, precision, recall, and F1 score. Additionally, we analyze the robustness and generalization capabilities of the ensemble learning approach. We utilized a publicly available heart disease dataset that contains a wealth of patient information, including demographic attributes, clinical measurements, and diagnostic test results. The dataset served as a valuable resource for training and evaluating our ensemble learning models, enabling us to explore the relationship between different input features and the occurrence of heart disease.

Throughout the project, we implemented rigorous preprocessing and feature engineering techniques to handle missing data, normalize features, and address any potential biases or outliers. Additionally, we conducted extensive model selection and hyperparameter tuning to optimize the individual models and ensemble learning algorithms, aiming to achieve the highest possible prediction accuracy. However, the performance of these models can be further enhanced to improve their accuracy, robustness, and generalization capabilities.

Algorithms like logistic regression and decision trees rely on a single model to make predictions, making them susceptible to overfitting or underfitting the data and unable to capture complex relationships between features and the target variable. On the other hand, ensemble learning methods combine the predictions of multiple models, resulting in more accurate and robust predictions.This mitigates the risk of overfitting or underfitting the data and enhances the model's ability to generalize to new, unseen data. Ensemble methods excel in handling imbalanced datasets, which is particularly relevant in medical applications such as heart disease detection. Therefore, utilizing an ensemble model improves the accuracy and robustness of heart disease detection compared to relying solely

on traditional machine learning algorithms.

In this project, we focus on utilizing ensemble learning to enhance heart disease prediction models by integrating four widely used base models: Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Each of these models has particular advantages and disadvantages, making them suitable candidates for ensemble learning.

1. Decision Tree : Decision Tree is a non-parametric supervised learning algorithm that builds a tree-like model for classification or regression. To generate a tree structure, the algorithm divides the dataset based on the input feature values. Each leaf node represents a class label or a value, and each interior node represents a decision based on a feature. Decision Trees can handle both categorical and numerical data, handle missing values, and capture complex relationships between features and the target variable. Due to their simplicity in interpretation and visualization, decision trees are frequently used to understand how a model makes decisions. Decision trees can, however, be sensitive to noisy or irrelevant data, and if the tree gets overly complicated, there is a risk of overfitting. To address these issues, techniques like pruning and ensemble methods like random forests and gradient boosting are often used in practice.

2. K-Nearest Neighbors (KNN): : K-Nearest Neighbors is a non-parametric algorithm used for classification and regression tasks. Given a new data point, KNN assigns it a class label based on the majority vote of its k nearest neighbors in the feature space. The K-Nearest Neighbours (KNN) technique determines the distance between each training point in the feature space and the new point in order to determine the k nearest neighbors of a new data point. A distance metric, such as the Manhattan distance or the Euclidean distance, is used to do this. KNN is effective at capturing local patterns and can adapt to different data distributions. It is relatively simple to implement and does not make strong assumptions about the underlying data distribution. KNN can be computationally expensive for large datasets because it necessitates calculating distances to every other data point, and its performance can be sensitive to the choice of the value of k.

3. Support Vector Machines (SVM): Support Vector Machines are effective supervised learning algorithms used for classification and regression tasks. SVMs seek to find an optimal hyperplane that maximizes the margin between different classes in the feature space.

SVMs can handle linear and non-linear relationships by using kernel functions. They are particularly effective in handling complex relationships and can capture intricate decision boundaries. SVMs perform well even in high-dimensional feature spaces and are known for their robustness. However, SVMs may be computationally expensive, especially for large datasets, and their performance is sensitive to the choice of hyperparameters such as the kernel type and regularization parameter.

The kernel trick is a method that SVM employs in non linear decision boundaries to transfer the input data into a higher-dimensional space where the data points are more separable by a linear hyperplane. The radial basis function (RBF), sigmoid, and polynomial functions are the most frequently employed kernel functions.In addition to binary classification problems, SVM can be used to resolve one-vs-all or one-vs-one challenges involving multi-class classification.

4. Logistic Regression: Logistic Regression is a widely used statistical model for binary classification tasks. It models the relationship between the input features and the probability of a particular outcome using a logistic function. The output is subsequently transformed using a logistic function, yielding a value ranging from 0 to 1. This value represents the probability assigned to the positive class. Logistic Regression is advantageous as it is interpretable, computationally efficient, and can handle both linear and non-linear relationships. It is particularly useful in situations where understanding the impact of individual features on the outcome is important. Logistic Regression estimates the coefficients of the features to compute the log odds of the outcome. These coefficients indicate the strength and direction of the relationship between the features and the target variable. However, Logistic Regression may not capture complex interactions between features as effectively as other models, and its performance may be limited when dealing with imbalanced datasets or non-linear relationships.

The following are the main contributions of our work:

1. We proposed an ensemble model which takes multiple algorithms and gives the maximum accuracy of disease prediction.

2. We applied the proposed model on the publicly available Heart Failure Prediction Dataset which contains records of 918 people.

3. We compared the performance of the proposed ensemble model over the base clas-

sification algorithms.

It is our belief that the utilization of ensemble learning techniques for heart disease prediction can significantly enhance the existing predictive models and provide valuable insights for healthcare professionals. By accurately identifying individuals at high risk of heart disease, we can facilitate early interventions, adopt preventive measures, and ultimately reduce the burden of heart disease on individuals and society as a whole. Given the potential of ensemble learning to enhance heart disease prediction models, this project aims to explore the application of ensemble learning techniques to improve the performance and reliability of heart disease prediction. By integrating multiple base models and leveraging their collective intelligence, we seek to overcome the limitations of individual models and create an ensemble system that offers superior predictive capabilities.

This project aims to contribute to the development of more accurate and robust heart disease prediction models, ultimately leading to better risk assessment, early detection, and personalized interventions for individuals at risk of heart disease. The significance of our project lies in its potential to assist healthcare professionals in identifying individuals at high risk of heart disease, allowing for timely interventions and targeted treatments. By leveraging the power of ensemble learning, we sought to develop a strong prediction model that outperforms individual models and offers more precise risk evaluations.

In the following sections of this report, we will detail the methodology employed, describe the dataset used, present the results obtained, and discuss the implications of our findings. We will also address the limitations of our study and suggest future research directions to further improve heart disease prediction using ensemble learning techniques.

Overall, this project contributes to the growing field of machine learning applications in healthcare and emphasizes the potential of ensemble learning in enhancing heart disease prediction. The outcomes of this research have the potential to significantly impact clinical decision-making, enabling healthcare professionals to make informed decisions and improve patient outcomes in the domain of heart disease prediction.

2 LITERATURE REVIEW

A literature review serves as a critical and analytical assessment of the existing body of research and academic writing related to a particular topic. Its primary objective is to offer a comprehensive overview of the current knowledge and understanding surrounding the subject matter by thoroughly examining and evaluating the relevant literature. In scientific research, a literature review holds significant importance as it plays a crucial role in research papers or dissertations. It serves as a means to position and situate the study within the existing corpus of knowledge. By reviewing the literature, researchers can identify and understand the key theories, concepts, methodologies, and findings that have already been established in the field. One of the main purposes of a literature review is to identify any gaps or limitations in the current research. By examining the existing literature, researchers can pinpoint areas where further investigation and exploration are needed. This identification of gaps helps to justify the rationale for conducting the research and contributes to the development of research questions or hypotheses. Moreover, a literature review enables researchers to recognize the different perspectives, arguments, and debates within the field. It allows them to synthesize and integrate diverse findings and viewpoints from various sources, contributing to a more comprehensive understanding of the topic. This synthesis of information helps researchers to develop a conceptual framework or theoretical foundation for their own study. The attached table provides concise and summarized information about the literature study. It likely includes details such as the title of the source, the author(s), the publication year, key findings or arguments, and any relevant notes or comments. This table serves as a handy reference for organizing and synthesizing the information gathered from the literature review process.

2.1 BATTINENI, GOPI, GETU GAMO SAGARO, NALINI CHINATALAPUDI, AND FRANCESCO AMENTA.

In order to accurately diagnose chronic diseases, this study will undertake a thorough evaluation of the numerous machine learning (ML) predictive model applications. The study examines various prediction models utilized in machine learning specifically for the diagnosis of chronic diseases. Determining the optimal learning method for disease prediction can be challenging, as it depends on factors such as data volume and user accessibility.

Since they are the most straightforward and simple to use in predictive modelling, supervised machine learning (SML) computations are the most common in the analysed research. The integration of these models into clinical practice has the potential to improve healthcare services and enhance specialist decision-making.

The review highlights that Support Vector Machines (SVM) and Logistic Regression (LR) models are widely implemented in numerous studies related to chronic disease diagnosis. Specifically, sixteen studies focused on using SVM and LR models for diseases such as hepatitis B, COPD, and diabetes. SVM models have proven effective in identifying early-stage COPD, potentially facilitating a stronger doctor-patient relationship. Numerous research have used Naive Bayes (NB) and Bayesian networks to predict the diagnosis of asthma. To recognise clinical symptoms and create connections through Bayesian networks to predict future symptoms, these models use past patient records. Additionally, using a variety of primary and secondary data sources, the K-Nearest Neighbours (KNN) algorithm has been used in five studies to identify and anticipate illness progression across multiple stages.

The prevalence of supervised models for disease prediction or classification in the literature reviewed is one limitation found in the current investigation. Future studies should investigate the use of unsupervised models, such clustering, and deep learning models, like neural networks, to overcome this problem.

The results of this study indicate that there is no standardised way to choose the optimum strategy for in-the-moment clinical practise because each method has pros and cons of its own. Support vector machines (SVM), logistic regression (LR), and clustering, however, stood out as the techniques most frequently used in the diagnosis of chronic diseases among the methodologies examined.

2.2 HAMSAGAYATHRI, P., AND S. VIGNESHWARAN.

This paper delves into the utilization of various machine learning (ML) techniques for the diagnosis of different diseases, such as heart diseases and diabetes. The effectiveness of these models is primarily attributed to their ability to effectively capture the unique characteristics and patterns associated with each disease. Notably, previous studies have demonstrated that Support Vector Machines (SVM) yield a significant improvement in performance, achieving an impressive accuracy rate of 94.60% for identifying heart diseases. On the other hand, Naive Bayes has proven to be particularly proficient in correctly diagnosing diabetes, exhibiting the highest classification precision of 95

In addition to highlighting the successes of these ML algorithms, the paper also sheds light on their inherent benefits and drawbacks. It provides a comprehensive survey that delves into the advantages and limitations of employing such algorithms in the medical domain. Furthermore, the paper presents a range of tools developed within the AI community, showcasing their practical applications and significance in addressing specific problems. These approaches not only offer valuable insights for the analysis of medical conditions but also hold immense potential for enhancing the decision-making process.

The field of evaluation has witnessed an influx of statistical prediction models that often fall short of generating high-quality outcomes. Statistical models tend to be less efficient when it comes to handling missing values and dealing with extensive data points, making them less capable of producing generalized knowledge. This is where the value of machine learning techniques truly emerges. Applications including image identification, data mining, natural language processing, and disease detection all rely heavily on machine learning.

Overall, this paper explores the wide-ranging benefits and applications of machine learning in medical diagnostics. It emphasizes the success of specific ML algorithms in disease identification, while also acknowledging their limitations. The survey presented in the paper offers valuable insights and serves as a valuable resource for researchers and practitioners in the field, contributing to improved decision-making processes and enhancing overall healthcare outcomes.

2.3 SHARATHCHANDRA, D., AND M. RAGHU RAM

In this paper, the focus lies on exploring different machine learning (ML) techniques employed for the diagnosis of various diseases, including heart disease and diabetes. The effectiveness of these techniques can be attributed to their ability to capture and analyze disease-specific characteristics. Previous studies have revealed noteworthy outcomes, such as Support Vector Machines (SVM) achieving a remarkable 94.60% improvement in heart disease identification. Similarly, Naive Bayes has demonstrated an accurate diagnosis of diabetes with the highest classification precision of 95

Expanding on these findings, the ML models discussed in the paper have showcased excellent results due to their specific characterization of disease patterns and features. The SVM algorithm, as observed in previous studies, has proven to be highly successful in identifying and diagnosing heart diseases. It has achieved a significant enhancement of 94.60% in performance, providing promising results for the accurate identification of heart-related conditions.

Furthermore, the Naive Bayes algorithm has been identified as a reliable tool for the diagnosis of diabetes. With a classification precision rate of 95%, it has demonstrated its capability to correctly identify and categorize instances of diabetes. This showcases the potential of Naive Bayes in assisting healthcare professionals in accurately diagnosing this chronic condition.

The mentioned ML techniques have exhibited their strengths in the field of disease diagnosis, leveraging their ability to capture and analyze the specific characteristics and patterns of different diseases. By employing these techniques, researchers and medical practitioners can enhance their diagnostic capabilities and contribute to more precise and effective disease identification.

Overall, the paper highlights the notable success achieved by ML techniques in diagnosing diseases like heart disease and diabetes. The specific characteristics and patterns identified by these models have paved the way for improved performance and higher classification precision. These findings open up opportunities for further research and the development of more advanced ML models in the domain of disease diagnosis.

2.4 SHAH, DEVANSH, SAMIR PATEL, AND SANTOSH KUMAR BHARTI

This research paper focuses on exploring various attributes associated with heart disease and proposes a supervised learning model based on algorithms such as Naïve Bayes, decision tree, K-nearest neighbor (KNN), and random forest. These algorithms are utilized to develop a predictive model for heart disease diagnosis. The paper delves into a comprehensive analysis of different attributes that are relevant to heart disease. These characteristics may include elements like age, sex, levels of cholesterol and blood pressure, smoking behaviours, and a parent's past of heart disease. By examining these attributes, researchers aim to gain a deeper understanding of the factors contributing to the occurrence and progression of heart disease. To create an effective heart disease diagnosis model, the paper employs supervised learning algorithms such as Naïve Bayes, decision tree, KNN, and random forest. These algorithms utilize labeled training data to learn patterns and relationships between the attributes and the presence or absence of heart disease. By leveraging these algorithms, the research aims to develop a predictive model that can accurately classify individuals as either having or not having heart disease based on their attribute values. Naïve Bayes is a probabilistic classifier that assumes independence between attributes and has been widely used in various domains, including healthcare. Decision trees are tree-based classifiers that use a series of decision rules to classify instances. KNN is a lazy learning algorithm that classifies instances based on their proximity to neighboring instances in the feature space. Random forest, on the other hand, is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. By employing these algorithms, the research aims to develop a robust and accurate predictive model for heart disease diagnosis. This model can then be applied to new, unseen instances to determine the likelihood of heart disease based on their attribute values. The ultimate goal of this research is to contribute to improved methods for early detection and diagnosis of heart disease, enabling timely interventions and better patient outcomes.

2.5 DAHIWADE, DHIRAJ, GAJANAN PATLE, AND EKTAA MESHRAM

In this research paper, the authors propose a general disease prediction system that leverages machine learning algorithms. Specifically, they utilize the K-nearest neighbor (KNN)

and convolutional neural network (CNN) algorithms to classify patient data and predict general disease risk. By inputting patients' records into the system, it generates accurate predictions regarding the level of disease risk.

The paper compares the performance of the KNN and CNN algorithms in terms of accuracy and processing time. The results indicate that the CNN algorithm outperforms the KNN algorithm in both aspects. The accuracy of the CNN algorithm is higher, meaning it achieves more precise predictions compared to the KNN algorithm. Additionally, the time required for classification using the CNN algorithm is less than that of the KNN algorithm, implying that it performs faster.

Based on these findings, it can be concluded that the CNN algorithm is superior to the KNN algorithm in terms of both accuracy and time efficiency. This suggests that the CNN algorithm is better suited for the general disease prediction system proposed in the research. By employing the CNN algorithm, healthcare professionals can obtain more accurate and timely predictions of disease risk for patients, facilitating early interventions and improved healthcare outcomes.

This research study proposes a comprehensive machine learning-based disease prediction system. The system analyses patient data to accurately forecast the risk of common diseases using the KNN and CNN algorithms. In terms of accuracy and processing speed, the comparison analysis shows that the CNN algorithm performs better than the KNN method, making it the go-to option for accurate and effective disease risk prediction.

2.6 KRISHNAMOORTHI, RAJA, SHUBHAM JOSHI, HATIM Z. ALMARZOUKI, PIYUSH KUMAR SHUKLA, ALI RIZWAN, C. KALPANA, AND BASANT TIWARI.

This study examines several existing machine learning classification models for predicting diabetic patients, with a specific focus on accuracy. The findings reveal that glucose and BMI exhibit a strong correlation with diabetes, as determined through association rule mining. Additionally, the study reports that the Receiver Operating Characteristic (ROC) value of Logistic Regression (LR) reaches 86%.

By discussing various ML classification models, the researchers contribute to the understanding of their effectiveness in predicting diabetes. The emphasis on accuracy underscores the significance of identifying reliable predictors and models for accurate diabetes

diagnosis. The study highlights the importance of glucose and BMI as influential factors in diabetes prediction, as indicated by their strong correlation with the disease. Moreover, the ROC value of 86% for the LR model suggests its potential for accurately discriminating between diabetic and non-diabetic individuals.

Overall, this research delves into the evaluation of existing ML classification models for diabetic patient prediction, emphasizing accuracy as a key performance metric. Through the application of association rule mining, the study identifies significant correlations between diabetes and factors such as glucose and BMI. Furthermore, the reported ROC value of 86% for the LR model indicates its promising predictive capabilities in distinguishing individuals with diabetes.

2.7 MUJUMDAR, AISHWARYA, AND V. VAIDEHI

This study focuses on the application of various machine learning algorithms on a dataset for classification purposes. Among these algorithms, Logistic Regression stands out with the highest accuracy rate of 96%. The study also includes a comparison of the accuracies achieved by different machine learning algorithms using two distinct datasets. The results demonstrate that the proposed model enhances the accuracy and precision of diabetes prediction when compared to the existing dataset.

Building upon these findings, future research could extend this work to investigate the likelihood of nondiabetic individuals developing diabetes within the next few years. By utilizing the developed model and analyzing relevant data, it would be possible to identify and assess the risk factors that contribute to the onset of diabetes in previously nondiabetic individuals. This would provide valuable insights for preventive measures and interventions, aiding in early detection and potential intervention to mitigate the risk of developing diabetes.

Overall, this study demonstrates the effectiveness of various machine-learning algorithms for diabetes prediction. The Logistic Regression algorithm emerges as the most accurate in this context. Moreover, the research highlights the potential for further exploration of nondiabetic individuals' likelihood of developing diabetes, paving the way for proactive healthcare measures and personalized interventions to reduce the prevalence and impact of diabetes.

2.8 MANIRUZZAMAN, MD, MD JAHANUR RAHMAN, BENOJIR AHAMMED, AND MD MENHAZUL ABEDIN

In this study, the researchers employ logistic regression (LR) as a tool to identify the risk factors associated with diabetes disease. This is achieved by analyzing the p-value and odds ratio (OR) of various factors. By examining these statistical measures, the study aims to determine which factors contribute significantly to the risk of developing diabetes.

Additionally, the researchers adopt three different classifiers, namely naïve Bayes (NB), decision tree (DT), and random forest (RF), to predict the occurrence of diabetes in patients. These classifiers utilize the collected data and the identified risk factors to generate predictions regarding the presence or absence of diabetes in individuals.

The naïve Bayes algorithm is a probabilistic classifier that calculates the probability of an instance belonging to a particular class based on the observed attribute values. Decision tree algorithms create a tree-like structure of decision rules to classify instances. Random forest, on the other hand, is an ensemble learning method that combines multiple decision trees to make predictions.

By leveraging these three classifiers, the study aims to enhance the accuracy of diabetes prediction. Each classifier brings its own unique approach to the task, utilizing different algorithms and methodologies. The comparison of these classifiers allows for an evaluation of their effectiveness in predicting the occurrence of diabetes in patients.

Overall, this research utilizes logistic regression to identify the risk factors associated with diabetes disease. It further extends the analysis by adopting three different classifiers (naïve Bayes, decision tree, and random forest) to predict diabetes in patients. This approach provides valuable insights into the identification of risk factors and the accurate prediction of diabetes, contributing to improved understanding and management of the disease.

2.9 KHANAM, JOBEDA JAMAL, AND SIMON Y. FOO

In their research, a system was designed to accurately predict diabetes. The data preprocessing stage involved utilizing the WEKA tool, where a feature reduction method was employed, resulting in the removal of three features. The PIMA dataset was used, with five input features (Glucose, BMI, Insulin, Pregnancy, and Age) and one output feature

(outcome).

The researchers explored the performance of seven different machine learning algorithms in predicting diabetes. These algorithms included Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Naïve Bayes (NB), AdaBoost (AB), Logistic Regression (LR), and Support Vector Machines (SVM). The evaluation of these models incorporated various performance measures such as accuracy, precision, recall, and F-measure.

The findings of the study indicated that all of the models demonstrated promising results across several evaluation metrics. Specifically, they achieved accuracies greater than 70%, highlighting their effectiveness in predicting diabetes. The models exhibited good performance in terms of accuracy, precision, recall, and F-measure, indicating their ability to accurately classify instances and capture relevant patterns in the data.

Overall, this research emphasized the design of a system for predicting diabetes with high accuracy. By leveraging feature reduction techniques and utilizing various machine learning algorithms, the study demonstrated the potential of accurately predicting diabetes based on the selected input features. The evaluation results confirmed the effectiveness of the proposed models, showcasing their ability to achieve reliable predictions for diabetes detection.

2.10 NAWAB, MOHAMMAD ALI, TAVISI MALPANI, TUSHAR KAUNDAL, AND MS DIVYA SONI

In the study, a comparison was conducted to assess the accuracies of different algorithms including Random Forest, Support Vector Machine (SVM), and Naive Bayes. The findings revealed that the accuracies of these algorithms were approximately 96 percent. Based on this analysis, the researchers concluded that their proposed method exhibited higher accuracy in disease prediction. Specifically, the comparison highlighted that SVM demonstrated slightly higher accuracy among the evaluated algorithms.

In terms of future work, the researchers aim to prioritize the delivery of medical support and appropriate dosage to patients in a timely manner. This objective is aligned with establishing a robust infrastructure and optimizing efficiency within the medical profession. By focusing on these aspects, the researchers intend to enhance the overall medical support system, ensuring prompt and effective care for patients.

3 Problem Definition

Heart disease is a significant health concern that affects a large portion of the population worldwide. Early and accurate prediction of heart disease can help in timely interventions and improved patient outcomes. Ensemble learning, a technique that combines multiple machine learning models to make predictions, has shown promise in improving the accuracy and robustness of predictive models.

The problem at hand is to develop an ensemble learning-based system for predicting heart disease. The goal is to leverage the strengths of different machine learning algorithms and create an ensemble model that outperforms individual models in terms of accuracy, sensitivity, specificity, and overall predictive performance. The core of our approach will involve constructing an ensemble model that combines the predictions of multiple individual models. We will explore different ensemble methods such as bagging, boosting, or stacking to create a diverse and robust ensemble. The individual models within the ensemble will be trained using the dataset, and their performance will be evaluated using various evaluation metrics such as accuracy, precision, recall, F1-score.

---

4 Proposed Model

Figure 1 shows the suggested method for applying ensemble learning to forecast heart disease. The dataset is split into three sections with a ratio of 7:1:2 each: training data, validation data, and testing data. The predictive framework receives the training data. The predictions from various classifiers are combined using an ensemble learning technique called a voting classifier, which chooses the class label with the most votes in the end. The final prediction is created by combining the results of the training data's n separate classification algorithms (classification algorithm 1, classification algorithm 2,..., and classification algorithm n) using the voting classifier. The prediction model is deemed completely trained and prepared for deployment on real-time or live data after it reaches the necessary degree of accuracy on the test data. A machine learning model can understand patterns and relationships within the data by using training data, which is the initial dataset used to train the model. Training data includes input features and corresponding target labels. The model's performance is evaluated using test data, a distinct dataset that contains fresh and unexplored data points. It ensures generalisation and prevents overfitting to the training data by estimating the accuracy of the model's predictions on fresh real-world data. A part of the training data known as validation data is used to adjust the model's hyperparameters, prevent overfitting, and ensure accuracy on unobserved data. In ensemble methods, classifier weights are used to rank the importance of each classifier according to how well it performs when combining predictions. From each of the individual predictions made by the classification algorithms, the voting classifier creates a final prediction by a majority vote.

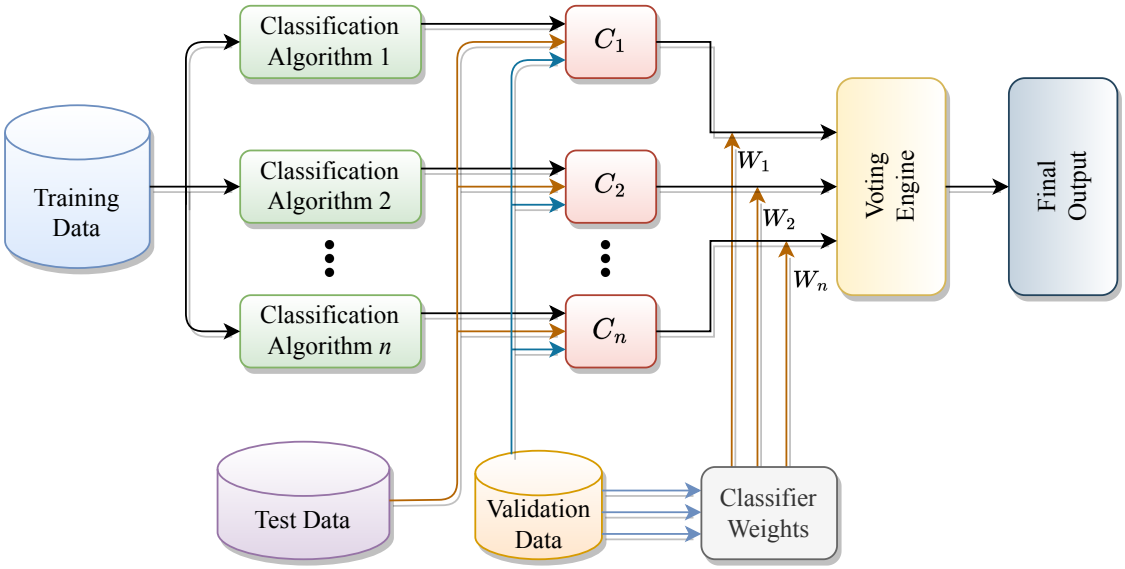


Figure 4.1: Flow Chart

4.1 DATA COLLECTION

4.1.1 Overview:

Our dataset, which we downloaded from Kaggle, is used to analyse various clinical and demographic aspects of patients in order to forecast the likelihood of heart failure. Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Old Peak, ST Slope, and Heart Disease are among the 12 parameters that make up the 918 records that make up the dataset. A patient's heart failure status can be determined by the binary indicator Heart Disease, which is the target variable.

4.1.2 Features of the dataset:

1. Age: It describes the age of the patient (in years).
2. Sex: It describes the sex of the patient, male or female.
3. Chest Pain type: It describes the type of chest pain from which the patient is suffering.
4. Resting Blood Pressure: It describes the blood pressure of the patient.
5. Cholesterol: It describes the cholesterol levels of the patient.
6. Fasting Blood Pressure: It describes the glucose levels in the blood of the patient.

7. Resting Electro Cardio Gram:
8. Maximum Heart Rate: It describes the maximum heart rate achieved by the patient.
9. Exercise Angina: It describes whether the patient experiences any kind of chest pain when he/she does exercise.
10. Old Peak: It describes the ST depression of the patient induced by exercise relative to rest.
11. ST Slope: It describes the ST segment shift relative to exercise-induced increments in the heart rate of the patient.
12. Heart Disease: It is a binary indicator representing whether the patient experienced heart failure or not (0 = No, 1 = Yes).

4.1.3 Data Format:

The dataset is displayed in CSV (comma-separated values), a popular file format for tabular data organisation. Each column in the dataset corresponds to a single feature or the target variable, and each row to a unique patient.

4.2 DATA PREPROCESSING:

Data preparation includes all necessary procedures and techniques used on raw data before it is used for modelling or analysis. To assure the data's quality, homogeneity, and applicability for the planned analytical or modelling aims, these steps comprise converting, purifying, and organising the data. Data preprocessing is a crucial step in the data workflow since it has a direct impact on the precision and effectiveness of future studies or models.

Steps in Data Preprocessing:

1. Data Cleaning: This involves handling missing data, dealing with outliers, and correcting errors in the data. Missing data can be addressed by either removing the corresponding entries or imputing values based on statistical methods. Outliers, which are extreme or erroneous data points, can be detected and treated by various techniques such as trimming, or removing them if they are genuine anomalies. Data errors, such as incorrect values or inconsistencies, need to be corrected or resolved.

2. Data Transformation: Sometimes, data needs to be transformed to meet specific assumptions or improve its distribution. Common transformations include normalization, standardization, logarithmic transformation, or power transformation. These transformations can help reduce the impact of different scales, make the data more suitable for certain algorithms, or improve the interpretability of the results.

3. Feature Selection or Extraction: When dealing datasets where not all features are important or relevant for the analysis or modelling process, feature selection or extraction is essential. By identifying and keeping the most useful features, feature selection approaches help to reduce dimensionality and improve efficiency. Additionally, innovative derived features that capture important information from the original features can be created using feature extraction techniques like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA).

4. Handling Categorical Variables: Categorical variables must be properly encoded in order to support machine learning methods, which often require numerical input. Depending on the type and features of the categorical variables, different strategies can be used. One-hot encoding, label encoding, and ordinal encoding are examples of common strategies. These techniques make it easier to convert category data into numerical formats compatible with machine learning algorithms.

5. Data Splitting: It is usual practise to separate the dataset into training, validation, and testing subsets in order to evaluate model performance effectively. This part permits model training, hyperparameter adjustment, and evaluation of the model's generalizability to new data.

5 Classification Algorithms

Classification algorithms like Logistic Regression, Decision Tree, K-Nearest Neighbour, and Support Vector

Machines are used in the ensemble model, which is described in further subsections.

5.1 LOGISTIC REGRESSION CLASSIFICATION ALGORITHM:

Within the field of machine learning, logistic regression is a widely used supervised learning technique, notably for the prediction of categorical dependent variables. To produce predictions, it uses a specified set of independent variables. When a dependent variable is categorical, logistic regression predicts the result by giving probabilistic values that range from 0 to 1, rather than precise values between 0 and 1. These numbers correlate to categories like true or false, 0 or 1, and so on. Except for the differences in their intended uses, logistic regression and linear regression are comparable. While linear regression is used for regression tasks, logistic regression is used to address classification issues. Instead of fitting a regression line, logistic regression uses a sigmoid-shaped logistic function that produces two maximum values (0 or 1). The logistic function's curve shows the chance of different outcomes, such as the possibility of cancerous cells or the likelihood that a mouse would be obese based on its weight. Due to its capacity to categorise new data using both continuous and discrete datasets, logistic regression is a major machine learning technique. When classifying observations using a variety of data sources, it proves useful in swiftly finding influential elements.

5.1.1 Logistic Function (Sigmoid Function):

The sigmoid function, a mathematical tool, is used in logistic regression to turn the projected values into probabilities. This function changes any real number between 0 and 1 into another number. As a result of its distinctive S-shaped curve, it is also known as the logistic function or the sigmoid function. Given that logistic regression produces proba-

bilities, the numbers it produces must lie between 0 and 1. To fulfil this limitation, the sigmoid function is essential. We present the idea of a threshold value in logistic regression, which establishes the likelihood of either 0 or 1. Values over the threshold are often categorised as 1, while values below it are categorised as 0.

5.1.2 Assumptions for Logistic Regression:

1. The nature of the dependent variable must be categorical.

2. There shouldn't be any multi-collinearity in the independent variable.

5.1.3 Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below: We know the equation of the straight line can be written as:

$z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$

where $b_0, b_1, b_2, \ldots, b_n$ are the coefficients and $x_1, x_2, \ldots, x_n$ are the predictor variables.

In Logistic Regression z can be between 0 and 1 only, so for this let's divide the above equation by (1-z):

$z/(1-z)$; 0 for z = 0 and Infinity for z= 1 .

But we need a range between -[infinity] to +[infinity], then take the logarithm of the equation it will become:

$Log[z/(1-z)] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$

5.1.4 Types of Logistic Regression:

Based on the classifications of the dependent variable, there are three different types of logistic regression:

1. 1. Binomial Regression: In this type of logistic regression, the dependant variable can only fall into one of two potential categories, such as 0 or 1.

2. 2. Multinomial Regression: When the dependent variable has three or more unordered categories, multinomial logistic regression is utilised. The dependent variable might, for instance, be "cat," "dog," "sheep," or any other unordered category.

3. 3. Ordinal Regression: When the dependent variable can be divided into three or more ordered classes, ordinal logistic regression is appropriate. The categories are arranged in a particular hierarchy, such as ”low,” ”medium,” and ”high” or ”small,” ”medium,” and ”large.”

Each type of logistic regression is suited for different scenarios depending on the nature of the dependent variable and its categories.

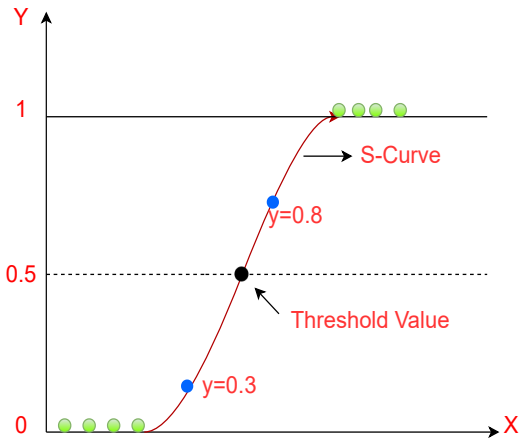


Figure 5.1: Logistic Regression

5.2 DECISION TREE CLASSIFICATION ALGORITHM:

Decision trees, a popular supervised learning technique, are frequently used to solve classification and regression issues. It is especially preferred for activities requiring classification. A decision tree is organised as a diagram that resembles a tree, with core nodes standing in for dataset attributes, branches for decision-making, and leaf nodes for the results of classification.Decision nodes and Leaf nodes are the two different sorts of nodes that make up a decision tree. While decision nodes are in charge of making decisions and have several branches, leaf nodes deliver the ultimate decision or outcome without any more branches. These decision nodes run tests and reach conclusions using the characteristics of the provided dataset.A visual depiction known as a decision tree seeks to identify every option or choice based on predetermined criteria. Given that it has a tree-like structure that grows with consecutive branches from the root node, it is called a decision tree.The classification and regression tree algorithm, or CART, is frequently used to build decision trees. The decision tree asks questions, and depending on the answers (Yes/No) it receives, it divides the tree into subtrees.A decision tree offers an organised method

for making decisions based on preset conditions, making it an effective tool for handling classification and regression issues.

The below diagram explains the general structure of a decision tree:

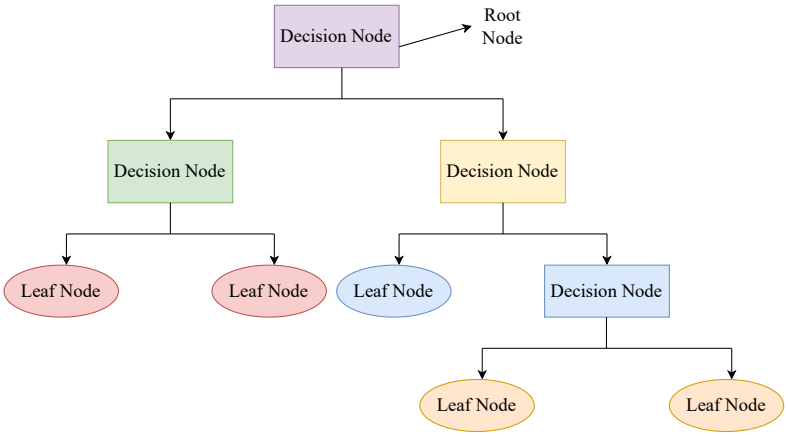


Figure 5.2: Decision Tree

5.2.1 Why use Decision Trees?

The most important thing to keep in mind while developing a machine learning model is to select the optimal method for the dataset and task at hand. The two rationales for employing the decision tree are as follows:

1. Decision trees are typically designed to resemble how people think when making decisions, making them simple to comprehend.

2. Because the decision tree displays a tree-like structure, the rationale behind it is simple to comprehend.

5.2.2 Decision Tree Terminologies:

1. Root node: The decision tree's root node is where it all begins. The full dataset is represented, which is then split into two or more homogeneous sets.

2. Leaf Node: Leaf nodes are the ultimate output nodes, after which the tree cannot be further divided.

3. Splitting: The division of the decision node/root node into sub-nodes in accordance with the specified conditions is known as splitting.

4. Branch/subtree: A branch or subtree is a tree created by slicing another tree.

5. Pruning: Pruning is the procedure of removing the tree's undesirable branches.

6. Parent/Child node: The root node of the tree is referred to as the parent node, while the other nodes are referred to as the child nodes.

5.2.3 Working of Decision Tree Algorithm:

To determine the class of the given dataset, a decision tree uses an algorithm that starts at the root node and moves up the tree structure. By comparing the values of the root attribute with the corresponding attribute values of the record (real dataset), the algorithm iterates through the tree. Based on this comparison, the algorithm moves on to the next node by following the branch that corresponds to the matched attribute value. The algorithm repeats the process at each succeeding node by comparing the attribute value to the sub-nodes related to that attribute. The method repeats this repeating comparison and traversal process until it reaches a leaf node in the tree.

The complete process can be better understood using the below algorithm:

Step 1: According to S, start the tree at the root node, which contains the entire dataset.

Step 2: Utilize Attribute Selection Measure (ASM) to identify the dataset's top attribute.

Step 3: Separate the S into subsets that include potential values for the best qualities.

Step 4: Create the decision tree node that holds the best attribute.

Step 5: Recursively generate new decision trees using the subsets of the dataset produced in step 3. Continue doing this until you reach a point when you can no longer categorize the nodes and you refer to the last node as a leaf node.

5.2.4 Advantages of a Decision Tree:

Because they closely resemble the method through which people arrive at decisions in the actual world, decision trees are simple to comprehend. Because of this, they are quite intuitable and interpretable.

Decision-making-intensive problem domains are where decision trees shine. They are capable of solving problems involving decisions and offering insightful solutions.

The ability of decision trees to take into account several situations and outcomes for a particular problem is one of their benefits. They analyse many options and potential solutions as they explore various paths and branches in the tree.

Decision trees need less preprocessing or data cleaning than other techniques. Without significantly affecting their performance, they can deal with missing values and outliers in the data.

5.2.5 Disadvantages of a Decision Tree:

Especially when working with datasets that contain a large number of traits or attributes, the decision tree can grow complex because it may have various tiers or layers.

The Random Forest technique can be used to solve potential overfitting problems in decision trees. An ensemble technique called Random Forest mixes various decision trees to produce forecasts. By combining the predictions of various trees and lowering the variance, it helps prevent overfitting.

When dealing with more class labels, a decision tree's computational complexity may rise. The building and evaluation of the decision tree become computationally more difficult as the number of class labels increases. This may affect the amount of time and materials needed to create and apply the decision tree model.

5.3 K-NEAREST NEIGHBOR CLASSIFICATION ALGORITHM:

The K-Nearest Neighbours (K-NN) algorithm is a straightforward supervised learning technique that functions on the presumption that new examples are comparable to those already present and should be classified into the category that most closely resembles those classifications. By comparing a new data point to previously stored data points, the K-NN algorithm classifies the new data point. This makes it possible to quickly and accurately categorise new data using the K-NN method.

K-NN can be utilised for regression jobs even though classification problems are where it's most frequently used. K-NN is a non-parametric method that makes no assumptions about the distribution of the underlying data. Because it saves the training material rather being actively learning from it, the technique is frequently referred to as a lazy learner algorithm. When new data needs to be categorised, it instead makes use of the dataset to do so. The K-NN method merely stores the data during the training phase,

and when new data is presented, it places it in a category that closely resembles the new data based on the stored data.

5.3.1 Why is a K-NN algorithm necessary?

Which category does the new data point, x1, belong in if there are two categories, Category A and Category B? A K-NN algorithm is necessary to handle this kind of problem. Finding the category or class of a given dataset is made simple by K-NN.

Take a look at the diagram below:

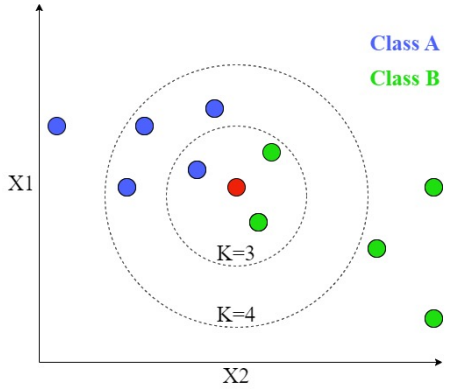


Figure 5.3: K-Nearest Neighbor

5.3.2 How does K-NN function?

The following algorithm can be used to describe how the K-NN works:

Step 1: Decide on the neighbors' K-numbers.

Step 2: Calculate the Euclidean distance between K neighbors in step two.

Step 3: Based on the determined Euclidean distance, select the K closest neighbors.

Step 4: Count the number of data points in each category among these k neighbors.

Step 5: Assign the fresh data points to the category where the neighbor count is highest.

Step 6: Our model is complete.

5.3.3 How should K be chosen in the K-NN Algorithm?

We discuss the following steps while utilising the K-Nearest Neighbours (K-NN) technique to identify the category or class of a fresh data item, x1:

1. Calculate the separation: Find the separation between the newly added data point, x1, and every other data point in the dataset. The type of data will determine which distance metric is used (such as Manhattan distance or Euclidean distance).

2. Choose the K closest neighbours: Determine the K data points that are closest to x1 in terms of distance. The category of x1 will be established based on these K neighbours.

3. Classify x1: Find the dominant class among the K closest neighbours. Give x1 to the class or category that has the greatest prevalence among its K closest neighbours.

5.3.4 Advantages of the KNN Algorithm:

The following benefits of the K-Nearest Neighbours (K-NN) method make it useful and efficient:

1. K-NN is a straightforward method to implement, making it available and user-friendly even for people who are unfamiliar with machine learning.

2. Robustness to noisy data: K-NN is capable of handling noisy training data. Outliers or incorrect data points are less likely to have a substantial impact on the classification or regression findings since the algorithm takes the proximity of the data points into account.

3. Performance with lots of training data: K-NN frequently performs better with more training data. The algorithm can produce more precise predictions when there are more data points available for comparison because it can more easily spot patterns and relationships.

5.3.5 Disadvantages of KNN:

There are a few disadvantages to the K-Nearest Neighbours (K-NN) algorithm that should be considered:

1. K selection: Choosing the appropriate K value—the number of nearest neighbours to take into account—can be difficult. Finding the ideal K value for a certain dataset necessitates extensive thought and testing.

2. Cost of computation: K-NN can be expensive to compute, particularly when working with big training datasets.

3. Sensitivity to irrelevant features: When calculating distances, K-NN treats all features equally. The algorithm's performance can suffer and predictions can be wrong if the dataset contains unimportant or noisy features. 4. Unbalanced data: K-NN can be sensi-

tive to datasets that include one class with a disproportionately high number of samples compared to other classes.

5.4 SUPPORT VECTOR MACHINE CLASSIFICATION ALGORITHM:

In the realm of machine learning classification, Support Vector Machine (SVM) is a frequently used supervised learning approach for tackling Classification and Regression issues. Finding an ideal decision boundary—often referred to as a hyperplane—that successfully divides classes inside an n-dimensional space is the main goal of the SVM method. Future efficient classification of additional data points is made possible by this decision boundary. SVM accomplishes this by locating and picking out extreme vectors and points that help to form the hyperplane. These chosen examples, referred to as support vectors, are crucial to the SVM process.

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

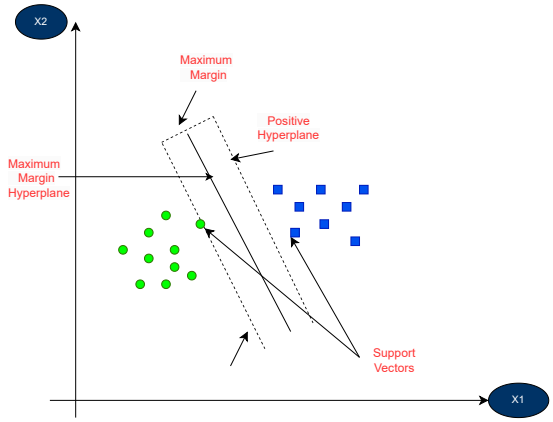


Figure 5.4: Support Vector Machine

5.4.1 Types of Support Vector Machines:

1. An algorithm for classifying data that can be divided into two groups by a single straight line is called linear SVM. Using a Linear SVM classifier, this kind of data—also referred to as linearly separable data—can be efficiently categorized.

2. However, Non-linear SVM is used when a straight line cannot be used to separate the data. A Non-linear SVM classifier must be used to accurately classify the

data points in non-linearly separated data, which is characterised by complicated patterns and relationships.

5.4.2 Hyperplanes in Support Vector Machines:

Multiple lines or decision boundaries may be used to divide various classes in a dataset in an n-dimensional space. Finding the best decision boundary that accurately categorises the data points is vital, though. The hyperplane is the term used in SVM to describe this ideal boundary. The features found in the dataset determine the hyperplane's dimensions. For instance, the hyperplane will be a straight line if the dataset contains only two features. Similar to this, the hyperplane will have two dimensions if there are three features. Building a hyperplane with the highest margin is the goal of SVM, assuring the biggest separation between the data points.

5.4.3 Support Vectors in Support Vector Machines:

The specific data points or vectors that are closest to the hyperplane and have a substantial impact on its position are referred to as support vectors. They are known as "support vectors" because of the support they provide for the hyperplane. These specific vectors, by providing critical details regarding the distribution and division of the classes, play a crucial role in choosing the best decision boundary. They play a key role in the proper classification of data points, which makes their presence in the Support Vector Machine algorithm essential.

5.5 ENSEMBLE LEARNING:

The phrase "ensemble learning" describes a machine learning technique where predictions from several models are integrated to increase the accuracy and resilience of the final forecast. The basic tenet of ensemble learning is to use the collective intelligence of many models to decrease any biases or errors that may exist in individual models, ultimately leading to more accurate predictions.

Some Ensemble learning techniques are as follows:

1. Max Voting: The max voting strategy is frequently used for categorization problems. With this approach, various models are used to forecast each data item. Pre-

dictions from each model are viewed as a ”vote.” The final projection is based on the majority of the forecasts from the models.

2. Averaging: Similar to the max voting method, many forecasts are made for each data point when averaging. In this method, the outcomes of all the models are averaged to provide the final prediction. Averaging can be used to compute probabilities for classification problems or make predictions in regression problems.

3. Weight Averaging: This is a development of the averaging approach. Different weights are assigned to each model, indicating the significance of each model for prediction.

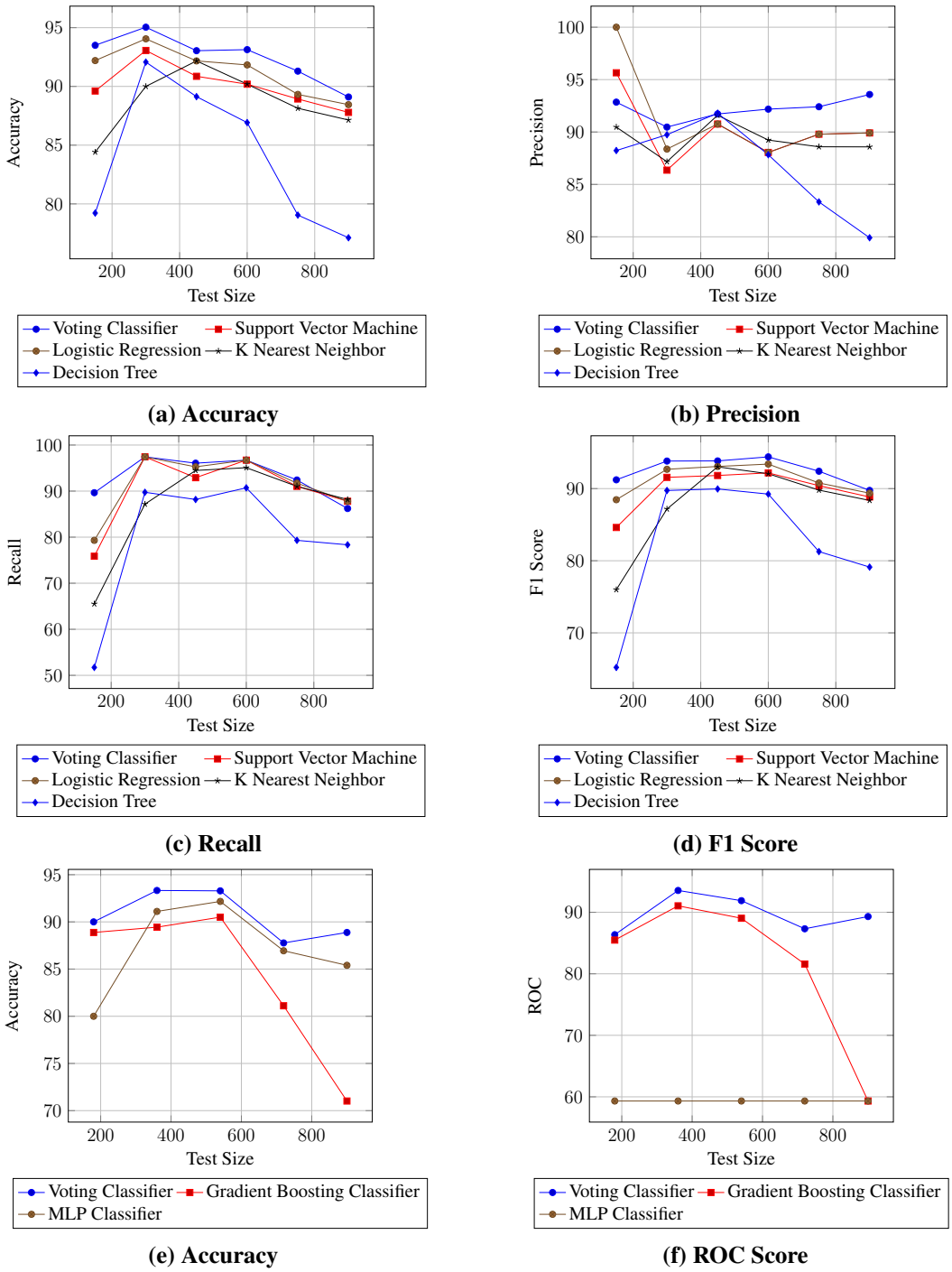
6 Results and Discussion:

6.1 EXPERIMENTAL SETUP AND DATASET:

The experiments were performed on a Ryzen 7 processor of 2.40 GHz clock speed having 16 GB memory. We opted for Python as a tool for implementation. We performed the experiments on 6 datasets which were subdivided from a single large dataset. In this paper, we referred to these datasets as D1, D2, D3, D4, D5, D6. These datasets consist of 12 attributes, among which, one was heart disease which was the result attribute, 1 means the person has heart disease and zero means the person does not have heart disease. Among 918 persons, 508 persons have heart disease and 410 persons do not have heart disease.

6.2 PERFORMANCE ANALYSIS:

The Figures display five machine-learning algorithms represented by different colored lines. We have taken the range from 0 to 900 for the dataset model with an interval of 150 between them. The accuracy of the voting classifier is the highest (92.81) within the test size of 300 among the five algorithms, whereas the decision tree has the lowest accuracy (79.22) in the same test size. The precision of logistic regression is the highest (100) within the test size of 150 and lowest (88.23) for the decision tree, whereas, for larger test sizes that are 900, the precision of the voting classifier is the highest (91.16) and lowest (82.09) for decision tree. The recall of the voting classifier is the highest (96.72) within the interval 600, while the recall of the decision tree is the lowest (51.72) within the test size of 150. The F1 score of the voting classifier is the highest (93.15) within the test size of 600, and the F1 score of the decision tree is the lowest (65.21) within the test size of 150.



(a) Accuracy **(b) Precision**

(c) Recall **(d) F1 Score**

(e) Accuracy **(f) ROC Score**

Figure 6.1: Evaluation parameters with respect to a number of iterations on the datasets

7 Conclusion and Future Scope

Heart disease is a major global health concern and is the leading factor in death globally. It covers a wide range of illnesses that have an impact on the heart and blood arteries, such as coronary artery disease, heart failure, arrhythmias, and valve issues. In order to increase the final model's accuracy and robustness, ensemble learning is a potent machine learning technique that integrates predictions from various classifiers. The ensemble learning model aggregates the outputs of individual classifiers to make a final decision, taking into account the strengths and weaknesses of each classifier. An ensemble model is only effective when the individual classifiers are diverse and complementary. If the classifiers are too similar, the ensemble model may not improve the accuracy of the model significantly. Additionally, the ensemble model may suffer from the problem of bias if one or more of the classifiers are biased. It is important to carefully select and evaluate the individual classifiers used in the voting ensemble to ensure that they are diverse and complementary. We can work on new better datasets and create a UI for this and try some different models.

Bibliography

[1] Battineni, Gopi, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. "Applications of machine learning predictive models in the chronic disease diagnosis." Journal of personalized medicine 10, no. 2 (2020): 21.

[2] Hamsagayathri, P., and S. Vigneshwaran. "Symptoms based disease prediction using machine learning techniques." In 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV), pp. 747-752. IEEE, 2021.

[3] Sharathchandra, D., and M. Raghu Ram. "ML Based Interactive Disease Prediction Model." In 2022 IEEE Delhi Section Conference (DELCON), pp. 1-5. IEEE, 2022.

[4] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1 (2020): 1-6.

[5] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019.

[6] Krishnamoorthi, Raja, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, and Basant Tiwari. "A novel diabetes healthcare disease prediction framework using machine learning techniques." Journal of Healthcare Engineering 2022 (2022).

[7] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[8] Maniruzzaman, Md, Md Jahanur Rahman, Benojir Ahammed, and Md Menhazul Abedin. ”Classification and prediction of diabetes disease using machine learning paradigm.” Health information science and systems 8 (2020): 1-14.

[9] Khanam, Jobeda Jamal, and Simon Y. Foo. ”A comparison of machine learning algorithms for diabetes prediction.” ICT Express 7, no. 4 (2021): 432-439.

[10] Nawab, Mohammad Ali, Tavisi Malpani, Tushar Kaundal, and Ms Divya Soni. ”DISEASE PREDICTION WEB APP USING MACHINE LEARNING.”