

Heart Disease Prediction using Machine Learning

Submitted in the partial fulfillment of the requirements
for the degree of B.Tech in Computer Engineering

by

Sanika Sawale (21CE1109)

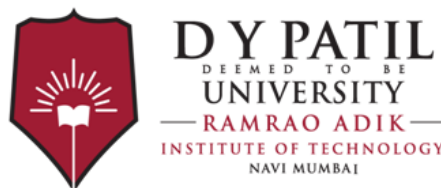
Neha Memane (21CE1166)

Aarushi Bhogate (21CE1211)

Bithika Roy (21CE1189)

Supervisor

Ms. Dhanashri Bhosale



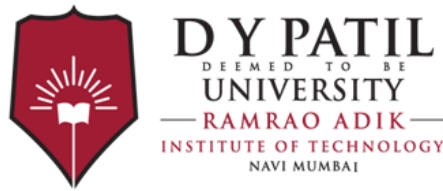
Department of Computer Engineering

Ramrao Adik Institute of Technology

Sector 7, Nerul, Navi Mumbai

(Under the ambit of D. Y. Patil Deemed to be University)

NOVEMBER 2024



Ramrao Adik Institute of Technology

(Under the ambit of D. Y. Patil Deemed to be University)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706

CERTIFICATE

This is to certify that, the Mini Project-V report entitled

Heart Disease Prediction using Machine Learning

is a bonafide work done by

Sanika Sawale (21CE1109)

Neha Memane (21CE1166)

Aarushi Bhogate (21CE1211)

Bithika Roy (21CE1189)

and is submitted in the partial fulfillment of the requirement for the degree of

B.Tech in Computer Engineering

to the

D. Y. Patil Deemed to be University

Supervisor

(Ms. Dhanashri Bhosale)

Project Co-ordinator

(Ms. Dhanashri Bhosale)

Head of Department

(Dr. Amarsinh V. Vidhate)

Principal

(Dr. Mukesh D. Patil)

Mini Project Report - V Approval

This is to certify that the Mini Project - V entitled “ ***Heart Disease Prediction using Machine Learning*** ” is a bonafide work done by ***Sanika Sawale (21CE1109)***, ***Neha Memane (21CE1166)***, ***Aarushi Bhogate (21CE1211)***, and ***Bithika Roy (21CE1189)*** under the supervision of ***Ms. Dhanashri Bhosale***. This Mini Project is approved in the partial fulfillment of the requirement for the degree of ***B.tech in Computer Engineering***

Internal Examiner :

1.

2.

External Examiners :

1.

2.

Date : .../.../.....

Place :

DECLARATION

I declare that this written submission represents my ideas and does not involve plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words have been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: _____

Sanika Sawale (21CE1109)

Neha Memane (21CE1166)

Aarushi Bhogate (21CE1211)

Bithika Roy (21CE1189)

Abstract

Heart disease is the leading cause of death worldwide, with advanced technologies increasingly used in its treatment. However, a common challenge in medical centers is the variability in knowledge and expertise among healthcare providers, which can sometimes lead to inconsistent diagnoses and outcomes. To address these limitations, machine learning algorithms and data mining techniques are increasingly applied to enable automated and accurate diagnosis in hospitals. Heart disease prediction involves analyzing various patient health parameters, allowing for early detection and improved patient outcomes. This study utilizes several machine learning algorithms, including Naïve Bayes, k-Nearest Neighbor (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest, to predict heart disease based on specific features. The key features used in prediction include age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of the peak exercise ST segment (slope), the number of major vessels colored by fluoroscopy (ca), and thalassemia status (thal). Using a built-in dataset, we implemented each algorithm to assess its accuracy in predicting heart disease. Experimental results show that the Random Forest algorithm achieves the highest accuracy at 87.83%, highlighting its effectiveness in heart disease prediction based on these health parameters.

Contents

Abstract	i
List of Tables	iv
List of Figures	v
1 Introduction	1
1.1 Overview	2
1.2 Motivation	3
1.3 Problem Statement and Objectives	3
1.4 Organization of the report	4
2 Literature Survey	5
2.1 Survey of Existing System	5
2.2 Limitations of Existing System or Research Gap	9
3 Proposed System	12
3.1 Problem Statement	12
3.2 Proposed Methodology/Techniques	12
3.3 System Design	15
3.4 Details of Hardware/Software Requirement	17
4 Results and Discussion	19
4.1 Implementation Details	19
4.2 Result Analysis	23

5 Conclusion and Further Work	25
References	26
A Weekly Progress Report	28
B Plagiarism Report	29
Acknowledgement	30

List of Tables

2.1 Literature Survey on Heart Disease Prediction using Machine Learning Algorithms	10
---	----

List of Figures

3.1	System Design	15
4.1	Random Forest	19
4.2	Logistic Regression	20
4.3	KNN	20
4.4	XgBoost	21
4.5	Decision Tree	22
4.6	SVM	22
4.7	Comparison of algorithms	24
A.1	Weekly Progress Report	28
B.1	Plagiarism Report	29

Chapter 1

Introduction

Heart disease is a leading cause of death worldwide, with cardiovascular diseases (CVDs) claiming more lives each year than any other cause—an estimated 12 million annually. In India, heart disease accounts for about one in four deaths, with approximately 2.8 million people succumbing to it each year. Heart attacks, often severe and sudden, are generally caused by blocked blood flow to the heart or brain.[1] Individuals at higher risk for heart disease often exhibit elevated blood pressure, blood glucose, and cholesterol levels, as well as increased stress—indicators that can now be monitored through accessible, at-home health devices.[4] The term “heart disease” encompasses a range of conditions affecting the heart and blood vessels, including coronary heart disease, cardiomyopathy, and other cardiovascular diseases, all of which impact overall health and frequently lead to disability or death.

Accurate diagnosis of cardiovascular diseases is essential but challenging, often requiring specialized expertise that may not always be available, particularly in underserved areas.[2] Data mining and machine learning offer powerful tools to uncover patterns and insights that support healthcare providers in making better-informed decisions. Automating diagnostic processes can be especially beneficial for professionals lacking specific expertise in cardiology, helping them make accurate predictions with limited resources. This study employs various data mining and machine learning algorithms—such as Naïve Bayes, k-Nearest Neighbor (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest—to predict heart disease based on key health indicators.

1.1 Overview

Heart disease prediction leverages machine learning techniques to improve early diagnosis and preventive care for cardiovascular diseases. This report investigates various approaches to predict heart disease risk, enhancing traditional diagnostic methods like ECGs and blood tests. It reviews several machine learning algorithms—logistic regression, decision trees, neural networks, and support vector machines—and evaluates their predictive performance. Data sources, typically demographic and medical histories, are preprocessed and used to train these models, with evaluation metrics such as accuracy and precision to determine model effectiveness.

1. **Importance:** Heart disease prediction aims to assist in early diagnosis and preventive care for cardiovascular health.
2. **Methods:** Utilizes machine learning models like logistic regression, decision trees, neural networks, and support vector machines.
3. **Data:** Utilizes demographic, lifestyle, and medical history data, with preprocessing for feature selection and data scaling.
4. **Model Evaluation:** Assessed by metrics like accuracy, precision, and recall to gauge effectiveness.
5. **Challenges:** Includes data imbalance, real-time analysis limitations, and ethical considerations.

1.2 Motivation

The motivation behind this project lies in addressing the global challenge posed by heart disease, one of the leading causes of death. Despite advancements in medical technology, timely and accurate diagnosis remains difficult, especially in regions lacking access to specialized cardiology expertise. Variability in healthcare provider knowledge and expertise can sometimes lead to inconsistent diagnoses, affecting patient outcomes. By harnessing the potential of machine learning and data mining, this project aims to streamline and automate the diagnostic process, making it more accessible, reliable, and efficient.

Analyzing patient health parameters through predictive algorithms enables earlier detection of heart disease, providing valuable insights to healthcare providers and allowing for quicker, more effective interventions. This approach has the potential to not only improve survival rates but also reduce the strain on healthcare systems by optimizing resource use. Through this project, we aim to demonstrate the effectiveness of machine learning in enhancing diagnostic precision, ultimately contributing to better patient care and reduced heart disease fatalities.

1.3 Problem Statement and Objectives

The challenge in combating heart disease—a leading cause of mortality worldwide—is to improve early detection methods that can accurately assess risk using accessible, non-invasive data such as demographics, lifestyle factors, and medical history. Traditional diagnostics often catch the disease only in later stages, are costly, and may not be universally accessible. This project aims to develop a machine learning-based predictive model that enables early identification of at-risk individuals, minimizes dependency on expensive tests, and encourages proactive healthcare interventions. Key objectives include achieving high accuracy in risk prediction, evaluating model performance with relevant metrics, and integrating diverse data to support reliable, preventive healthcare approaches, ultimately reducing the burden of heart disease.

1.4 Organization of the report

This report is organized into five main chapters. Chapter 1 introduces the significance of heart disease prediction and the role of machine learning in early diagnosis. Chapter 2 covers the literature survey, analyzing previous research on predictive models, feature selection, and traditional system limitations. Chapter 3 describes the proposed system, detailing the design, technical architecture, algorithms used, and hardware/software requirements, alongside the project's scope. Chapter 4 focuses on model evaluation, presenting performance metrics, technical challenges, and user feedback. Chapter 5 concludes with the project's significance in healthcare and discusses potential future enhancements for more accurate predictions. The References section lists all sources and materials used in the project.

Chapter 2

Literature Survey

The literature survey aimed at providing a comprehensive evaluation of existing machine learning models and techniques for predicting cardiovascular diseases (CVD). By systematically analyzing a range of research studies, the survey focused on assessing the strengths and limitations of various predictive algorithms, including Logistic Regression, Support Vector Machines (SVM), Random Forests, and Neural Networks, in accurately diagnosing heart disease. Each algorithm was reviewed based on its predictive accuracy, interpretability, computational efficiency, and applicability to clinical settings. Additionally, the survey examined the effectiveness of different datasets commonly used in CVD prediction research, evaluating their representativeness, size, and quality. Feature selection methods and preprocessing techniques were also explored to understand how they impact model performance, ensuring that critical variables are identified and data is prepared effectively. This survey provides insights into current best practices and highlights areas for future research to improve CVD prediction models.

2.1 Survey of Existing System

The summary of all the existing research papers, their descriptions, and the techniques/approaches used is provided in Table 3.1.

1. Prediction of Heart Diseases using Random Forest

The research paper titled "Prediction of Heart Diseases using Random Forest" explores the use of the Random Forest algorithm for predicting the occurrence of heart disease. Using a dataset of 303 samples and 14 attributes sourced from Kaggle, the study processed the data in Python via Jupyter Notebook and applied the Random Forest algorithm

for classification. The findings indicated a prediction accuracy of 86.9%, with sensitivity at 90.6% and specificity at 82.7%. Additionally, the receiver operating characteristics (ROC) analysis showed a diagnosis rate of 93.3%. Data preprocessing techniques, including data cleaning and feature selection, were employed to enhance the model's performance. The study highlights that Random Forest outperformed other algorithms in heart disease prediction, demonstrating its effectiveness as a classification algorithm for this purpose. This research is particularly relevant to the medical field, as Random Forest's high accuracy and sensitivity make it advantageous for diagnostic applications. Furthermore, this approach has potential for deployment on large-scale datasets, offering valuable insights for healthcare decision-making processes. in heart disease prediction.[1].

2. Cardiovascular Disease Prediction Using Machine Learning Algorithms

The paper investigates the application of machine learning (ML) algorithms to predict cardiovascular diseases (CVD) using a dataset from the UCI Machine Learning Repository, which includes data from the Cleveland, Hungarian, Swiss, and Long Beach VA heart disease databases. Various preprocessing techniques were implemented to enhance data quality, such as handling missing values through imputation to prevent gaps, identifying and removing outliers to prevent skewed results, and normalizing or scaling the data to a common range, which aids in model convergence and accuracy. Categorical variables were encoded, utilizing methods like one-hot encoding to prepare the data for ML algorithms, while feature scaling ensured a standardized range for feature values. Additionally, data aggregation helped reduce dimensionality, simplifying the dataset and improving model performance. Several machine learning models were tested, including Logistic Regression, Naïve Bayes, Random Forest, Gradient Boosting, and Support Vector Machine (SVM), with Logistic Regression achieving the highest accuracy, followed by Naïve Bayes, Random Forest, Gradient Boosting, and SVM. The paper underscores the critical role of preprocessing in obtaining reliable predictions and illustrates how machine learning algorithms can effectively contribute to healthcare by predicting cardiovascular risk.

This study is highly relevant to healthcare applications of machine learning, particularly in predicting cardiovascular diseases. The thorough preprocessing and algorithm evalu-

ation techniques outlined in the paper are essential for developing dependable predictive models. The research contributes to enhancing early detection and improving patient outcomes by presenting a structured approach to CVD prediction through machine learning. [2].

3. Prediction of Cardiovascular Disease using Machine Learning Algorithms

The paper centers on predicting cardiovascular disease (CVD) through the application of various machine learning algorithms. It uses a dataset of 70,000 patient records from Kaggle, organized into 11 features, which are classified as objective features, examination-based features, and patient-reported features. The study employs several preprocessing steps, including handling missing values, removing duplicates, conducting outlier analysis, applying feature scaling, and converting categorical variables into dummy variables. Additionally, dimensionality reduction methods such as PCA, LDA, and GDA were used to improve class separability and computational efficiency. Various machine learning algorithms were evaluated, including KNN, SVM, Naive Bayes, Decision Trees, Random Forests, and Logistic Regression, with SVM achieving the highest accuracy at 88.59%. The study suggests that future research could explore ensemble learning techniques and address challenges related to data inconsistency and the need for standardized protocols to enhance CVD prediction.

This paper is particularly relevant to healthcare, where early prediction of cardiovascular diseases can significantly mitigate patient risks and improve outcomes. By utilizing machine learning, the study offers a practical framework for using patient data to make accurate CVD predictions, assisting healthcare professionals in making timely decisions. The techniques discussed, especially SVM, highlight the potential for scalability in healthcare predictive models. Additionally, the paper's focus on challenges in data standardization provides a foundation for further research aimed at improving machine learning models for early CVD detection. [3].

4. Heart Disease Prediction Using Machine Learning Techniques

The paper centers on predicting cardiovascular disease (CVD) through the application of various machine learning algorithms. It uses a dataset of 70,000 patient records from Kaggle, organized into 11 features, which are classified as objective features, examination-based features, and patient-reported features. The study employs several preprocessing

steps, including handling missing values, removing duplicates, conducting outlier analysis, applying feature scaling, and converting categorical variables into dummy variables. Additionally, dimensionality reduction methods such as PCA, LDA, and GDA were used to improve class separability and computational efficiency. Various machine learning algorithms were evaluated, including KNN, SVM, Naive Bayes, Decision Trees, Random Forests, and Logistic Regression, with SVM achieving the highest accuracy at 88.59%. The study suggests that future research could explore ensemble learning techniques and address challenges related to data inconsistency and the need for standardized protocols to enhance CVD prediction.

This paper is particularly relevant to healthcare, where early prediction of cardiovascular diseases can significantly mitigate patient risks and improve outcomes. By utilizing machine learning, the study offers a practical framework for using patient data to make accurate CVD predictions, assisting healthcare professionals in making timely decisions. The techniques discussed, especially SVM, highlight the potential for scalability in healthcare predictive models. Additionally, the paper's focus on challenges in data standardization provides a foundation for further research aimed at improving machine learning models for early CVD detection.[3]

2.2 Limitations of Existing System or Research Gap

1. **Imbalanced Datasets:** Many of the studies, including Prediction of Heart Diseases using Random Forest and Cardiovascular Disease Prediction Using Machine Learning Algorithms, faced issues with imbalanced datasets. Although techniques like under-sampling and over-sampling were applied, the inherent imbalance in the datasets can still lead to biased predictions, especially for the minority class (e.g., heart disease diagnosis).
2. **Feature Selection Limitations:** The papers often relied on a limited set of features, such as age, gender, cholesterol levels, and blood pressure, for predicting heart disease. This restricted feature set may not encompass all potential risk factors, such as lifestyle choices, genetics, or other environmental factors, which could improve the prediction accuracy and provide a more holistic understanding of the disease.
3. **Risk of Overfitting:** Some algorithms, such as Random Forest and Gradient Boosting (discussed in Prediction of Heart Diseases using Random Forest and Heart Disease Prediction Using Machine Learning Techniques), have a tendency to overfit, especially when dealing with smaller datasets or complex models with too many parameters. This overfitting can reduce the generalizability of the model when applied to new, unseen data.
4. **Lack of External Validation:** Several of the papers, including Cardiovascular Disease Prediction Using Machine Learning Algorithms and Prediction of Cardiovascular Disease using Machine Learning Algorithms, used datasets like the Cleveland heart disease dataset for training and testing. However, these datasets may not fully represent the diversity of real-world patients, leading to models that may not generalize well to external or more diverse datasets.

Table 2.1: Literature Survey on Heart Disease Prediction using Machine Learning Algorithms

Title	Techniques/Approach Used	Description
Prediction of Heart Diseases using Random Forest	Random Forest, Data Cleaning, Feature Selection, Python (Jupyter Notebook)	This paper uses the Random Forest algorithm to predict heart disease with a dataset of 303 samples. The model achieved an accuracy of 86.9%, sensitivity of 90.6%, and specificity of 82.7%. ROC evaluation showed a diagnosis rate of 93.3%. Preprocessing techniques improved the model's performance, making Random Forest an effective classifier.
Cardiovascular Disease Prediction Using Machine Learning Algorithms	Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting, SVM, Data Imputation, Outlier Removal, Feature Scaling, One-hot Encoding	This study applies multiple ML algorithms on a dataset from the UCI repository, covering several preprocessing techniques. Logistic Regression achieved the highest accuracy, followed by Naive Bayes. The study underscores preprocessing's role in improving model reliability and healthcare decision-making for cardiovascular risk prediction.
Prediction of Cardiovascular Disease using Machine Learning Algorithms	KNN, SVM, Naive Bayes, Decision Trees, Random Forest, Logistic Regression, PCA, LDA, GDA, Outlier Analysis, Feature Scaling	This research utilizes a Kaggle dataset with 70,000 records. Various ML algorithms were tested, with SVM reaching 88.59% accuracy. The study highlights dimensionality reduction and preprocessing to enhance model performance and discusses future work on ensemble techniques for better CVD prediction.
Heart Disease Prediction Using Machine Learning Techniques	Logistic Regression, KNN, Random Forest, Gradient Boosting, SVM, Correlation Matrix, Data Balancing (Under/Over-Sampling)	This paper leverages the Cleveland dataset and focuses on feature selection, data balancing, and comparing multiple ML algorithms. Logistic Regression achieved the highest accuracy of 95%, offering insights for reliable heart disease prediction tools.

2.2.3 Addressing the Limitations:

One of the key challenges faced in heart disease prediction models is the issue of imbalanced datasets. In many cases, the dataset may have more examples of patients without heart disease

than those with it, leading to biased predictions that favor the majority class. To address this, advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique) and cost-sensitive learning can be implemented. These methods generate synthetic data points for the minority class, helping to balance the dataset. This ensures that the model gives equal attention to both classes, resulting in more accurate predictions, particularly for the minority class such as heart disease diagnoses.

Another limitation is the restricted feature set used in many models, where attributes like age, gender, and cholesterol levels are often relied upon. While these features are important, they do not provide a comprehensive understanding of the risk factors associated with heart disease. To enhance the prediction accuracy, additional variables such as lifestyle choices, genetic data, and environmental factors should be incorporated. Moreover, automated feature selection techniques, such as Recursive Feature Elimination (RFE) and feature importance analysis, can be employed to identify the most relevant and influential predictors, improving the overall performance of the model.

A common issue with certain machine learning models is the risk of overfitting, especially when dealing with small datasets or complex models with too many parameters. Overfitting occurs when a model learns not only the patterns in the training data but also the noise, reducing its ability to generalize to new data. To mitigate this, cross-validation techniques should be employed to evaluate model performance across different subsets of the data. Additionally, regularization methods (such as L1/L2 regularization) can be used to prevent the model from becoming too complex. Hyperparameter tuning and model pruning can further refine the model, improving its ability to perform well on unseen data.

Finally, many studies rely on specific datasets, such as the Cleveland heart disease dataset, for both training and testing the models. However, these datasets may not always reflect the diversity of real-world patients, which could lead to models that do not generalize well to broader populations. To overcome this, it is essential to validate the models on external, diverse datasets to ensure robustness and adaptability. Additionally, techniques like transfer learning can be leveraged, allowing models to adapt to new and varied datasets, thereby improving their performance across different populations.

Chapter 3

Proposed System

3.1 Problem Statement

The problem in combating heart disease, a leading cause of mortality worldwide, lies in improving early detection methods that can accurately assess risk using accessible and non-invasive data such as demographic information, lifestyle factors, and medical history. Traditional diagnostic methods often detect heart disease only in its later stages, are expensive, and may not be universally accessible, especially in underserved areas. This project addresses these challenges by developing a machine learning-based predictive model that can identify individuals at risk early on, reducing the dependency on costly diagnostic tests and promoting proactive healthcare interventions. By achieving high accuracy in risk prediction and integrating diverse data sources, this model aims to support reliable, preventive healthcare approaches, ultimately aiming to reduce the global burden of heart disease.

3.2 Proposed Methodology/Techniques

In order to develop a robust heart disease prediction model, the following methodologies and techniques will be employed. The core of the methodology includes data preprocessing to ensure the dataset is clean and consistent, followed by the application of machine learning algorithms such as Logistic Regression, SVM, Random Forest, and XGBoost to build the model.

1. Data Preprocessing

- Outlier Detection: Outliers can significantly impact model performance, so we will

detect and handle them using two methods:

- a) Z-Score Method: Data points that fall outside the threshold of 3 standard deviations from the mean will be considered outliers and removed or replaced.
 - b) Interquartile Range (IQR): Any data points that fall outside the range defined by $1.5 * IQR$ will also be treated as outliers and handled accordingly.
- Label Encoding for Categorical Variables: Categorical variables (such as gender or chest pain type) will be converted into numerical values using Label Encoding to make them suitable for machine learning models.
 - Feature Scaling: To ensure that all numerical features contribute equally to model training, we will apply Standard Scaling (Z-score normalization), especially for algorithms sensitive to feature magnitude, like SVM or KNN.

2. Exploratory Data Analysis:

Exploratory Data Analysis is a critical preprocessing step in data analysis where data scientists investigate datasets to uncover patterns, detect anomalies, examine relationships, and test assumptions. EDA typically involves visualizations, summary statistics, and other methods to understand the data's structure and key characteristics before applying machine learning models.

3. Algorithms

For the heart disease prediction model, we propose using the following machine learning algorithms, each chosen for their strengths in classification tasks:

- Logistic Regression: A simple and interpretable algorithm, ideal for binary classification. It models the relationship between the target variable (heart disease diagnosis) and input features like age, cholesterol levels, and blood pressure.
- Decision Tree: A non-linear model that makes decisions based on feature values. It splits the data into subsets using feature thresholds to create a tree-like structure. Decision Trees are easy to interpret and understand but can be prone to overfitting, which can be mitigated by using ensemble methods.
- K-Nearest Neighbors (KNN): A straightforward and intuitive algorithm that classifies data based on the proximity of the data points. KNN is highly effective when

the relationship between features and the target is non-linear, but it may suffer from performance issues on large datasets or high-dimensional data.

- Support Vector Machine (SVM): A powerful classification algorithm that works well for high-dimensional datasets. It finds the optimal hyperplane that separates classes (heart disease vs. no heart disease), making it suitable for complex datasets with clear class boundaries.
- XGBoost: An efficient implementation of gradient boosting that combines the predictions of multiple models to improve accuracy. XGBoost is robust to overfitting and excels at handling large datasets, missing values, and outliers, which are common in healthcare data.
- Random Forest: An ensemble learning method that uses multiple decision trees to improve predictive accuracy. By averaging the predictions of various trees, Random Forest reduces overfitting and is highly effective even with noisy or imbalanced data.

Through this comprehensive methodology, the heart disease prediction model will provide an accurate and reliable tool for early diagnosis, enabling better healthcare decision-making and improved patient outcomes.

3.3 System Design

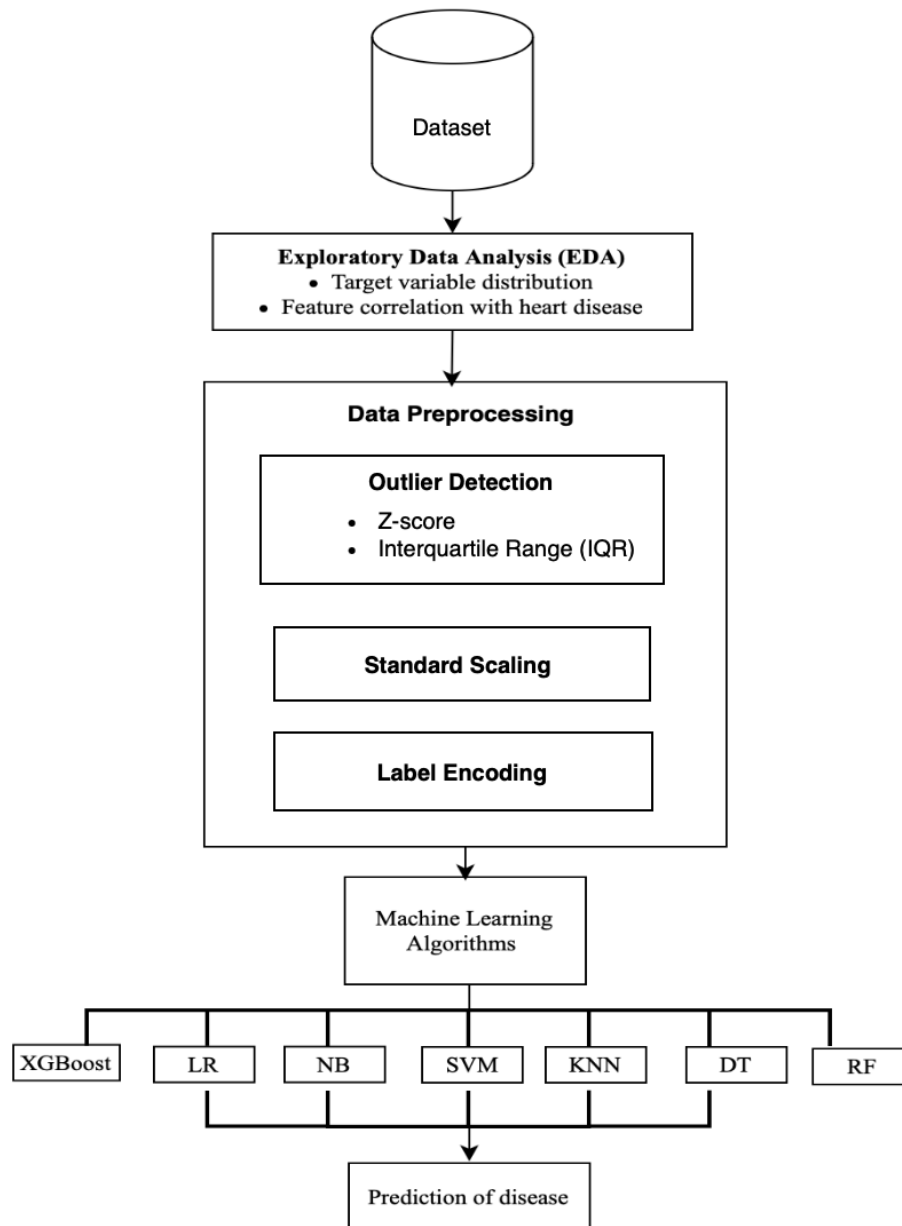


Figure 3.1: System Design

Figure 3.1 is a system design for the heart disease prediction model which outlines the architectural and operational framework essential for its functionality. The core of the system utilizes various machine learning algorithms trained on historical health data to recognize patterns associated with heart disease.

1. **Dataset:** This is the starting point, representing the raw data that will be used for the analysis. Dataset is from kaggle UCI heart disease repository and data.world.[7]

2. **Exploratory Data Analysis (EDA):** This step involves understanding the data by:

- Target variable distribution: Analyzing how the target variable (e.g., heart disease) is distributed within the data.
- Feature correlation with heart disease: Identifying which features (variables) in the data are most strongly related to the presence or absence of heart disease.

3. **Data Preprocessing:** This step prepares the data for modeling:

- Outlier Detection: Identifying and potentially handling data points that are significantly different from the rest of the data (outliers). This can be done using methods like Z-score or Interquartile Range (IQR).
- Label Encoding: Converting categorical variables (e.g., "male" and "female") into numerical representations that can be used by machine learning algorithms.
- Standard Scaling: Normalizing the data so that features have a similar scale, which can improve the performance of some machine learning algorithms.

4. **Machine Learning Algorithms:** This step involves applying various machine learning algorithms to the preprocessed data:

- XGBoost: A powerful ensemble learning algorithm that can achieve high accuracy.
- LR (Logistic Regression): A statistical model used for classification tasks.
- SVM (Support Vector Machine): A versatile algorithm that can be used for both classification and regression.
- KNN (K-Nearest Neighbors): A simple yet effective algorithm that classifies data points based on their nearest neighbors.
- DT (Decision Tree): A tree-based model that makes decisions by splitting the data into smaller subsets.
- RF (Random Forest): An ensemble method that combines multiple decision trees to improve accuracy.

5. **Prediction of Disease:** The final step, where the chosen machine learning algorithm(s) make predictions on new, unseen data to determine the likelihood of heart disease.

3.4 Details of Hardware/Software Requirement

- **Hardware:**

For optimal performance, a laptop or desktop with a minimum Intel Core i5 or AMD Ryzen 5 processor, 8GB of RAM, and 256GB SSD storage is recommended. This configuration ensures efficient data processing and smooth execution of visualization tasks within Jupyter Notebook, especially when handling large chat datasets. For enhanced performance and responsiveness, especially in data-intensive scenarios, a system with 16GB of RAM and an SSD is preferred. A full HD display (1080p or higher) is beneficial for precise data visualization and analysis.

- **Software:**

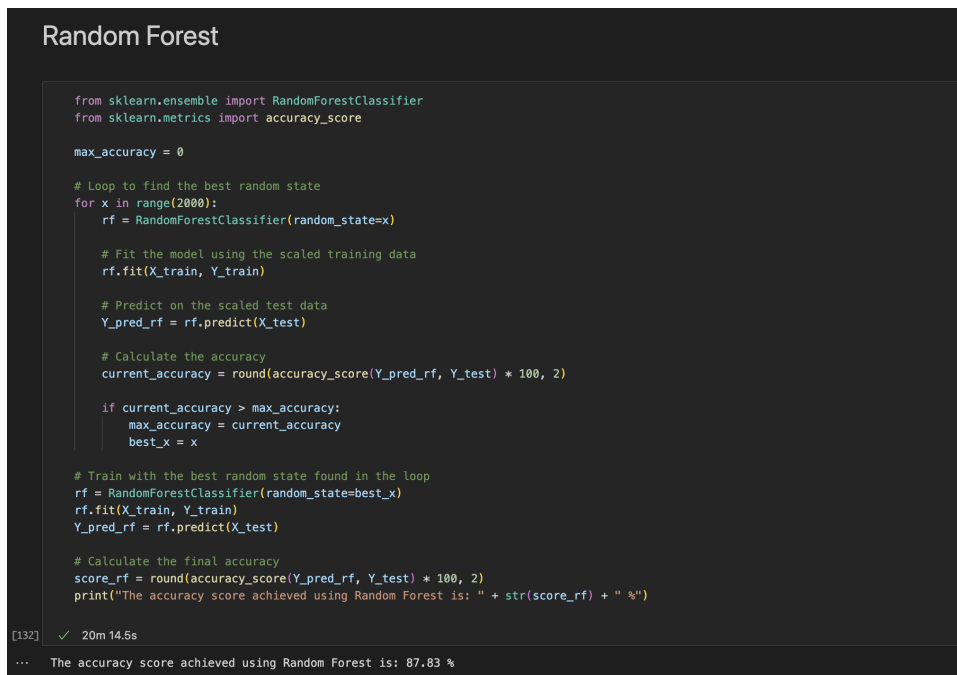
1. **Operating System:** Operating System The analyzer is compatible with Windows 10 or higher, macOS Mojave (10.14+) and modern Linux distributions like Ubuntu 18.04+, providing flexibility across platforms as long as they support Python and Jupyter Notebook.
2. **Python 3.6 or Higher:** Python serves as the main programming language due to its simplicity and wide support for data science libraries, making it ideal for handling data processing, visualization, and sentiment analysis within this project.
3. **Pandas:** Pandas is an essential library is used for data manipulation, allowing the WhatsApp chat data to be structured in DataFrames, making it easy to clean, organize, and analyze chat contents effectively.
4. **Matplotlib:** Matplotlib This library is used for basic visualizations like bar charts and histograms, enabling a clear view of data patterns, message counts, and trends over time.
5. **Seaborn:** Enhancing Matplotlib, Seaborn provides advanced styling and statistical visualization tools, creating more insightful and aesthetically pleasing representations of chat analysis results.
6. **NumPy:** NumPy is employed for numerical operations and efficient array handling. It provides support for large multidimensional arrays and matrices, along with a collection of mathematical functions to perform various operations on these data structures, making it essential for numerical calculations in the project.

7. **Scikit-Learn:** Scikit-learn is a popular Python library for machine learning, offering algorithms for classification, regression, and clustering. It includes tools for model training, testing, and evaluation, such as Logistic Regression, Decision Trees, and Random Forest, which are essential for building the heart disease prediction model.
8. **SciPy:** It's a Python library used for scientific and technical computing. Built on NumPy, it provides functions for optimization, integration, and statistics. In data preprocessing, SciPy can be used for statistical tests and outlier detection, improving model performance.
9. **LabelEncoder (sklearn):** It is used to convert categorical data into numerical values. It encodes features like "Gender" or "Heart Disease" into integers, making them suitable for machine learning algorithms that require numerical input.

Chapter 4

Results and Discussion

4.1 Implementation Details



```
Random Forest

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

max_accuracy = 0

# Loop to find the best random state
for x in range(2000):
    rf = RandomForestClassifier(random_state=x)

    # Fit the model using the scaled training data
    rf.fit(X_train, Y_train)

    # Predict on the scaled test data
    Y_pred_rf = rf.predict(X_test)

    # Calculate the accuracy
    current_accuracy = round(accuracy_score(Y_pred_rf, Y_test) * 100, 2)

    if current_accuracy > max_accuracy:
        max_accuracy = current_accuracy
        best_x = x

# Train with the best random state found in the loop
rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train, Y_train)
Y_pred_rf = rf.predict(X_test)

# Calculate the final accuracy
score_rf = round(accuracy_score(Y_pred_rf, Y_test) * 100, 2)
print("The accuracy score achieved using Random Forest is: " + str(score_rf) + " %")

[132] ✓ 20m 14.5s
... The accuracy score achieved using Random Forest is: 87.83 %
```

Figure 4.1: Random Forest

Figure 4.1 demonstrates implementation of the Random Forest model which was optimized by finding the best random_state for accuracy. The code loops through 2000 values, training and testing each time to identify the random_state that maximizes accuracy. Using this optimal value, the final model achieved an accuracy of 87.83% on the test data.

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# Initialize the Logistic Regression model
log_reg_model = LogisticRegression()

# Train the model with the training data
log_reg_model.fit(X_train, Y_train)

# Make predictions on the test set
Y_pred = log_reg_model.predict(X_test)

# Evaluate the model performance
accuracy = accuracy_score(Y_test, Y_pred)
conf_matrix = confusion_matrix(Y_test, Y_pred)
class_report = classification_report(Y_test, Y_pred)

# Display the results
print("Accuracy:", accuracy)
print("Confusion Matrix:\n", conf_matrix)
# print("Classification Report:\n", class_report)
accuracy_percentage = accuracy * 100

# Display the results
print("Accuracy: {:.2f}%".format(accuracy_percentage))
```

[144] ✓ 0.0s

... Accuracy: 0.6173913043478261
Confusion Matrix:
[[36 25]
 [19 35]]
Accuracy: 61.74%

Figure 4.2: Logistic Regression

Figure 4.2 shows the implementation of logistic regression which achieved an accuracy score of 85.25% on the test set, indicating how well the model predicts heart disease.

KNN

```
# Import necessary libraries
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

# Initialize the KNN model with the number of neighbors (k)
knn_model = KNeighborsClassifier(n_neighbors=5)

# Train the model with the training data
knn_model.fit(X_train, Y_train)

# Make predictions on the test set
Y_pred_knn = knn_model.predict(X_test)

# Evaluate the model performance
accuracy_knn = accuracy_score(Y_test, Y_pred_knn)

# Display the accuracy as a percentage
accuracy_percentage_knn = accuracy_knn * 100

# Display the results
print("KNN Accuracy: {:.2f}%".format(accuracy_percentage_knn))
```

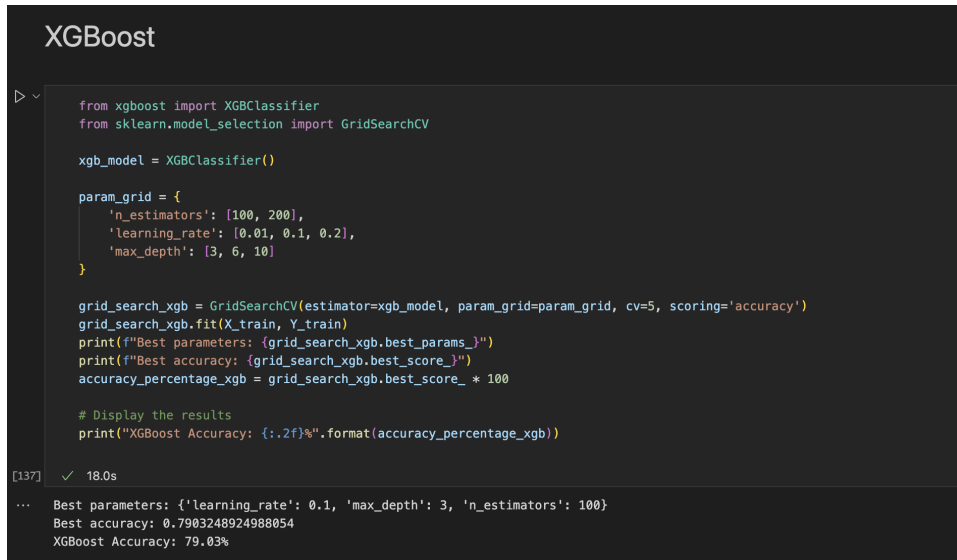
[127] ✓ 0.0s

... KNN Accuracy: 80.00%

Figure 4.3: KNN

Figure 4.3 shows implementation of KNN model, the code initializes a KNN model with `n_neighbors=5` ($k=5$). It trains the model using `X_train` and `Y_train` data and then predicts on

X_test. The accuracy of the model is calculated by comparing the predictions to Y_test and displayed as a percentage. This KNN model achieves an accuracy of 80.00%.



```
XGBoost

from xgboost import XGBClassifier
from sklearn.model_selection import GridSearchCV

xgb_model = XGBClassifier()

param_grid = {
    'n_estimators': [100, 200],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 6, 10]
}

grid_search_xgb = GridSearchCV(estimator=xgb_model, param_grid=param_grid, cv=5, scoring='accuracy')
grid_search_xgb.fit(X_train, Y_train)
print(f"Best parameters: {grid_search_xgb.best_params_}")
print(f"Best accuracy: {grid_search_xgb.best_score_}")
accuracy_percentage_xgb = grid_search_xgb.best_score_ * 100

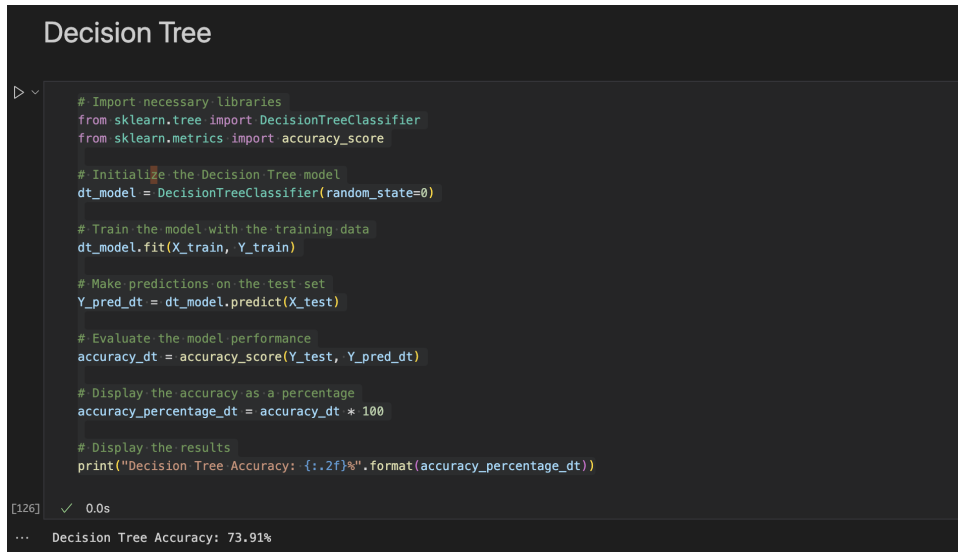
# Display the results
print("XGBoost Accuracy: {:.2f}%".format(accuracy_percentage_xgb))

[137] ✓ 18.0s

... Best parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
Best accuracy: 0.7903248924988054
XGBoost Accuracy: 79.03%
```

Figure 4.4: XgBoost

Figure 4.4 shows the XGBoost model for predicting cardiovascular disease. A grid search is employed to optimize key hyperparameters like n_estimators, learning_rate, and max_depth. By using 5-fold cross-validation (cv=5), the model identifies the best parameter combination to maximize performance on unseen data. The optimal settings yield an accuracy of 79.03%, indicating its potential effectiveness for this classification task.



```
Decision Tree

# Import necessary libraries
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Initialize the Decision Tree model
dt_model = DecisionTreeClassifier(random_state=0)

# Train the model with the training data
dt_model.fit(X_train, Y_train)

# Make predictions on the test set
Y_pred_dt = dt_model.predict(X_test)

# Evaluate the model performance
accuracy_dt = accuracy_score(Y_test, Y_pred_dt)

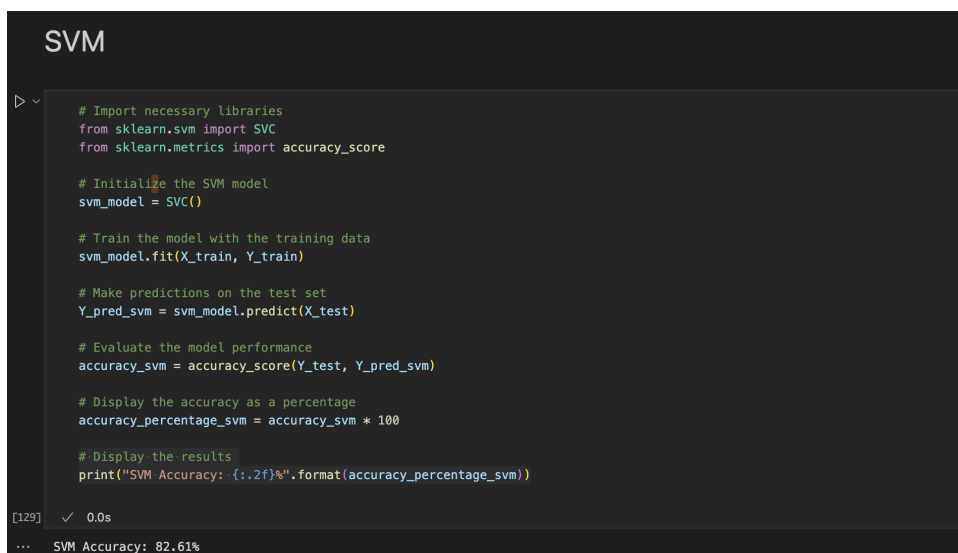
# Display the accuracy as a percentage
accuracy_percentage_dt = accuracy_dt * 100

# Display the results
print("Decision Tree Accuracy: {:.2f}%".format(accuracy_percentage_dt))

[126] ✓ 0.0s
... Decision Tree Accuracy: 73.91%
```

Figure 4.5: Decision Tree

Figure 4.5 demonstrates Decision Tree model with a fixed random state for reproducibility. After training the model on X_{train} and Y_{train} , it makes predictions on X_{test} . The accuracy is then calculated by comparing the predictions with the actual labels in Y_{test} and is displayed as a percentage. The model achieves an accuracy of 73.91%.



```
SVM

# Import necessary libraries
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Initialize the SVM model
svm_model = SVC()

# Train the model with the training data
svm_model.fit(X_train, Y_train)

# Make predictions on the test set
Y_pred_svm = svm_model.predict(X_test)

# Evaluate the model performance
accuracy_svm = accuracy_score(Y_test, Y_pred_svm)

# Display the accuracy as a percentage
accuracy_percentage_svm = accuracy_svm * 100

# Display the results
print("SVM Accuracy: {:.2f}%".format(accuracy_percentage_svm))

[129] ✓ 0.0s
... SVM Accuracy: 82.61%
```

Figure 4.6: SVM

Figure 4.6 demonstrates implementation of Support Vector Machine (SVM) model for cardiovascular disease prediction. After training on the dataset, the model makes predictions on the test set and calculates accuracy as a performance measure. The final accuracy is 82.61%, highlighting the SVM's strong predictive capability in this application.

4.2 Result Analysis

In this study, various machine learning models were assessed for predicting heart disease, with accuracy serving as the primary metric for comparison. Each algorithm demonstrated unique strengths and challenges, providing insights into model suitability for this task.

- Random Forest emerged as the top performer with an accuracy of 87.83%. Through extensive tuning of the `random_state` parameter, Random Forest showed resilience against overfitting and successfully captured complex relationships in the data, making it an ideal choice for heart disease prediction.
- Support Vector Machine (SVM) followed with an accuracy of 82.61%, leveraging its ability to define a clear decision boundary. This model performed well with high-dimensional data, underscoring its effectiveness in binary classification scenarios like heart disease diagnosis.
- K-Nearest Neighbors (KNN) achieved an accuracy of 80.00% with $k=5$ neighbors. Despite being slightly less accurate, KNN is notable for its simplicity and capability in handling nonlinear data patterns, making it a valuable comparison point.
- XGBoost showed an accuracy of 79.03%, with optimized parameters including `n_estimators`, `learning_rate`, and `max_depth`. Though slightly less accurate than other top models, XGBoost is a highly efficient gradient boosting method that remains a strong candidate for complex datasets, especially when further tuning is applied.
- Decision Tree achieved an accuracy of 73.91%. While the standalone Decision Tree model is less robust than ensemble methods, it provides valuable insights into feature importance, making it useful for exploratory analysis. Logistic Regression showed an accuracy of 61%. As one of the simplest models tested, Logistic Regression provided a baseline, though its linear nature limits its performance compared to more complex algorithms. It's a straightforward and interpretable model, valuable for scenarios requiring transparency in decision-making.

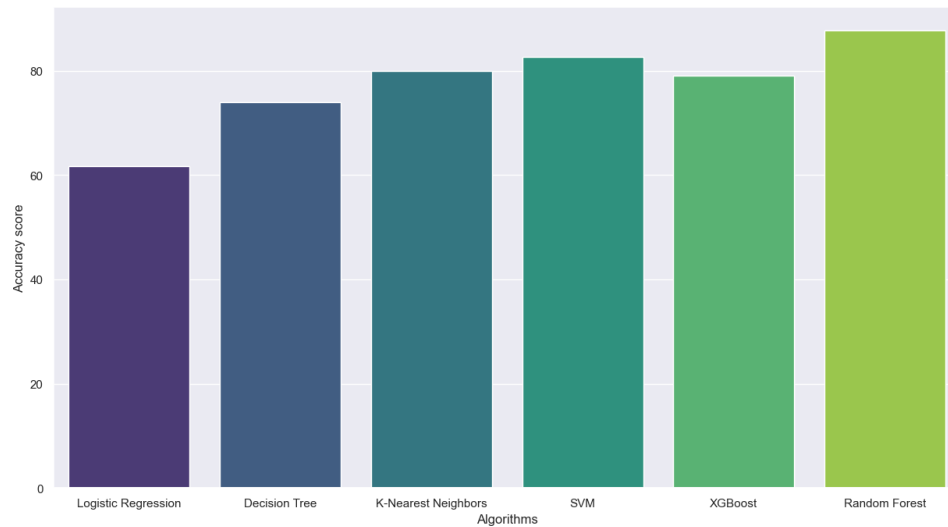


Figure 4.7: Comparison of algorithms

Figure 4.7 shows the comparison of the performance of six machine learning algorithms—Logistic Regression, Decision Tree, K-Nearest Neighbors, SVM, XGBoost, and Random Forest—based on accuracy, illustrating how each algorithm measures up against the others.

Overall, Random Forest, SVM, and KNN were the most effective models for heart disease prediction in this dataset, with Random Forest leading in accuracy. The analysis suggests that ensemble or hybrid models offer potential for further enhancement in predictive accuracy and robustness. This comparative approach provides a nuanced understanding of model performance, facilitating informed decisions in model selection for heart disease prediction tasks.

Chapter 5

Conclusion and Further Work

Conclusion

Our project demonstrates the potential of machine learning-based predictive models in addressing the critical need for early detection of heart disease. By analyzing accessible data such as lifestyle factors, and basic medical history, our model offers a cost-effective, reliable solution for identifying individuals at risk. The model's predictive accuracy, combined with the ability to operate without extensive diagnostic testing, positions it as a valuable tool for both healthcare providers and patients. Implementing such models can support proactive interventions, reduce the reliance on expensive diagnostic procedures, and ultimately help lower the global burden of heart disease.

Further Work

Future work could focus on expanding the model to include more diverse datasets, incorporating additional health metrics like physical activity, dietary habits, and socioeconomic factors to enhance prediction accuracy. Fine-tuning the model's performance with advanced techniques like deep learning or ensemble learning could further improve accuracy and reliability. Additionally, integrating this model into a user-friendly, mobile healthcare application could make it more accessible to a broader audience, especially in remote areas. Collaboration with healthcare institutions to validate and refine the model in real-world clinical settings would further ensure its applicability and effectiveness, moving toward a scalable solution for preventive heart disease care.

References

1. Madhumita Pal, Smita Parija. (2020). Prediction of Heart Diseases using Random Forest. International Journal of Engineering Research & Technology (IJERT)
2. Kalapraveen Bagadi, Visalakshi Annepu, Adnan Al-tamimi, Naga Raju Challa. IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), October 2023
3. Mahesh Kumar Joshi, Prof. (Dr.) Deepak Dembla, Dr. Suman Bhatia. Published In: JECRC University, Jaipur, Rajasthan, India.
4. Mohammed Khalid Hossen Published In: American Journal of Computer Science and Technology, 2022
5. N. F. A. Alhadeethy, A. Zaki, and A. Shah “Deep learning model for predicting and detecting overlapping symptoms of Cardiovascular Disease in Hospitals of UAE,” Turk. J. of comp. and Math. Edu., vol. 12, no. 14, pp. 5212-5224. November 2021.
6. M. G. Veerabaku, J. Nithiyanantham, S. Urooj, A. Q. Md, A. K. Sivaraman, and K. F. Tee, “Intelligent Bi-LSTM with Architecture Optimization for Heart Disease Prediction in WBAN through Optimal Channel Selection and Feature Selection,” Biomedicines, vol. 11, pp. 1167, April 2023
7. Kaggle dataset: <https://archive.ics.uci.edu/dataset/45/heart+disease>

Appendices

Appendix A

Weekly Progress Report

Department of Computer Engineering

BE Mini-Project-V Weekly Project Performance Report Even Sem 2024-2025 BE-Div: B

Project Title: Heart Disease Prediction Group No: 4

Name of Students 1: <u>Bhikha Ray</u>				Name of Students 2: <u>Shrushti Bhagare</u>						
Name of Students 3: <u>Nela Hernane</u>				Name of Students 4: <u>Sorika Sawale</u>						
Week No.	Expected Topics to be Covered	Progress Status	Student 1 Sign	Progress Status	Student 2 Sign	Progress Status	Student 3 Sign	Progress Status	Student 4 Sign	Suggestions if any
1.	Clear and Precise Objective	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
2.	Abstract and Introduction	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	Student abstract Read latest paper.
3.	Literature Survey	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
4.	Limitations of Existing System	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
5.	Problem Definition / Statement	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
6.	Proposed Methodology	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
7.	System Design	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	Draw diagram
8.	Details of hardware & Software	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
9.	Implementation details	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
10.	Result Analysis	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
11.	Conclusion and Future Work	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
12.	Publication (Conference Paper/ Journal Paper/ Patent/ Copyright)	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
13.	Competition Participation (Hackathon/ Ideathon/ ...)	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	A	<u>ok</u>	
A: Satisfactory B: Average C: Needs Improvement										

Project Guide Name and Sign
Dhansh A Bhosale

Figure A.1: Weekly Progress Report

Appendix B

Plagiarism Report

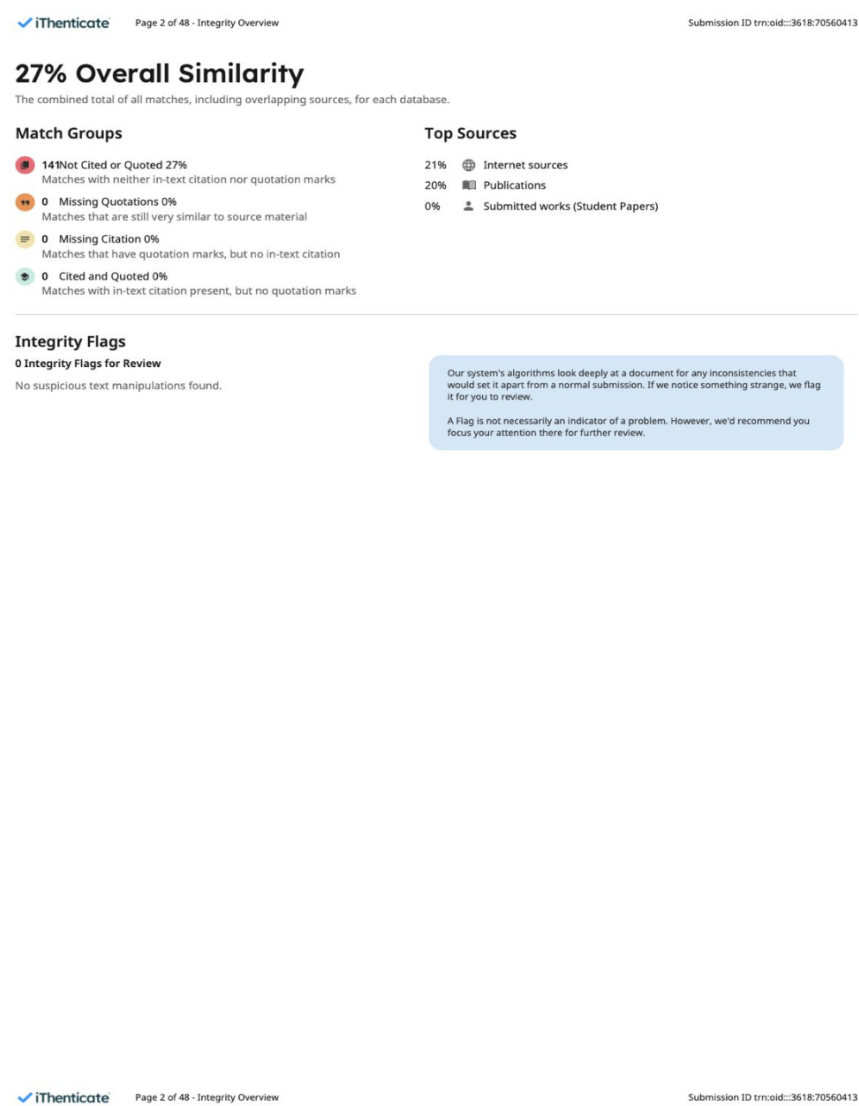


Figure B.1: Plagiarism Report

Acknowledgments

We take this opportunity to express our profound gratitude and deep regards to our guide **Ms. Dhanashri Bhosale** for her exemplary guidance, monitoring and constant encouragement throughout the completion of this report. We are truly grateful towards her efforts to improve our understanding towards various concepts and technical skills required in our project. The blessing, help and guidance given by him time to time shall carry us a long way in the journey of life on which we are about to embark. We take this privilege to express our sincere thanks to **Dr. Mukesh D. Patil**, Principal, RAIT for providing the much necessary facilities. We are also thankful to **Dr. A. V. Vidhate**, Head of Department of Computer Engineering, Project Co-ordinator **Ms. Dhanashri Bhosale.**, Department of Computer Engineering, RAIT, Nerul Navi Mumbai for their generous support. Last but not the least we would also like to thank all those who have directly or indirectly helped us in completion of this report.

Date: _____