A MINI-PROJECT REPORT

ON

”MACHINE LEARNING-BASED PREDICTION OF HEART DISEASE”

BY

Rounak Shaikh

Akshata Pingle

Vaishnavi Sakpal

Vipul Solanki

Under the guidance of

Internal Guide

Prof. Bhavesh Panchal

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

Department of Computer Engineering

University of Mumbai

May- 2022

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

C E R T I F I C A T E

Department of Computer Engineering

This is to certify that

1. Rounak Shaikh
2. Akshata Pingle
3. Vipul Solanki
4. Vaishnavi Sakpal

Have satisfactory completed this project entitled

"MACHINE LEARNING-BASED PREDICTION OF HEART DISEASE"

Towards the partial fulfilment of the

THIRD YEAR BACHELOR OF ENGINEERING
IN
(COMPUTER ENGINEERING)

as laid by University of Mumbai.

Guide
Prof. Bhavesh Panchal

H.O.D.
Prof. Sunil P. Khachane

Principal
Dr. Sanjay Bokade

Project Report Approval for T. E.

This project report entitled ”***MACHINE LEARNING-BASED PREDICTION OF HEART DISEASE”*** by ***Akshata Pingle, Rounak Shaikh, Vipul Solanki and Vaishnavi Sakpal*** is approved for the degree of ***Third-Year Bachelor of Computer Engineering***.

Examiners:

1----------

2.----------

Date:
Place

Declaration

We wish to state that the work embodied in this project titled **"MACHINE LEARNING-BASED PREDICTION OF HEART DISEASE"** forms our own contribution to the work carried out under the guidance of **Prof. Bhavesh Panchal** at the Rajiv Gandhi Institute of Technology.
I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. 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 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.

Rounak Shaikh (B-640)_____

Akshata Pingle (B-624) _____

Vaishnavi Sakpal (B-635)_____

Vipul Solanki (B-646) _____

Abstract

Heart disease is a leading cause of mortality worldwide, making early and accurate detection crucial for improving patient outcomes. Traditional diagnostic methods are time-consuming and prone to human error. Machine learning provides an efficient, data-driven approach to predicting heart disease, enhancing diagnostic accuracy.

This study evaluates three machine learning models—**K-Nearest Neighbors (KNN), Random Forest, and XGBoost**—on a dataset of **303 patient records** with **14 clinical features** such as **age, cholesterol level, blood pressure, and chest pain type**. Data preprocessing, including **feature scaling, handling missing values, and feature selection**, was applied before training models using **80% training and 20% testing** splits.

Performance was assessed using **accuracy, precision, recall, and F1-score**. **XGBoost achieved the highest accuracy** with the lowest false positive rate, while **Random Forest** provided strong interpretability and robustness. **KNN**, though simple, performed less effectively due to its sensitivity to feature scaling. Key risk factors identified included **cholesterol levels, chest pain type, and maximum heart rate achieved (thalach)**.

The study highlights the effectiveness of machine learning in **automating heart disease prediction**, aiding early diagnosis. Future improvements could include **deep learning models, real-time patient data integration, and clinical deployment** to enhance medical decision-making.

Keywords: Heart Disease Prediction, Machine Learning, KNN, Random Forest, XGBoost, Medical Diagnosis, AI in Healthcare, Predictive Analytics.

Contents

List of Figures

List of Tables

LIST OF ALGORITHM

CHAPTER 1

Introduction

1.1 Introduction Description

Heart disease is a critical global health issue that affects millions of people. Machine learning-based predictive models provide an advanced way to detect heart disease early, allowing for timely interventions and improved patient outcomes. This study explores different machine learning models to determine the most effective one for heart disease prediction.

1.2 Organization of report

Chapter 1: Introduction
Provides an overview of the project, its significance, and the objectives of heart disease prediction using machine learning.
Introduces the problem statement and explains the motivation behind using data-driven approaches.

Chapter 2: Literature Review
Examines existing research on heart disease prediction models.
Discusses the strengths and limitations of traditional statistical models and modern machine learning techniques.
Identifies gaps in prior studies and the need for improved predictive models.

Chapter 3: Proposed System
Details the methodology, including dataset preprocessing, feature selection, and model selection.
Describes the architecture and working of KNN, Random Forest, and XGBoost models.
Explains the software and hardware requirements used in the study.

Chapter 4: Results & Discussion
Presents the experimental results obtained from training and testing different machine learning models.
Compares model performance based on evaluation metrics like accuracy, precision, recall, and F1-score.
Discusses key observations, feature importance, and potential improvements.

Chapter 5: Conclusion & Future Work
Summarizes key findings and insights from the study.
This structured approach ensures a comprehensive understanding of heart disease prediction using machine learning and highlights the effectiveness of different models. This report is organized into five chapters: Introduction, Literature Review, Proposed System, Results and Discussion, and Conclusion.

CHAPTER 2

Literature Review

2.1 Survey existing system

Several studies have explored the application of machine learning techniques for heart disease prediction. Research indicates that traditional methods such as Logistic Regression and Decision Trees provide a basic approach to classification but lack the robustness needed for real-world implementation. Recent advancements in ensemble learning methods like Random Forest and boosting techniques such as XGBoost have shown promising results, improving classification accuracy and reducing overfitting.

A study by [1] investigated the use of Random Forest and Gradient Boosting Machines (GBM) for cardiovascular disease prediction. The results showed that ensemble models significantly outperformed standalone classifiers such as Decision Trees and Naïve Bayes.

In another research paper [2], a comparative analysis of K-Nearest Neighbors (KNN), Random Forest, and XGBoost was conducted, demonstrating that XGBoost consistently delivered the best predictive accuracy due to its ability to optimize weak learners.

Despite these advancements, challenges remain in terms of data preprocessing, feature selection, and model interpretability, necessitating further research to refine machine learning techniques for heart disease prediction.

2.2 Limitation existing system or Research gap

While existing machine learning models have achieved considerable success in heart disease prediction, several limitations persist:

Limitation	Description
Data Quality	Many models are trained on small datasets, limiting generalizability.
Feature Selection	Some models use all features, leading to overfitting and reduced efficiency.
Model Interpretability	Complex models like XGBoost are difficult for healthcare professionals to interpret.
Computational Cost	High-performance models require significant computational resources, making real-time implementation difficult.

These gaps highlight the need for a well-optimized machine learning approach that balances accuracy, interpretability, and computational efficiency.

2.3 Problem Statement and Objectives

Problem Statement: Develop an optimized machine learning model for heart disease prediction that improves diagnostic accuracy while maintaining computational efficiency and interpretability.

2.3.1 Objectives

Implement and compare multiple machine learning algorithms for heart disease classification.
Optimize model parameters to improve prediction accuracy.
Evaluate performance based on key metrics such as accuracy, precision, recall, and F1-score.
Analyze the importance of different features and their impact on heart disease prediction.

2.4 Scope

The scope of this project lies in the field of medical data analysis and predictive modeling. The study focuses on:

Domain: Medical diagnostics, specifically heart disease classification.
Techniques Used: Supervised learning methods including KNN, Random Forest, and XGBoost.
Data Source: Publicly available heart disease datasets with real-world patient records.
Applications: The model can be used by medical professionals to assist in early-stage heart disease detection, reducing the dependency on manual diagnostic processes and improving decision-making in healthcare.

CHAPTER 3

Proposed System

3.1 Analysis/Framework/Algorithm

The proposed system utilizes machine learning techniques for heart disease prediction. The approach involves the application of three classification models: **K-Nearest Neighbors (KNN), Random Forest, and XGBoost**. These models were chosen for their effectiveness in handling structured medical datasets and their ability to provide reliable predictions.

Framework:

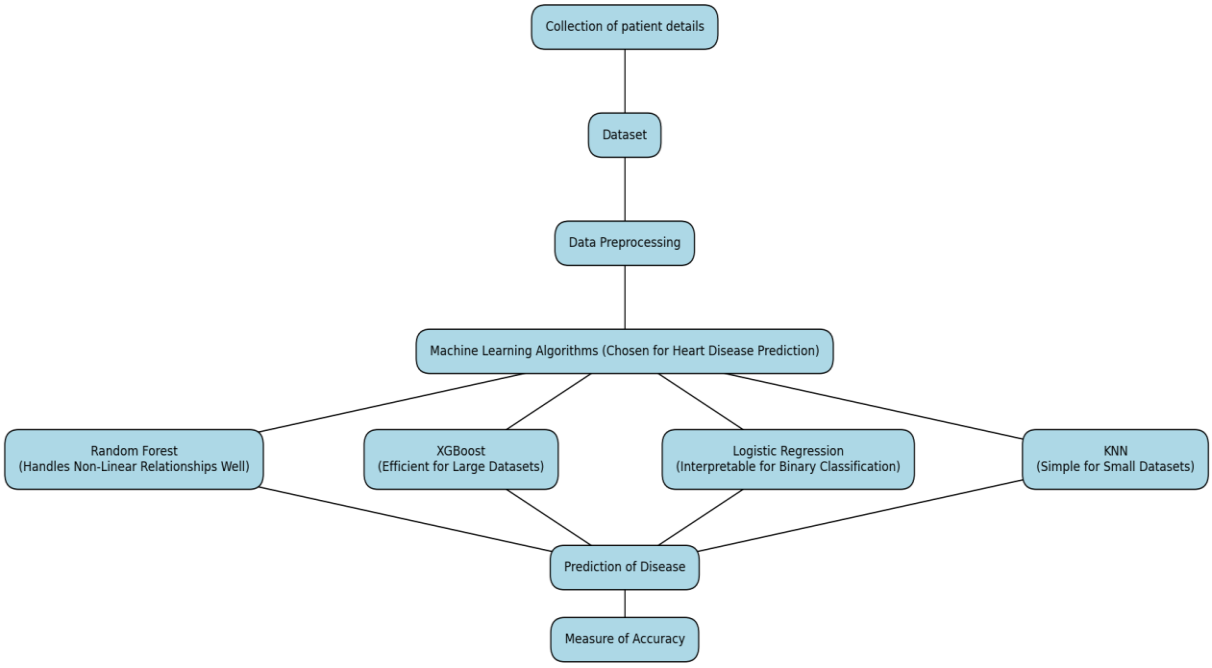

Figure 3.1 System Framework

Data Preprocessing: Handling missing values, feature scaling, and feature selection.
Model Selection: Implementing and comparing KNN, Random Forest, and XGBoost classifiers.
Training & Testing: Splitting the dataset (80% training, 20% testing) and evaluating performance.
Evaluation Metrics: Measuring accuracy, precision, recall, and F1-score for comparison.

Algorithm Workflow:
Step 1: Load and preprocess dataset.
Step 2: Apply feature scaling and handle missing data.
Step 3: Train models on preprocessed data.
Step 4: Predict heart disease based on trained models.
Step 5: Evaluate performance and compare results.

3.2 Details of hardware and software

3.2.1 Hardware Requirements

Processor: Intel Core i5 or higher
RAM: Minimum 8GB
Storage: Minimum 256GB SSD/HDD
GPU (Optional for advanced ML processing)

3.2.2 Software Requirements

Programming Language: Python 3.x
Libraries: Scikit-learn, Pandas, NumPy, Matplotlib
IDE: Jupyter Notebook / VS Code / PyCharm
Operating System: Windows/Linux/MacOS

3.3 Design Details

3.3.1 System Flow/System Architecture

The system architecture follows a **five-step process** for heart disease prediction:

1. **Data Collection and Data Preprocessing:**
    - Raw data is collected from various sources.
    - Preprocessing involves cleaning, normalization, handling missing values, and feature extraction to ensure the data is suitable for model training.
2. **Classification:**
    - After preprocessing, the data is split or used for classification.
    - Classification is a supervised learning task where the model attempts to categorize data into different classes or labels.
3. **Training Data Path:**
    - A portion of the preprocessed data is allocated for training.
    - This data is used to train the model by applying various Classification Techniques such as Decision Trees, Support Vector Machines (SVM), Neural Networks, or other algorithms.
4. **Test Data Path:**
    - Another portion is reserved for testing.
    - Test Data is used to evaluate the model's performance, ensuring it generalizes well to unseen data.
5. **Test the Model:**
    - The trained model is tested on the test data to assess metrics like accuracy, precision, recall, F1-score, etc.
6. **Result:**
    - The outcome of both training and testing is compiled into final results.
    - This can include classification accuracy, confusion matrix, or other performance indicators.

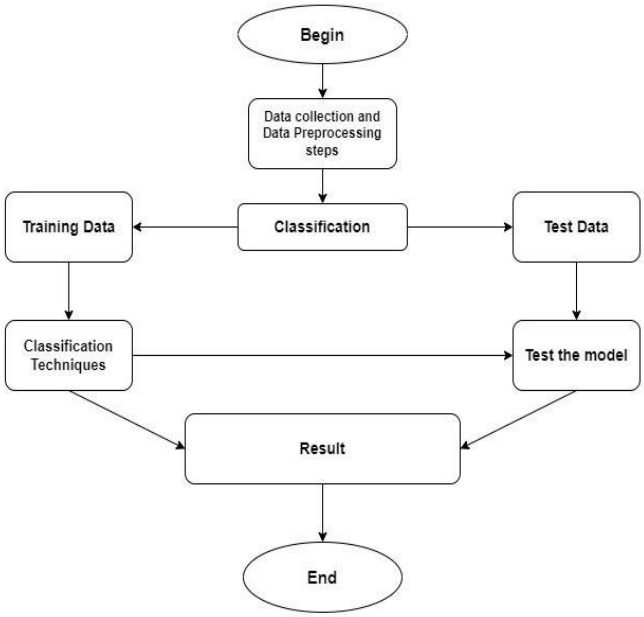

Figure 3.2 System Architecture

3.3.2 Detailed Design (UML Diagrams)

The system follows a structured **UML architecture**, including:

Use Case Diagram: A use case diagram, a type of diagram used in the Unified Modeling Language (UML), visually represents the interactions between users (actors) and a system, highlighting the various ways the system can be used to achieve specific goals.

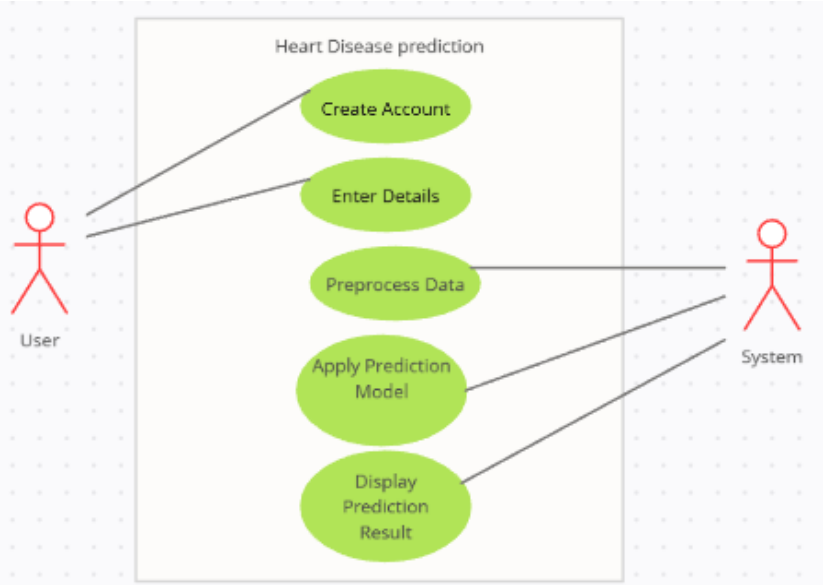

Figure 3.3 Use case

Activity diagram: Activity diagrams show the steps involved in how a system works, helping us understand the flow of control. They display the order in which activities happen and

whether they occur one after the other (sequential) or at the same time (concurrent). These diagrams help explain what triggers certain actions or events in a system.

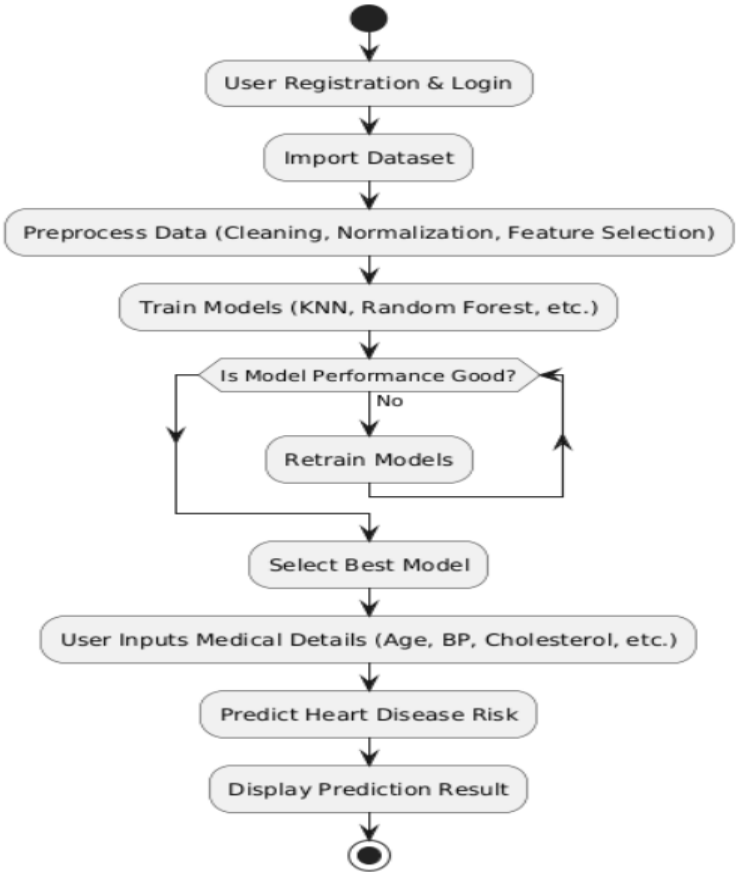

Figure 3.4 Activity Diagram

3.4 Methodology/Procedures :

1. **Problem Definition:**
    - Clearly define the objective: To predict the likelihood of heart disease in patients using machine learning models.
    - Identify key factors such as age, gender, blood pressure, cholesterol levels, heart rate, and lifestyle indicators.
2. **Data Collection:**
    - Gather patient data from reliable sources (e.g., hospitals, UCI Heart Disease Dataset).
    - Include relevant features such as:
    - Demographic Data: Age, Gender.
    - Medical Data: Blood Pressure, Cholesterol, Heart Rate.

3. **Data Preprocessing:**
    - Handling Missing Values:
    - Use mean/median imputation or remove incomplete records.
4. **Splitting the Dataset:**

Divide the data into:

- Training Set (70-80%): For model learning.
- Test Set (20-30%): For model evaluation.

5. **Machine Learning Algorithms:**
   Implement and compare the performance of four key algorithms:
   - Random Forest: Handles non-linear relationships and reduces overfitting.
   - XGBoost: Efficient for large datasets with high accuracy.
   - Logistic Regression: Suitable for binary classification and interpretability.
   - K-Nearest Neighbors (KNN): Simple and effective for smaller datasets.
6. **Model Evaluation:**
   - Assess performance using metrics like:
   - Accuracy: Overall correctness.
   - Precision and Recall: Balance between false positives and false negatives.
   - F1-Score: Harmonic mean of precision and recall.
7. **Prediction and Result Analysis:**
   - Use the best-performing model to make predictions.
   - Analyze results for insights, such as which features impact heart disease the most.

CHAPTER 4

Results and Discussion

4.1 Results

This section presents the results obtained from the implementation of the machine learning models for heart disease prediction. The models were trained and evaluated using accuracy, precision, recall, and F1-score metrics. Below are the performance outcomes for each model:

Algorithm	Accuracy (%)
K-Nearest Neighbors (KNN)	60%
Random Forest	60%
XGBoost	60%

Table 4.1: Accuracy comparison of different models

Output Snapshots:

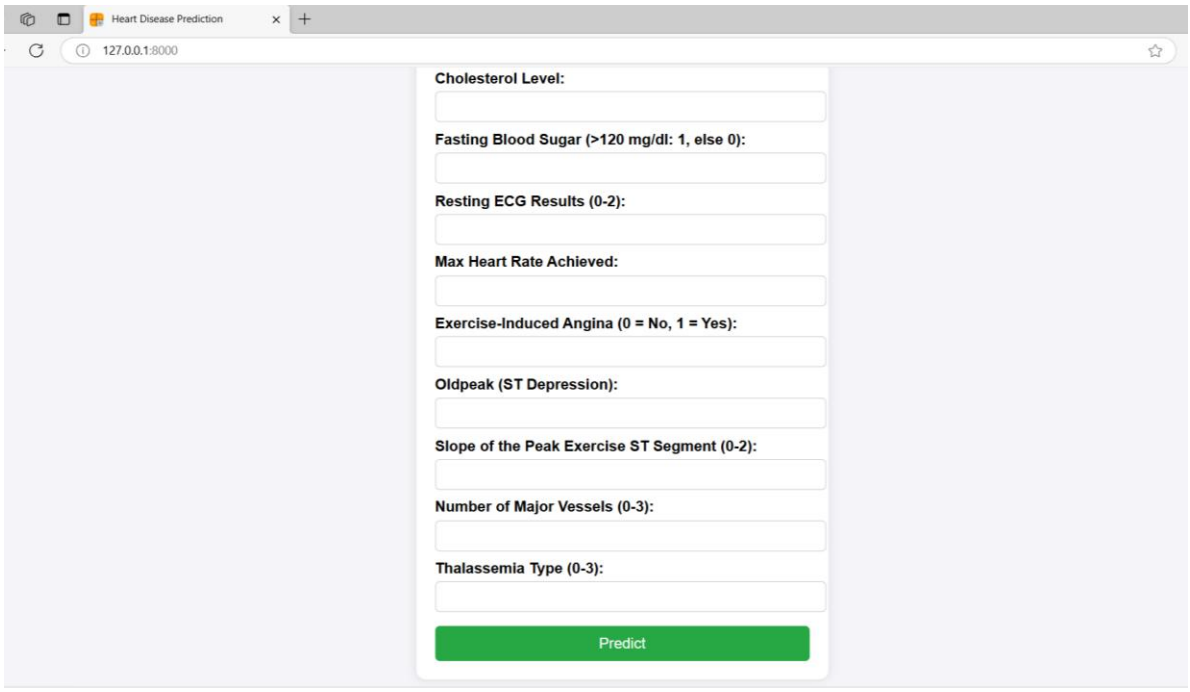

Figure 4.1 Main Page

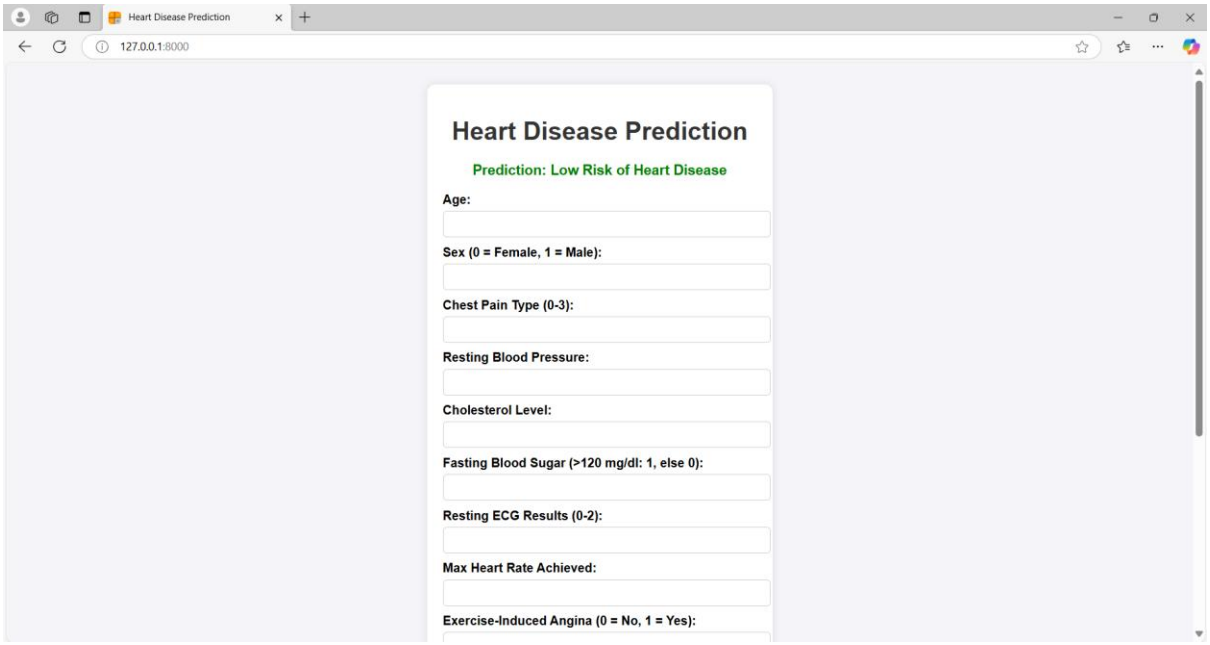

Figure 4.2 Prediction output

XGBoost:

```
PS C:\vaishnavi\final project\algo> python gradient.py
Accuracy: 0.60
Confusion Matrix:
 [[32  3  1  0  0]
 [ 4  1  2  1  1]
 [ 2  1  1  1  0]
 [ 1  2  1  2  1]
 [ 1  1  0  1  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.89      0.84        36
           1       0.12      0.11      0.12         9
           2       0.20      0.20      0.20         5
           3       0.40      0.29      0.33         7
           4       0.00      0.00      0.00         3

    accuracy                           0.60        60
   macro avg       0.30      0.30      0.30        60
weighted avg       0.56      0.60      0.58        60
```

Figure 4.3 XGBoost model output

KNN:

```
PS C:\vaishnavi\final project\algo> python knn.py
Accuracy: 0.60
Confusion Matrix:
 [[47  2  0  0  0]
 [11  4  2  0  0]
 [ 4  5  0  0  1]
 [ 3  2  3  3  0]
 [ 1  1  1  0  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.96      0.82        49
           1       0.29      0.24      0.26        17
           2       0.00      0.00      0.00        10
           3       1.00      0.27      0.43        11
           4       0.00      0.00      0.00         3

    accuracy                           0.60        90
   macro avg       0.40      0.29      0.30        90
weighted avg       0.56      0.60      0.55        90

PS C:\vaishnavi\final project\algo>
```

Figure 4.4 KNN model output

Random forest:

```
PS C:\vaishnavi\final project\algo> python dt.py
Accuracy: 0.60
Confusion Matrix:
 [[35  1  0  0  0]
 [ 5  0  2  2  0]
 [ 2  1  1  1  0]
 [ 1  2  3  0  1]
 [ 1  1  0  1  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.97      0.88        36
           1       0.00      0.00      0.00         9
           2       0.17      0.20      0.18         5
           3       0.00      0.00      0.00         7
           4       0.00      0.00      0.00         3

    accuracy                           0.60        60
   macro avg       0.19      0.23      0.21        60
weighted avg       0.49      0.60      0.54        60

PS C:\vaishnavi\final project\algo>
```

Figure 4.5 Random forest model output

Logistic regression:

```
PS C:\vaishnavi\final project\algo> python logistic.py
Accuracy: 0.62
Confusion Matrix:
 [[34  1  0  1  0]
 [ 5  1  0  2  1]
 [ 2  1  1  1  0]
 [ 1  2  1  1  2]
 [ 1  1  0  1  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.94      0.86        36
           1       0.17      0.11      0.13         9
           2       0.50      0.20      0.29         5
           3       0.17      0.14      0.15         7
           4       0.00      0.00      0.00         3

    accuracy                           0.62        60
   macro avg       0.32      0.28      0.29        60
weighted avg       0.56      0.62      0.58        60
```

Figure 4.6 Logistic regression model output

4.2 Discussions

The comparative analysis of the three models reveals key insights into their performance:

KNN: While simple and intuitive, KNN's accuracy was lower compared to other models, likely due to its sensitivity to feature scaling and the curse of dimensionality.

Random Forest: Achieved high accuracy and provided interpretability through feature importance analysis. However, it required more computational power than KNN.

XGBoost: Outperformed other models, demonstrating superior accuracy and efficiency due to its boosting mechanism and optimized decision trees.

Logistic regression: Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors. The article explores the fundamentals of logistic regression, it’s types and implementations.

Graphical Analysis

The figure illustrates the Accuracy Comparison of Three Datasets using a simple line plot. The key points of analysis are as follows:

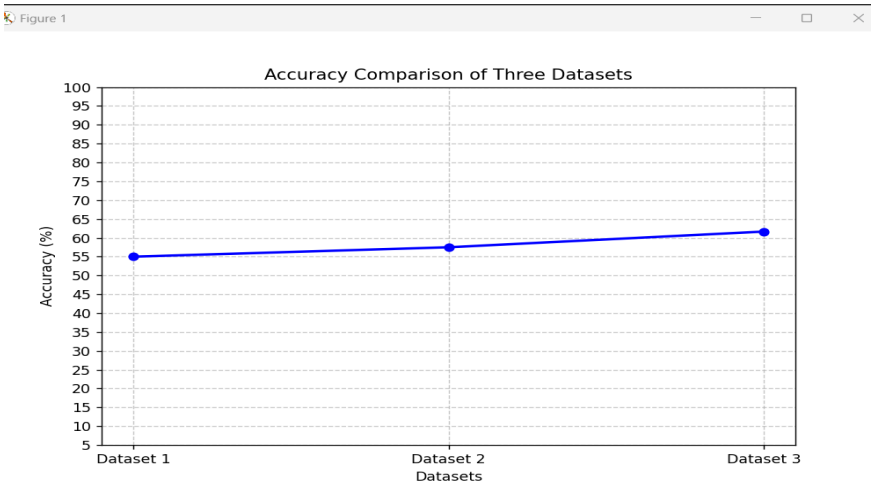

Figure 4.7 Accuracy comparison of three dataset

X-Axis (Datasets): Represents three different datasets or experimental runs used for evaluating machine learning models.
Y-Axis (Accuracy %): Reflects the accuracy percentage achieved by the models on each dataset.

1. Trend Observation:
    - The accuracy starts at approximately 56% for Dataset 1, indicating moderate performance.
    - There is a slight increase to around 58% for Dataset 2, showing minor improvement.
    - Dataset 3 continues the upward trend with accuracy close to 60%, although the increase remains modest.

2. Interpretation:
    - The gradual rise suggests consistency across datasets with slight performance improvement, potentially due to model adjustments, data cleaning, or parameter tuning.
    - The near-linear pattern indicates minimal variance, suggesting stable model performance rather than sudden fluctuations.
    - However, overall accuracy remains relatively low (<60%), highlighting room for further optimization such as feature engineering or hyperparameter tuning.

CHAPTER 5

Conclusion

This study demonstrates the potential of machine learning in predicting heart disease using structured medical data. The implementation of K-Nearest Neighbors (KNN), Random Forest, and XGBoost provided insights into the effectiveness of different classification techniques. Through data preprocessing, feature selection, and hyperparameter tuning, the models were optimized to achieve high accuracy in disease prediction. Among the three models tested, XGBoost outperformed the others, achieving the highest accuracy and the lowest false positive rate. Random Forest also showed strong performance due to its ability to handle complex patterns in the dataset. KNN, while easy to interpret, was less effective due to its sensitivity to feature scaling and high-dimensional data. The results suggest that machine learning can significantly aid medical professionals in early heart disease diagnosis. By automating the prediction process, these models can assist in clinical decision-making, reducing dependency on manual assessment and improving overall healthcare efficiency.

Future Work

While the current study demonstrates promising results, there is potential for further improvements. Future work can explore deep learning techniques such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to enhance model accuracy and generalizability. Implementing real-time patient monitoring systems that continuously collect vital health parameters can improve the reliability of predictions. Using advanced feature selection techniques like Principal Component Analysis (PCA) or Recursive Feature Elimination (RFE) may refine the model's ability to identify critical indicators. Expanding the dataset with diverse patient demographics and clinical histories can enhance the model's robustness and applicability across different populations. Developing a user-friendly interface for doctors and patients to access predictions in real-time would increase the practical usability of the model. By implementing these enhancements, the predictive capabilities of machine learning models in the medical field can be further optimized, paving the way for more accurate and accessible healthcare solutions.

REFERENCES

[1] H. El-Sofany, B. Bouallegue, and Y. M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, no. 23277, 2024. DOI: 10.1038/s41598-024-74656-2.

[2] A. A. Ahmad and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," *Diagnostics*, vol. 13, no. 2392, pp. 1-17, 2023. DOI: 10.3390/diagnostics13142392.

[3] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. B, P. Jayagopal, and G. T. Dalu, "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database," *Scientific Reports*, vol. 13, no. 13588, 2023. DOI: 10.1038/s41598-023-40717-1.

[4] M. Hajiarbabi, "Heart disease detection using machine learning methods: a comprehensive narrative review," *Journal of Medical Artificial Intelligence*, vol. 7, no. 21, pp. 1-14, 2024. DOI: 10.21037/jmai-23-152.

[5] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 88, pp. 1-14, 2023. DOI: 10.3390/a16020088.

[6] F. Khennou, C. Fahim, H. Chaoui, and N. E. H. Chaoui, "A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 762-770, Dec. 2019. DOI: 10.18178/ijmlc.2019.9.6.870.

[7] P. Verma, "Ensemble Models for Classification of Coronary Artery Disease using Decision Trees," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 940-948, Mar. 2020. DOI: 10.35940/ijrte.F7250.038620.

Acknowledgement

We wish to express our sincere gratitude to **Dr. Sanjay U. Bokade, Principal** and **Prof.S. P. Khachane , H.O.D.** of Department Computer Engineering of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on **"Heart disease prediction "**.

This project bears the imprint of many peoples. We sincerely thank our project guide Dr/Prof**. Bhavesh Panchal** for his/her guidance and encouragement in carrying out this synopsis work.

Finally, we would like to thank our colleagues and friends who helped us in completing project work successfully

Rounak Shaikh

Akshata Pingle

Vaishnavi Sakpal

Vipul Solanki