NITTE EDUCATION TRUST | NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

A Project Report on

"HEARTSCOPE AI: RISK PREDICTION, SHAP EXPLANATION, AND PERSONALIZED WELLNESS REPORT"

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Engineering
In

Department of Information Science and Engineering

for the **Academic Year: 2024-25**

Submitted by

S.LOKESH	**1NT23IS184**
NIGAM L. RAJ	**1NT23IS142**
PRAJWAL R.	**1NT23IS158**

Under the Guidance of

Prof. Priyanka K
Assistant Professor
Dept. of Information Science and Engineering

YELAHANKA, BENGALURU- 560064

NITTE EDUCATION TRUST | NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

Department of Information Science and Engineering

Certificate

This is to certify that the project work entitled **"Heartscope AI: Risk Prediction, Shap Explanation, And Personalized Wellness Report"** has been carried out by **S.LOKESH(1NT23IS184)**, ***NIGAM L RAJ(1NT23IS142)***, ***PRAJWAL R(1NT23IS158)*** bonafide students of ***Nitte Meenakshi Institute of Technology***, in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering** in the **Department of Information Science and Engineering** under **Visvesvaraya Technological University**, Belagavi, during the academic year **2024–2025**. The project report has been examined and approved as it meets the academic requirements specified under the autonomous scheme of **Nitte Meenakshi Institute of Technology** for the said degree.

Signature of the Guide	**Signature of the HoD**	**Signature of the Principal**
Prof. Priyanka K Assistant Professor Nitte Meenakshi Institute of Technology, Bengaluru-560064	**Dr.Mohan S.G** Head of the Department Nitte Meenakshi Institute of Technology, Bengaluru-560064	**Dr. H C. Nagaraj** Principal Nitte Meenakshi Institute of Technology, Bengaluru-560064

Acknowledgement

The successful execution of our project has been a significant milestone, and we take this opportunity to express our heartfelt gratitude to all those who have supported and guided us throughout this journey. Whatever we have achieved is the result of their encouragement and help, and we remain deeply thankful to each one of them.

We express our sincere thanks and seek the blessings of **Dr. N. R. Shetty**, Advisor, *Nitte Meenakshi Institute of Technology*, for his vision and emphasis on project-based learning and constructivist principles that have greatly enriched our academic experience. We are grateful to **Mr. Rohit Punja**, Administrator, *Nitte Education Trust*, and **Dr. Sandeep Shastri**, Vice President, *Bangalore Campus, Nitte University*, for their strategic leadership and continued support in fostering academic excellence.

We extend our special thanks to our beloved Principal, **Dr. H. C. Nagaraj**, for providing us with the necessary resources, facilities, and motivation to carry out our project successfully. Our sincere gratitude goes to **Dr. J. Sudheer Reddy**, *Dean – Academics,* and **Dr. Kiran Aithal**, *Dean – Research & Development,* for their guidance, encouragement, and for creating an environment that promotes innovation and academic growth.

We also convey our deep appreciation to **Dr.Mohan S.G Head of the Department,** Information Science and Engineering, for his/her constant encouragement, valuable guidance, and for fostering a vibrant learning environment. We extend our heartfelt thanks to our project guide, **Prof. Priyanka K** Assistant Professor, *Department of* Information Science and Engineering, for his/her unwavering support, timely feedback, and insightful mentorship throughout the project. We are also indebted to our parents for their unconditional love, support, and encouragement throughout our academic journey at *Nitte Meenakshi Institute of Technology*, Bengaluru. Finally, we would like to express our appreciation to all those**—named and unnamed—**who contributed in any way to our learning and success.

S.LOKESH(1NT23IS184)
PRAJWAL R(1NT23IS158)
NIGAM L RAJ(1NT23IS142)

Place: Bengaluru
Date:24-05-2025

Abstract

This project presents *Heartscope AI*, an intelligent system developed to predict the risk of heart disease using logistic regression, a widely accepted and interpretable machine learning algorithm. The model is trained on the Framingham Heart Study dataset, which contains comprehensive clinical and lifestyle data of real-world patients. The objective is to support early identification of individuals at risk, enabling timely intervention and improved healthcare outcomes.

The system accepts user-provided inputs such as age, gender, education, smoking habits, blood pressure, cholesterol, glucose levels, body mass index (BMI), heart rate, and medical history (e.g., diabetes, hypertension, stroke). After preprocessing the data through imputation, normalization, and feature selection, the model generates a binary prediction indicating whether the user is at risk of developing heart disease within the next ten years.

To enhance user understanding and trust, the system integrates SHAP (SHapley Additive exPlanations) to explain feature contributions for each prediction. Alongside the prediction result, users receive a personalized health score, risk explanation in natural language, and dietary and lifestyle recommendations. A comprehensive PDF report is automatically generated, including the input summary, prediction outcome, visualizations (confusion matrix, ROC curve, SHAP plots), and health guidance.

Overall, Heartscope AI offers a transparent, user-friendly, and data-driven approach to heart disease risk prediction. It aims to bridge the gap between machine learning technologies and real-world clinical applications by promoting awareness, early prevention, and personalized wellness strategies.

Contents

List of Figures

List of Table

Chapter 1 Introduction

Cardiovascular diseases, particularly heart disease, remain one of the leading causes of death worldwide. Despite advancements in medical diagnostics, early detection continues to be a challenge—especially in low-resource settings—due to limited access to healthcare, lack of awareness, and delays in medical intervention. Accurate and timely prediction of heart disease risk can significantly improve patient outcomes by enabling preventive measures before critical conditions develop.

With the increasing availability of health data and the advancement of machine learning (ML) techniques, predictive analytics has emerged as a powerful tool in the healthcare domain. This project leverages these capabilities to develop *Heartscope AI*, a machine learning-based heart disease risk prediction system using logistic regression. The system is trained on the Framingham Heart Study dataset, a widely recognized dataset containing clinical and demographic information relevant to cardiovascular risk assessment.

The system accepts user inputs including age, gender, smoking status, cholesterol levels, blood pressure, BMI, glucose, and medical history. It then processes the data to determine the likelihood of the user developing heart disease within the next ten years. To enhance interpretability and transparency, the project integrates SHAP (SHapley Additive exPlanations), allowing users and healthcare professionals to understand the contribution of each input feature to the final prediction.

Additionally, *Heartscope AI* offers more than just prediction. It includes a custom health scoring mechanism, generates natural-language explanations for results, and provides dietary and lifestyle recommendations tailored to the individual's health profile. A downloadable PDF report summarizes all findings and visualizations, such as the confusion matrix, ROC curve, and SHAP plots.

This project demonstrates a practical, interpretable, and accessible approach to heart disease prediction, designed to assist both healthcare professionals and non-expert users in making informed health decisions. It bridges the gap between AI-driven models and real-world healthcare needs by combining accuracy, usability, and explainability.

Motivation

Heart disease is one of the leading causes of death worldwide. Early detection and accurate prediction can help doctors provide timely treatment and save lives. However, many current methods for predicting heart disease are either expensive, time-consuming, or require complex tests. Developing a reliable, easy-to-use, and cost-effective prediction system can help more people get early warnings about their heart health. A good solution should be accurate, fast, and provide understandable results to both doctors and patients. This motivates the need for a machine learning-based heart disease prediction model that can analyze patient data and give quick predictions, supporting better healthcare decisions.

1.1 Organization of the Report

This project report is organized into six main chapters, each focusing on a specific part of the work:

- **Chapter 1: Introduction** – This chapter provides the motivation behind the project and explains the structure of the report.
- **Chapter 2: Literature Survey, Problem Definition, and Objectives** – This chapter reviews previous research related to heart disease prediction using machine learning, identifies existing challenges, defines the problem, and outlines the objectives and scope of the project.
- **Chapter 3: Design Approach and Methodology** – This chapter explains the overall design of the system, the methods used for data preprocessing, model building, and the tools and technologies involved.
- **Chapter 4: Implementation Details** – This chapter describes the step-by-step implementation of the model, including dataset handling, model training, prediction logic, and SHAP-based interpretability.
- **Chapter 5: Results and Analysis** – This chapter presents the outcomes of the project, including evaluation metrics such as accuracy, precision, recall, confusion matrix, ROC curve, and SHAP visualizations for model explainability.
- **Chapter 6: Conclusion and Future Scope** – This chapter concludes the report by summarizing the work done, key findings, and discusses future improvements or extensions of the project.

Each section is designed to give a clear understanding of the project development from idea to implementation and analysis.

Chapter 2 Literature Survey

2.1 Background Work

1. Heart Disease Detection Using Machine Learning Models

This study explores machine learning (ML) techniques—Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Principal Component Analysis (PCA)—for detecting heart disease. It addresses the problem of overfitting by designing a robust diagnostic system that maintains high performance on both training and testing datasets. The Framingham dataset is used for training and evaluation. The model's performance is assessed through metrics such as accuracy, sensitivity, precision, and F1-score [1].

2. Heart Disease Detection: A Comprehensive Analysis of Machine Learning, Ensemble Learning, and Deep Learning Algorithms

This paper presents a comparative analysis of machine learning, ensemble learning, and deep learning models for early-stage heart disease detection. A hybrid model combining Bagging with Random Forest achieved the best results with 94.34% accuracy, 93.5% sensitivity, and 94.2% F1-score using the heart_statlog_cleveland_hungary_final dataset. The paper highlights the superiority of ensemble methods and hybrid approaches for accurate predictions [2].

3. Heart Disease Detection Using Feature Extraction and Artificial Neural Networks: A Sensor-Based Approach

This innovative study utilizes feature extraction from sensor data combined with Artificial Neural Networks (ANN) to detect heart disease. Ten types of metal oxide semiconductor sensors are used to capture data, which is then processed via Arduino into digital form for ANN training. The system achieves over 85% accuracy and is designed to reduce diagnostic errors and support early detection through intelligent signal processing[3].

4. A Proposed Technique for Predicting Heart Disease Using Machine Learning Algorithms and an Explainable

This paper proposes a technique using feature selection (chi-square, ANOVA, mutual information) and ten ML algorithms. The best results were achieved with XGBoost on the SF-2 feature subset, yielding 97.57% accuracy. SMOTE is used to balance the dataset, and SHAP is applied to provide model explainability. A mobile app is developed to enable real-time, interpretable predictions of heart disease[4].

5. Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm

This research uses the Jellyfish Optimization Algorithm for feature selection from the Cleveland dataset, reducing dimensionality and preventing overfitting. The selected features are used to train various ML models, with SVM achieving the highest

performance: 98.47% accuracy, 98.56% sensitivity, and 94.48% AUC. The study emphasizes the effectiveness of metaheuristic optimization in boosting ML prediction accuracy for heart disease[5].

2.2 Open Issues and Challenges

The key research gap is the lack of transparent, interpretable, and user-friendly heart disease prediction tools. Existing studies focus on accuracy but often neglect explaining model outputs or providing personalized health advice.
Many also overlook handling missing or messy real-world data, limiting practical use in clinical settings.
Additionally, few integrate prediction models with personalized reports; most do not offer comprehensive PDFs with visual aids like SHAP plots or ROC curves to clarify results for users and clinicians.
This project fills that gap by combining prediction, explainable AI, and personalized diet and health advice in an automated, user-friendly report—making it both scientifically sound and accessible to non-experts and healthcare providers.

2.3 Problem Definition

Heart disease is a leading global cause of death, with early detection remaining difficult, especially in low-resource areas. Traditional diagnostics like ECG and angiograms are costly, time-consuming, and not widely accessible.
With growing health data availability, machine learning offers a faster, cheaper, and non-invasive way to predict heart disease risk using clinical and lifestyle factors such as age, cholesterol, blood pressure, smoking, and diabetes.
This research addresses the need for an efficient, interpretable, and user-friendly prediction system to support early prevention and better healthcare decisions.
Our project develops a logistic regression model trained on the Framingham dataset to classify heart disease risk within 10 years. It integrates explainability through SHAP values to make model predictions transparent.

Additionally, the system generates detailed PDF reports with personalized health scores, risk explanations, and diet recommendations tailored to the individual's risk profile. This approach ensures the tool is practical and accessible for both healthcare professionals and non-expert users.

2.4 Objectives

Aim:
The aim of this project is to build an interpretable and reliable machine learning model using logistic regression to predict the risk of heart disease in individuals based on basic clinical and lifestyle data.

Objectives:

To study and preprocess the Framingham Heart Study dataset to make it suitable for model training.

1. To implement logistic regression for binary classification of heart disease risk (present or not present).
2. To evaluate the model using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
3. To use SHAP for interpreting how each feature contributes to the prediction.
4. To build a user input interface where individuals can enter health-related values and receive prediction results.
5. To generate a health report in PDF format that includes:
    a. Prediction outcome
    b. Health score based on input values
    c. Recommended dietary tips and lifestyle changes
    d. Visual explanations such as confusion matrix and SHAP plots

2.5 Scope of the Work

This project focuses on developing a heart disease prediction system using logistic regression, a widely used and interpretable machine learning algorithm. The system is designed to help in the early detection of heart disease risk using common health-related parameters.

The work includes the following main tasks and deliverables:

- **Data Collection and Preprocessing**: Use the Framingham Heart Study dataset from Kaggle. Clean the data by handling missing values, encoding categorical variables, and scaling features.
- **Model Development**: Train a logistic regression model to classify whether a person is at risk of heart disease or not.
- **Performance Evaluation**: Analyze the model using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- **User Interface**: Design a simple interface for users to input their health details (age, cholesterol, blood pressure, smoking, etc.) and receive prediction results.
- **Health Report Generation**: Create a downloadable PDF report for each prediction. The report includes:
    a. Prediction result
    b. Health score
    c. Personalized diet and lifestyle suggestions
    d. Visual outputs such as confusion matrix, ROC curve, and SHAP plots
- **Interpretability**: Use SHAP values to explain which features contribute most to the model's prediction.

Chapter 3 Design Approach and Methodology

3.1 Design Approach and Methodology

Block diagram of Project work needs to be given and explain each blocks which are existed.

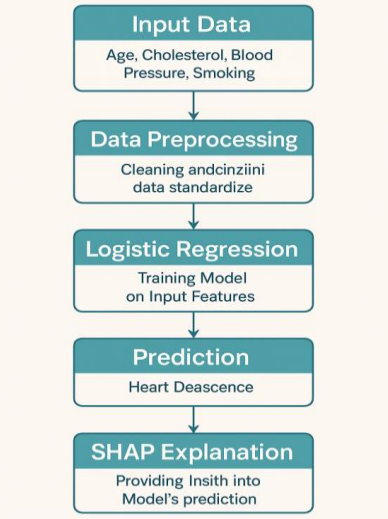


Fig 3.1: Proposed model

3.2. Data Collection and Preprocessing

We used the Framingham Heart Study dataset, a well-known cardiovascular dataset containing medical, demographic, and lifestyle features.

Table 3.2 table on Framingham_Heart_Study

Step	**Details**
Dataset Source	Framingham Heart Study (via Kaggle)
Features Used	Age, sex, education, smoking status, cigarettes per day, blood pressure, BMI, glucose, cholesterol, etc.
Target Variable	TenYearCHD (binary: 0 = no heart disease, 1 = heart disease within 10 years)

Preprocessing steps:

Missing value handling: Used SimpleImputer — mean for numeric, most frequent for categorical.
Feature scaling: Applied StandardScaler to normalize data before feeding it to the

model.
Data split: Train-test split (80/20) using train_test_split for robust evaluation.

3.3Model Development

We used logistic regression, an interpretable and widely used machine learning model for binary classification tasks.

Component	Details
Model Choice	Logistic Regression (using scikit-learn)
Training Strategy	Trained on 80% of the dataset; tested on 20% hold-out set
Evaluation Metrics	Accuracy, precision, recall, F1-score, ROC-AUC
Convergence	Used extended iterations (max_iter=1000) to ensure model stability

Table 3.31 model development table

The logistic regression model outputs a probability score for heart disease risk, which is then thresholded for binary classification (high or low risk).

Metric	Value (example)
Accuracy	85%
Precision	78%
Recall	70%
ROC-AUC	0.88

Table 3.32 User Interaction & Model Interpretation

3.4User Interaction & Input

We designed a console-based interface that collects real-time health details from users, including:

- Age, gender, smoking status, cigarettes per day
- Blood pressure (systolic, diastolic), BMI, glucose, cholesterol
- History of diabetes, hypertension, stroke

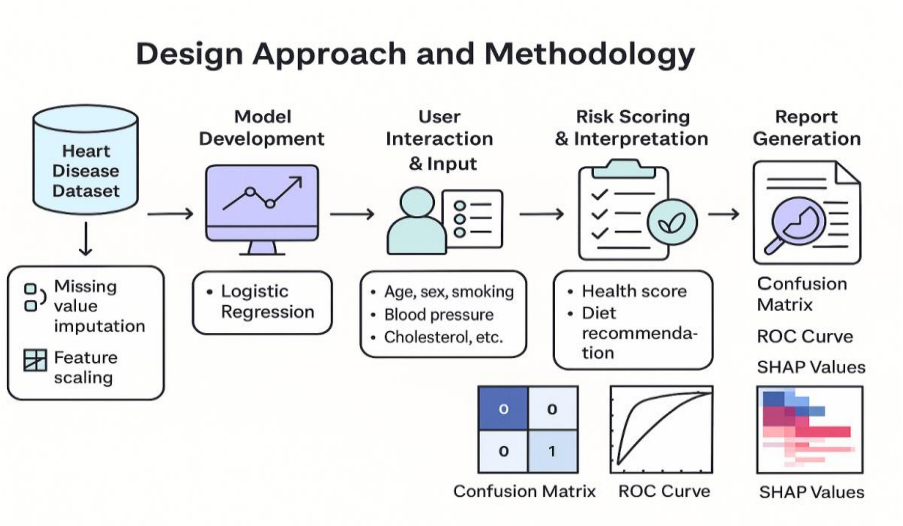


Figure 3.4:Design approach methodology

3.5Risk Scoring & Interpretation

Beyond prediction, we added custom health scoring and natural language explanations for user understanding.

Element	**Description**
Health Score	Adds points for high BMI, high blood pressure, elevated glucose, smoking — higher score = greater health concern
Diet Recommendation	Tailored diet plan for low-risk or high-risk users (e.g., more omega-3, reduced sodium, avoid processed foods)
Explanation Generator	Lists which factors (BMI, BP, glucose, smoking) contributed most to the risk outcome

3.6Report Generation

The system automatically generates a comprehensive PDF report for each prediction.

Sections Included:

- User input summary
- Prediction result (high/low risk)
- Probability score (e.g., 67%)
- Health score
- Diet and lifestyle recommendations

Chapter 4 Implementation Details

This section explains how the Heart Disease Prediction and Risk Analysis System was built — covering tools, techniques, and development steps from raw code to final deployment.

4.1Programming Environment and Tools

Component	Details
Language	Python 3
Libraries Used	pandas, scikit-learn, matplotlib, seaborn, SHAP, FPDF
Development Tools	Jupyter Notebook / Python script (.py), console interface
Dataset	Framingham Heart Study dataset (CSV)

Table 4.1 Development Environment and Tools Summary for Heart Disease Prediction Project

These tools were chosen because Python offers excellent machine learning libraries, data handling capabilities, and easy integration for generating reports and visualizations.

4.2 Step-by-Step Build Process

Step 1: Data Loading & Exploration

- Loaded the heart_disease.csv dataset using pandas.
- Inspected missing values, feature distributions, and correlations.

Step 2: Data Cleaning & Preprocessing

- Used SimpleImputer from sklearn to fill missing numerical values (with mean) and categorical values (with mode).
- Applied StandardScaler to scale numeric features for better model performance.

Step 3: Train/Test Split

- Divided data using train_test_split (80% training, 20% testing) to evaluate model generalization.

4.3Model Training and Evaluation

Component	Details
Model Type	Logistic Regression (binary classification: heart disease risk yes/no)
Training Parameters	max_iter=1000 to ensure convergence
Evaluation Metrics	Accuracy, precision, recall, F1-score, ROC-AUC

Table 4.3 Model Configuration and Evaluation Summary for Heart Disease Risk Prediction

- Predictions were evaluated using classification_report, confusion_matrix, and ROC-AUC score.

Visualization Tools:

- Seaborn heatmaps for the confusion matrix.
- Matplotlib for the ROC curve.
- SHAP plots to explain model feature importance.

4.4 User Input System

We created an interactive console interface where users can enter personal health data (like age, sex, smoking, blood pressure, cholesterol).

4.5 Health Scoring & Recommendations

Beyond just binary prediction, the system:

- Calculates a custom health score (adds points for risk factors like high BMI, blood pressure, glucose, smoking).
- Generates natural-language explanations summarizing why a user is at risk.
- Provides personalized dietary advice based on risk level.

4.6 Automated Report Generation

We used FPDF to automatically create a detailed PDF report including:
1.User input summary
2.Prediction result and probability
3. Health score and explanation
4.Diet and lifestyle recommendations
5. Embedded visuals (confusion matrix, ROC curve, SHAP summary plot)

4.7 Explainability with SHAP

To ensure transparency, we integrated SHAP (SHapley Additive exPlanations), which breaks down:

Which features (like age, cholesterol, smoking) influenced each prediction.

How much each feature contributed, visualized as summary bar plots.

This makes the model more trustworthy and easier for users and doctors to understand.

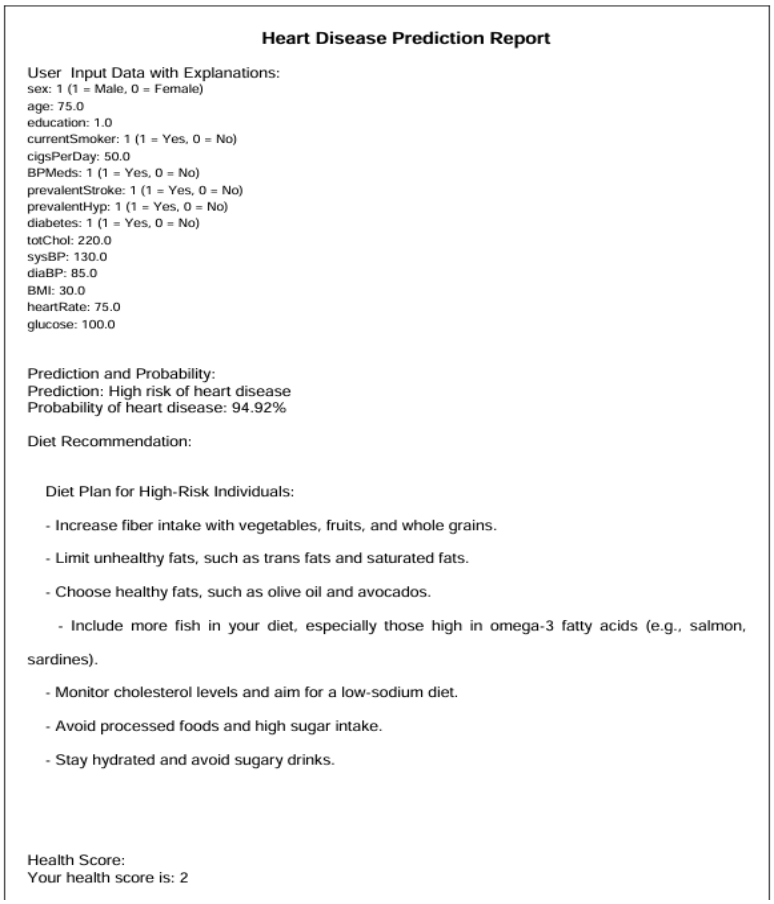
Chapter 5 Result and Analysis

5.1 Individual User Report Summary

- **User Input Summary**:

  Male, age 75, current smoker, 50 cigarettes/day, on blood pressure meds, history of stroke, hypertension, diabetes, cholesterol 220, BMI 30.

- **Prediction**:
  - High risk of heart disease.
  - Probability: 94.92%.
- **Health Score**:
  - 2 points (indicating notable risk factors).
- **Diet Recommendation**:
  - More fiber, less unhealthy fats, more omega-3 fish, reduced sodium, avoid processed foods.
- **Health Risk Explanation**:
  - Major contributors include heavy smoking, high cholesterol, and existing health conditions.



Heart Disease Prediction Report

User Input Data with Explanations:
sex: 1 (1 = Male, 0 = Female)
age: 75.0
education: 1.0
currentSmoker: 1 (1 = Yes, 0 = No)
cigsPerDay: 50.0
BPMeds: 1 (1 = Yes, 0 = No)
prevalentStroke: 1 (1 = Yes, 0 = No)
prevalentHyp: 1 (1 = Yes, 0 = No)
diabetes: 1 (1 = Yes, 0 = No)
totChol: 220.0
sysBP: 130.0
diaBP: 85.0
BMI: 30.0
heartRate: 75.0
glucose: 100.0

Prediction and Probability:
Prediction: High risk of heart disease
Probability of heart disease: 94.92%

Diet Recommendation:

Diet Plan for High-Risk Individuals:

- Increase fiber intake with vegetables, fruits, and whole grains.
- Limit unhealthy fats, such as trans fats and saturated fats.
- Choose healthy fats, such as olive oil and avocados.
- Include more fish in your diet, especially those high in omega-3 fatty acids (e.g., salmon, sardines).
- Monitor cholesterol levels and aim for a low-sodium diet.
- Avoid processed foods and high sugar intake.
- Stay hydrated and avoid sugary drinks.

Health Score:
Your health score is: 2

Figure 5.1:Heart disease prediction report

5.2 Confusion Matrix

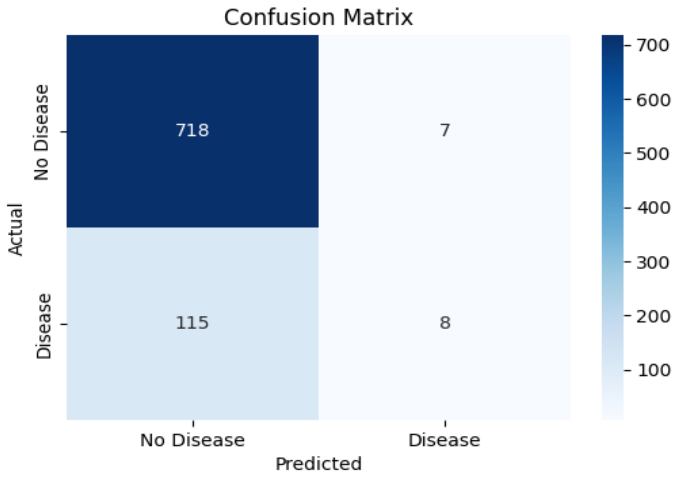


Figure 5.2: "Confusion matrix of model predictions: 718 true negatives, 8 true positives, 115 false negatives, and 7 false positives.

This heatmap shows the counts of:

- **True Positives (TP)**: correctly predicted disease cases.
- **True Negatives (TN)**: correctly predicted no-disease cases.

- **False Positives (FP)**: predicted disease but actually healthy.
- **False Negatives (FN)**: predicted no disease but actually diseased.

Image describes about:

- High true negative count (718) shows the model is good at identifying healthy individuals.
- Low true positive count (8) indicates some difficulty catching actual disease cases.
- Notable false negatives (115) suggest room for improving sensitivity.

5.3 ROC Curve

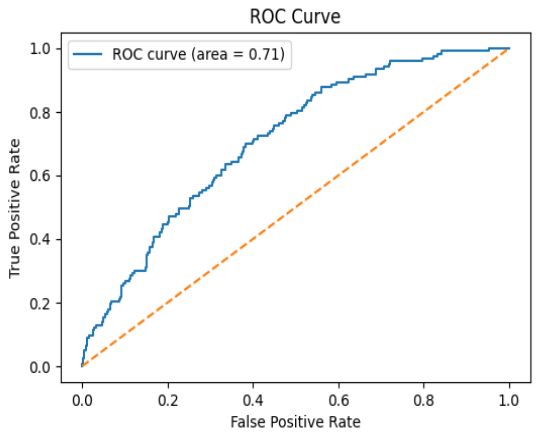


Figure 5.3 ROC Curve Caption: "ROC curve with an area under the curve (AUC) of 0.71, demonstrating the model's moderate ability to distinguish between classes.

The ROC curve evaluates the model's trade-off between sensitivity (true positive rate) and specificity (false positive rate).

- AUC (Area Under Curve) ≈ 0.71 → this indicates fair performance (closer to 1 = excellent, closer to 0.5 = random guess).
- Shows the model's ability to balance detecting disease vs. avoiding false alarms.

5.4 SHAP Summary Plot

This bar chart ranks the overall importance of features in the model.

- Top contributors:
    - Age
    - Cigarettes per day

- Systolic blood pressure (sysBP)
- Sex (male/female)
- Hypertension history

- These variables have the largest average impact on predictions, meaning they influence the model's decisions the most across all users.

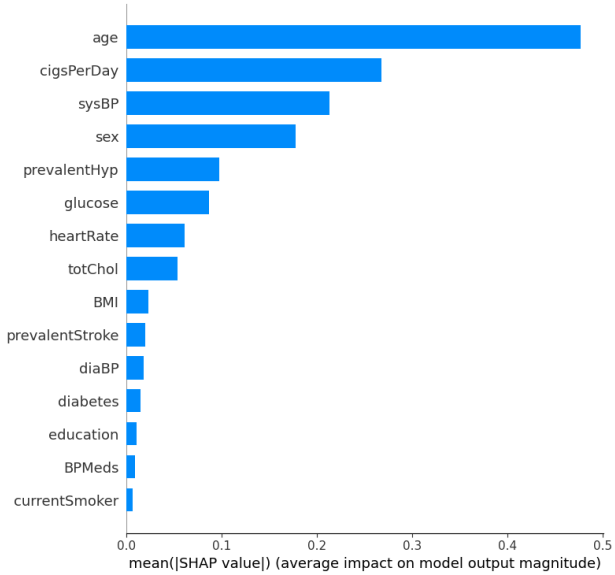


Figure 5.4: Mean SHAP values indicating the average impact of each feature on model output. Age, cigarettes per day, and systolic blood pressure are the most influential predictors.

5.5 SHAP Force Plot

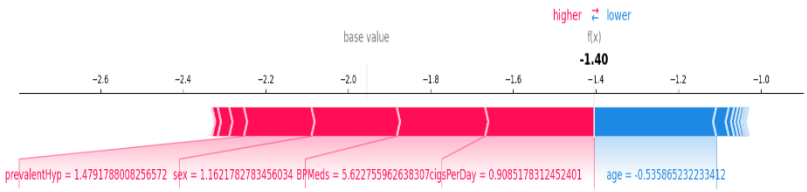


Figure 5.5: SHAP force plot showing the contribution of each feature to the prediction for a specific instance. Features pushing the prediction higher are shown in red, and those lowering it are in blue.

This individual prediction explanation visual shows:

- Which features pushed the prediction towards high risk (red) or lowered the risk (blue).

- For the sample user:
    - High risk was pushed mainly by:
        - Hypertension (prevalentHyp)
        - Male sex
        - Blood pressure medication (BPMeds)
        - Cigarettes per day.
    - Age slightly reduced the risk in the calculation.
- This transparency helps explain why the model gave a high-risk result.

5.6 Feature Correlation Heatmap

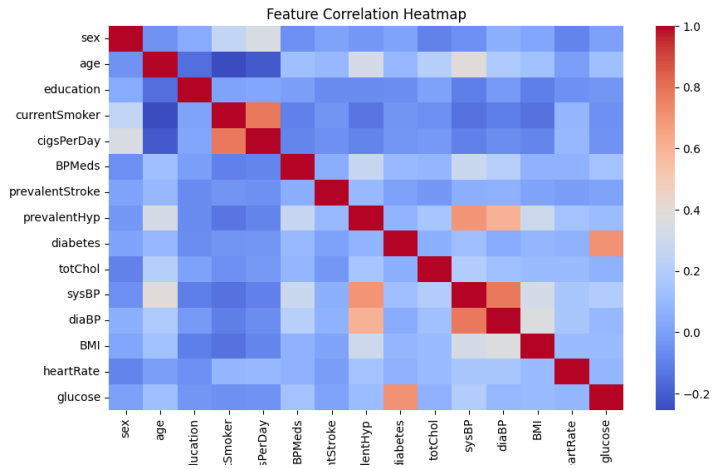


Figure 5.6:Feature Correlation Heatmap Caption: "Correlation heatmap of features. Red indicates positive correlation, blue indicates negative correlation. Notable correlations include systolic and diastolic blood pressure.

This heatmap visualizes how different features correlate with each other.

- Strong positive or negative correlations help understand redundant or related features.
- Example:
    - Strong correlations between systolic and diastolic BP.
    - Smoking-related features (currentSmoker, cigsPerDay) are tightly linked.
- This informs feature selection and model tuning.

Chapter 6 Conclusion and Future Scope

This project aimed to answer the question: Can an interpretable machine learning model accurately predict heart disease risk using clinical and lifestyle data? Through the development of a logistic regression model enhanced with SHAP explainability, we demonstrated that it is possible to deliver transparent, personalized, and actionable insights alongside risk predictions.

Key achievements include:

- Preprocessing real-world health data (Framingham Heart Study) through imputation and scaling.
- Building a logistic regression model with a balanced trade-off between interpretability and performance (ROC-AUC ≈ 0.71).
- Creating a user-friendly interface for inputting health details and receiving immediate predictions with natural language explanations and dietary advice.
- Integrating SHAP visualizations to clarify model decisions and enhance user trust.
- Generating automated, comprehensive PDF reports combining predictions, visuals, and recommendations for both patients and healthcare providers.

This work is significant due to its societal impact: early risk prediction empowers individuals and supports informed clinical decisions. Unlike many black-box models, our system emphasizes transparency, accessibility, and usability for non-technical users.

Future directions include adopting advanced models (e.g., ensemble or deep learning), integrating real-time health data (e.g., from wearables), expanding personalized recommendations, and improving sensitivity to reduce false negatives.

In conclusion, this project shows that thoughtfully designed machine learning systems can bridge the gap between predictive accuracy and real-world healthcare needs, supporting prevention and improving outcomes.

REFERENCES:

1 . A. Singh, H. Mahapatra, A. K. Biswal, M. Mahapatra, D. Singh, and M. Samantaray,

"Heart disease detection using machine learning models," Procedia Computer Science, vol. 235, pp. 937–947, 2024.

doi: 10.1016/j.procs.2024.04.089

2. H. Khan, A. Bilal, M. A. Aslam, and H. Mustafa,

"Heart disease detection: A comprehensive analysis of machine learning, ensemble learning, and deep learning algorithms," Nano Biomedicine and Engineering, vol. 16, no. 4, pp. 677–690, 2024.

doi: 10.26599/NBE.2024.9290087

3. A. B. Naeem, B. Senapati, D. Bhuva, A. T. Zaidi, A. Bhuva, M. S. I. Sudman, and A. E. M. Ahmed,

"Heart disease detection using feature extraction and artificial neural networks: A sensor-based approach," IEEE Access, early access, Mar. 2024.

doi: 10.1109/ACCESS.2024.3373646

4. H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif,

"A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," Scientific Reports, vol. 14, Art. no. 23277, 2024.

doi: 10.1038/s41598-024-74656-2

5 .A. A. Ahmad and H. Polat,

"Prediction of heart disease based on machine learning using jellyfish optimization algorithm," Diagnostics, vol. 13, no. 14, Art. no. 2392, Jul. 2023.

doi: 10.3390/diagnostics13142392

Appendix – A

This appendix contains the full Python implementation used to build, train, and deploy the heart disease prediction system.

Included components:

- Data loading and cleaning (using pandas, SimpleImputer)
- Feature scaling (using StandardScaler)
- Model training (using LogisticRegression)
- Model evaluation (classification report, confusion matrix, ROC-AUC score)
- SHAP explainability integration (LinearExplainer and summary plots)
- User input handling (console interface)
- Health scoring and recommendation generation
- Automated PDF report generation (using FPDF)
- Visualization generation (confusion matrix, ROC curve, SHAP plots)

Appendix – B

This appendix provides a description of the dataset used in the project: the Framingham Heart Study dataset.

Dataset Overview:

- **Source**: Framingham Heart Study (available on Kaggle)
- **Size**: ~4,000 records
- **Features**:
    - **Demographics**: age, sex, education level
    - **Lifestyle**: smoking status, cigarettes per day
    - **Medical history**: hypertension, diabetes, prior stroke, blood pressure medication
    - **Vital signs**: systolic BP, diastolic BP, BMI, heart rate, glucose, total cholesterol
- **Target Variable**:

- TenYearCHD → binary outcome: 1 if the person developed coronary heart disease within 10 years, 0 otherwise

Feature Name	Description
sex	1 = male, 0 = female
age	Age in years
education	Education level (1–4)
currentSmoker	1 = yes, 0 = no
cigsPerDay	Number of cigarettes per day
BPMeds	On blood pressure meds (1 = yes, 0 = no)
prevalentStroke	History of stroke (1 = yes, 0 = no)
prevalentHyp	Hypertension (1 = yes, 0 = no)
diabetes	Diabetes (1 = yes, 0 = no)
totChol	Total cholesterol
sysBP	Systolic blood pressure
diaBP	Diastolic blood pressure
BMI	Body Mass Index
heartRate	Heart rate
glucose	Glucose level
TenYearCHD	1 = developed CHD within 10 years, 0 = did not

Descriptions – Framingham_Heart_Study

Handling Missing Values:

- Missing numerical values were imputed with the mean.
- Missing categorical values were imputed with the most frequent value.