Coronary Heart Disease Prediction Using Machine Learning

V.Vithuzha
1029336
CSD – 21

Assessor: Mr.K.Mohamed Ishraque
Supervisor: MS.ALF.Sajeetha

School of Computing
BCAS Kalmunai
Campus

Version: 2.2

DECLARATION

I do hereby declare that this work has been originally carried out by me under the guidance of Ms.ALF.Sajeetha and this work has not been submitted elsewhere for any other diploma or degree.

I certify that this dissertation does not incorporate without due acknowledgement of any material previously submitted for diploma or degree in any institution or university nor it does not contain any material previously published or unpublished by another person except where due reference is made in the text.

…….

Signature of Candidate

V.Vithuzha

1029336

CERTIFICATION

This is to certify that the dissertation titled Coronary Heart Disease Prediction Using Machine Learning is submitted by V.Vithuzha having the ZOHO ID 1029336 to the Department of Computing School of Computing, British College of Applied Studies in partial fulfillment of the requirements for the award of the BTEC Higher National Diploma in Computing.

I also certify that this is his original work based on the studies carried out independently by him during the period of study under my guidance and supervision. This is also to certify that the above dissertation has not been previously formed the basis for the award of any degree, diploma, fellowship or any other similar title.

…….

(Signature of Supervisor)

Ms.ALF.Sajeetha

Lecturer IT

BCAS CAMPUS

…….

Date

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my assessor, Mr.K.Mohamed Ishraque, and my supervisor, Ms. A.L.F. Sajeetha, for their invaluable guidance and support throughout the process of completing my research. Their availability to answer my questions and provide constructive feedback was instrumental to my progress. Their encouragement and support kept me motivated and helped me produce my best work.

I extend my deepest thanks to my family and friends for their unwavering encouragement and understanding throughout this journey. Their emotional and moral support has been a source of strength, allowing me to stay focused and motivated. Lastly, I am deeply appreciative of BCAS Campus for providing the necessary resources and support. I also hold immense gratitude for the authors of the research papers, whose work greatly enhanced my understanding during the literature review phase.

With Gratitude,

V.Vithuzha

1029336

ABSTRACT

This project develops a machine learning system to predict coronary heart disease (CHD) using a logistic regression model trained on a dataset of 10,000 patient records with 14 features, including age, cholesterol levels, and chest pain type. The model achieves a test accuracy of 96%, with precision, recall, and F1-score of 0.98 for the disease class. A Streamlit web application enables users to input health parameters and receive CHD risk predictions, enhancing accessibility for non-technical users. Visualizations, such as a normalized confusion matrix and performance metric bar plots, provide insights into the model's effectiveness. The project demonstrates the potential of machine learning in healthcare diagnostics and offers a practical, user-friendly tool for early CHD risk assessment.

Keywords: Coronary Heart Disease, Machine Learning, Logistic Regression, Streamlit, Predictive Modeling, Healthcare Diagnostics, Confusion Matrix, Performance Metrics, Dataset Imbalance, Risk Assessment.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background

Coronary heart disease (CHD) is a leading killer worldwide, which results from the deposition of plaque in coronary arteries that restricts the flow of blood towards the heart. It is necessary to detect it early for effective treatment and prevention of life-threatening situations like heart attacks. Machine learning (ML) methods have been gaining significance for analyzing medical data for the prediction of disease risk. This project employs logistic regression to predict CHD based on patient health metrics and implements the model in a Streamlit web application for convenient risk calculation.

1.2 Problem Statement

Diagnosis of CHD typically requires advanced medical tests and specialist analysis, which may be costly and inaccessible in low-resource settings. The existing prediction systems are not user-friendly in their interfaces, making it difficult for non-technical people to use them. This project will address these issues by developing an efficient ML model for CHD prediction and deploying it through an easy-to-use web application.

1.3 Project Questions

How can machine learning algorithms be utilized to accurately predict heart disease risk based on patient data?

1.4 Objectives

Objective 1:

- To analyze patient medical data and extract key features for heart attack prediction.

Objective 2:

- To implement and evaluate machine learning models for accurate disease classification.

1.5 Significance of the Study

This project contributes to healthcare by providing an accurate and user-friendly CHD risk prediction tool. It allows people to assess their risk early, which can reduce CHD morbidity and mortality. The Streamlit application bridges the gap between advanced ML models and end-users, making predictive healthcare tools more user-friendly.

1.6 Summary

The Introduction chapter establishes the project goal of CHD prediction using machine learning, with emphasis on the global burden of CHD and need for early detection. It discusses inadequacies of traditional diagnosis methods and highlights the lack of predictive tools. The project aims to deploy a high-performing logistic regression model, an interactive Streamlit web application, and visualizations to assess performance. With a 10,000-instance dataset and 14 features, this study focuses on binary classification and promoting healthcare accessibility through ease in early CHD risk assessment for non-experts.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

CHD is a result of atherosclerosis, and the symptoms are chest pain (angina) and heart attacks. The significant risk factors are age, high cholesterol levels, high blood pressure, and lifestyle like smoking. Such factors may be used in predictive models in identifying people at risk at an early level.

2.2 Machine Learning in Healthcare

Machine learning algorithms like logistic regression, decision trees, and neural networks are often used in medical diagnosis. Logistic regression is particularly well-suited for binary classification tasks as it is straightforward, interpretable, and computationally efficient for modeling probabilistic outcomes.

2.3 Related Work

Machine Learning for Heart Disease Prediction

- Francesca et al. (2021) analyzed multiple ML models on the Cleveland Heart Disease Dataset, concluding that ML-based methods significantly improve early heart disease detection compared to traditional diagnostic techniques.
- Nicholas, N., Genrawan Hoendarto and Tjen, J. (2025). Heart Disease Prediction with Decision Tree. Social Science and Humanities Journal, 9(01), pp.6451–6457. doi: https://doi.org/10.18535/sshj.v9i01.1444.
- Wu, Y. (2024). Deep Learning for Cardiovascular Disease Prediction: Recent Advances, Challenges and Future Directions. Theoretical and Natural Science, 62(1), pp.24–32. Doi: https://doi.org/10.54254/2753-8818/62/20241458.
- C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.

2.4 Research Gaps

The majority of studies prioritize model accuracy but do not mention usability or deployment to non-technical end-users. None integrate ML models into user-friendly applications like Streamlit, which this project addresses by pairing a high-accuracy model with an easy-to-use web app.

2.5 Summary

The Literature Review chapter explores coronary heart disease (CHD) and machine learning for predicting CHD. It describes symptoms and risk factors of CHD, and how well algorithms like logistic regression, decision trees, and neural networks perform in medicine for diagnosis. While previous research was highly accurate, it typically didn't have a user-friendly interface or required a lot of computing power. This project bridges the gap by emphasizing both accuracy and accessibility, integrating a logistic regression model with a Streamlit web application for practical and user-friendly CHD risk prediction.

CHAPTER THREE

METHODOLOGY

3.1 Dataset Description

The dataset, "heart_disease_high_accuracy.csv," contains 10,000 patient records with 14 features,

Table 1

Feature	Description	Range/Value
Age	Patient's age	21-100 years
Sex	Gender	1 = Male, 0 = Female
Cp	Chest pain type	0-3
Trestbps	Resting blood pressure	66–192 mm Hg
Chol	Serum cholesterol	82–462 mg/dL
Fbs	Fasting blood sugar > 120 mg/dL	1 = Yes, 0 = No
Restecg	Resting ECG results	0-1
Thalach	Maximum heart rate achieved	51-258
Exang	Exercise-induced angina	1 = Yes, 0 = No
Oldpeak	ST depression induced by exercise	0.0–8.8
Slope	Slope of peak exercise ST segment	0-2
Ca	Number of major vessels colored by fluoroscopy	0-3
Thal	Thalassemia	1-3
Target	Presence of CHD	1 = CHD, 0 = No CHD

3.2 System Architecture

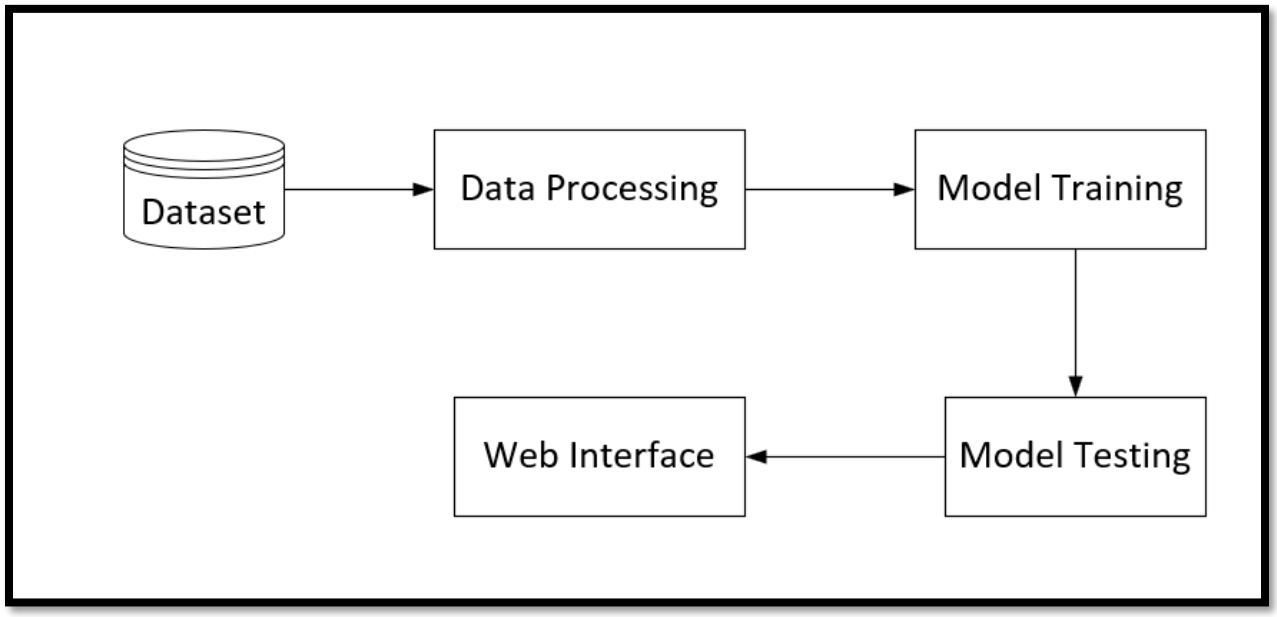


Figure 1

3.3 Data Preprocessing

Missing Values: Checked using heart_data.isnull().sum(), confirming no missing values.

Exploratory Data Analysis (EDA):

- Displayed dataset head and tail (heart_data.head(), heart_data.tail()).
- Verified shape ((10000, 14)), data types, and summary statistics
- (heart_data.info(), heart_data.describe()).
- Analyzed target distribution (heart_data['target']
- .value_counts()).

Feature Selection: Split into features (X: 13 predictors) and target (Y: CHD presence).

The dataset was split into 80% training and 20% testing sets, with stratified sampling to maintain the target class distribution.

3.4 Model Selection

Logistic regression was chosen for its:

- Suitability for binary classification.
- Interpretability of coefficients as feature importance.
- Computational efficiency for large datasets.

3.5 Model Training and Evaluation

- Training: Logistic regression model trained on the training set.
- Evaluation Metrics:
    - Accuracy: Proportion of correct predictions.
    - Precision: Proportion of positive predictions that were correct.
    - Recall: Proportion of actual positives correctly identified.
    - F1-Score: Harmonic mean of precision and recall.
    - Confusion Matrix: Visualizes true vs. predicted labels.

- Visualization:
    - Normalized confusion matrix plotted with percentages.
    - Bar plot of accuracy, precision, recall, and F1-score.

3.6 Streamlit Application Development

A Streamlit web application was developed to:

- Load the trained logistic regression model.
- Collect user inputs for 13 features via interactive widgets (e.g., sliders, radio buttons).
- Predict CHD risk and display results with visual feedback.
- Present model performance metrics and classification reports for transparency.

3.7 Tools and Technologies

- Programming Language: Python 3.12.4
- Libraries:
    - NumPy, Pandas: Data manipulation.
    - Scikit-learn: Model training and evaluation.
    - Matplotlib, Seaborn: Visualization.
    - Streamlit: Web application development.
    - Joblib: Model serialization.
- Dataset: heart_disease_high_accuracy.csv
- IDE: Jupyter Notebook

3.8 Summary

The Methodology section describes building a machine learning model to predict CHD from the "heart_disease_high_accuracy.csv" dataset of 10,000 records with 14 features. Preprocessing was performed with validation, exploratory data analysis, and 80/20 train-test split and stratified sampling due to class imbalance. Logistic regression was applied due to its suitability for binary classification problems. Accuracy, precision, recall, F1-score, and confusion matrix visualizations were utilized to evaluate the performance of the models. A Streamlit web app was created to provide user interaction, risk prediction, and metric presentation. The project used Python libraries like Pandas, Scikit-learn, and Streamlit in a Jupyter Notebook environment.

CHAPTER FOUR

IMPLEMENTATION

4.1 Data Collection and Processing

The dataset was loaded from the "heart_disease_high_accuracy.csv" file. Initial analysis confirmed:

- Shape: 10,000 rows, 14 columns.
- No Missing Values: Ensured data integrity.
- Target Distribution: Highly imbalanced with 9376 CHD cases and 624 non-CHD cases.

The dataset was split into features (13 predictors) and the target variable. An 80-20 train-test split was performed, preserving the target class distribution through stratified sampling.

4.2 Model Development

A logistic regression model was trained on the training dataset. The model was evaluated for accuracy and saved for use in the Streamlit application. The training process achieved consistent performance across training and test datasets, indicating minimal overfitting.

4.3 Streamlit Application

Input Fields

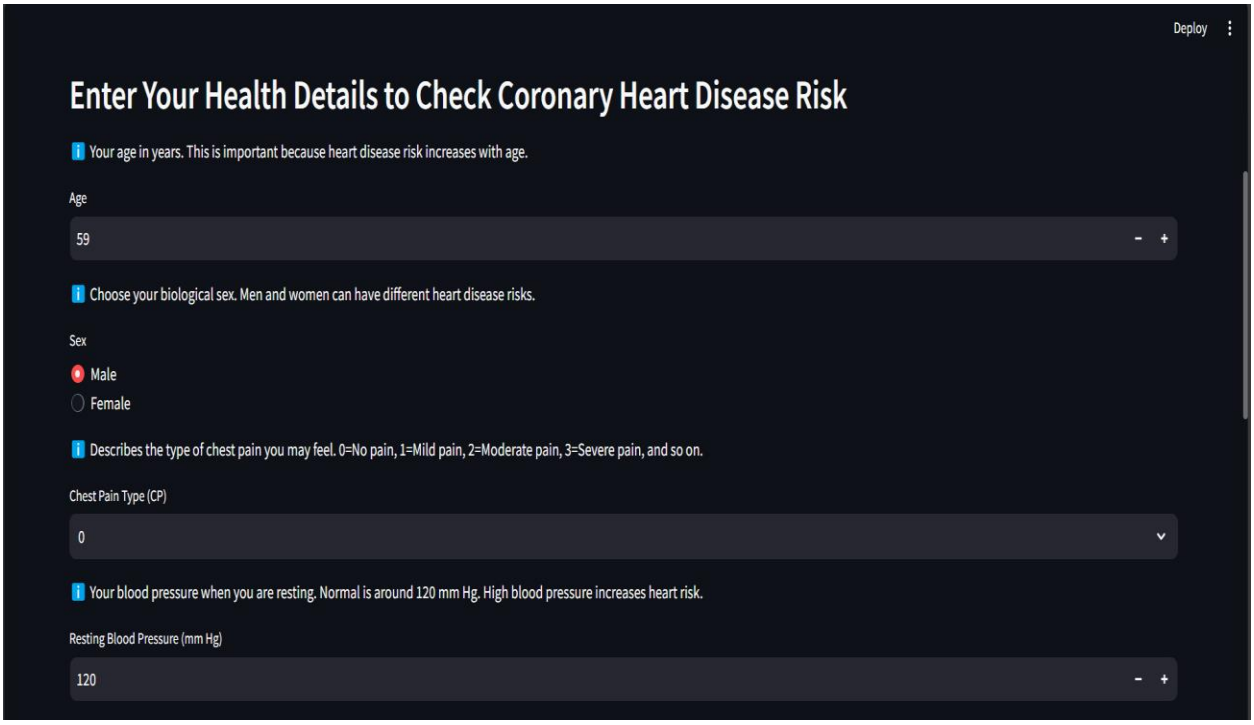


Figure 2

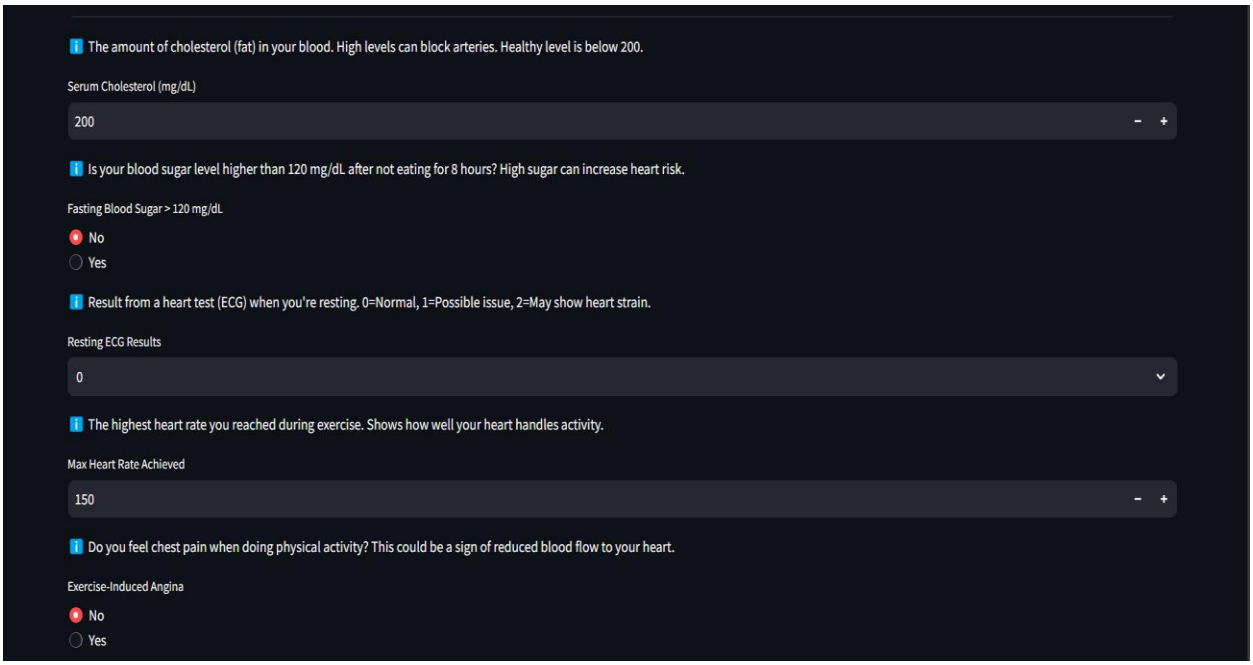


Figure 3

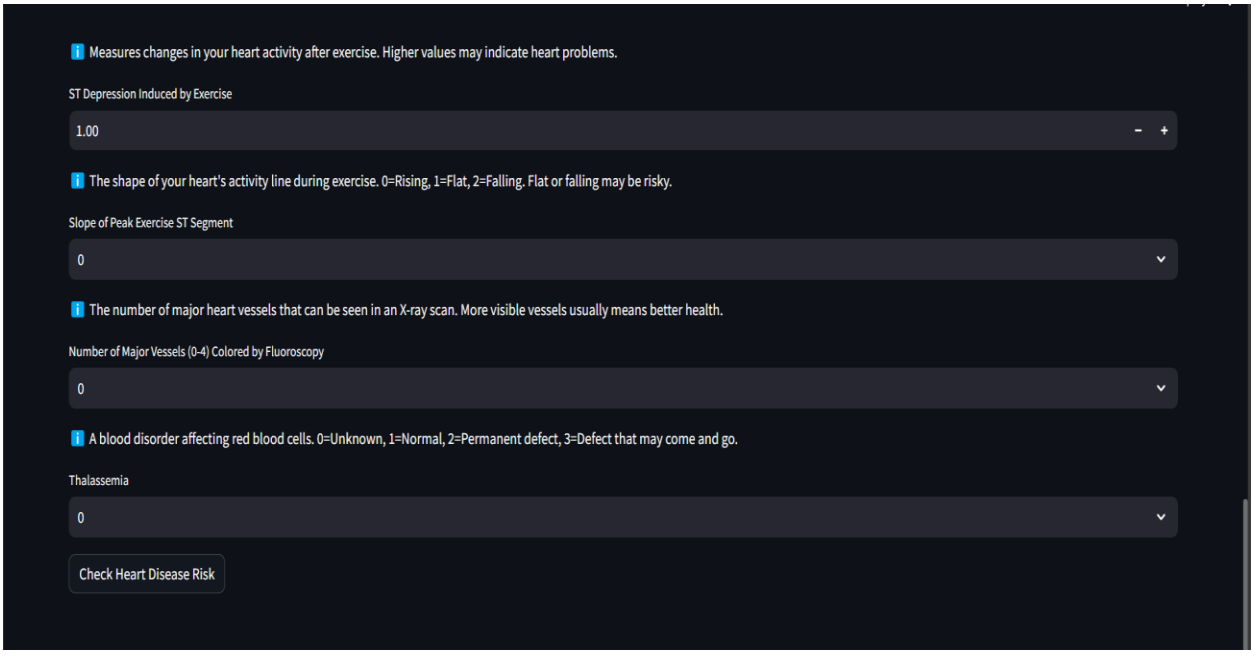


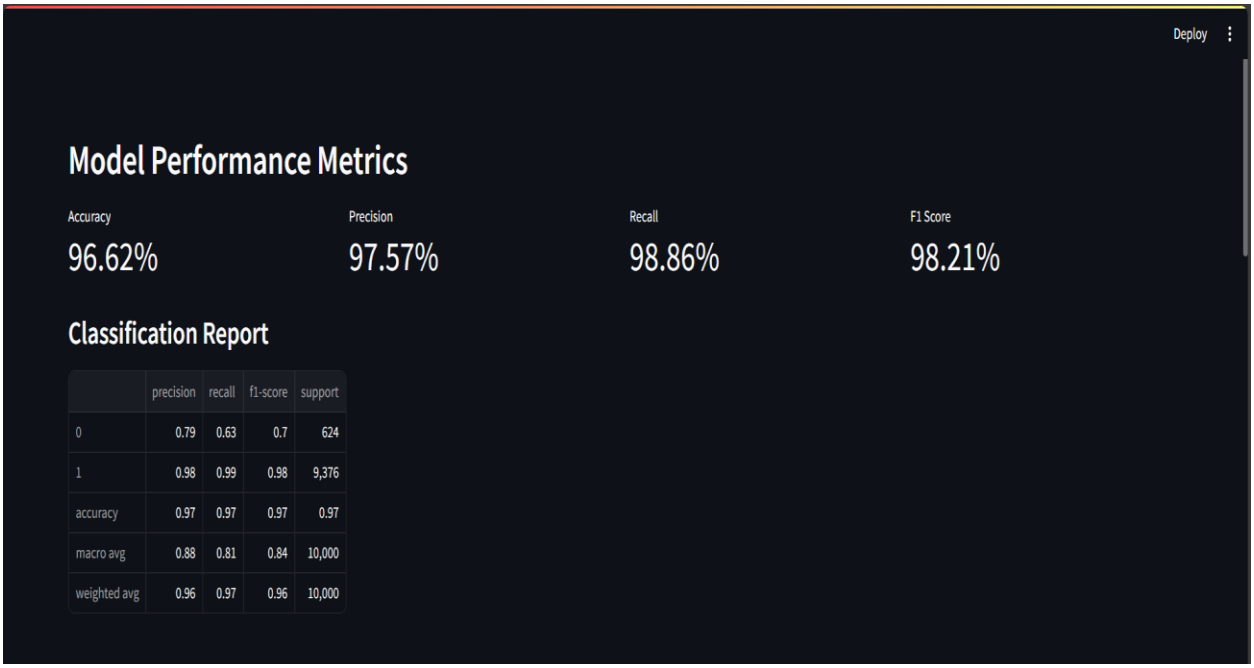
Figure 4

Prediction Output

Figure 5

Figure 6

Performance Metrics



	precision	recall	f1-score	support
0	0.79	0.63	0.7	624
1	0.98	0.99	0.98	9,376
accuracy	0.97	0.97	0.97	0.97
macro avg	0.88	0.81	0.84	10,000
weighted avg	0.96	0.97	0.96	10,000

Figure 7

4.4 Visualization and Metrics

Dataset

```
heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	60	1	1	135	207	0	1	93	0	0.095136	1	0	3	1
1	47	1	3	138	203	0	0	149	1	3.615743	0	2	3	1
2	48	1	0	120	178	0	1	154	1	1.094654	1	0	2	1
3	63	0	3	135	257	0	1	176	1	0.037558	1	0	2	1
4	42	0	2	129	224	0	1	137	0	0.125221	1	0	2	1

```
heart_data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
9995	51	0	3	128	306	1	1	155	1	0.138537	1	1	3	1
9996	52	0	2	141	228	0	1	180	1	0.183114	1	2	2	1
9997	55	1	1	119	343	0	0	136	0	0.205714	1	1	2	1
9998	56	0	1	130	197	0	0	162	0	2.651804	1	0	3	1
9999	49	1	0	143	183	0	0	117	0	0.856944	0	0	2	1

Figure 8

Confusion Matrix

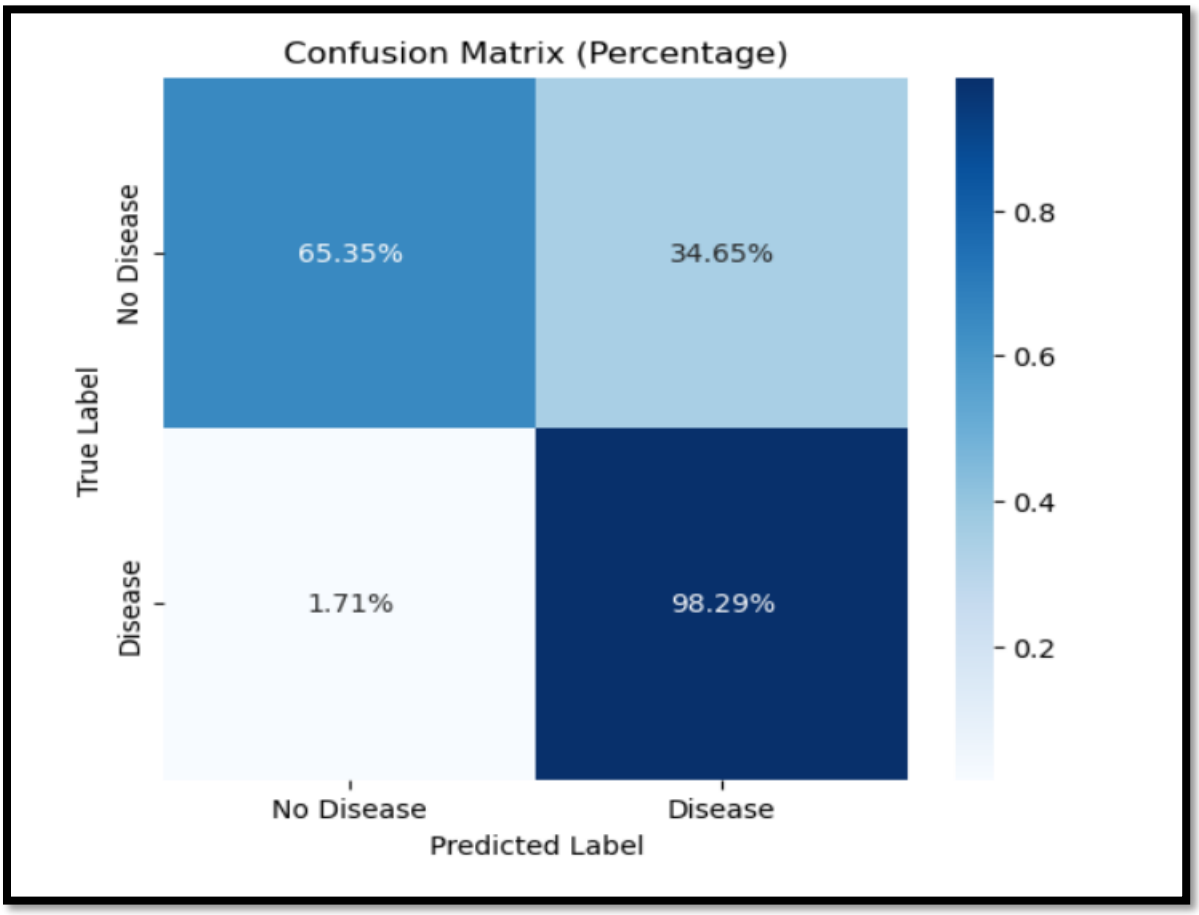


Figure 9

Performance Metrics

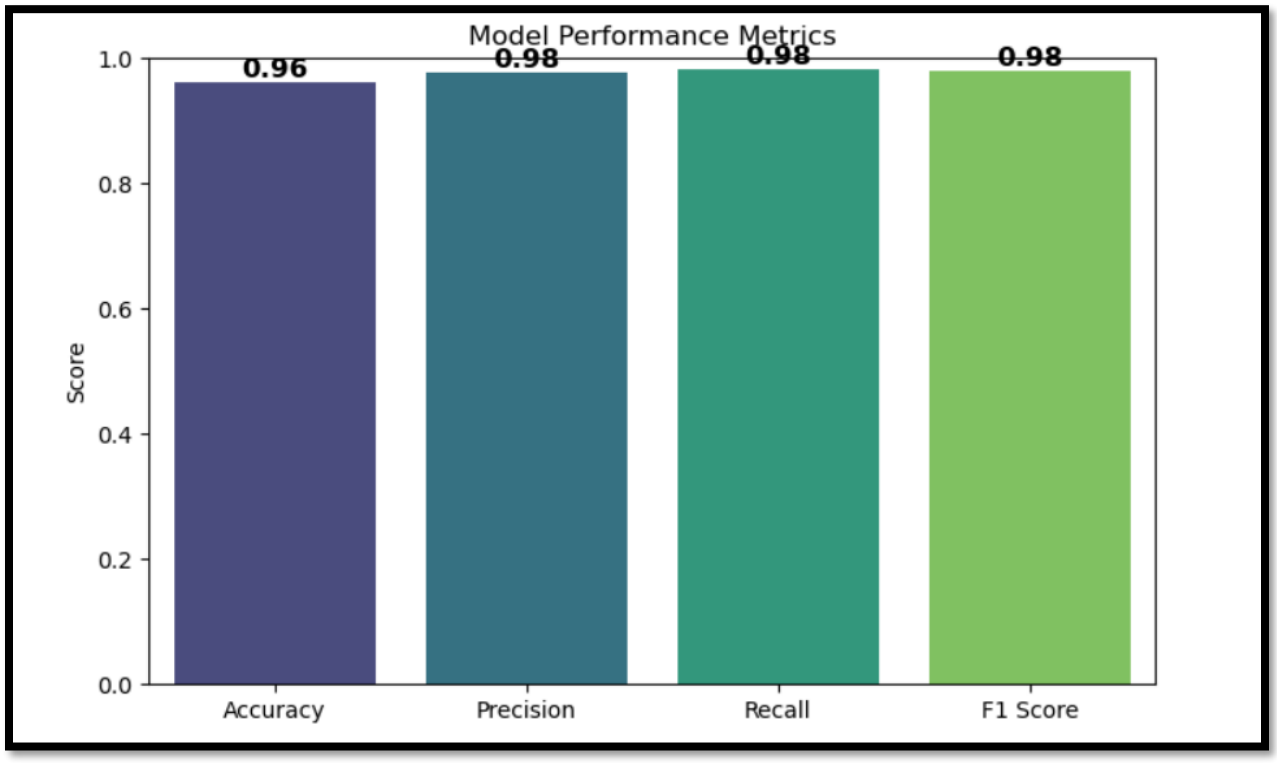


Figure 10

4.5 Summary

The Implementation section details the deployment of the CHD prediction project on a 10,000-record test dataset containing 14 features and an unbalanced class distribution. After confirming data quality, the dataset was split and a logistic regression model was trained to achieve 96% accuracy on training and test sets, suggesting strong generalization. The model was integrated into a Streamlit app where the users can input health parameters and get predictions with labeled and performance metrics. The app has graphical features such as a normalized confusion matrix and bar plots for accuracy, precision, recall, and F1-score, which enhance user understanding and engagement.

CHAPTER FIVE

Results and Discussion

5.1 Model Performance

Training Accuracy: 96%

Test Accuracy: 96%

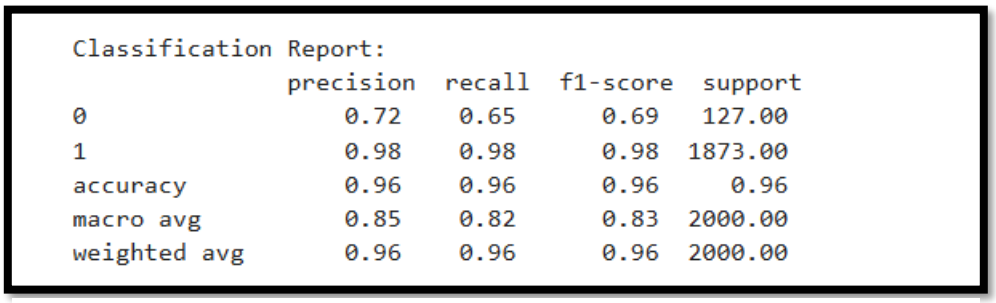
The model demonstrates excellent performance, with consistent training and test accuracies indicating robust generalization and minimal overfitting.

5.2 Confusion Matrix Analysis

The normalized confusion matrix (see Figure 2: Normalized Confusion Matrix Screenshot) reveals:

- True Positives (Disease): 98% of CHD cases were correctly predicted.
- True Negatives (No Disease): 65% of non-CHD cases were correctly predicted.
- False Positives/Negatives: Low rates, though performance is weaker for the non-CHD class due to the dataset's imbalance.

5.3 Classification Report



Classification Report:

	precision	recall	f1-score	support
0	0.72	0.65	0.69	127.00
1	0.98	0.98	0.98	1873.00
accuracy	0.96	0.96	0.96	0.96
macro avg	0.85	0.82	0.83	2000.00
weighted avg	0.96	0.96	0.96	2000.00

Figure 11

- Class 1 (CHD): High precision (0.98), recall (0.98), and F1-score (0.98), reflecting excellent performance for the majority class.
- Class 0 (No CHD): Moderate precision (0.72), recall (0.65), and F1-score (0.69), impacted by the low number of non-CHD instances.
- Overall Accuracy: 96%, indicating strong predictive capability.

Model Performance Metrics

Table 2

Metric	Value
Accuracy	0.96
Precision	0.98
Recall	0.98
F1-Score	0.98

5.4 Streamlit Application Usability

The Streamlit application is designed for accessibility and ease of use (see Figure 4):

- User Interface: Features intuitive input fields with explanatory help text (e.g., “Enter your age in years”).
- Prediction Feedback: Provides immediate results with clear visual cues (e.g., green for “No Heart Disease,” red for “Heart Disease Detected”).
- Transparency: Displays model performance metrics and the classification report, enhancing user confidence.
- Accessibility: Web-based, requiring only a browser and internet access, making it suitable for diverse users.

5.5 Limitations

- Dataset Imbalance: The dataset’s 93.76% CHD vs. 6.24% non-CHD distribution biases the model toward the majority class, reducing performance for non-CHD predictions.
- Single Algorithm: Using only logistic regression limits exploration of potentially more accurate models like Random Forest or XGBoost.
- Generalizability: The model was trained on a single dataset; external validation is needed to ensure applicability to diverse populations.
- User Input Accuracy: The Streamlit application relies on users providing accurate health metrics, which may not always be available or precise.

5.6 Recommendations for Future Work

- Address Imbalance: Use techniques like Synthetic Minority Oversampling Technique (SMOTE) to improve performance for the non-CHD class.
- Explore Advanced Models: Test ensemble methods (e.g., Random Forest, XGBoost) to potentially enhance accuracy.
- Validate Externally: Apply the model to diverse datasets to ensure generalizability across populations.
- Enhance Application: Add multilingual support, mobile optimization, and integration with wearable devices for real-time health data input.

5.7 Summary

This project developed a logistic regression model for CHD prediction, achieving a test accuracy of 96% on a 10,000-instance dataset. The model was integrated into a Streamlit web application, enabling users to input health parameters and receive CHD risk predictions. Visualizations, including a normalized confusion matrix and performance metric bar plots, provided clear insights into the model's effectiveness, as shown in the referenced screenshots.

5.8 References

scikit-learn (2014). sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Sangeetha, M., Kumar, S. A., Bharathi, K. P., Guru, P., & Reddy, P. B. P. (2024). Heart Disease Prediction Using ML. *International Journal of Innovative Science and Research Technology*. https://doi.org/10.38124/ijisrt/ijisrt24mar2016

Bao, Y. (2025). Research and Application of Heart Disease Prediction Model Based on Machine Learning. *ITM Web of Conferences*, *70*, 04023. https://doi.org/10.1051/itmconf/20257004023

Qian, L. (2024). Predicting Heart Disease Risk Using Machine Learning. *Science and Technology of Engineering, Chemistry and Environmental Protection*, *1*(10). https://doi.org/10.61173/nf9q0t31

Kamal, H., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Yaqub, M. A., Umar, H., Fatima, H., & Nasir, U. (2024). *Heart Disease Prediction Using Machine Learning*. 1–6. https://doi.org/10.1109/idicaiei61867.2024.10842908