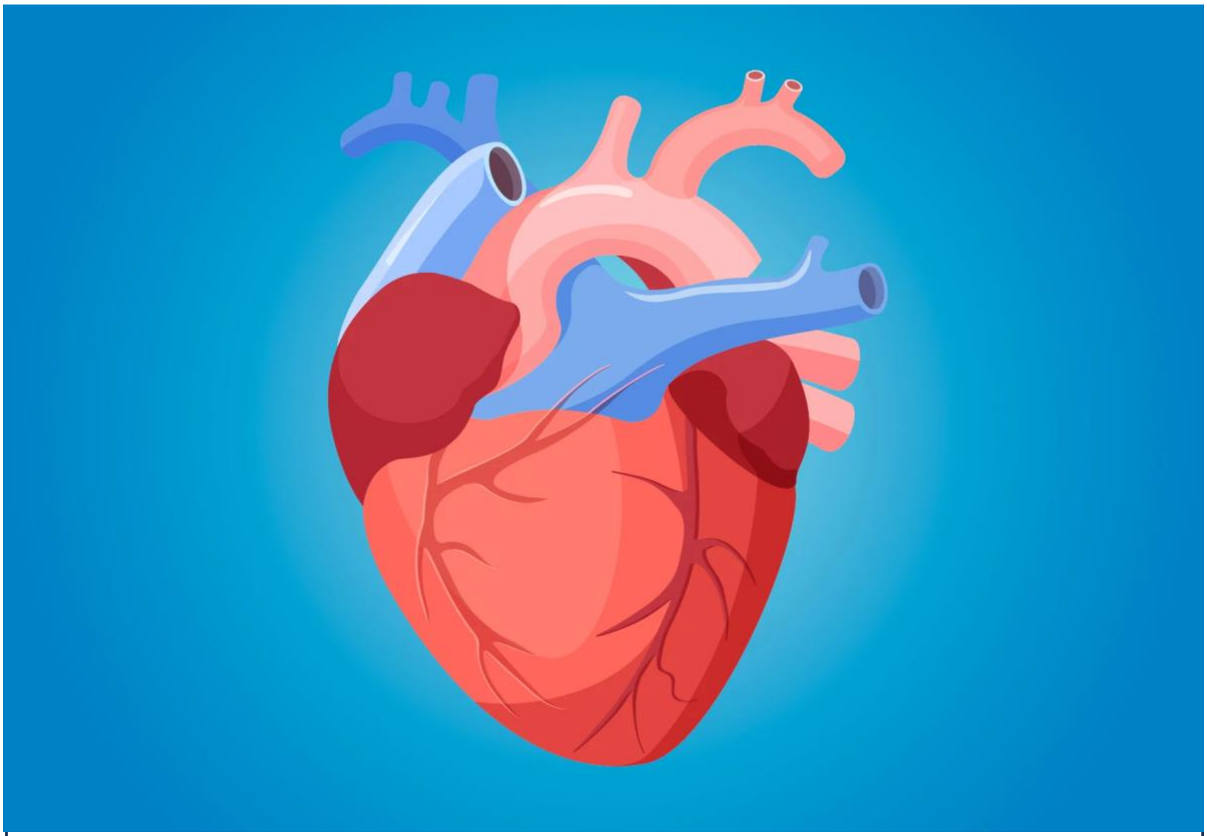

TECHNICAL DOCUMENTAION

CORONARY HEART RISK PREDICTION

Table of Contents

1. Introduction

Purpose of the Document

This technical documentation provides a detailed overview of the cardiovascular risk prediction model developed using machine learning techniques. The document covers data preprocessing, exploratory data analysis (EDA), model selection, and model performance evaluation. The goal is to explain the methodology, present key findings, and justify the selection of the final model.

Background Information

Cardiovascular diseases, including Coronary Heart Disease (CHD), are a significant global health concern. Early detection and accurate risk assessment are crucial for preventive healthcare. Machine learning models offer the potential to predict CHD risk based on various health-related factors. This document focuses on the development of such a model.

2. Data Overview

Dataset Description

The dataset used for this project contains information on 3,390 individuals and includes the following columns:

1. `id`: An identifier for each individual.

2. `age`: Age of the individual.

3. `education`: Education level of the individual.

4. `sex`: Gender (encoded as 0 for female and 1 for male).

5. `is_smoking`: Smoking status (encoded as 0 for non-smoker and 1 for smoker).

6. `cigsPerDay`: Number of cigarettes smoked per day.

7. `BPMeds`: Use of blood pressure medications.

8. `prevalentStroke`: History of stroke.

9. `prevalentHyp`: Prevalent hypertension.

10. `diabetes`: Diabetes status.

11. `totChol`: Total cholesterol levels.

12. `sysBP`: Systolic blood pressure.

13. `diaBP`: Diastolic blood pressure.

14. `BMI`: Body Mass Index.

15. `heartRate`: Heart rate.

16. `glucose`: Glucose levels.

17. `TenYearCHD`: Target variable, indicating CHD risk (0 for no risk, 1 for risk).

18. `pulsePressure`: Pulse pressure.

Data Preprocessing

Data preprocessing steps included handling missing values (if any), encoding categorical variables, and scaling numerical features. The dataset had no missing values, and categorical variables were appropriately encoded. Numerical features were scaled to ensure consistent model performance.

---

3. Exploratory Data Analysis (EDA)

Key Findings from EDA

The EDA phase revealed several important insights:

1. Age is a pivotal factor influencing the risk of CHD, with risk increasing with age.

2. Men exhibit a higher propensity for CHD compared to women.

3. Smoking emerged as a risk factor, with the intensity of smoking playing a role in risk determination.

4. Individuals with high blood pressure, a history of stroke, or diabetes are at elevated CHD risk.

5. The presence of both a history of stroke and prevalent hypertension is associated with a particularly high CHD risk.

6. Elevated cholesterol levels are observed among individuals at risk for CHD.

7. Certain variables, such as age and systolic blood pressure, exhibit positive associations, indicating interrelationships.

---

4. Model Implementation

Model Selection

Six different machine learning models were tested for predicting CHD risk using the preprocessed dataset. These models included:

- Random Forest Classifier

- XGBoost Classifier

- K-Nearest Neighbors (KNN) Classifier

- Support Vector Classifier (SVC)

- ... (additional models, if any)

Model Performance Metrics

Model performance was evaluated using the following metrics:

- Accuracy

- Precision

- Recall

- ROC AUC

Top Performing Models

1. Random Forest Classifier: High accuracy, precision, and recall scores.

2. XGBoost Classifier: High accuracy, precision, recall scores, and superior ROC AUC score.

3. KNN Classifier: Relatively high recall score but lower accuracy and precision.

4. SVC Classifier: Lower accuracy and ROC AUC scores.

Model Selection and Justification

The XGBoost Classifier was selected as the optimal model for cardiovascular risk prediction. It achieved an accuracy of 89.67%, outperforming other models. The decision was based on its well-rounded performance across accuracy, precision, and recall, as well as its superior ROC AUC score. This model is recommended for use in predicting CHD risk.

---

5. Conclusion

Summary of Key Findings

In summary, this document has outlined the development of a cardiovascular risk prediction model. Key findings from the EDA phase revealed the importance of age, gender, smoking, and various health conditions in influencing CHD risk. Among the models tested, the XGBoost Classifier was identified as the optimal choice, offering high accuracy and predictive power.

Model Selection and Recommendations

The XGBoost Classifier, with an accuracy of 89.67%, is recommended for predicting cardiovascular risk based on the dataset. Implementing this model could aid in early identification of individuals at risk of CHD, allowing for targeted preventive measures and healthcare interventions.

This technical documentation serves as a comprehensive guide to the model development process and findings, enabling users to understand and utilize the CHD risk prediction model effectively for healthcare applications.