# REPORT ON

# HEART DISEASE PREDICTION PROJECT

(Predict the Coronary Heart Disease Risk of Patients)

BY

SAMUEL OSEI

AND

JOANA LAWER

## Introduction

This report presents a predictive modeling project aimed at assessing the risk of coronary heart disease (CHD) using the Framingham dataset. The project involves data preprocessing, model building, hyperparameter tuning, and evaluation to develop a robust machine learning model for CHD prediction.

## 1. Description of the Dataset

The Framingham dataset is a well-known dataset in the field of cardiovascular research. It contains various health-related features collected from the Framingham Heart Study, a long-term, ongoing cardiovascular cohort study on residents of the town of Framingham, Massachusetts. The dataset includes demographic information, lifestyle factors, and clinical measurements for each participant.

Key features in the dataset include:

- Sex: Male or Female
- Age: Age of the participant
- Education: Level of education
- CurrentSmoker: Smoking status
- CigsPerDay: Number of cigarettes smoked per day
- BPMeds: Blood pressure medication usage
- PrevalentStroke: History of stroke
- PrevalentHyp: History of hypertension
- Diabetes: Diabetes status
- TotChol: Total cholesterol level
- SysBP: Systolic blood pressure
- DiaBP: Diastolic blood pressure
- BMI: Body mass index
- HeartRate: Heart rate
- Glucose: Glucose level
- TenYearCHD: Target variable indicating the risk of CHD within ten years

From our analysis, the dataset consists of 4238 entries on 16 features being the following: age, sex, education, smoking status, cigarettes per day, BP medications, stroke history, hypertension, diabetes, cholesterol, systolic and diastolic blood pressure, BMI, heart rate, glucose, and Ten-Year Coronary Heart Disease (CHD) prediction.

## 2. Overview of the Machine Learning Algorithm

The machine learning algorithm used in this analysis is the Random Forest Classifier. Random Forest Classifier is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is known for its robustness and high accuracy.

## 3. Overview of the Metrics Used to Evaluate the Performance

Several metrics are used to evaluate the performance of the machine learning model:

1. Accuracy: The proportion of correctly classified instances out of the total instances.

Accuracy = $(TruePositives\ +\ True\ Negatives)/TotalInstances$

2. Confusion Matrix: A table used to describe the performance of a classification model. It shows the counts of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) classifications.

3. Classification Report: Provides a detailed summary of the precision, recall, and F1 score for each class.

## 4. Results and Interpretation

The results of the model are summarised below:

Accuracy = $(715\ +\ 10)/848$ = 0.8549528301886793

This means the model correctly predicts about 85.5% of the instances.

Confusion Matrix: The confusion matrix is as follows:

| | Predicted No CHD | Predicted CHD |
|---|---|---|
| Actual No CHD | 715 | 9 |
| Actual CHD | 114 | 10 |

**True Positives (TP):** 10 (instances correctly predicted as class 1)

**True Negatives (TN):** 715 (instances correctly predicted as class 0)

**False Positives (FP):** 9 (instances incorrectly predicted as class 1)

**False Negatives (FN):** 114 (instances incorrectly predicted as class 0

### Classification Report

Class 0 (No CHD)

- Precision: 0.86

  Out of all the instances predicted as class 0, 86% were correct.
- Recall: 0.99

  The model correctly identified 99% of the actual class 0 instances.
- F1-Score: 0.92

The harmonic mean of precision and recall, indicating a strong performance for class 0.

Class 1 (CHD)

- Precision: 0.53
  Out of all the instances predicted as class 1, 53% were correct.
- Recall: 0.08
  The model correctly identified only 8% of the actual class 1 instances.
- F1-Score: 0.14
  The harmonic mean of precision and recall, indicating poor performance for class 1.

Based on these results, we can interpret that the model performs well in predicting No CHD cases (high precision) but struggles with CHD cases (low recall). This means it correctly identifies most No CHD cases but misses many actual CHD cases. To improve performance, consider:

- Adjusting the decision threshold

- Handling class imbalance

- Enhancing feature engineering

- Exploring alternative models

- Fine-tuning hyperparameters

Improving recall for CHD is essential to ensure more accurate identification of actual CHD cases.

## Plot Explanation

### Age Distribution

This plot shows the distribution of ages in the dataset. It helps to understand the age range and the frequency of each age group among the participants. The histogram with a KDE (Kernel Density Estimate) curve provides insights into the central tendency and dispersion of ages.

### Total Cholesterol Distribution

This plot illustrates the distribution of total cholesterol levels among the participants. By examining this plot, one can observe the range and frequency of different cholesterol levels, which is crucial for understanding the overall cardiovascular health of the population.

### Count of Ten Year CHD

This count plot shows the number of participants who are at risk of developing coronary heart disease (CHD) within ten years versus those who are not at risk. It provides a clear visual representation of the class imbalance in the dataset, which is important for model evaluation and improvement strategies.

### Correlation Matrix

The correlation matrix heatmap displays the correlation coefficients between different features in the dataset. Positive correlations are shown in warm colors, while negative correlations are shown in cool colors. This plot helps to identify relationships between variables, which can be useful for feature selection and understanding multicollinearity.

### Confusion Matrix

The confusion matrix plot depicts the performance of the classification model. It shows the counts of true positive, false positive, true negative, and false negative predictions. This plot is essential for evaluating the accuracy and reliability of the model, highlighting areas where the model performs well and where it needs improvement.