**PROJECT**

**Cardiovascular (Heart) Disease Prediction**

**Framingham Heart Study Dataset**

**Description**

The Framingham Heart Study is a long-running epidemiological study that has played a significant role in advancing our understanding of cardiovascular disease. The primary goal of the Framingham Heart Study is to identify and understand the risk factors associated with cardiovascular diseases, such as heart disease and stroke. Researchers collect data on various aspects of participants' health, including their medical history, lifestyle, diet, physical activity, and more. This data is used to identify patterns and risk factors associated with the development of heart diseases.

**Data Dictionary**

**Sex:** The gender of the participant (typically coded as 0 for female and 1 for male).

**Age:** The age of the participant at the time of data collection.

**Education:** The "education" column typically records information about the educational attainment of the study participants.

1: Less than high school education

2: High school graduate or equivalent

3: Some college or vocational training

4: College graduate

5: Postgraduate education (e.g., master's degree, doctorate)

**Current Smoker:** Information on whether the participant is a current smoker or non-smoker.

**CigsPerDay:** If the participant is Smoker, then it shows number of Cigarettes per day.

**BPMeds:** Information on the use of medications, such as blood pressure medication.

**Prevalent Stroke:** The "Prevalent Stroke" column in the dataset records whether or not each participant has had a prevalent stroke at the time of data collection. A prevalent stroke refers to the occurrence of a stroke before or at the beginning of the study period.

Values:

1: Participants with a value of 1 in this column have had a stroke before or at the beginning of the study. This indicates a history of stroke.

0: Participants with a value of 0 in this column have not had a stroke before or at the beginning of the study. This indicates the absence of a stroke history at the time of data collection.

**PrevalentHyp:** The "PrevalentHyp" column represents the presence or absence of prevalent hypertension (high blood pressure) among study participants at the time of data collection.

Values:

1: Participants with a value of 1 in this column have prevalent hypertension. This indicates that they have high blood pressure.

0: Participants with a value of 0 in this column do not have prevalent hypertension. This indicates that they do not have high blood pressure at the time of data collection.

**Diabetes:** Information on whether the participant has diabetes

**TotChol:** Cholesterol Levels: Typically includes measurements of total cholesterol, HDL cholesterol (high-density lipoprotein), and LDL cholesterol (low-density lipoprotein).

**SysBP:** The "sysBP" column contains numerical values representing the systolic blood pressure of each participant in the study. Systolic blood pressure is the higher of the two blood pressure values and is associated with the force exerted on blood vessel walls when the heart contracts.

Measurement Unit: Systolic blood pressure is typically measured in millimeters of mercury (mm Hg).

**DiaBP:** The "diaBP" column contains numerical values representing the diastolic blood pressure of each participant in the study. Diastolic blood pressure is the lower of the two blood pressure values and is associated with the force exerted on blood vessel walls when the heart is at rest between beats.

Measurement Unit: Diastolic blood pressure is typically measured in millimeters of mercury (mm Hg).

**BMI (Body Mass Index):** A measure of body weight relative to height, calculated as weight (kg) divided by height (m^2).

**Heart Rate:** The "heartrate" column contains numerical values representing the heart rate (number of heartbeats per minute) of each participant in the study. It provides information about the rate at which the heart is pumping blood throughout the body.

Measurement Unit: Heart rate is typically measured in beats per minute (bpm).

**Glucose:** The "Glucose" column contains numerical values representing blood glucose levels, typically measured in milligrams per deciliter (mg/dL), for each participant in the study. Blood glucose levels indicate the concentration of sugar (glucose) in the blood.

Measurement Unit: Blood glucose is typically measured in milligrams per deciliter (mg/dL).

**TenYearCHD:** The "TenYearCHD" column records binary values indicating whether each participant is at risk of developing coronary heart disease (CHD) within the next ten years. It serves as an outcome variable in the study, with "1" typically indicating a higher risk and "0" indicating a lower risk.

Values:

1: Participants with a value of 1 in this column are considered at risk of developing CHD within the next ten years.

0: Participants with a value of 0 in this column are considered not at risk of developing CHD within the next ten years.

**Business Problem Statement**

1. How can we develop an accurate and scalable heart disease prediction model using the Framingham Heart Study data to proactively identify individuals at high risk of developing coronary heart disease (CHD) and implement targeted interventions for prevention and early management?
2. Which features in the dataset have the strongest correlations with the presence or absence of heart disease?
3. What features (biomarkers, clinical data) are the most important in predicting heart disease?
4. Are there significant differences in heart disease risk factors and prediction between males and females?
5. Is it possible to detect heart disease at an early stage, and what are the key indicators?

***More Problem Statements will be added as we proceed further with this Project.***

**Tools that will be used in Project.**

**Excel**: Data Understanding.

**SQL**: Data Cleaning & Data Preprocessing.

**Python**: Data Analysis & Machine Learning.

**Power BI**: Data Visualization.

First, we started with importing the Packages

A screenshot of a computer

Description automatically generated

Then We imported the dataset

A screenshot of a computer

Description automatically generated

Then to understand the Dataset we viewed the dimension of the dataset and its variables.

A screenshot of a computer

Description automatically generated

Then We started with the data cleaning and data Preprocessing steps. We checked the missing values first.

A screenshot of a computer

Description automatically generated

Then we visualized the missing values in the dataset.

A screenshot of a graph

Description automatically generated

Then We dealt with the missing values in the dataset

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generated

Then We checked the Duplicated values in the dataset

A screenshot of a computer

Description automatically generated

There was no duplicated values.

Then We started with Exploratory Data Analysis (EDA) of the Dataset.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA graph of different sizes and shapes

Description automatically generated with medium confidenceA graph of a number of blue bars

Description automatically generated with medium confidence

**Thank You**