Heart Disease Prediction

Data Analysis and Visualization Project Synopsis

Submitted By- Sanchit Gangwar
2K21/CS/96 (21013570089)

Course- BSc(Honours) Computer Science

Semester- V

Submitted to- Dr. Geetika Vashisht

College- College Of Vocational Studies(University Of Delhi)

Synopsis: Heart Disease Prediction

Problem Statement:

Heart diseases remain a significant global health concern, with an estimated 12 million deaths annually, as per the World Health Organization. In developed countries, like the United States, half of the deaths are attributed to cardiovascular diseases. The early identification of cardiovascular diseases can empower individuals to make informed lifestyle changes, potentially mitigating complications. This project aims to identify relevant risk factors for heart disease and predict overall risk using logistic regression.

Solution Approach:

The project utilizes a Logistic Regression Model for predicting Coronary Heart Disease (CHD) risk. Logistic Regression is a statistical and machine-learning technique that classifies records based on input field values, predicting a dependent variable (CHD risk) using one or more independent variables. It is suitable for both binary and multi-class classification.

Dataset:

The dataset, sourced from the Framingham Heart study and available on Kaggle, comprises over 4,000 records with 15 attributes. These attributes include demographic factors (gender, age, education), behavioral risk factors (smoking), medical history risk factors (blood pressure medication usage, prevalent stroke, prevalent hypertension, diabetes), and physical examination metrics (cholesterol level, blood pressure, BMI, heart rate, glucose level). The target variable is the 10-year risk of coronary heart disease (CHD).

Sanchit Gangwar, [30-11-2023 16:29]

METHODOLOGY

The project followed the following steps to accomplish the desired objectives and deliverables.

- **Obtain and review raw data :** in this section we upload a file which we will use to analyse the data.
- **Data preprocessing : we** do notice missing values using the info() method. our data preprocessing steps will do:
    - Remove columns not useful for our analysis.
    - count missing values

- **Dealing with missing values :** we can fill in the missing values with an average value. This process is called mean imputation.
- **Plot running data:** in this we plot the running data .

- **Detailed summary report :** With all this data cleaning, analysis, and visualization, we create detailed summary tables

Libraries Used:

1. Pandas: For data manipulation.
2. Numpy: For Computation
3. Matplotlib: For data visualization.
4. Seaborn: For enhanced data visualization.

Project Workflow:

1. Importing Libraries: Begin by importing the necessary libraries.
2. Importing and Reading Dataset: Load the Framingham Heart study dataset.
3. Exploratory Data Analysis (EDA): Explore and analyze the dataset to gain insights.
4. Data Preprocessing: Handle missing values, encode categorical variables, and perform any necessary data transformations.
5. Data Visualization: Utilize visualizations such as correlation matrices, pair plots, and count plots to understand the data distribution and relationships.

Conclusion:

The Heart Disease Prediction project provides a comprehensive analysis of the Framingham Heart study dataset, employing logistic regression to predict the 10-year risk of coronary heart disease. By identifying significant risk factors, the model contributes to early prognosis and informed decisionmaking for individuals at high risk, potentially reducing the incidence and complications of heart diseases.