Project Name: Risk Factors for Cardiovascular Heart Disease

1. Introduction

I used “Heart Data” dataset from Kaggle datasets. This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits, and alcohol consumption of over 70 thousand individuals. This white paper is intended to provide an overview of data mining on the Heart Data dataset, including data preparation, exploratory data analysis, modeling, and evaluation.

1. Dataset

The Heart Data dataset is available in CSV file and can be downloaded from Kaggle.

1. Data Preparation

The first step in data mining on the Heart Data dataset is to prepare the data. This involves cleaning and transforming the data to ensure that it is in a format that can be easily analyzed. Once the dataset has been downloaded, it is important to check for missing values, outliers, and inconsistencies. Any missing values should be imputed, and outliers should be removed or treated appropriately.

1. Exploratory Data Analysis

After the data has been prepared, the next step is to perform exploratory data analysis (EDA).

EDA involves visualizing and summarizing the data to gain insights into the relationships between the variables. For the Iris dataset, this may involve plotting the distribution of each variable, creating scatterplots to explore pairwise relationships, and calculating summary statistics for each variable. EDA can help identify any patterns or relationships that may exist in the data and guide the selection of appropriate modeling techniques.

1. Modeling

The next step is to select a modeling technique. I will use some algorithms for the Heart Data dataset, including Principal Component Analysis (PCA), Factor Analysis (FA), Kernel PCA, t-distributed stochastic neighbor embedding (t-SNE). The goal of modeling here is dimensional reduction on the Heart Data.

1. Evaluation

Model evaluation is an important step in data mining on the Iris dataset. This involves calculating performance metrics, such as accuracy, precision, recall, and F1 score, to determine how well the model performs. The metrics will depend on the modeling technique used, but in general, a higher accuracy indicates a better-performing model.