Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease

Group 2: Shwetha, Teshani, Komalpreet, Rachel

Objective: The goal of this study is to explore the Heart Disease Dataset and answer the following questions:

1. How many principal components are required to explain most of the variance in the dataset?
2. What are the underlying latent factors associated with the variables in the dataset?
3. How can categorical data in the study be represented and analyzed to identify patterns using Multiple Correspondence Analysis (MCA)?
4. How effective is Multivariate Discriminant Analysis (MDA) in classifying patients into heart disease and non-heart disease categories based on multiple clinical and demographic variables?
5. Which independent variables are the most significant predictors of heart disease, and how can stepwise regression techniques be applied to optimize the predictive model?

Population of interest and size of dataset:

The study population consists of 303 individuals with heart disease-related symptoms who have undergone diagnostic testing. Specifically, the population is made up of all patients from the Cleveland Database with information related to heart disease, including clinical and diagnostic data. It contains 13 medical attributes from these patients, such as age, sex, and results from various diagnostic tests, all related to heart disease prediction along with one target variable.

Define the variables you use in your project and provide the source of your dataset.

Source: https://archive.ics.uci.edu/dataset/45/heart+disease

In this project we are using the Cleveland Heart Disease dataset.

Quantitative Variables:

Variable Name	Description	Unit
Age	The patient's age	years
Resting Blood Pressure	patient's level of blood pressure at resting mode	mm/HG
Cholesterol (chol)	The serum cholesterol level	mg/dl
Maximum Heart Rate	measures the highest heart rate reached during physical exertion	
Oldpeak	Exercise induced ST-depression in relative with the state of rest	

Qualitative Variables:

Variable Name	Description
Chest Pain Type (cp)	Type of chest pain experienced by patient
Resting Electrocardiographic Results (restecg)	Measures heart function during rest
Exercise-Induced Angina (exang)	Measures whether physical exertion induces chest pain
Major Vessels (ca)	Counts the number of major vessels (0–3)
Thalassemia (thal)	A blood disorder measured
Target	Classify individuals as having heart disease or not.