Cardiovascular Disease Prediction

Brief Project Report

Hung-Jui Chen
Team Leader
CHEN1391@e.ntu.edu.sg

Nov 2021

Contents

1 Introduction

Cardiovascular diseases (CVDs) are among the leading causes of mortality worldwide. Detecting individuals at risk through early, data-driven interventions can significantly reduce health complications. This project aims to predict cardiovascular disease using the 2018 Kaggle Cardiovascular Disease Dataset, focusing on features such as blood pressure, cholesterol levels, weight, and lifestyle indicators.

2 Dataset Overview

2.1 Data Source

The dataset used in this project originates from the Cardiovascular Disease Kaggle Dataset (2018). It contains 70,000 records of individuals with multiple health-related features.

2.2 Features

The dataset includes the following variables:

- **id**: Unique patient identifier.
- **age**: Age in days.
- **gender**: 1 = Female, 2 = Male.
- **height**: Height in centimeters.
- **weight**: Weight in kilograms.
- **ap_hi**: Systolic blood pressure.
- **ap_lo**: Diastolic blood pressure.
- **cholesterol**: Cholesterol levels (1: normal, 2: above normal, 3: well above normal).
- **gluc**: Glucose levels (1: normal, 2: above normal, 3: well above normal).
- **smoke**: Whether the individual smokes (0 = No, 1 = Yes).
- **alco**: Alcohol consumption (0 = No, 1 = Yes).
- **active**: Physical activity (0 = No, 1 = Yes).
- **cardio**: Cardiovascular disease indicator (0 = No, 1 = Yes).

3 Methodology

3.1 Data Preprocessing

- **Cleaning and Validation**: Checked for missing or invalid values and addressed any identified outliers.
- **Feature Engineering**: Encoded categorical variables (e.g., gender, cholesterol, gluc) and performed normalization or standardization as needed.
- **Train-Test Split**: Partitioned the dataset into training and testing subsets to evaluate model performance on unseen data.

3.2 Exploratory Data Analysis

- Analyzed distributions of key variables such as blood pressure, weight, and cholesterol.
- Created correlation plots to identify potential relationships among features.
- Observed that **systolic blood pressure (ap_hi)** appeared to be a key factor in predicting CVD.

3.3 Machine Learning Models

Decision Tree Classifier

- Built a classification tree to predict whether an individual has CVD.
- Used Gini impurity or entropy for splitting.
- Examined feature importance and discovered that systolic blood pressure played a central role.

Random Forest Classifier

- Trained an ensemble of decision trees to improve model generalization.
- Tuned hyperparameters such as the number of estimators and maximum depth.
- Achieved higher accuracy and better performance compared to a single decision tree.

4 Results and Discussion

- **Model Performance**: The Random Forest model outperformed the Decision Tree in terms of accuracy and precision.
- **Key Predictor**: Systolic blood pressure (`ap_hi`) was consistently identified as the most influential feature.
- **Insights**: Lifestyle factors like smoking, alcohol consumption, and physical activity also showed correlation with the presence of CVD.

5 Conclusions and Future Work

5.1 Conclusions

- The use of ensemble methods (Random Forest) significantly improved predictive performance.
- Systolic blood pressure emerged as a critical variable for early detection of cardiovascular disease.

5.2 Future Improvements

- **Additional Models**: Explore Support Vector Machines, Gradient Boosting, or Deep Learning architectures.
- **Feature Engineering**: Incorporate additional health metrics or demographic data.
- **Deployment**: Develop a web application or API for real-time CVD risk assessment.

6 References

- **Kaggle Dataset**: `https://www.kaggle.com/sulianova/cardiovascular-disease-dataset`
- Standard textbooks and research articles on cardiovascular disease prediction and machine learning.