Project Report: Exploratory Data Analysis and Machine Learning for Cardiovascular Disease Prediction

Introduction

The objective of this project is to analyze the "healthcare.csv" dataset provided by Simplilearn to predict the occurrence of cardiovascular disease (CVD) in individuals. This dataset contains information about various health factors for a group of people, with a binary classification target variable (0 = Negative, 1 = Positive) indicating the presence or absence of CVD. In this report, we will outline the key findings and insights obtained during the data exploration and analysis phase, as well as the performance of machine learning models in predicting CVD.

Data Cleaning

Data Integrity: The dataset is mostly categorical, consisting of integer and float data types. It was found to be clean, with no missing values. One duplicate row was identified and removed to ensure data integrity.

Preliminary Data Analysis

Descriptive Statistics: Preliminary statistics did not reveal any standout insights.

Data Overview

Gender Distribution: The dataset comprises approximately 100 females and 200 males.
Fasting Blood Sugar*:* Fewer than 50 patients had fasting blood sugar levels below 120 mg/dl.

Age and CVD

Age Distribution: Positive CVD cases were observed predominantly among individuals aged between 42 to 58 years.

Blood Pressure and CVD

Resting Blood Pressure (trestbps): Analysis of resting blood pressure showed that:
- Normal Blood Pressure (<120 mmHg): 61.7% tested positive for CVD out of 60 individuals.
- Elevated Blood Pressure (120-139 mmHg): 57.2% tested positive for CVD out of 145 individuals.
- High Blood Pressure (>=140 mmHg): 47.6% tested positive for CVD out of 84 individuals.

Cholesterol and CVD

Cholesterol Levels(chol): Cholesterol levels in the dataset were categorized as:
- Normal (<200 mg/dl): 59.2% tested positive for CVD out of 49 individuals.
- Borderline High (200-239 mg/dl): 60.2% tested positive for CVD out of 98 individuals.
- High (>239 mg/dl): 48.7% tested positive for CVD out of 150 individuals.

Maximum Heart Rate and CVD

Maximum Heart Rate (thalach): In all positive CVD cases, most individuals achieved maximum heart rate from 150 to 175. While in negative cases most individuals achieved maximum heart rate from 135 to 165.

Machine Learning Models

Logistic Regression Model: The logistic regression model performed well when trained on 80% of the data. It achieved an accuracy of 87% on the training data and 80% on the test data. Furthermore, it correctly identified 76% of all positive CVD cases.
Random Forest Model: The random forest model outperformed logistic regression when trained on 80% of the data. With specific hyperparameters (max_depth = 3, criterion='entropy', n_estimators = 65), it achieved an accuracy of 89% on the training data and 84% on the test data. Model performance decreased when the criterion was changed to "entropy." It also identified 81% of all positive CVD cases.
Logistic Regression with Statsmodels: Logistic regression was also performed using Statsmodels, revealing features with p-values less than 0.05 as significant predictors of CVD. The significant features include 'sex', 'cp' (chest pain type), 'thalach'

(maximum heart rate achieved), 'exang' (exercise-induced angina), 'oldpeak', 'ca' (number of major vessels colored by fluoroscopy), and 'thal' (thalassemia).

Conclusion

In this project, we conducted a comprehensive exploratory data analysis (EDA) of the "healthcare.csv" dataset to understand the relationships between various health factors and the occurrence of cardiovascular disease. We also built and evaluated machine learning models, including logistic regression and random forests, to predict CVD. The random forest model with specific hyperparameters showed the best performance.

This analysis provides valuable insights for healthcare professionals and researchers, aiding in the identification of significant risk factors for cardiovascular disease. Further research and refinement of predictive models could contribute to more accurate CVD risk assessment and prevention strategies.

References

https://www.cedars-sinai.org/health-library/diseases-and-conditions/h/high-blood-pressure-hypertension.html#:~:text=Normal%20blood%20pressure%20is%20systolic,diastolic%20is%2080%20to%2089.
https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel#:~:text=Here%20are%20the%20ranges%20for,or%20above%20240%20mg%2FdL
https://www.healthline.com/health/RPE#scale-comparison