CS-433 Machine Learning: Project 1

Ali Elkilesly, Selim Sherif, Roy Turk

Abstract

Heart disease remains one of the leading causes of death worldwide, presenting a persistent challenge to healthcare systems. As its prevalence rises, the importance of early detection and risk assessment grows. Recent advancements in machine learning are revolutionizing predictive medicine by allowing healthcare professionals to identify at-risk individuals with unprecedented accuracy. This report presents a machine learning approach to developing a classification model that predicts coronary heart disease susceptibility using a dataset of diverse health and lifestyle features.

1 Introduction

This report aims to outline the comprehensive steps taken in developing a machine learning model to predict susceptibility to coronary heart disease. We begin with a detailed data analysis, scrutinizing the dataset for quality and integrity. Following this, data cleaning procedures are implemented to address any inconsistencies, missing values, or anomalies that may impact model performance.

Next, the focus shifts to feature engineering, a crucial phase where raw data is transformed into meaningful inputs for the model. The report then delves into model building, where an algorithm is carefully selected to determine the best fit for our problem. Finally, the implementation phase details the deployment of the chosen model and the evaluation of its effectiveness.

2 Data analysis and feature engineering

The dataset from the Behavioral Risk Factor Surveillance System (BRFSS) contains 321 features. Initially, an examination of the dataset is essential to gain an understanding of the organization of the features, their types, and their significance. This lead to the following pre-processing.

2.1 Missing values

A substantial number of features in the dataset contain a high proportion of missing values. To address this, a custom function is developed to exclude features with over 60% missing values. This threshold improves data quality by removing features that contribute insufficient information, thus enhancing model reliability.

2.2 Insignificant features

To improve dataset relevance for modeling, the variance of each feature is computed, and a threshold is applied to eliminate those with low variance. Features with minimal variance remain nearly constant across samples, contributing little to the differentiation between instances. Removing these low-variance features reduces redundancy, allowing the dataset to focus on variables with greater potential to provide meaningful insights into coronary heart disease prediction.

2.3 Mean difference and ratio of standard deviation

To enhance feature relevance, a mean difference technique is applied by calculating the mean of each feature with respect to the two target classes $\{-1, 1\}$. Features with an absolute mean difference below a specified threshold are eliminated, as they contribute to minimal discriminatory power between classes.

For each feature X, the mean difference D is given by

$$D = |\mu_1 - \mu_{-1}| \tag{1}$$

where μ_1 and μ_{-1} denote the mean values of X for each target group.

In an analogous way, features are eliminated using the ratio of standard deviation technique to assess their discriminatory power relative to the binary target variable.

For each feature X, the ratio of standard deviation R is defined as

$$R = \frac{\sigma_{-1}}{\sigma_1} \tag{2}$$

where σ_{-1} and σ_1 denote the standard deviations of feature X for each target group.

2.4 Feature types

Our dataset comprises three distinct types of features: continuous, categorical, and ordinal. Each type requires different preprocessing methods, prompting us to manually classify the features accordingly.

2.4.1 Continuous features

The continuous variables in the dataset include missing values and specific codes indicating refusals or uncertainties. To address this, these entries are replaced with the mean of the respective feature. Following this imputation, we apply Z-score normalization, defined by the formula

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

where Z is the Z-score, X is the original value, μ is the feature mean, and σ is the standard deviation.

The normalization standardizes the features to a common scale with a mean of 0 and a standard deviation of 1. This process is essential for ensuring that all features contribute equally to model performance, preventing those with larger ranges from disproportionately influencing results, and improving the convergence of optimization algorithms in machine learning.

An additional step involves computing the correlation matrix of the features to identify and eliminate continuous features that are interdependent. Features exhibiting high correlation (above a specified threshold) are removed to reduce redundancy.

2.4.2 Categorical features

In an analogous way, the values containing specific codes or missing values are replaced with a new category of 0.

To eliminate interdependent categorical features, Cramér's V is utilized to measure the strength of association between pairs of features. This statistic ranges from 0 (no association) to 1 (perfect association). By calculating Cramér's V for each pair, features with high correlation, indicated by values above a specified threshold, can be identified. One of these redundant features is then removed, ensuring the dataset retains only the most informative categorical variables for modeling.

Cramér's V is defined as

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\left(k_r - 1, k_c - 1\right)}} \tag{4}$$

where χ^2 is the chi-squared statistic, n is the total number of observations, k_r and k_c are the number of categories in each variable.

One-hot encoding is then applied to convert the numerical categories into a suitable format for the model. This process involves creating binary columns for each unique numerical category, where a value of 1 indicates membership in that category and 0 indicates non-membership.

2.4.3 Ordinal features

Coded values and blanks in the ordinal features are replaced with 0 as well. Additionally, the inter-dependencies among the features are assessed in a manner similar to that used for continuous features, allowing us to identify and eliminate redundant variables from the dataset.

3 Model building

The objective of this project is to predict the likelihood of an individual having coronary heart disease, framing it as a binary classification problem. The loss function that has to be minimized is the negative-log likelihood:

$$L(w) = \frac{1}{N} \sum_{n=1}^{N} \left[-y_n x_n^T w + \log\left(1 + e^{x_n^T w}\right) \right] \tag{5}$$

where N is the total number of observations, y_n is the binary label for the n-th observation, x_n is the feature vector for the n-th observation, and w is the weight vector.

To utilize the function correctly, the labels vector is converted from the $\{-1, 1\}$ format to the $\{0, 1\}$ format.

A regularization term $\lambda \|w\|^2$ can be added to the function to prevent overfitting by penalizing large weights. This approach helps ensure that the model generalizes well to unseen data by discouraging overly complex models that might fit the training data too closely. By promoting simpler weight configurations, regularization enhances the model's ability to capture essential patterns without being misled by noise or outliers in the training set.

The function is then optimized using gradient descent following the formula

$$w^{(t+1)} = w^{(t)} - \gamma \nabla L(w^{(t)}) \tag{6}$$

where $\nabla L(w^{(t)})$ is the gradient of the loss function, and γ is the step-size.

4 Results

In this project, predictions are evaluated using the AI Crowd competition arena, where performance metrics include both the F1 score and accuracy. The F1 score combines precision and recall to provide a balanced measure of model performance, particularly valuable in cases with class imbalance. It ranges from 0 to 1, with higher values indicating better model performance in distinguishing between the two classes.

$$F1 = \frac{TP}{TP + \frac{1}{2}\left(FP + FN\right)} \tag{7}$$

Accuracy, on the other hand, measures the proportion of correct predictions among all predictions, reflecting the overall effectiveness of the model in classifying patients at risk for coronary heart disease versus those who are not. These metrics together offer a comprehensive view of the model's predictive quality.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Logistic regression	
F1 score	Accuracy
0.439	0.875