Machine Learning Approaches for Coronary Heart Disease Prediction

Salim Ameziane, Charaf Eddine Dahbi
EPFL

***Abstract*—This report presents the preprocessing and feature engineering pipeline designed for the EPFL CS-433 Project 1, which aims to predict Coronary Heart Disease (CHD) using the 2015 BRFSS dataset. We implemented a systematic data cleaning and ..**

I. Introduction

The aim of this project [1] is to predict the likelihood of coronary heart disease (CHD) based on individual health data, undertaken as part of the Machine Learning course (CS-433) at EPFL [2]. We carried out data cleaning and feature engineering, followed by binary classification using models tuned for accuracy and F1-score. Beyond achieving predictive accuracy, this study sheds light on the potential of machine learning for CHD risk assessment, offering healthcare professionals tools for early identification of at-risk patients.

II. MODELS AND METHODS

A. Implementing Machine Learning Algorithms

To meet project requirements, we implemented six machine learning algorithms: Linear Regression using Gradient Descent (GD), Least Squares and Ridge Regression, and both standard and regularized Logistic Regression using GD.

B. Exploratory Data Analysis

To understand the dataset, we conducted an exploratory data analysis, assessing feature structure, quality, and distribution.

leftmargin=1.5em

- The dataset includes 321 features, with 328,136 training samples and 109,380 testing samples.
- Features are a mix of continuous and categorical types, depending on the survey questions.
- Some features showed inherent correlations, as survey questions were conditionally asked based on prior responses.
- About half of the features had over 50% missing values (cf. Figure 2).

After reviewing the data's codebook [**?**], we chose automated feature selection over manual methods to avoid bias and redundancy, applying a consistent approach for categorical and continuous features.

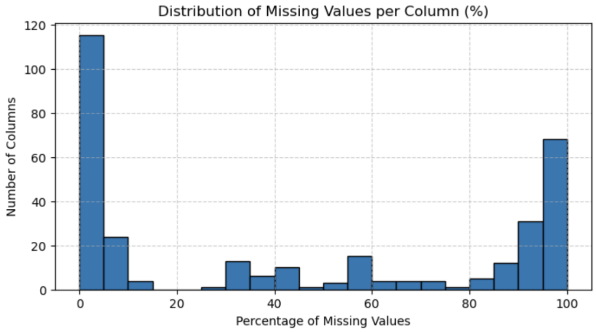

Fig. 1. Distribution of missing values across features

III. Preprocessing Pipeline

1) **Removing Features with Excessive Missing Values:** Columns with over 50% missing values were removed, keeping only those with at least 50% filled entries. This reduced the feature count from 321 to 175, retaining informative columns.
2) **Imputing :** Filling missing values with column mean.
3) **Removing zero-Variance Features:** Zero–variance features were removed. This likely reflects features concentrated within specific classes.
4) **Remove multicellularity:** Remove highly correlated features, higher than 0.5.
5) **Keep relevant features:** Remove features with low correlation with target, less than 0,1.
6) **Standardization:** Adjusted features to a mean of 0 and a standard deviation of 1, balancing differences in scale. We standarize train set and test set independently.

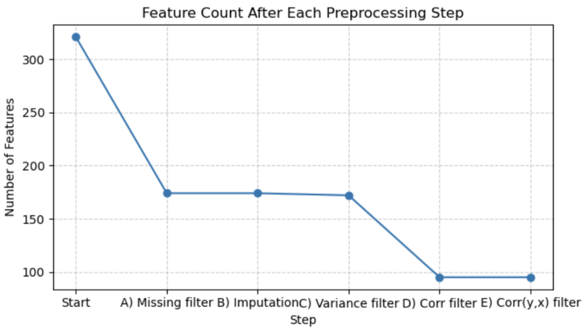

Fig. 2. Distribution of missing values across features

IV. Model process

A. Cross-Validation

A 5-fold cross-validation scheme was implemented to ensure robust model evaluation. The dataset was randomly split into 5 folds, ensuring reproducibility through a fixed seed. At each iteration, one fold was used for testing while the others formed the training set. The cross validation function trained and evaluated models across folds, averaging losses to estimate generalization performance.

B. Parameter Tuning

A grid search procedure was implemented to optimize model hyperparameters. For each regression method, regularization strengths (λ), was systematically tested. Each configuration was evaluated on the training set using accuracy and F1 score as performance metrics. The obtained results are summarized in Table I The best model parameters were selected based on the highest F1 score, balancing predictive accuracy and robustness across methods.

V. result

TABLE I
Optimal Thresholds and Best Hyperparameters for Models

Model	**Optimal Threshold**	**Best Hyperparameters**	
		λ	γ
Least Squares	-0.57	—	—
Ridge Regression	-0.57	1e-09	—
Logistic Regression	-	—	0.01
Regul. Logistic Regression	-	0.01	0.01

TABLE II
Model Performance

Model	**CV- Tests**	
	Accuracy	**F1 Score**
Least Squares	0.8628	0.404
Ridge Regression	0.8628	0.404
Logistic Regression	0.8513	0.3861
Regul. Logistic Regression	0.8451	0.3762

References