Early detection of cardiovascular diseases using machine learning

Nino Avetikovi, François Mendiburu, Malo Olszewski
Machine Learning course, EPFL, Switzerland

***Abstract*—Throughout our first project of this course, we chose to use some basic methods of machine learning to develop a model which enables to detect cardiovascular diseases from numerous factors. We adressed this subject by using penalized logistic regression and we tried to find the best hyperparameters in order to maximize both the F1-score and the accuracy of the model.**

I. Introduction

As per the World Health Organization, Cardiovascular Diseases (CVD), including heart attacks, are increasingly prevalent as a leading cause of global mortality. As the adult population ages, heart and circulatory conditions are on the rise, necessitating advanced technologies like machine learning for early detection and preventive measures against CVDs. This project focuses on leveraging machine learning techniques to assess an individual's risk of developing CVD based on their lifestyle factors. The dataset used in this study is sourced from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey collecting data on U.S. residents' health behaviors, chronic conditions, and preventive service usage. In this survey, respondents were classified as having coronary heart disease (MICHD) if diagnosed by a healthcare provider or if they reported having experienced a heart attack or angina. The primary objective of the project is to develop a predictive model using BRFSS dataset, aimed at estimating the probability of developing MICHD based on clinical and lifestyle factors. The model will be tasked with analyzing a set of health-related features and determining the likelihood of an individual developing MICHD. The following report presents the different techniques used to perform the classification of healthy individuals and individuals at risk of developping cardiovascular deseases. This approach will ultimatly enable to perform a predictive analysis focused on early disease detection and prevention strategies.

II. Models and Methods

A. Dataset description

The dataset under study comprises two files: xtrain.csv containing various features and ytrain.csv containing corresponding labels. The labels indicate whether individuals were classified as healthy (0) or not (1). The dataset's size is substantial, containing 320 features and 328,125 participants.

The features within this dataset are categorized into three types:

- Variables for Data Stratification and Weighting
- Intermediate Variables: These variables were derived from survey question responses and were used in calculating other variables or risk factors.
- Variables for Respondent Categorization or Classification: These variables, mostly starting with an underscore, help categorize respondents. Some calculated variables group continuous variables (like weight, age, or body mass index) into categories, while others simplify analyses by regrouping non-continuous variables. These variables predominantly focus on health behaviors associated with a risk of illness or injury.

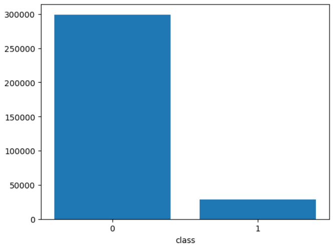

Fig. 1. Distribution of labels

For our classification task, we made a deliberate choice to exclusively utilize the variables falling under the third category (classification variables). These variables are primarily concerned with health behaviors associated with the risk of illness or injury, aligning closely with our goal of identifying individuals with coronary heart disease.

Further analysis of the dataset showed imbalances in data distribution across the two calsses

The following plot illustrates the disparity in participant distribution across the two labels: label 1, denoting individuals exhibiting a risk of developing MICHD, and label 0, representing participants categorized as not being at risk.

Indeed only 8.8% of the dataset corresponds to participants at risk (label 1).

B. Preprocessing the data

After exploring the data, a structured data preprocessing pipeline was established to refine the dataset, essential for subsequent analysis and model development.

At first the missing values were handled by replacing by the mean of each column (each feature), this enabled us to conserve the overall mean of the dataset.

To address the issue of missing values, a pragmatic approach was implemented. Missing values, a common occurrence in a dataset of this magnitude, were imputed using a simple yet effective method—replacing them with the arithmetic mean of their respective columns. This allowed us to preserve the central tendency of the dataset, maintaining the overall statistical properties of the dataset, making it more representative of the actual data distribution.

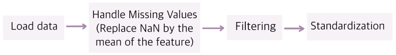
Load data
Handle Missing Values (Replace NaN by the mean of the feature)
Filtering
Standardization

Fig. 2. Data preprocessing pipeline

A crucial subsequent step in this pipeline was data filtering. With a plethora of features (320) available, feature selection techniques were employed to isolate the most pertinent features for the classification task. The aim was to alleviate the curse of dimensionality and focus on features directly relevant to cardiovascular diseases. Specifically, features related to participants' health factors—such as physical activity, dietary habits,access to healthcare and so on—were retained. This strategic filtering process aimed to refine the dataset for optimal model performance.

Standardization of the data followed the feature selection process. This step ensured that all features were on a uniform scale, mitigating the dominance of certain features during model training due to differences in scale. By standardizing the data, the aim was to foster fair and accurate model performance across all features, contributing to a more robust model for the classification task.

C. Training methodologies

To train a classification model with this dataset, we chose to go for the logistic regression algorithm using the negative log likelihood cost function :

$$\mathrm{w}_* = argmin\ L(w) := \frac{1}{N}\sum_{n=1}^{N} -y_n x_n^T w + \log(1+e^{x_n^T w})$$

The logistic regression solution allows us to represent straightforwardly the two possible outcomes (i.e 0 for no heart disease and 1 for heart disease) and to use an adaptative threshold to fit the data correctly. The negative log likelihood cost function is robust to outliers, which seems interesting in a high dimensional dataset.

D. Validation of the model

We are facing a classification task, so we considered logistic regression to tackle the dataset. We decided to optimize the model by tuning the hyperparameters which are lambda, the degree of the feature expansion polynomial and the threshold of the regression task. In fact, we first saw that the model had a very low F1-score with the usual threshold of 0.5, because the dataset is unbalanced. So we tried to balance it and we went for regularization to avoid overfitting and feature expansion to increase precision by reducing dimensionality and capturing complex relationships in the data.

Using cross validation with 4 folds allowed us to determine what were the best lambda-degree combination.

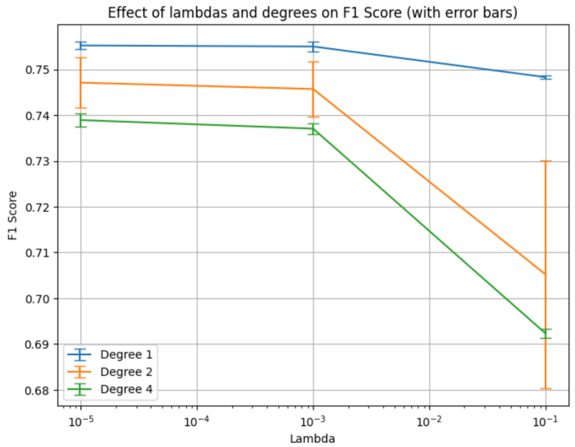

Fig. 3. Plot of F1-score vs lambda and degree

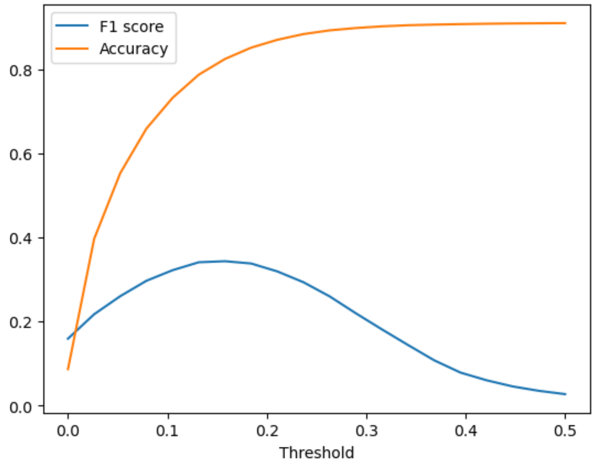

Fig. 4. Plot of the impact of threshold on F1-score and accuracy

By not expanding the training data and not using a regularization factor $\lambda = 0$, we found the highest F1-score. in fact, using regularization gives us poor results in both accuracy and F1-score. Then we looked for the best threshold which takes into account the unbalanced data we have (Fig. 4) and obtain an optimal around 0.14.

III. Model Evaluation

To evaluate our model, one can use 2 different metrics to characterize its efficiency. Those quantities being the accuracy and the F1 score, the last one emphasising the balance between a balanced and a robust model. We saw that an important lambda tends to lower the F1-score and the accuracy.

We got a F1-score of 0.376, which is not optimal, but it can be justify because of the recall that can be very high due to the unbalanced data and a sufficient accuracy of 0.847.

To conclude, the dataset is pretty hard to handle, and we can see the limitations of our simplistic approach in such a high dimensional problem.