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 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 primary objective of the project is to develop a predictive model using, aimed at estimating the probability of developing coronary heart disease 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 coronary heart disease. The following report presents the different techniques used to perform the classification of healthy individuals and individuals at risk of developing cardiovascular deseases.

II. Models and Methods

A. Dataset description

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 if diagnosed by a healthcare provider or if they reported having experienced a heart attack or angina.

The dataset used for training comprises the answers of 328,125 participants, characterized by 320 features. The label vector indicate whether individuals were classified as healthy (0) or not (1)

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, allowing respondent categorization. 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.

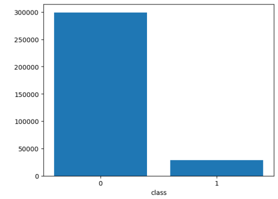

Fig. 1. Distribution of labels

For our classification task, we made a the 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.

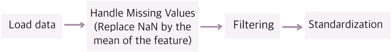
Load data
Handle Missing Values (Replace NaN by the mean of the feature)
Filtering
Standardization

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. This strategic filtering process aimed to refine the dataset for optimal model performance.

The following step, regularization, 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 perform the binary classification of the data, logistic regression algorithm was implemented which models the probability for datapoint to belong to a particular class. This is achieved by the use of the sigmoid functions which maps the model output between 0 and 1. By default, the threshold allowing us to decision about the classification of a datapoint was set to 0.5, meaning that if the probability is greater than or equal to 0.5, the model will classify the participant under the label 1: at risk of developping cardiovascular desease, while if it's less than 0.5, it's classified into the other class. In order to maximize the probability to classify the data correclty, we minimized the negative log likelihood cost function :

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

Therefore by minimizing the logistic loss, the probability of misclassify a datapoint is also minimized. This optimization was tackled using gradient descent algorithm.

$$\mathrm{w}^{(t+1)} := w^{(t)} - \gamma \nabla_w L(w)$$

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.

III. Validation of the Model and Results

A. Evaluation metrics

The evaluation metric that was used to perform cross validation was the F1 score, and allow us to quantize the trade off between recall and precision, which emphasise respectively the rate of false negatives and false positives. Maximizing it aims to develop a robust model with a high rate of true positives and true negatives (i.e a high accuracy) which is the second metric use to evaluate our model.

B. Cross validation

In order to improve the model, cross validation technique was used. The goal of cross-validation is to evaluate the performance and generalizability of a machine learning model. It is a resampling technique used to assess how well a model will perform on new, unseen data. We decided to optimize the model by tuning the hyperparameters which are the regularization cofficient lambda, which prevents the model to be too complex and therefore overfit the data, and the effect of degree of the feature expansion polynomial on model performance was also evaluated, capturing complex relationships in the data. To perform cross validation the data was balanced using undersampling technique: random samples were selected from the majority class. The threshold was kept to 0.5. The Fig 3 represents the obtained results.

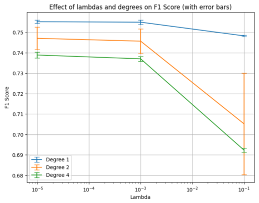

Fig. 3. Plot of F1-score vs lambda and degree

The figure shows that the feature expansion and model regularization decreases the model performances. In fact, using regularization gives us poor results in both accuracy and F1-score meaning that the model doesn't overfit the data.

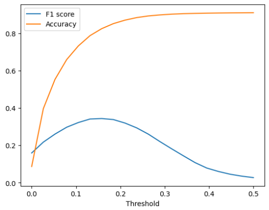

Fig. 4. Plot of the impact of threshold on F1-score and accuracy

Then we analysed the effect of the threshold on the model performances with the original data. This allows us to account for the data imbalanced without having to downsample the data, as more data means better model performances. The accuracy and f1 score ware plotted on (Fig. 4) as a function of different threshold. We obtained an optimal of around 0.14.

IV. Conclusion

The implemented techniques finally led us to generate a model with an F1-score of 0.376, which is not optimal, but it can be justified by high recall value due to the unbalanced data and an accuracy of 0.847. In a nutshell, the model performances were acceptable in the context of the project, however it is limited by our simplistic approach and further improvements are necessary.