Project 1 - A Machine Learning Analysis of MICHD Predictions

Mohamed Aziz Dhouib, Hyeongdon Moon and Ahmed Reda Seghrouchni
CS-433: Machine Learning
October 30, 2023

***Abstract*—Cardiovascular Diseases (CVD), notably Myocardial Ischemic Coronary Heart Disease (MICHD), pose a growing global health risk, especially among the expanding older adult population. With advancements in technology, machine learning algorithms stand out as potential tools for early detection and prevention of these diseases. This study utilizes the Behavioral Risk Factor Surveillance System (BRFSS) data, containing 321 distinctive features, to assess the risk of developing MICHD based on individual lifestyle factors. The research is segmented into three focal areas: data preprocessing, comparison of six foundational machine learning methodologies, and hyperparameter searching. Following a meticulous analysis and optimization process, our final model achieves an accuracy of 86.3% and an F1 score of 38.5, showcasing the potential of machine learning in predicting and combating the onset of CVDs.**

I. Introduction

Cardiovascular Diseases (CVD) emerge as a significant global health challenge. As older adult populations grow, so does the prevalence of heart and circulatory vessel diseases. While the problem is evident, leveraging new technological advancements presents a promising solution. Specifically, machine learning algorithms, celebrated for their pattern recognition capabilities, offer potential in the early detection and prevention of developing CVDs. Using data from the Behavioral Risk Factor Surveillance System (BRFSS), this report delves into the application of various machine learning techniques to gauge the risk of developing Myocardial Ischemic Coronary Heart Disease (MICHD) based on individual lifestyle factors.

Given the intricate nature of the provided dataset comprising 321 features, this investigation revolves around three primary sections: data preprocessing, a comparative analysis of six foundational ML methodologies, and hyperparameter searching. Each section plays a crucial role in refining the accuracy of the prediction model. By exploring various preprocessing methods and comparing different machine learning techniques, the goal is to optimize the model for the highest accuracy and precision in predicting the likelihood of an individual developing MICHD. Through methodical exploration, research, and analysis, this report provides a comprehensive overview of our approach and findings in developing a predictive model for MICHD. The given data was observed and cleaned. Training and testing data sets were used to implement an algorithm and optimize the results. The competition arena on AIcrowd was used as a platform to evaluate the performance of our model, through accuracy and F1 score.

II. Data Analysis

A. Data Exploration & Preprocessing

The training data set is composed of a table of 321 columns, representing the features, and 328'135 rows, representing the different data points.
Following our exploration of the given data set, we have noticed that several features have skewed distribution. Skewed data can make it difficult to apply certain statistical techniques or machine learning algorithms that assume a normal distribution. We therefore applied min-max clipping, as well as log transform to skewed feature values. In order to do that, we calculated skewness for each feature, and if |skewness|>1, we apply log transform for that column. Skewness s is defined as

$$s = \frac{\sum_i^N (X_i - \bar{X})^3}{N\sigma^3}$$

where $\bar{X}$ is mean of the feature and σ is standard distribution.

Next, we have noticed the presence of multiple outliers in our data, which are probably due to experimental errors: multiple features have their values distributed with high variance, which suggests the need to standardize our data before training.

To compare the effectiveness of preprocessing methods, we run linear regression with Mean Absolute Error on four preprocessing settings.

- baseline: raw train data
- scaled: apply min-max scaling within 0,1 to all features
- standardized: apply standardization to all features
- log transform: apply log transform to the features whose skewness is greater than 1.

We run 5-fold cross validation for four data preprocessing mode. As shown in 1, standardization shows stable and highest f1 score, so we choose standardization as our preprocessing logic.

III. Models and Results

We trained a machine learning model on our training set and used the results to compute predictions for the test set and finally merge all the predictions. In total, we have tried

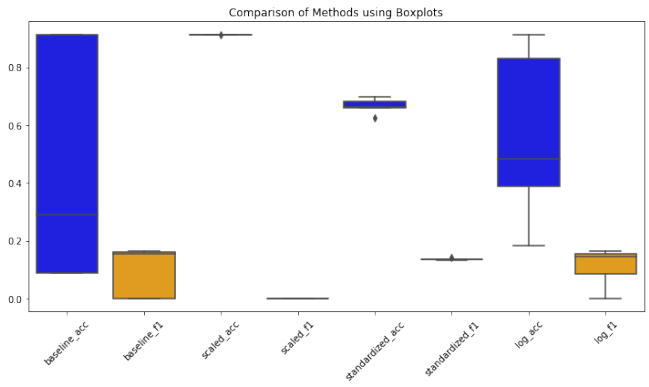

Figure 1: Comparison of Preprocessing methods

six machine learning models using 5-fold cross validation to validate the hyperparameters:

- degree : Degree of polynomial expansion
- thresh : Threshhold of the binary classification
- lambda : Regularization coefficient

We computed the accuracy and F-score of the predictions using different variations of the data set. Results for the ”final” data are summarized in the following Tables, where the F-score is rounded up to two decimal places the accuracy is rounded up to no decimal place.

Model	Final Data F-score
Linear Regression	0.39
Mean Squared Error, GD	0.12
Mean Squared Error, SGD	0.14
Ridge Regression	0.41

Table I: Used models and their respective F-score

Model	Final Data Accuracy
Linear Regression	86%
Mean Squared Error, GD	60%
Mean Squared Error, SGD	60%
Ridge Regression	70%

Table II: Used models and their respective Accuracy

IV. Discussion

Linear Regression was the model which gave us the best performance, with an accuracy of 86.3%. We also observed that our models are not over fitting on the training set since we get a very similar train and test accuracy, which explains why regularization did not significantly improve our predictions for this data set.

Due to the imbalance of the data label, Ridge regression shows both the best F-score and poor accuracy. In order to handle label imbalance, we adjusts the classification boundary, but performance could be higher if we applied other methods that address this problem.

V. Summary

This project was a good opportunity to have a more practical understanding of the machine learning methods discussed throughout the lectures and to compare the performance of different models applied to our data set. It was also the occasion to work on raw, real-world data, and see the importance and effects of pre-processing and cleaning, which required a deep dive into the different features of the data set. This highlights the importance of making a good data analysis and having a good understanding of the set overall. In addition to developing the six models, we tried to optimize the results in order to get the best prediction, and we were able to reach 86.3% of accuracy with our best model.