Machine Learning Methods for Predicting MICHD

Huiyun Zhu, Junjia Kang, Yuekai Yan
CS-433, EPFL
October 30, 2024

***Abstract*—Cardiovascular Diseases (CVD) has become prevailing in modern society with growing population of older people, among which coronary heart disease (MICHD) is considered dominate. Therefore, prediction, prevention and treatment of MICHD based on individual's health-related risk behaviors as well as chronic health conditions requires the development of efficient technologies. With the related MICHD data from the Behavioral Risk Factor Surveillance System (BRFSS) in America, we apply five machine learning algorithms to forecast the likelihood of developing MICHD for people with different lifestyle and clinical situations.**

I. INTRODUCTION

As a significant global health challenge, Cardiovascular Diseases (CVD) account for cause of death especially regarding old people. While the situation becomes severer along with the development of the age, the appropriate application of emerging technologies provides promising coping measures with reliable results. For example, the binary classification techniques involved in machine learning methodologies could help predict the probability of people developing coronary heart disease (MICHD) upon comprehensive analysis of available health-related data.

Using data from the Behavioral Risk Factor Surveillance System (BRFSS) which aims to prevent risk behaviors and health problems in the United States[1], we implement and compare various machine learning techiniques to gauge the tendency of people developing MICHD considering personal lifestyle factors. In the following sections, corresponding data preprocessing and model establishment are illustrated in detail in order to refine the accuracy of prediction and discussion is thereby drawn accordingly for public review.

II. DATA PREPROCESSING

Given the complex nature of the provided dataset involving 321 features [2], we conduct selected data preprocessing techniques to decrease data dimensions so as to minimize the possible interference caused by irregular data in the raw training data. Corresponding features of the test dataset are also restrained for further model prediction.

A. Data Selection and Substitution

In order to fully utilize the information contained in the dataset, strictly unnecessary variables in relation to the questionnaire itself (e.g. 'IDATE', 'IMONTH', etc.) are eliminated. Features like eye care health insurance ('VIINSUR2') which may imply health consciousness and family financial situation are retained for this stage. Considering the large amount of data, we then remove features with more than 10% missing values (e.g. 'INSULIN', 'FEETCHK2', etc.) for the sake of accuracy. For vectors with less than 10% missing values, we fill the blank with the relevant rounded average value. In the end, there are 119 features left and they are categorized into two types: continuous and discrete.

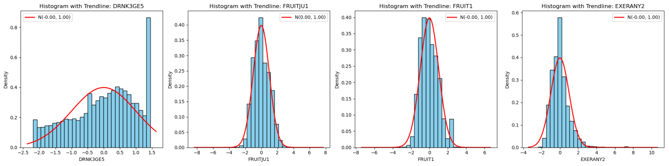

Figure 1: Distribution of selected features after standardization

B. Normalization

Since feature values have different scales and units, input eigenvectors are **normalized** for all continuous variables to ensure a balanced contribution of eigenvalues to the model and to speed up the convergence of the optimization algorithm. Standardization centers the data at 0 with a standard deviation of 1, thus preserving the shape of the original distribution and providing certain robustness. Fig. 1 shows the distribution of selected features after data processing. While DRNK3GE5 has a slight positive bias, the distribution of FRUITJU1 is close to the standard normal distribution. This treatment helps the model better adapt to the characteristic of each variable during the training process and improves the prediction accuracy.

C. One-Hot Coding

One-Hot coding is applied to transform discrete data into numerical data that can be processed by machine learning models, avoiding the potential misleading sequence or size relationships between categories. By creating separate binary features for each category, One-Hot coding improves the model's ability to recognize category differences, making it particularly suitable for algorithms like logistic regression that require category-independent inputs.

III. METHODS AND MODELS

In this section, we elaborate on the classification methods applied in the project and evaluate the prediction of models through F1 score and accuracy rate. Five machine learning methodologies are employed for data classification, which accordingly facilitate the improvement of initial data preprocessing methods in order to better adapt to the characteristics of the dataset.

A. Baseline Classification Models

Given the objective of binary classification, it is desirable to begin with the classic classification models, and we start with linear regression, support vector machine (SVM) and logistic regression.

To test our classification methods, we sample 10% of the training data as our test data. As shown in Table 1, logistic regression has the best accuracy and F1 score among classic classification models. However, F1 scores of linear regression and SVM almost approach zero, possibly resulting from the imbalance and skewness of the training dataset where only 28975 people (8.8%) have MICHD compared to 299160 healthy people (91.2%). That is because F1 score is the harmonic mean of precision and recall which focuses on the accuracy of identifying diseased people and it can be strictly influenced by the skewness of the data.

Model	Accuracy	F1 Score
Linear Regression	0.912	0.009
Support Vector Machine (SVM)	0.912	0.009
Logistic Regression	0.915	0.247

Table 1: Classic model performance on training dataset

In linear regression, the threshold is set as 0.5 manually and neglect the skewness of the dataset. In SVM, since the data is not necessarily linearly separable, it might be difficult to find a hyperplane that separates positive and negative data points correctly. Regarding the logistic regression, due to the large number of features in the dataset and One-Hot coding involved in data preprocessing, the final training data may have issues with multicollinearity, affecting classification performance of the model.

B. Ridge Regression and l_2 Regularized Logistic Regression

In order to resolve the excessive skewness and potential multicollinearity issues in the dataset, we consider adding a penalty term to baseline models, i.e., the ridge regression and l_2 regularized logistic regression. The prediction results under both models have shown significant improvements: the F1 score and the accuracy reach 0.394 and 0.905 in l_2 regularized logistic regression model ($\lambda = 0.1$, threshold $\theta = 0.3$), and the F1 score and the accuracy reach 0.414 and 0.856 respectively in the ridge regression model which produces the best prediction in our analysis on the above testing data.

C. Hyperparameter Tuning

In our ridge regression model, there are two hyperparameters: regularization parameter λ and threshold value θ. λ controls the sparsity of features, which helps to address multicollinearity issues, while θ mitigates the effect of skewness in the original dataset if chosen appropriately.

As illustrated in Figure 2, we can choose λ as 0.1 to achieve a relatively small loss value in ridge regression.

Additionally, when $\lambda = 0.1$, we would like to find the best threshold θ that maximizes F1 score. As visualized in Figure 3, the optimal value found is approximately 0.191. Under these two hyperparameters, the F1 score and the accuracy achieve 0.414 and 0.856.

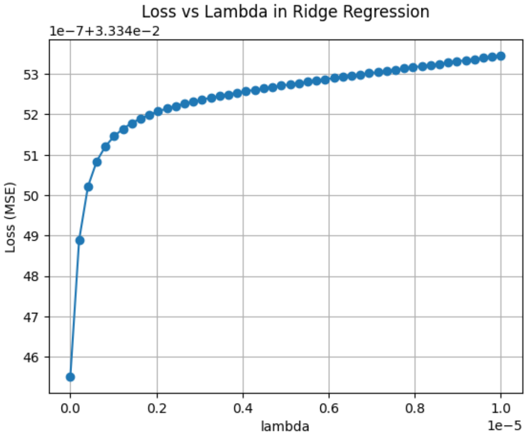

Figure 2: Loss vs. Lambda in Ridge Regression

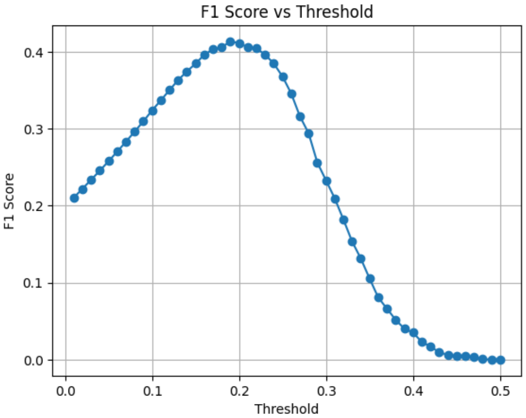

Figure 3: F1 score vs. Threshold in Ridge Regression

IV. Conclusion

In this report, we propose an effective method to identify MICHD from given dataset. We apply various data preprocessing techniques, including one-hot encoding, NaN value replacements, normalization. We also study five classification models and perform hyperparameter tuning. We ultimately selected l_2 regularized logistic regression as the final predictive model, which achieves the best result with an F1 score of 0.403 and an accuracy of 0.906 on the test set by AIcrowd.

References

[1] U.S. Department of Health & Human Services, "CDC – 2015 BRFSS Survey Data and Documentation," 2015. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html

[2] U.S. Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System 2015 Codebook Report," 2016. [Online]. Available: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf