Exploratory Data Analysis and Predictive Modeling for Heart Disease

MASMOUDI Aicha, CHAABANE Sarra, BENCHEKROUN Ali
CS-433: Machine Learning, EPFL, Group MMLT

***Abstract*—We present our Machine Learning (ML) approach for predicting the likelihood of Myocardial Ischemia and Coronary Heart Disease (MICHD) using a comprehensive dataset of health and lifestyle factors. Early detection of cardiovascular diseases (CVDs) is vital for improving patient outcomes and reducing mortality rates. Our binary classification model demonstrates an accuracy of 0.85 and an F1 score of 0.407, providing a valuable tool for the early detection and prevention of CVDs.**

I. Introduction

Early identification of CVDs is essential for enhancing patient outcomes and decreasing mortality rates. Nowadays, leveraging ML algorithms to serve this purpose and predict patient's outcome is a growing trend [1], [**?**]. Our focus is on predicting the likelihood of developing MICHD based on individual health and lifestyle data. Using data from The Behavioral Risk Factor Surveillance System (BRFSS) dataset from 2015 [2], we aim to develop a binary classification model that facilitates early diagnosis. This model has the potential to lower healthcare costs, helping in early diagnosis and deliver data-driven insights for preventive care.

II. Models and Methods

A. Dataset

The dataset $\mathbf{X}$ is a matrix where $\mathbf{X} \in R^{N \times D}$. It consists of $D = 321$ features and $N = 328,135$ data entries.

B. Data Cleaning and Feature Engineering

*1) **Irrelevant features**:* We manually examine the dataset's column names and assess their relevance to our project's objectives. Based on this, we remove 29 columns deemed non-essential for predicting MICHD, mainly demographic indicators like state code and cell phone information.

*2) **Handling NaNs**:*

(a) **Identifying NaNs:** We replace all placeholder values (e.g., 777777, 999999 etc. in the dataset) representing responses such as "don't know" or "refused to answer" with NaN.

(b) **Removing features with too many NaNs:** We set the threshold to 80% missing values and eliminated features exceeding this limit. Visual analyses, such as histograms and line charts, guided this threshold choice.

(c) **Identifying Categorical and Continuous features:** In order to appropriately handle NaNs, we define a threshold to classify features into categorical and continuous types. Features with more than 10 unique values are considered as continuous, and those with less as categorical. We chose this threshold based on exploratory data analysis on the number of entries per feature.

(d) **Imputing NaNs:** To ensure data integrity, we employed imputation techniques. For continuous features, they were replaced with the median of the respective feature. This approach is advantageous as the median is robust to outliers and provides a central tendency measure that is less influenced by extreme values, in opposition to overall mean imputation which leads to bias [3], [4]. For categorical features, we used the mode to impute NaNs.

*3) **Correlation Analysis**:* Highly correlated variables are redundant in the sense that little to no additional information is gained by using them simultaneously[5]. To enhance the predictive power of our model and reduce redundancy, we analyzed the correlation matrix of the training dataset and removed one variable from each pair with a correlation coefficient over 75%.

*4) **Low variance features**:* Features with very low variance provide minimal information for model prediction. We set a threshold of 0.1 above which features were retained, a common practice in handling low variance features.

*5) **Standardization**:* We applied standardization to the feature sets, making the features centered around their means and scaled to their standard deviation. Properly scaling the features is essential for many ML algorithms, as it ensures that each feature contributes equally to the distance calculations and model performance. [6].

*6) **Bias Term**:* We add a bias term to the model to enable it to shift the decision boundary away from the origin, allowing for better fitting of the data and improved predictive performance.

*7) **Label Modification (for Logistic regression)**:* The target $\mathbf{y} \in R^N$ use $y_i = -1$ for individuals without MICHD, and $y_i = 1$ for individuals with MICHD. For convenience, we set $y_i = -1$ to $y_i = 0$.

With this pre-processing and feature engineering pipeline, we refined the dataset, reducing the initial 321 features to the 101 most significant ones for predicting MICHD.

C. ML Models

We run several ML models to tackle the classification task of MICHD. These models include Gradient Descent (GD), Stochastic Gradient Descent (SGD), Least Squares (LS) and Logistic Regression (LR).

Upon analyzing the performance metrics, F1-score, Accuracy, Recall and Precision, on the training set, we observe that the results for GD and SGD are significantly inferior to those achieved by LS and LR. This leads us to prioritize these two models for further analysis.

When testing the model obtained with LR and LS on the train set, we obtain a high accuracy of 0.9035 and 0.8926 respectively, and a null F1 score, suggesting that the models are extremely biased towards the majority class. Indeed, after examining the distribution of the target variable, we notice that the class representing individuals without the condition (-1) is significantly over-represented, with over 91% of the data. To tackle this imbalance, we implement the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the under-represented class[7], [8].

Table I compares LR and LS performances on the train set, for both balanced and imbalanced datasets.

Table I
COMPARISON OF MODEL PERFORMANCE ON TRAIN SET

Dataset	**Model**	**F1 Score**	**Accuracy**
Balanced	LR	0.2993	0.9035
	LS	0.3472	0.8926
Imbalanced	LR	0.0002	0.9117
	LS	0.0003	0.9117

We observe that both LR and LS models yield promising results on the balanced dataset, leading us to pursue further analysis with this dataset.

The performance metrics obtained indicate that LS is the model with the highest F1-score of 0.3472. Given its strong performance, we focus on this model for further optimization and refinement.

To optimize LS, we use Ridge Regression with feature expansion. We perform a grid search both over the regularization parameter (λ) and the polynomial feature expansion degree (D), using a train-validation split done on the train set. Training and validation splits enabled cross-validation, where each λ-D combination was evaluated using F1-score. The configuration with the highest F1-score on the validation set was then selected for final predictions on the test set.

III. RESULTS

Figure 1 visualizes the F1 scores obtained from various combinations of λ and D in the Ridge Regression model. Cells with a score of zero suggest configurations that did not contribute positively to model performance.

The best performance is obtained using Ridge Regression with $\lambda = 10^{-5}$ and degree D = 1, as shown in Table II. The model yields an F1-score of 0.407 and an accuracy of 0.850 on the test set.

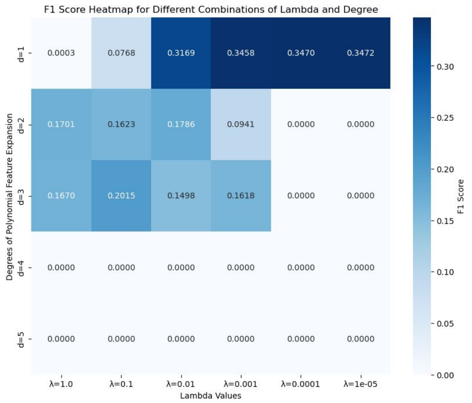

Figure 1. F1 Score Heatmap for Different Combinations of Lambda and Degree

Table II
PERFORMANCE OF RIDGE REGRESSION MODEL ($\lambda = 10^{-5}$, D = 1)

Dataset	**F1 Score**	**Accuracy**
Train Error	0.3472	0.8926
Test Error	0.4070	0.8500

IV. DISCUSSION

Our best model achieves an F1-score of 0.407 and an accuracy of 0.850 on the test set. These results may be satisfying in some contexts, but they may not be sufficient when predicting a condition like MICHD. The feature engineering techniques employed, including the way NaNs were handled, correlation analysis, etc., may have led to the exclusion of relevant features, compromising the model's performance [5]. Additionally, the use of SMOTE to balance the dataset may have introduced artificial patterns that do not reflect real-world variability, potentially leading to overfitting on synthetic data [9]. Lastly, our best model was achieved with minimal feature expansion. This suggests that introducing higher-order features may have added unnecessary complexity without contributing additional predictive power. Such behavior can occur when the complexity of the data is not adequately matched by higher-order interactions.

V. SUMMARY

This study employed machine learning techniques to predict MICHD using health data from the BRFSS dataset. Ridge Regression with regularization parameter $\lambda = 10^{-5}$ and polynomial degree D = 1 was identified as the best model, achieving an F1-score of 0.407 and accuracy of 0.850 on the test set. Although these results are promising, they may not meet clinical requirements for predicting MICHD. Future work should focus on refining feature selection to enhance predictive accuracy.

REFERENCES

[1] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1–4.

[2] C. for Disease Control and Prevention, "Behavioral risk factor surveillance system (brfss): Annual data 2015," https://www.cdc.gov/brfss/annual_data/annual_2015.html, 2015, accessed: 2024-10-29.

[3] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895435606001971

[4] J. L. Peugh and C. K. Enders, "Missing data in educational research: A review of reporting practices and suggestions for improvement," *Review of Educational Research*, vol. 74, no. 4, pp. 525–556, 2004.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[6] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494619302947

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 20–29, Jun. 2004. [Online]. Available: https://doi.org/10.1145/1007730.1007735

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.