Machine Learning - Project 1: Early Detection of Cardiovascular Disease

AI Huasca Team: Henry Schultz 342151, Louis Tschanz 315774,
Majandra Garcia 347470

November 1, 2024

1 Introduction

Cardiovascular diseases (CVD) are among the leading causes of death globally, affecting a growing proportion of the aging population [1]. This project aims to leverage machine learning to estimate the risk of coronary heart disease (MICHD) in individuals based on various health and lifestyle factors collected through the Behavioral Risk Factor Surveillance System (BRFSS) [2]. Our goal was to develop a model capable of predicting the likelihood of a CVD event, focusing on binary classification tasks.

Our methodology encompasses comprehensive data preprocessing, model selection, and feature engineering steps, ultimately optimizing predictive accuracy. By balancing rigorous baseline comparisons, regularization, and cross-validation techniques, we provide insights into model performance and generalizability on unseen data.

2 Data Preprocessing

The BRFSS dataset used in this project comprises health survey responses, including features such as physical activity, smoking habits, and pre-existing health conditions [3]. Each feature represents either a quantitative or categorical variable, with substantial missing values in some columns, posing specific preprocessing challenges.

2.1 Handling Missing Values

Features with over 30% missing values were removed, as their limited data would add noise rather than information. For the remaining missing values, the following imputation was used:

- **Mode Imputation for Categorical Data:** Categorical features with missing values were imputed using the mode, maintaining the integrity of the original distribution.
- **Mode Imputation for Continuous Data:** Similarly, missing values of quantitative values were imputed by dividing the range of values into 200 bins and computing the mode value of the bins.

2.2 Identifying Categorical Features

We discriminated features containing only integer values based on their number of unique values :

- **Over 5 unique values:** Identified quantitative feature
- **5 unique values or less:** Identified as categorical feature

2.3 Feature Engineering

To prepare the features, we applied transformations:

- **Binary Encoding:** Features identified as categorical (e.g., general health status) were transformed into binary vectors.
- **Standardization:** Features identified as quantitative variables were standardized to have a mean of zero and a standard deviation of one.

3 Model Selection and Implementation

We experimented with multiple regression and classification models:

- **Linear Regression:** Initially, we implemented linear regression with gradient descent and MSE as loss. This model was primarily useful for feature analysis rather than the final binary classification task.
- **Logistic Regression:** Due to the binary nature of MICHD prediction, logistic regression became the primary model.
- **Regularized Logistic Regression, final model:** An $L2$ regularization term was added to logistic regression, controlling for overfitting by penalizing large coefficients.

3.1 Parameters tuning

5-Fold Crossvalidation: Model parameters (γ and λ) were optimized using 5-fold crossvalidation with grid search.

4 Evaluation and Results

We assessed models using cross-validation (CV) on accuracy and F1 metrics. Cross-validation provided insights into model stability and generalizability:

- **Ablation Study:** Performance comparisons showed regularized logistic regression consistently outperformed basic logistic regression in generalization, reducing overfitting and improving accuracy. The $L2$ penalty, when optimized, improved the model's ability to handle data sparsity without losing essential information.
- **Final Model Selection:** Regularized logistic regression emerged as the best-performing model, showing both high accuracy and stability across CV folds ($\gamma = 0.3$, $\lambda = 0.1$, 1000 iterations).

5 Conclusion

This project demonstrates the applicability of machine learning to health prediction tasks, specifically in cardiovascular disease risk assessment. Regularized logistic regression proved the most effective for this binary classification task. Future research could explore ensemble methods or deep neural networks to capture complex interactions among features, potentially increasing predictive accuracy on unseen data.

Our model pipeline is reproducible and well-documented, allowing further extensions or comparisons. This work underscores the importance of robust preprocessing, thoughtful model selection, and the utility of regularization when dealing with high-dimensional, sparse healthcare data.

References

[1] World Health Organization. Cardiovascular Diseases (CVDs) Fact Sheet. *WHO*. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System (BRFSS). *CDC*. Available at: https://www.cdc.gov/brfss/index.html

[3] Centers for Disease Control and Prevention. 2015 Codebook Report: BRFSS. *CDC*, 2016.

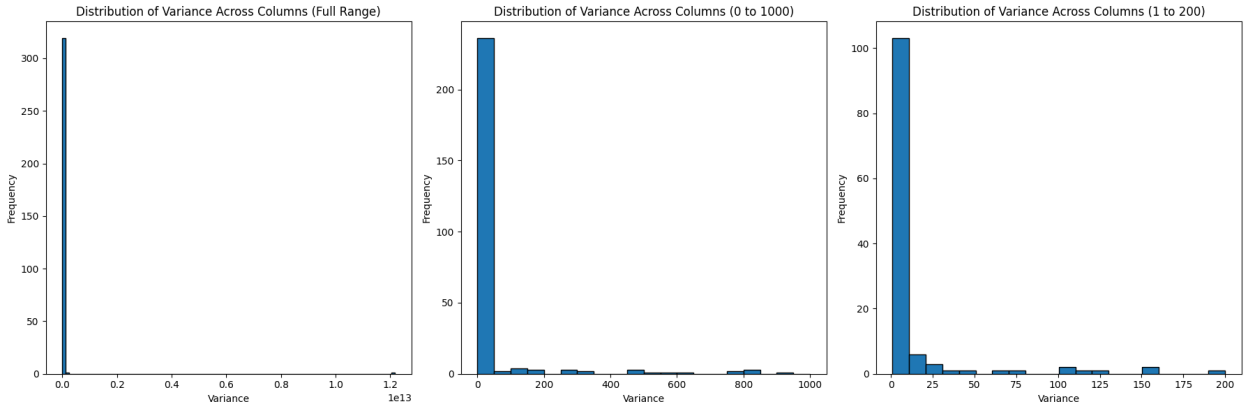

Figure 1: Variance analysis across features. Many features exhibit low variance, making them potential candidates for removal. However, given the obvious presence of categorical features given the name of some features, we retained low-variance features to avoid discarding meaningful categorical features.

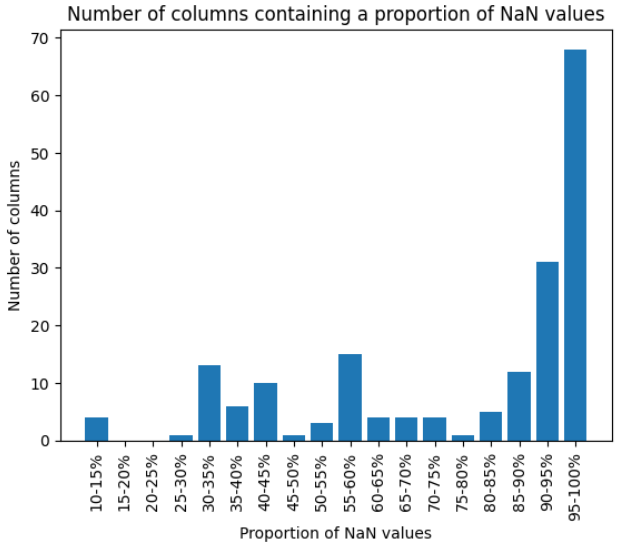

Figure 2: Proportion of missing values across features. We notice a particularly high number of features containing a high (more than 80 percent) proportion of NaN value. While we started by removing these features only, some research on the topic and manual tuning led us to remove all features with more than 30 percent of NaN values.

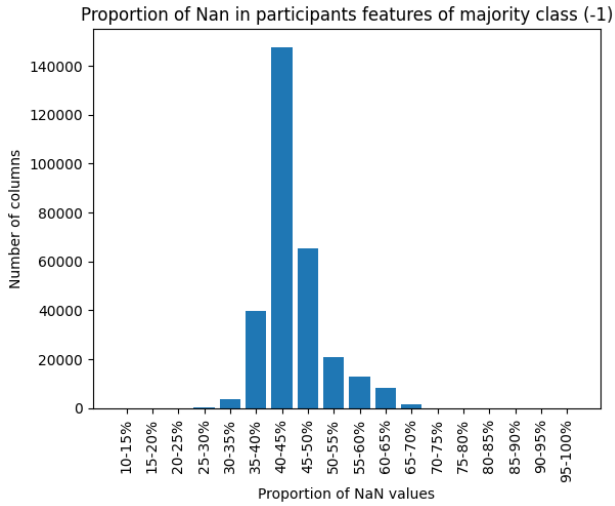

Figure 3: Addressing Class Imbalance by Under-Sampling. Initial predictions yielded an F1 score of 0, as the model classified all samples under the majority class due to severe imbalance. To mitigate this, we under-sampled the majority class starting by excluding samples with the highest proportion of missing values.

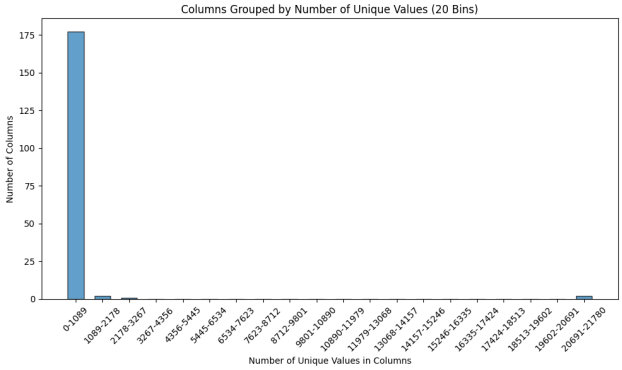

Figure 4: Distribution of the amount of unique values across integer features, full range. This figure shows the distribution of the amount of unique values across integer features We notice the majority is contained in the first bin (0-1089).

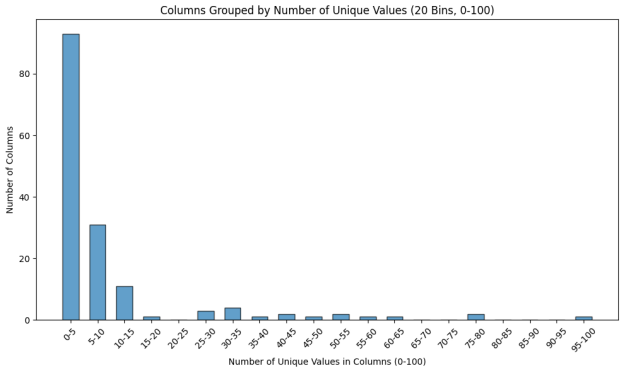

Figure 5: Distribution of the amount of unique integer values across integer features (range 0-100). Similarly to Figure 4, the majority of the amount of unique values in integer features is contained within the first bin (0-5).

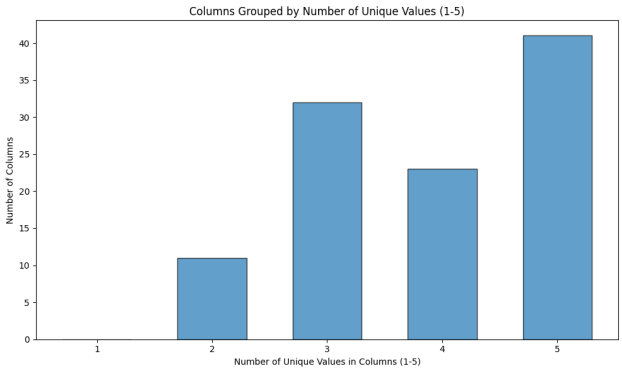

Figure 6: Distribution of the amount of unique integer values across integer features (range 0-5). Focusing on the first bin of Figure 5, we notice in fact that most integer features have a number of unique values between 1 and 5. This leads us to formulate the conservative hypothesis that identifying such features as categorical allows to correctly identify most categorical features while limiting the risk of misidentifying a quantitative feature

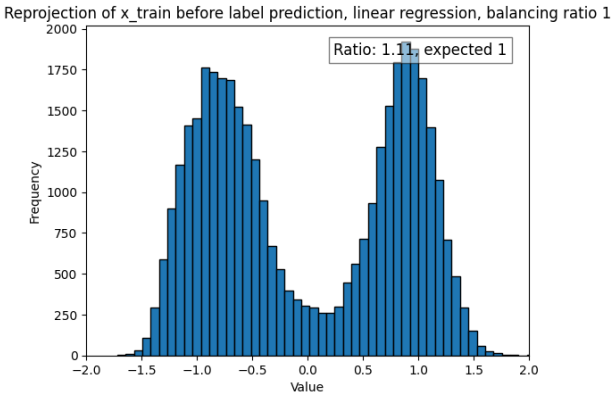

Figure 7: Training set reprojection based on linear regression weights (γ = 0.01, 300 iterations). The projection of the training data based on linear regression weights revealed an ideal ratio at 1 to balance class distributions. The real ratio is 1.11, which means the majority class (-1) is still overpredicted. However given the risk of overfitting, we consider this close enough to the theoretical ratio.

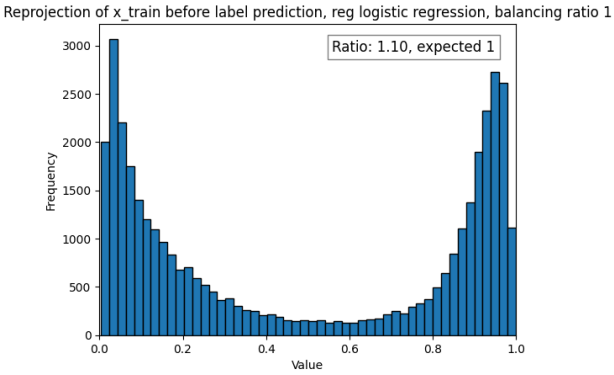

Figure 8: Training set reprojection based on regularized logistic regression weights (best model, γ = 0.3, 150 iterations). The projection of the training data based on regularized logistic regression weights also revealed an ideal ratio at 1 to balance class distributions.