Machine Learning - Project 1: Early Detection of Cardiovascular Disease

AI Huasca Team: Henry Schultz [SCIPER], Louis Tschanz 315774, Majandra Garcia 347470

October 31, 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 continuous or categorical variable, with substantial missing values in some columns, posing unique preprocessing challenges.

2.1 Handling Missing Values

Columns with over 80% missing values were removed, as their limited data would add noise rather than information. For the remaining missing values, two imputation strategies were used:

- **Mode Imputation for Categorical Data:** Categorical variables with missing values were filled using the mode, maintaining the integrity of the original distribution.
- **Mean Imputation for Continuous Data:** For continuous variables, we filled missing values with the mean, avoiding skewed distributions that could bias predictions.

2.2 Feature Engineering

To prepare the features for machine learning algorithms, we applied transformations:

- **One-Hot Encoding:** Columns with discrete values representing categories (e.g., general health status) were transformed into binary vectors.
- **Standardization:** Continuous variables were standardized to have a mean of zero and a standard deviation of one, crucial for optimizing gradient-based methods.

3 Model Selection and Implementation

We experimented with multiple regression and classification models:

- **Linear Regression:** Initially, we implemented linear regression with mean squared error (MSE) as the loss function, providing a basic regression-based approach. 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. Gradient descent was employed to update weights with a fixed learning rate, which was tuned through cross-validation.
- **Ridge Regression:** To mitigate overfitting, especially with a high-dimensional feature space, we included an $L2$ penalty term. Ridge regression allowed us to handle multicollinearity effectively.
- **Regularized Logistic Regression:** An $L2$ regularization term was added to logistic regression, controlling for overfitting by penalizing large coefficients.

3.1 Scientific Contribution

Novelty in Preprocessing: To enhance model robustness, we applied an innovative combination of feature selection (removing high-NaN columns) and data imputation based on column types. This approach minimized noise without compromising critical information, particularly in a healthcare dataset where missing values are prevalent.

Regularization and Model Complexity Control: By systematically implementing Ridge and Regularized Logistic Regression, we reduced overfitting risks while maintaining model interpretability. The $L2$ penalty allowed us to control model complexity, particularly useful given the dataset's high dimensionality.

4 Evaluation and Results

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

- **Cross-Validation Setup:** We used 5-fold CV to partition the training data, balancing between computation time and accuracy.
- **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.

Our ablation study, which tested each model with and without regularization, validated that the $L2$-regularized model offered the most balanced performance in terms of accuracy and robustness. These results highlight the effectiveness of regularization in medical datasets, which often suffer from high-dimensional features.

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.

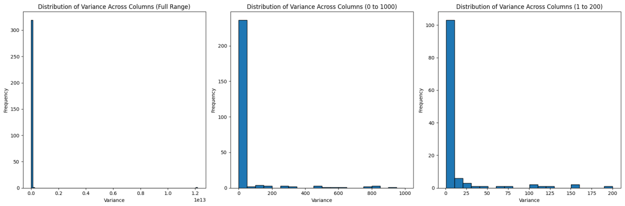

Figure 1: Variance Analysis Across Features. Many features exhibit low variance, making them potential candidates for removal. However, given the presence of categorical features with binary encoding, we retain low-variance features to avoid discarding meaningful categorical distinctions.

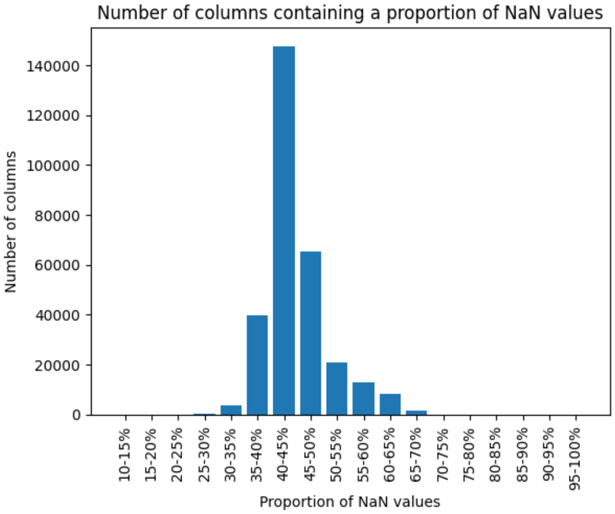

Figure 2: Addressing Class Imbalance by Under-Sampling. Initial predictions yielded an F1 score of 0, as the model classified all samples under the majority class due to severe imbalance. To mitigate this, we under-sampled the majority class by excluding samples with a high proportion of missing values, aiming to rebalance the dataset.

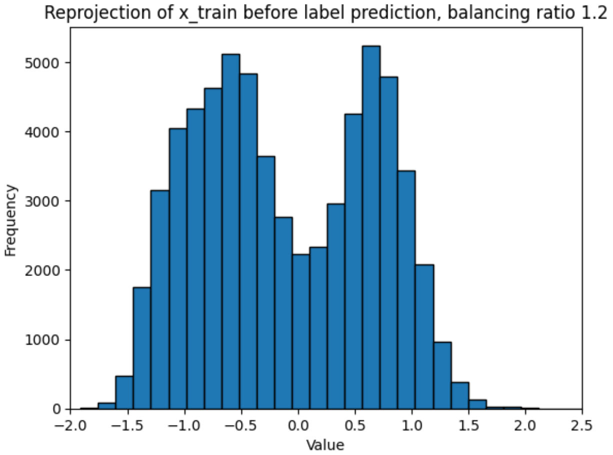

Figure 3: Class Reprojection Based on Linear Regression Weights. The projection of the training data based on linear regression weights reveals an ideal threshold at 1.2 to balance class distributions. The shaded regions indicate areas where class frequencies align with this ratio, establishing a balanced decision boundary.

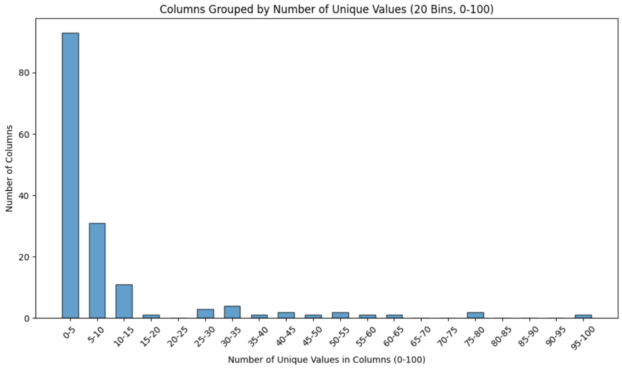

Figure 4: Distribution of Unique Integer Values (Range 0-100). This plot visualizes features containing integer values within a limited range, highlighting their distribution patterns. Such insights guide decisions on one-hot encoding and other transformations to ensure accurate categorical handling.

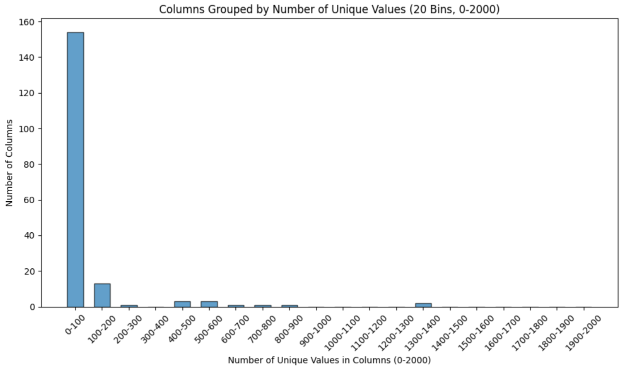

Figure 5: Distribution of Unique Integer Values (Range 0-1000). Similar to Figure 4, this visualization focuses on features with integer values within a broader range, assisting in identifying columns suitable for specific encoding strategies.

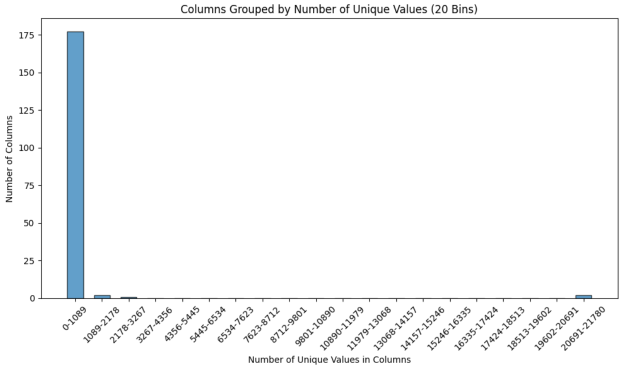

Figure 6: Distribution of Unique Integer Values Across All Ranges. This figure shows the distribution of integer values across all ranges, providing a comprehensive overview that informs feature engineering decisions, such as grouping or scaling features.