EPFL ML Course Class Project 1 Report

Sabri Yiğit Arslan, Eren Akçanal, Efe Tarhan
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—This project investigates machine learning approaches for predicting coronary heart disease (CHD) risk using health data from a diverse population of over 300,000 individuals. Comprehensive data preprocessing, including handling missing values, encoding categorical variables, cleaning the dataset, subsampling, and addressing multicollinearity, prepares the data for analysis. Classical machine learning models, such as least squares estimation, linear regression, logistic regression, and their L2-regularized counterparts, are employed to compare predictive performance. Regularized logistic regression emerged as the best-performing model, with optimal hyperparameters identified through grid search. Experimental results highlight challenges related to class imbalance and high dimensionality, which can hinder training efficiency and increase the likelihood of convergence to local minima. This project provides insights into the complexities of large-scale health data and the potential of machine learning in CHD risk prediction.**

I. Introduction

Heart disease remains a leading global cause of death, affecting millions annually. Predicting CHD risk through health and lifestyle data offers promise for early intervention and better health outcomes. This project harnesses machine learning to predict CHD likelihood using data from the Behavioral Risk Factor Surveillance System (BRFSS), which compiles information on lifestyle, chronic conditions, and preventive health behaviors[1].

While advanced methods like deep learning and gradient boosting are increasingly favored for classification tasks due to their ability to model complex, nonlinear relationships, classical machine learning techniques such as linear regression, least squares estimation, and logistic regression remain valuable. These traditional models can still deliver strong, interpretable results, making them useful for tasks where clarity and simplicity are prioritized alongside performance.

This report is organized as follows: Section II details the data analysis and preprocessing steps, including exploratory data analysis and feature transformations to optimize model inputs. In Section III, we present results from multiple models, utilizing hyperparameter tuning and k-fold cross-validation to select the best-performing model and parameters, comparing these predictions to benchmark standards. Finally, Section IV concludes with a summary of our findings and a discussion of key insights and potential implications.

II. Data Analysis and Preprocessing

The CDC data is obtained from phone surveys from over 300.000 people in US where it contains highly sparse and raw data like most of the real world data. Because of this we have developed a preprocessing pipeline for this dataset where each step and their short explanations can be seen below:

1) **Handling non-associated values in features:** The dataset, collected via phone survey, includes numerical values, with certain entries (e.g., 77, 99, 9999) representing "not answered" or "don't know"; these have been replaced with NaN values.
2) **One-hot encoding categorical values:** Categorical features coded as integers are one-hot encoded to represent their true, non-ordinal nature.
3) **Removing irrelevant features:** Irrelevant features such as ID numbers, phone numbers, and landline indicators are removed to enhance model performance and reduce dataset dimensionality.
4) **Handling sparsity:** NaN values in the dataset are filled with values sampled from a Gaussian distribution, using the column's mean and standard deviation to reflect population frequency.
5) **Standardization:** Features are standardized to enhance training stability and ensure consistency among numerical results, particularly in regularization, which can be sensitive to feature scales.
6) **Handling multicollinearity:** Features with a correlation above a set threshold of 0.95 are removed from the dataset to reduce multicollinearity, which can negatively impact model accuracy and training stability.
7) **Applying PCA:** Principal Component Analysis (PCA) is applied to the continuous features to identify the main variance directions in the dataset. A subset of principal components is then selected to represent the data, optimizing feature dimensionality for model training.
8) **Balancing the dataset:** To address the dataset's heavy imbalance, oversampling was used to replicate minority class samples, equalizing the class distributions.

III. Methods and Analysis

Due to algorithm limitations, simpler methods such as linear regression, logistic regression, and least squares estimation were applied to the dataset, with each model run for sufficient iterations to ensure convergence. While linear regression and least squares are not ideal for binary classification, all models were evaluated using k-fold cross-validation with k = 3 to assess their performance on the dataset. The comparative results of these models are presented in Table I, facilitating the selection of the most suitable model for the task.

Table I
Model Performance Comparisons Under Convergence

Model	Accuracy	F1-Score
Ridge Regression	75.7%	0.765
Linear Regression w/ MSE	69.8%	0.706
Logistic Regression	77.8%	0.783

As shown in the comparison of baseline models in Table I, logistic regression outperforms linear regression, as expected for a binary classification task. Consequently, logistic regression was selected as the primary model for the remainder of the project.

After selecting logistic regression as the appropriate model for this task, a hyperparameter search was conducted by tuning the regularization parameter (lambda) and learning rate to optimize performance. The results of this search are presented in Table II.

Table II
F1 Scores with Different Learning Rates (Gamma) and Regularization (Lambda)

Learning Rate (Gamma)	Lambda	F1 Score (Mean)	Std. Deviation
0.001	0.0001	0.5306	0.1055
0.001	0.001	0.5530	0.0083
0.001	0.01	0.3606	0.2554
0.001	0.1	0.4805	0.0865
0.01	0.0001	0.6552	0.0035
0.01	0.001	0.6550	0.0065
0.01	0.01	0.6645	0.0095
0.01	0.1	0.6850	0.0067
0.1	0.0001	0.7184	0.0008
0.1	0.001	0.7207	0.0017
0.1	0.01	0.7348	0.0027
0.1	0.1	0.7304	0.0010
1	0.0001	0.7363	0.0007
1	0.001	0.7391	0.0009
1	0.01	0.7372	0.0016
1	0.1	0.6816	0.0522

Due to the oversampling approach, the model performed well with higher learning rates without oscillating around convergence points. Although the F1 score achieved with a learning rate of 1 was the highest, we observed several instabilities in the optimization process at this rate. Therefore, we selected a learning rate of 0.1 and a regularization value of 0.001 for the final predictive model.

While we successfully implemented k-fold cross-validation and established separate validation and training sets, the F1 scores obtained were somewhat skewed due to changes in the validation set's distribution. Although the model was trained and validated on balanced subsets, these scores do not account for the imbalance present in the actual test set. As a result, our AIcrowd scores appear lower than the F1 scores obtained during validation.

As the final step in the processing pipeline, we considered reducing the number of continuous features using PCA, which appeared to be a promising approach for this project. We evaluated the model performance by applying the optimal parameters to different sets of principal components, selected based on their total explained variance (TEV). The results of testing the model across various numbers of principal components are presented in Table III.

Table III
Total Explained Variance and

TEV of PCA Components	# of PC	Accuracy	F1-Score
98% of Components	217	77.9%	0.7362
95% of Components	182	75.8%	0.7341
90% of Components	179	77.5%	0.731

The results indicate that the model can operate efficiently with significantly fewer parameters while retaining much of its performance.

IV. Conclusion

In this project, we implemented a range of machine learning algorithms to predict coronary heart disease risk based on large-scale health data. Through rigorous data preprocessing and feature engineering, we improved model accuracy by addressing challenges like class imbalance, high dimensionality, and overfitting. Using k-fold cross-validation, we selected optimal model parameters that best represent the data, ultimately achieving meaningful predictions for CHD risk with the provided dataset.

References

[1] U. D. of Health Human Services, "Cdc - 2015 brfss survey data and documentation," 2015, accessed on Oct 31, 2024. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html