Detection of Cardiovascular Disease with predictive models

Hugo Heinkele, Nils Manni, Philippine Laroche
CS-433 Machine Learning, EPFL

***Abstract*—Cardiovascular diseases (CVD) are among the leading causes of mortality worldwide. In this project, we apply and evaluate several machine learning algorithms to predict CVD risk using data from the Behavioral Risk Factor Surveillance System (BRFSS). Focusing on both lifestyle and clinical factors, our goal is to evaluate each model's effectiveness in predicting the likelihood of coronary heart disease. The analysis provides a comparative analysis of each model's performance for this binary classification task.**

I. Introduction

Heart disease is an important public health challenge globally, where early and accurate diagnosis is essential for effective treatment and prevention. This project aims to better understand the causes and predict heart disease using cutting-edge technologies such as machine learning models. By implementing and assessing a range of models, we aim to identify the most effective approaches for heart disease detection.

In order to have meaningful result, the creation of a comprehensive dataset is essential before training and testing any machine learning model. This study detailed the key stages of the machine learning pipeline including : data preprocessing, feature selection, model implementation and evaluation. The objective is as much to obtain accurate predictive models as to provide insights into the performance trade-offs among different algorithms. Different models are compared such as gradient descent, stochastic gradient descent, least squares regression, ridge regression, logistic regression, and regularized logistic regression to analyze the unique strengths and limitations of each approach.

II. Data Preprocessing and features selection

A. Data Cleaning

The objective of data cleaning was to establish consistency across parameter values and enhance the dataset's quality, providing a reliable input for our machine learning models. A major issue that required addressing was missing values, which can bias models and lead to inaccurate predictions. We applied two consecutive strategies to handle this. First, we removed parameters with at least 25% missing values, as this amount of missing data would result in biased estimates. Next, we used mean imputation to fill remaining missing values. Although advanced methods like k-nearest neighbors (KNN) imputation were considered, we opted for mean imputation due to its simplicity and effectiveness in maintaining the dataset's distribution. Additionally, we converted categorical variables to binary format through one-hot encoding and standardized binary responses to ensure consistent machine learning integration and improve data interpretability. This meticulous preparation lays a solid foundation for the next stages of the project.

B. Balancing the Data

With only about 9% of cases representing individuals with heart disease, the dataset was significantly imbalanced. Such skewed data distribution poses a risk of model bias toward the majority class, potentially limiting the accurate identification of heart disease cases. We addressed this imbalance using an "undersampling" strategy, reducing the size of the majority class to match that of the minority class, thereby achieving equal representation without increasing dataset size. After balancing, the training dataset was split into training and validation sets following an 80/20 ratio. It's important to note that the test set retains the real-world class distribution, ensuring an authentic evaluation of model performance. This approach allows our future models to learn effectively with balanced representation from both classes.

C. Features Selection

The final preprocessing step focused on selecting the most relevant features from an initial list of 321. Guided by common knowledge and data from the Behavioral Risk Factor Surveillance System [1], we retained relevant features such as tobacco use, physical activity, and blood pressure. To further enhance this selection, we implemented a feature selection pipeline with three main steps:

Variance Threshold: This step removes features with low variance, under the assumption that features with little variability do not contribute significantly to predictive power. By setting a variance threshold, we eliminated near-constant features, which are less informative.

Correlation Analysis: Features were evaluated for their correlation with both the target variable and each other. Features with low correlation to the target were removed due to their limited predictive value, while those highly correlated with others were also removed to avoid redundancy. This reduced multicollinearity, improved model stability, and simplified the dataset.

Statistical Relevance: Statistical tests were used to assess each feature's significance concerning the target variable. For continuous features, an F-test was applied, while binary features underwent a Chi-square test. Only features meeting the statistical relevance threshold were retained, ensuring a strong predictive relationship with the target variable.

This feature selection process should minimize the risk of overfitting by reducing noise and irrelevant information, ultimately crafting a dataset that is not only comprehensive but also rich in quality and relevance. This ensures that our machine learning models are learning from the best possible information, leading to accurate and reliable predictions for heart disease analysis.

III. Machine Learning models

A. Model Implementations

In order to predict as best as possible coronary heart disease, various machine learning algorithm have been implemented. The model implementation began with a straightforward approach by using mean squared error optimized through gradient descent, followed by its stochastic variant to achieve faster convergence. Then, a similar approach is explored with least squares regression and its variant ridge regression to prevent over fitting thanks to the L2 regularization term. Finally, it was essential to include logistic regression and its regularized counterpart as they are usually better suited for binary classification problems.

B. Hyperparameters Tuning and Model Training

Several hyperparameters, such as the number of iterations, learning rates (γ) and regularization term (λ) can be adjusted depending on the method in order to get the best results. To get the best results, the following general training pipeline has been set up (for all the model except for the least squares).

- **Find the best hyperparameters (γ, λ, number of iterations) -** Each model is trained on several values of hyperparameters, then optimality is judged on the basis of the best F1 score (III-C) calculated on the validation set.
- **Score metrics Analysis -** The F1 score, validation accuracy, and training and validation loss are plotted against the hyperparameters γ, λ, to analyze the evolution of the model complexity and potential overfitting or underfitting.
- **Find the best weights -** : if convergence was not reached during the hyperparameter tuning, the optimal values for γ, λ are used to train the model with a higher number of iterations. The goal is for the training loss to reach convergence (regarding a defined tolerance of 10^{-5}), in order to have the best weights.

In the least squares case, there are no hyperparameters as it is a direct method that captures the linear relationship within the data, so the model is simply trained on the preprocessed data. Some performance metrics (F1 score, accuracy, and loss) were then calculated on the validation set (III-C) to compare it with the other models.
Finally, once satisfied with the parameter settings and performance, the best model (regarding the F1 score) was applied to the test set to assign a label to each of the inputs.

C. Model Evaluation

To assess the performance of the model, different metric scores are used : accuracy, F1-score and loss function. Loss measures the error between actual values and predicted ones, and is always useful to evaluate model optimization as lower loss are researched. However it does not necessarily well represent the general performance of the model. Accuracy is a classification metric that measures the proportion of correctly predicted instances out of the total instances, providing a general overview of the model's correctness it is a good indicator of the general performances of the model given a balanced dataset . Moreover, the F1-score describes the harmonic mean of precision and recall :

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

It is particularly useful and reliable in imbalanced scenarios. F1-score close to 1 depicts the best performance. (TP: True Positive, FP: False Positive, FN: False Negative)

D. Comparative Analysis

The data is split between training and validation test to ensure a robust evaluation. The models are trained on the training set and then their performance are assessed on the validation set. The score obtained are resume in the Table I. In order to chose the best model, the objective is to have the higher F1 Score associated to small difference between training loss and validation loss to minimize overfitting.

Model	F1 score	Accuracy	Val. Loss
Gradient Descent with MSE	0.679	0.531	0.469
Stochastic Gradient Descent with MSE	0.665	0.498	0.502
Least Squares Regression	0.657	0.525	0.475
Ridge Regression	0.665	0.498	0.104
Logistic Regression	0.736	0.500	0.500
Reguralized Logistic Regression	0.735	0.500	0.499

TABLE I: Comparison of Model Performance

Regarding the Table I, Ridge Regression emerges as the best model for prediction as it has high F1-score combined to a low loss validation (compared to the other model). Ridge Regression adds a regularization term that is supposed to reduce overfitting by penalizing large coefficients. But it appears to show low performance on the testing set with a testing F1 score of 0.162 and an accuracy of 0.088 suggesting high overfitting.

IV. Conclusion

In the project, multiple machine learning models were explored to predict cardiovascular disease (CVD). After comprehensive data preprocessing—including missing value handling, dataset balancing, and feature selection. Ridge Regression emerged as the most effective model, achieving the highest F1 score and lowest validation loss. Despite these promising outcomes, more rigorous hyperparameter tuning could potentially improve the model further. In future work, deeper exploration into advanced preprocessing and feature engineering methods may yield even more accurate results.

References

[1] Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System (BRFSS) 2015 Codebook*. Accessed: 2024-11-01. 2015. URL: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.