Machine Learning - Project 1

Beatrice Campo Nathan Mutin Valentin Planes

***Abstract*—This project applies machine learning techniques to predict the risk of Cardiovascular Diseases (CVDs). Using health and lifestyle data from telephone surveys, we built and evaluated binary classification models, demonstrating how machine learning methods can be used to address real-world health challenges and support early disease detection.**

The implementation of the work described here can be found online (https://github.com/nathanmutin/MilaLyon/)

I. Introduction

Cardiovascular diseases (CVDs) are a major global health concern, emphasizing the need for early detection and prevention. This project applies Machine learning (ML) methods to predict the risk of Myocardial Infarct or Coronary Heart Disease (MICHD) using the Behavioral Risk Factor Surveillance System (BRFSS) dataset [1], a large collection of health and lifestyle information from adults across the United States. The final model classifies individuals into two categories: -1 for low risk and 1 for high risk of developing CVD. By applying techniques learned in the course, we aim to build a classification model that supports early identification of at-risk individuals.

II. Models and methods

A. Dataset Preprocessing

The BRFSS dataset is a large-scale survey containing 321 features per individual, encompassing various health-related aspects such as smoking habits, nutrition, physical activity, and general health status.

However, the dataset poses several challenges for machine learning applications:

- A substantial proportion of missing values;
- Numerous highly correlated features, often representing overlapping information;
- Occurrences of inconsistent entries;
- A pronounced class imbalance, with an approximate 10:1 ratio between the majority and minority classes.

These issues can bias model training and reduce predictive reliability. Therefore, an extensive data preprocessing phase was required to clean and prepare the dataset for modeling.

First, we reviewed the official BRFSS documentation [1] to better understand each feature and added metadata, such as indicating feature types (*categorical, binary, continuos…*). This information are stored in the project file *"features_description.csv"* for reference during preprocessing. Next, we addressed special codes and missing values:

- Empty responses such as *"I don't know"* or *"I prefer not to answer"* were replaced with missing values;
- Values incorrectly representing zero were converted to 0;
- Certain features with non-standard encodings were reformatted into meaningful numeric or categorical types;
- Features with an excessive proportion of missing values were dropped;
- Remaining missing values were imputed using the mean (for continuous features) or mode (for categorical features).

We then performed feature encoding and selection:

- One-hot encoding was applied to non-ordinal categorical variables;
- Features showing low correlation with the target were removed;
- Highly correlated features were dropped to reduce redundancy arising from overlapping survey questions.

Finally, we handled outliers and normalized feature scales:

- Outliers exceeding three standard deviations from the mean were clipped;
- All features were normalized using *Min–Max scaling* to constrain their values within the [0,1] range.

This comprehensive preprocessing pipeline ensured that the data was cleaner, more consistent, and better suited for subsequent modeling, thereby improving the interpretability and reliability of the results.

B. Model Selection

We selected Logistic Regression as our model due to its simplicity, efficiency, and interpretability, making it particularly suitable for binary classification tasks. The model estimates the probability that an observation belongs to a given class (0 or 1) based on its features. Model parameters w^* were optimized using gradient descent, which iteratively minimizes the log-loss cost function by updating the weights in the direction opposite to the gradient, scaled by the learning rate γ. Because the dataset was highly imbalanced, with class 1 being approximately ten times less frequent than class 0, a line search was performed on the validation set to identify the optimal classification threshold for each trained model. Predicted labels were then mapped back to $\{-1,1\}$ to comply with the required submission format.

Several variants of logistic regression were implemented and evaluated to maximize the F1-score, as further discussed in the *Discussion* section.

C. Parameter Optimization

Optimal parameter selection is essential to balance the bias–variance trade-off and ensure robust generalization. Hyperparameter optimization was conducted through line-search over predefined value ranges. For each candidate value, 5-fold cross-validation was performed, and models were compared based on their F1-scores to identify the optimal configuration.

The first parameter tuned was the learning rate γ, which determines the step size in the gradient descent update:

$$w_{t+1} = w_t - \gamma \nabla_w L(w_t) \tag{1}$$

Larger values of γ accelerate convergence but can cause mild oscillations in the training loss. Although $\gamma = 0.8$ produced these minor oscillations, cross-validation showed it consistently yielded higher F1 scores than lower, smoother learning rates. Therefore, it was chosen to prioritize validation performance while maintaining overall training stability.

Next, we assessed the impact of L2 regularization, which penalizes large weight magnitudes and prevents overfitting. The regularization strength (λ) was selected via 5-fold cross-validation.

To further enhance the model's expressive power, polynomial feature expansion was applied to continuous variables, allowing the model to capture nonlinear relationships. The degree of expansion was also optimized through cross-validation.

Finally, to address the pronounced class imbalance, a weighted logistic regression [2] variant was implemented, assigning higher weights to the minority class within the loss function. Several class-weight ratios were tested to achieve the best balance between sensitivity to the minority class and overall model performance.

D. Performance Evaluation

The F1-score was selected as the main evaluation metric, as it provides a balanced measure between precision and recall, accounting for both false positives and false negatives. Given the strong class imbalance in the dataset, overall accuracy would not be a reliable indicator of performance, since a model predicting predominantly the majority class could still achieve high accuracy.

The F1-score therefore serves as a more meaningful measure of the model's effectiveness in correctly identifying high-risk individuals while maintaining a balanced performance across both classes.

III. Results

The performance of the different models is summarized in the following table. The dataset was split into 90% for training and 10% for validation and both F1-score and accuracy were used to assess predictive performance.

TABLE I
Models parameters and performances

Model	γ	λ	degree	F1	Accuracy	α
LR[1]	0.8	0.0	0.0	0.423	0.877	-
Ridge LR	0.8	10^{-6}	0.0	0.423	0.877	-
FE[2] and LR	0.8	0.0	2.0	0.424	0.863	-
Weighted LR	0.8	0.0	0.0	0.423	0.877	1[3]
OS[4] and LR	0.3	0.0	0.0	0.422	0.868	0.2[5]

IV. Discussion

All models were trained and evaluated on the preprocessed dataset, consisting of 255 cleaned and normalized features. Gradient descent was used as the optimization algorithm, with a maximum of 1500 iterations across all models.

The baseline model implemented standard Logistic Regression with the optimized γ.
The second model introduced L2 regularization. After cross-validation, lower regularization strengths $\lambda = 10^{-6}$ yielded slightly better results. However, the improvement remained marginal, possibly because prior preprocessing had already mitigated overfitting effects.

The third and fifth models incorporated data augmentation to address nonlinearity and class imbalance. In particular, the polynomial feature expansion improved model accuracy, indicating that capturing nonlinear feature interactions enhanced predictive performance. Conversely, the oversampling model exhibited higher oscillations in the training loss when using $\gamma = 0.8$, which was subsequently reduced to 0.3 to stabilize training convergence.

A weighted logistic regression approach was also tested, modifying the loss function to assign greater importance to the minority class. Although theoretically improving sensitivity to rare cases, the resulting performance gains were limited.

Overall, all models achieved comparable performance levels, with the logistic regression model including a polynomial expansion of degree 2 achieving the best results on the test dataset (*F1: 0.437, Accuracy: 0.878*)

After identifying this best-performing configuration, the same model was retrained on the raw dataset without preprocessing, resulting in worse performance (*F1: 0.397 ± 0.006, Accuracy: 0.845 ± 0.009*) after cross-validation, highlighting the crucial role of preprocessing in achieving reliable predictive performance.

V. Conclusions

In this project, we developed and optimized a machine learning pipeline while addressing key challenges related to data quality, modeling, and class imbalance.

A systematic line search was conducted to identify the best-performing hyperparameters configurations, and multiple logistic regression variants were compared to enhance model robustness. Furthermore, the classification threshold was fine-tuned for each model to balance precision and recall.

These steps produced a binary classifier with decent performance and highlighted the importance of thorough preprocessing prior to model selection and hyperparameter tuning. Future work could explore alternative optimization algorithms (e.g., Adam, Newton's method) or different classification models (e.g., Support Vector Machines, Random Forests, Neural Network) to further enhance predictive accuracy.

[1] Logistic Regression

[2] Polynomial Feature Expansion

[3] Class Weight

[4] Oversampling

[5] Oversampling Ratio

References

[1] CDC, “Behavioral Risk Factor Surveillance System,” 2016, doi: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.

[2] M. Maalouf, “Weighted logistic regression for large-scale imbalanced and rare events data.” doi: https://www.sciencedirect.com/science/article/pii/S0950705114000239.