CS-433: Machine Learning Project 1

Anas Himmi - Antoine Cornaz - Zeineb Mellouli
Department of Computer Science, EPFL, Switzerland

***Abstract*—This paper represents the first project of the Machine Learning course taught in EPFL [1]. We explore and compare different supervised learning algorithms and how they deal with a dataset from Behavioral Risk Factor Surveillance System (BRFSS). The goal is to train and test a model to predict Cardiovascular Diseases (CVD) based on lifestyle and clinical data.**

I. Introduction

The dataset, obtained from **BRFSS**, contains data on health behaviors and chronic health conditions among U.S. residents. The aim is to develop a binary classification model to predict the likelihood of developing **(CVD)**. We have used methods from exploratory data analysis, feature engineering, hyper-parameter estimation through cross validation, visualization and implementation of six basic machine learning algorithms to obtain the results that we will describe in the following sections of this paper.

II. Models and Methods

A. Data Preprocessing

The provided dataset contains **328'135 samples** and **322 features** of numerical data. After reviewing the dataset codebook [3],[4], we exclude certain features that were unrelated to our training objectives. Given that most features are categorical, we convert those that weren't such as the weight and height into be categorical bins using quantiles for approximately equal distribution. Missing values are handled separately by assigning them a distinct category. We also map the target variable to $\{0, 1\}$ to be suitable for logistic regression.

B. Class Imbalance Correction

The target variable y is highly imbalanced, with only 8.83% of samples labeled positive. To address this, we apply an oversampling technique to increase the number of samples in the minority class. This approach aims to balance the class distribution and to mitigate any bias towards the majority class during model training. We over-sample the minority class by duplicating randomly selected instances from the CVD class.

C. Encoding Categorical Features

We implement a one hot encoding to transform categorical variables into binary columns, creating separate columns for each unique value to ensure that no ordinal relationships are inferred. We use it for logisitic regression in particular because it is not naturally suited to categorical data.

D. Feature Selection and Optimization

1) Normalized Mutual Information: To identify features most relevant for predicting cardiovascular disease (CVD), we calculate the normalized mutual information (NMI) between each feature and the target variable. Features with an NMI above 0.009 are retained to balance relevance and model complexity.

2) Point-Biserial Correlation: For binary target variables, we use Point-Biserial Correlation to assess feature importance. This method compares the mean feature values across target classes, adjusting for class proportions and standard deviation. Due to the categorical nature of features, this metric is less informative than with continuous variables. Features with an absolute correlation above 0.05 are retained.

E. Cross-validation

To evaluate the generalizing capabilities of the model, we use k-fold Cross-Validation which consists on dividing the training set into k parts and using each part as a validation set while the remaining $k-1$ parts are used as the training set. This process is repeated k times, with each fold taking a turn as the validation set. The model's performance is then averaged across these k iterations. This technique helps to reduce overfitting and gives a better indication of how the model will perform on unseen data.

F. Algorithms Used

We consider three classification algorithms that follows certain criteria that makes them suitable for this task : They can handle (binary) classification, they can handle categorical features, they can run in a reasonable time and they can be easily implemented.

- Regularized Logistic Regression

This algorithm models the probability of the binary outcome as a linear function of the features, using a regularization term to prevent overfitting. It finds optimal parameters by minimizing the log-loss function. To adapt it for categorical features, we apply a one-hot encoder, which converts each categorical variable into a set of binary variables, enabling the algorithm to process non-numeric data.

- Categorical Naive Bayes

A probabilistic classifier applying Bayes' theorem under the assumption of feature independence. It estimates class probabilities based on categorical feature values.

- Random Forest

An ensemble learning method that builds multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data and features, adding robustness to the model.

G. Evaluation Metrics

To evaluate the performance of the classification algorithms we use the accuracy, recall, precision and F1-score. These metrics allow us to evaluate not only the model's accuracy, which might be misleading because of class imbalance, but also its capacity to balance false positives and false negatives, a critical aspect in CVD classification where the consequences of incorrect predictions can impact patient outcomes.

III. Results

A. Initial Model Performance

When we didn't fix the class imbalance, all models learned to always predict negative because it guarantees an accuracy of 91%. However this is not useful for CVD prediction and the model gets an F1-score of 0. This is why we need a trade-off between the overall accuracy and the F1-score.

B. Effect of Feature Selection

We compare the model performances using NMI and Point-Biserial Correlation, as summarized in Table I.

Model	Feature Selection	Acc	F1
Categorical Naive Bayes	NMI	0.74	0.33
	Point Biserial	0.74	0.34
Logistic Regression	NMI	0.74	0.35
	Point Biserial	0.76	0.36
Random Forest	Smaller Set	0.72	0.34

Table I
Cross-Validation of NMI vs. Point-Biserial Feature Selection on Test Set (RF on Reduced Set because of computation time)

C. Greedy Feature Selection

NMI and PB-correlation help in feature selection but may lead to suboptimal results if features are highly correlated. To refine selection, we use a greedy approach that adds features iteratively, based on their improvement to the F1-score until no further improvement is observed. Due to its computational cost, we apply this method only with logistic regression on 1% of the training set which led to overfitting.

D. Hyperparameter Optimization

To further optimize the model, we perform hyperparameter tuning using grid search with cross-validation. This technique explores a range of hyperparameters, allowing us to identify the best combination for a given model. Due to its computational expense, we apply this method only to the best-performing model, evaluating two sets of hyperparameters.

E. Final Model Retained

For the final submission, we use the logistic regression on the features selected by the PB-correlation and with parameters $\gamma = 0.2$ and $\lambda_{=}0.001$. The scores obtained on AIcrowd are : F1=0.369 and acc=0.761 (id=#274795)

IV. Discussion

A. Interpretation of Model Performances

First we can observe that most models have very similar results. We also try ridge regression (by converting regression outputs to binary predictions with using a threshold 0.5), it leads to the same results (Acc:0.74, F1:0.35). This is probably due to the highly imbalanced nature of the dataset and the relatively limited predictive power of the available features. Since many features have low normalized mutual information and Point-Biserial correlation with the target, it's likely that they do not strongly differentiate between those with and without cardiovascular disease (CVD). Additionally, the oversampling strategy, while helpful in increasing the F1-score, may have also introduced some noise, as the duplicated samples do not contribute new information and could lead to overfitting on the minority class.

B. Importance of Data Preprocessing and Balancing

Data preprocessing and balancing are critical steps to enhance model performance. Given an initial training dataset with only 8.83% positive cases, the model showed a tendency to predict negative outcomes exclusively. By employing random oversampling of the positive class, we enabled the model to identify positive cases in specific instances, thereby improving its predictive balance.

C. Limitations and future improvement

Our models, though simple to prevent overfitting, may have underfit the data, as indicated by similar training and validation scores. Limited computational power prevented comprehensive grid searches, which may have hindered performance, as did using handmade models without optimizations from established libraries. Additionally, using AUROC as a performance metric could have improved our ability to assess model discrimination and optimize thresholds for our medical application.

V. Summary

This project focuses on predicting cardiovascular disease (CVD) using BRFSS data. We applied data preprocessing, class balancing, and various classification algorithms. Key improvements were achieved through addressing class imbalance and optimizing feature selection. Logistic Regression emerged as the best-performing model, highlighting the potential of reliable classification models to support clinical decision-making in CVD prediction.

REFERENCES

[1] A. Martin Jaggi and B. Nicolas Flammarion, *CS-433: Machine Learning* , [online] Available: https://github.com/epfml/ML_course.

[2] AIcrowd Machine Learning Project 1, [online] Available: https://www.aicrowd.com/challenges/epfl-machine-learning-project-1.

[3] Behavioral Risk Factor Surveillance System, *2015 Codebook Report: Land-Line and Cell-Phone Data*, Centers for Disease Control and Prevention, August 23, 2016. [Online]. Available: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.

[4] "Summary matrix of calculated variables," Respiratory Care, 2015. [online]. Available: https://www.cdc.gov/brfss/annual_data/2015/Summary_Matrix_15_version12.html.