Predicting Health Outcomes from the BRFSS 2015 Dataset

Julien Erbland Max Henrotin Francesca Fino

October 2025, EPFL Lausanne

1 Introduction

In this project, we apply fundamental machine learning techniques to predict the risk of cardiovascular disease based on health and lifestyle data. Our goal is to build and evaluate classification models using methods such as linear and logistic regression, while performing data exploration, preprocessing, and feature engineering to improve performance.

2 Data Preprocessing

The original dataset had 321 features, many of which were irrelevant or badly formatted. To avoid noisy data, we manually kept only the useful ones based on the official documentation [1]. The selected features and their corresponding invalid values (e.g., `999.99kg`) and zero equivalents were stored in a JSON file for cleaning.

We removed columns with more than 60% missing values. For the rest, missing data was filled with the most common value (for small categorical features) or the mean (for continuous ones). After cleaning, we reduced it to **49 features**.

To improve performance, we applied three transformations: **polynomial expansion** (to capture nonlinear patterns), **outlier clipping** (to reduce extreme values), and **interaction terms** (to model dependencies between features).

The table below shows the results of training regularized logistic regression models with the Newton optimization method on different configurations. We measured performance using the F1-score on a test split.

Transformation	F1-score
Baseline	0.4311
Norm	0.4366
Norm and bias	0.4392
Poly expansion (deg=3)	0.4452
Interactions	0.4467
Outlier Clipping	0.4465
All Steps	**0.4481**

Table 1: Impact of each preprocessing step on F1-score

The results show that all features steps lead to small improvements in F1-score. While normalization had a major positive impact on the performance of other models, its isolated effect was smaller for the regularized logistic regression trained with Newton's method. The optimal polynomial degree was determined empirically. The best configuration combined all preprocessing steps, achieving the highest overall F1-score.

Final Preprocessing Pipeline :

Feature selection → Replace NaNs/0s → Drop high-missing → Imputation → Outlier clipping → Polynomial expansion → Normalization → Add bias

3 Methodology

Models: To find the best prediction model, we explored and implemented different machine learning methods: linear regression (GD, SGD), least squares, ridge regression, logistic regression, and regularized logistic regression. Each algorithm outputs final weights and the loss $\mathcal{L}(w)$, mean-squared error for linear regression, log-loss for logistic regression, and a penalized version of these for regularized variants. For the regularized logistic regression, we added the **Newton variant**, where the weight update uses second-order information from the Hessian of the loss:

$$w^{(t+1)} = w^{(t)} - H^{-1}(w^{(t)})\,\nabla\mathcal{L}(w^{(t)}),$$

with $H(w^{(t)}) = \nabla^2\mathcal{L}(w^{(t)})$ the Hessian matrix. This allows faster convergence than gradient descent but has higher computational complexity.

Evaluation: Because the dataset is highly imbalanced (∼90% negative class), we used the **F1-score** as our main evaluation metric. It balances precision and recall and is therefore less affected by class imbalance than accuracy.

The F1-score is defined as :

$$F_1 = \frac{2PR}{P+R}, \quad P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}.$$

with P the precision and R the recall.

Unlike accuracy, which can stay high even when the model ignores the minority class, the F1-score reflects how well positive (rare) cases are actually detected.

Linear models mostly predicted the majority class and thus achieved $F_1 \approx 0$, whereas logistic models captured minority cases more effectively. Table 2 show the importance to rely more on the F1-score than the accuracy. We relied solely on logistic models thereafter.

Model tuning and validation: We optimized the regularization term λ and the learning rate γ using a grid search combined with **stratified k-fold cross-validation**.

Formally, the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ is partitioned into K disjoint folds $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$ such that class proportion is preserved across folds:

$$P(y = 1 \,|\, \mathcal{D}_k) \approx P(y = 1 \,|\, \mathcal{D}), \quad \forall k \in \{1, \ldots, K\}. \quad (1)$$

This ensures that each fold reflects the same class distribution as the full dataset, leading to more stable and representative validation results. We also introduced an **adaptive decision threshold** chosen to maximize the F1-score on each fold, rather than using the default 0.5 threshold, in order to better handle class imbalance.

Model	Acc.	F1
Lin. Reg. (GD)	0.914	0.000
Lin. Reg. (SGD)	0.914	0.000
Least Squares	0.915	0.098
Ridge (λ=10^{-3})	0.914	0.066
Log. Reg. (γ=0.3)	0.880	0.429
Reg. Log. (λ=10^{-4})	0.879	0.439
Reg. Log. (Newton)	0.873	**0.442**

Table 2: Accuracy and F1-score of all models. Linear models reach high accuracy but zero F1; logistic ones capture the minority class.

Final predictions: For each fold, we recorded the F1-score, accuracy, loss, and predicted labels. *The final prediction for each sample was obtained by taking the most frequently predicted label across folds.* By averaging metrics over all folds, we got a more stable and reliable estimate of model performance, reducing the randomness coming from a single train–test split.

4 Results

The dataset was very imbalanced (about 90% negative cases), and only logistic models could handle this properly. At first, F1-scores were below 0.3, showing that the models failed to detect most positive cases. After applying preprocessing steps, the F1-score improved to around 0.4, but results still changed a lot depending on how the data was split. To make the evaluation more stable, we used the stratified k-fold method defined in Eq.1. This made the results more reliable, with F1-scores staying consistently in a $\Delta = 0.03$ range.

Table 3, show how the mean F1-score across folds also closely match the test set F1.

Metric	F1-score
Mean F1-score (cross-validation)	0.435
F1-score (hold-out test set)	0.429

Table 3: Comparison between cross-validation and hold-out test performance. The small gap ($\Delta = 0.006$) shows minor overfitting probably due to the adaptative threshold introduced in the k-folding.

In addition to providing a more stable evaluation, k-folding allows us to compute the mean of the prediction vectors across folds, which can give better results than the mean F1-score, particularly when we trained on bigger datasets because no overfitting occured
Overall, the final results are consistent, showing low variance across samples: with Newton's method over $K = 10$ folds, we could consistently get $F_1 \in [\mathbf{0.41}, \mathbf{0.44}]$ with convergence considered satisfactory when the loss variation fell below 10^{-8}, with $n_{\text{iter}} < 50$.

Hyperparameter tuning We chose $\gamma \approx 0.3$ and $\lambda \approx 10^{-4}$. Figures 1 and 2 show the cross-validation results for our two regularized models.

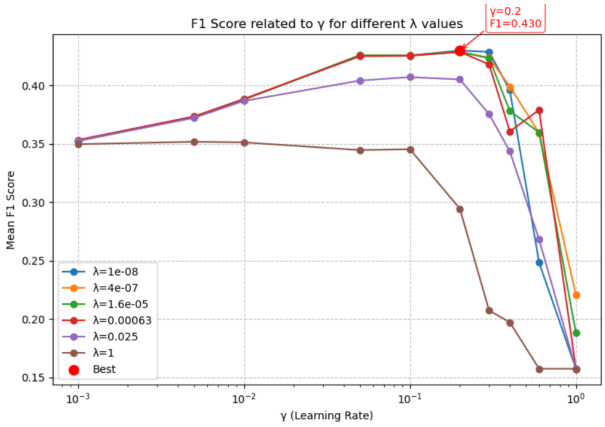

Figure 1: Fold-wise F1-scores for the regularized logistic model (Gradient Descent).

We can see that, in the gradient descent version, the learning rate is a crucial parameter, whereas the regularization strength has a smaller impact (Fig.1).

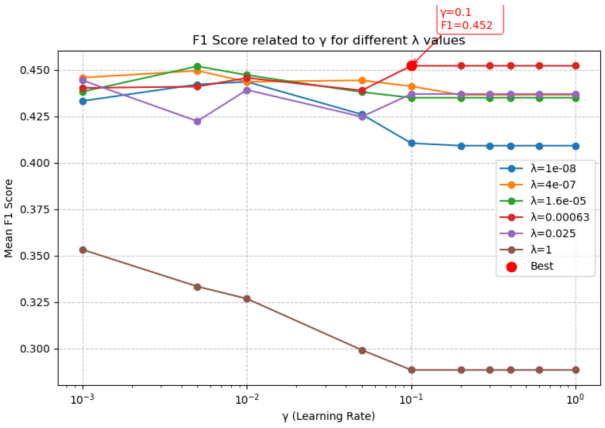

Figure 2: Fold-wise F1-scores for the regularized logistic model (Newton's Method).

In contrast, for Newton's method, λ mainly influences performance, while the learning rate mostly affects convergence speed, as this method converges rapidly and consistently across a wide range of γ values near 1(Fig.2).

5 Conclusion

After testing several models, **regularized logistic regression with the Newton method** gave the best results. Even if its performance was comparable to the gradient-based variant, we selected it as the best solution for its faster convergence. With all preprocessing steps applied (normalization, polynomial features, and outlier clipping), it reached a **F1-score of 0.429** on AICrowd.

References

[1] Centers for Disease Control and Prevention. *2015 BRFSS Codebook*. https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf