Diabetes Risk Screening Using Survey-Based Health Indicators

Author: Santiago Boccardo
Date: December 30th 2025
Dataset: The dataset used in this project is publicly available and open for educational and non-commercial use. It was obtained from Kaggle and can be accessed here: Diabetes dataset

Executive Summary

This project is part of my data science portfolio and focuses on the analysis of a large, publicly available population health survey dataset. The goal is to assess whether self-reported health, lifestyle, and demographic indicators can be used to identify individuals at increased risk of diabetes, framing the problem as a binary risk screening task rather than a clinical diagnosis.

The analysis includes data cleaning, exploratory data analysis, and supervised machine learning, combining an interpretable baseline model with a more flexible non-linear approach. Model performance is evaluated with an emphasis on sensitivity and robustness, highlighting the trade-offs between interpretability and predictive power in health risk screening applications.

Objective

- Evaluate the predictive value of survey-based health and lifestyle indicators for diabetes risk screening.
- Compare an interpretable linear model with a more flexible non-linear model.
- Analyze the trade-offs between model interpretability, sensitivity, and predictive performance.

Dataset Description

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC), collecting self-reported information on health behaviors and chronic conditions from over 400,000 U.S. adults each year.

This project uses a cleaned subset of the BRFSS 2015 data, originally comprising 441,455 responses and 330 features, made available on Kaggle. The final dataset includes 253,680 survey responses and 21 selected features capturing demographic, lifestyle, and health-related indicators. The original target variable contained three classes (non-diabetic, prediabetic, and diabetic), which were merged into a binary outcome during this analysis to frame the task as a diabetes risk screening problem. As expected for population-based survey data, the resulting target variable is highly imbalanced.

Methodology

Tools used:

- **Google Colab** with Python for data analysis and modeling
- **GitHub** for hosting and project presentation

Data preparation:

1. **Data loading and initial review**
2. **Data Cleaning and Preprocessing**:
    - Verification of missing values and duplicated observations.
    - Renaming of variables and categories to improve readability and interpretability.
    - Exploratory visualization of numerical and categorical variables to assess distributions and class-specific patterns.

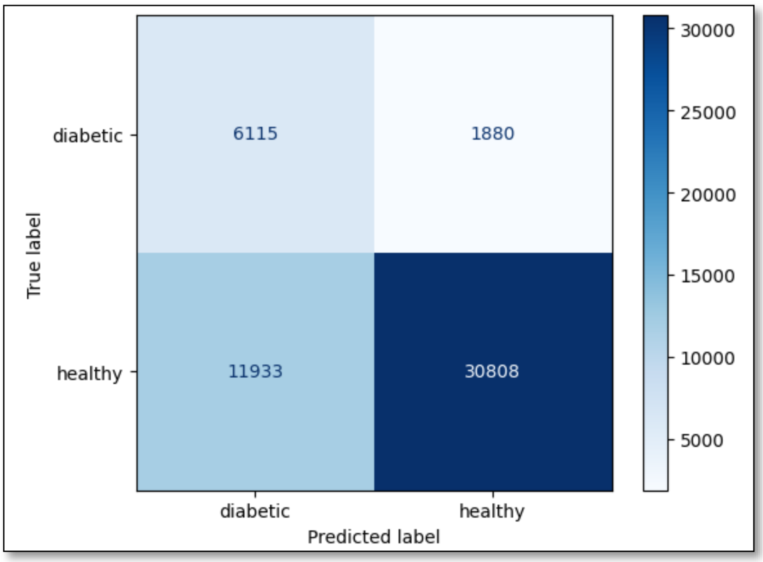
Data modeling

Two supervised classification models were trained and evaluated primarily using ROC-AUC and recall for the diabetic class, reflecting the priorities of a screening-oriented use case.

- **Logistic Regression:**

Used as an interpretable baseline model, incorporating class-weighted loss functions and regularization. Model coefficients were analyzed to identify the most influential risk factors.

Confusion matrix:



After hyperparameter tuning with cross-validation and class-weighted loss functions to address class imbalance, the model achieved a ROC-AUC of approximately 0.82, indicating good overall ability to distinguish between individuals with and without diabetes risk.

As illustrated by the confusion matrix, the model was optimized to favor recall for the diabetic class, meaning it prioritizes correctly identifying individuals who are diabetic or prediabetic. In practical terms, recall measures the proportion of truly at-risk individuals that the model successfully detects (0.76 in this case). This results in fewer missed diabetic cases (false negatives), at the expense of a higher number of healthy individuals being incorrectly flagged as at risk (false positives).

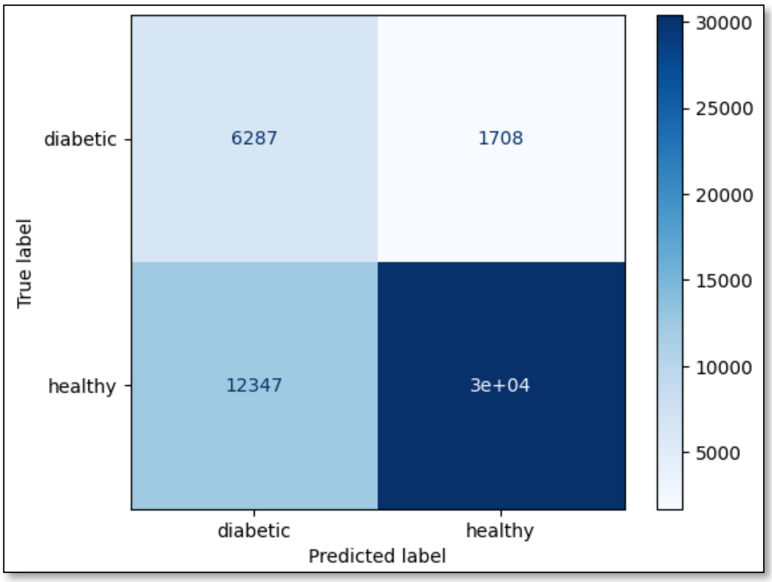
This behavior is well aligned with a screening-oriented use case, where failing to identify individuals who may require medical follow-up is more costly than recommending additional testing to some healthy individuals.

To support interpretability, model coefficients were analyzed and ranked by their relative importance. The most influential predictors were associated with older age groups, poorer self-reported general health, and known clinical risk factors, reinforcing the suitability of logistic regression as a transparent and robust baseline model for diabetes risk screening.

- **CatBoost Classifier:**

A gradient boosting model capable of capturing non-linear interactions between categorical and numerical features. Hyperparameters were tuned using cross-validation, and feature importance scores were used to assess variable relevance.

Confusion matrix:



CatBoost achieved a slightly higher ROC-AUC (≈ 0.83) and recall for the diabetic class (≈ 0.79) compared to logistic regression, resulting in a modest reduction in false negatives. As shown by the confusion matrix, this improvement reflects a shift toward higher sensitivity rather than an increase in overall accuracy or precision.

Similar to the logistic regression model, this behavior aligns well with screening-oriented objectives, where identifying as many at-risk individuals as possible is prioritized over minimizing false positives. However, the observed performance gains remain limited in magnitude.

Feature importance analysis identified self-reported general health status, age group, body mass index, and cardiovascular risk factors as the most influential predictors. These findings are consistent with both epidemiological evidence and the feature patterns observed in the linear model, suggesting that CatBoost captures limited additional non-linear structure beyond what is already modeled by logistic regression.

Model comparison and conclusions

Both models achieved similar ROC-AUC values (≈ 0.82) on both training and test datasets, indicating no strong evidence of overfitting in any case. The comparable performance between logistic regression and CatBoost suggests that introducing higher-order nonlinearities does not provide a substantial performance gain, and that much of the predictive signal in the available feature space can be captured by a linear decision boundary. Given this context, logistic regression emerges as the most appropriate model for this dataset, offering strong interpretability, stable performance, and ease of deployment—key requirements in medical and public health settings where transparency and auditability are essential. CatBoost provides a modest improvement in recall for the diabetic class, reducing false negatives at the cost of a comparable increase in false positives. While this trade-off may be advantageous in screening-oriented scenarios—where missing at-risk individuals is particularly costly—it comes with reduced interpretability and greater model complexity. Importantly, in a screening context, model behavior can be further adapted by calibrating the classification threshold to prioritize sensitivity for the diabetic class, depending on operational constraints and downstream capacity. Such calibration allows balancing recall and false positives without requiring changes to the underlying model. Overall, these results reinforce logistic regression as the preferred approach when robustness, transparency, and practical deployability are prioritized, with CatBoost representing a viable alternative for recall-focused screening use cases.

An important implication of this analysis is that meaningful diabetes risk screening can be achieved using a relatively small subset of survey questions. Despite the original BRFSS questionnaire comprising several hundred variables, the reduced set of 21 features analyzed in this project achieved strong discriminative performance. This finding suggests that a short-form survey may retain much of the predictive signal while substantially reducing respondent burden.

While incorporating additional features from the full BRFSS questionnaire could potentially yield incremental performance improvements, the observed results suggest diminishing returns when weighed against increased model complexity, reduced interpretability, and higher data collection costs. Future work could formally evaluate this trade-off through feature selection strategies or incremental model expansion, depending on the intended application and operational constraints.

In a practical screening setting, model outputs can be expressed as predicted probabilities rather than binary labels, allowing cases to be prioritized based on confidence. This enables downstream calibration of decision thresholds to balance sensitivity and false positive rates according to operational capacity.

About the author of this report

Santiago Boccardo

Data Analyst / Data Scientist
Biochemist & PhD in Chemical Sciences

Health sciences researcher with 8+ years in immunology and data analysis. Currently transitioning into data science to support evidence-based decision-making.

LinkedIn | GitHub