Predicting diabetes from health indicators and lifestyle factors

Shi Fan Jin, Amanpreet Binepal, Ian Gault, Vy Phan

2025-12-05

Table of contents

Data Reference: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Summary

This project attempts to predict diabetes status using the Logistic Regression and LinearSVC models, against a baseline DummyClassifier on an imbalanced dataset. All models achieved similar accuracy on the test set (approximately 0.86), which highlights a key issue: accuracy alone is not a reliable performance metric.

These findings motivate deeper exploratory data analysis, evaluation with additional metrics (precision, recall, F1), and exploration of alternative models and threshold tuning to get a more robust assessment of the model's predictability.

Introduction

Diabetes is a chronic disease that prevents the body from properly controlling blood sugar levels, which can lead to serious health problems including heart disease, vision loss, kidney disease, and limb amputation (Teboul, 2020). Given the severity of the disease, early detection

can allow people to make lifestyle changes and receive treatment that can slow disease progression. We believe that machine learning models using survey data can offer a promising way to create accessible, cost-effective screening tools to identify high-risk individuals and support public health efforts.

Research Question

Can we use health indicators and lifestyle factors from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) survey to accurately predict whether an individual has diabetes?

We are looking to : 1. Build and evaluate classification models that predict diabetes status based on 21 health and lifestyle features 2. Compare the performance and efficiency of logistic regression and support vector machine (SVM) classifiers 3. Assess whether survey-based features can provide sufficiently accurate predictions for practical screening applications

Methods & Results

This analysis uses the diabetes_binary_health_indicators_BRFSS2015.csv dataset, a cleaned and preprocessed version of the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey data, made available by Alex Teboul on Kaggle (Teboul, 2020).

For this analysis, we split the dataset into training (80%) and testing (20%) sets using a fixed random state (522) to ensure reproducibility. We implemented two classification algorithms:

1. Logistic Regression: A linear model appropriate for binary classification that estimates the probability of diabetes based on a linear combination of features.
2. Linear Support Vector Classifier (SVC): A classifier that finds an optimal hyperplane to separate diabetic from non-diabetic individuals.

Both models were implemented using scikit-learn pipelines that include feature standardization (StandardScaler) to normalize the numeric features to comparable scales. Binary categorical features were already processed in the dataset and were set to pass through the column transformer. We evaluated model performance using cross-validation on the training set and final accuracy assessment on the held-out test set.

Our results show that both models achieve approximately 86% accuracy, with logistic regression demonstrating slightly faster training time.

Discussion

The exploratory data analysis revealed several important patterns in the dataset. First, the target variable (diabetes_binary) is highly imbalanced (see Figure 1). This imbalance also appears across ordinal variables such as age, education, and income (see Figure 2).

Additionally, individuals with diabetes (diabetes_binary = 1) tend to have higher BMI values on average (see Figure 3).

Other binary health and lifestyle factors—such as high blood pressure, smoking status, and physical activity—also show clear differences between diabetic and non-diabetic groups (see Figure 4).

Together, these patterns suggest that several features are related to diabetes status, but the dominance of the majority class may obscure these relationships when evaluating model performance.

The baseline DummyClassifier achieves an accuracy score of about 0.86, derived from assigning the most frequent class (non-diabetic) to all patients. This highlights how approximatey 86% of the dataset is non-diabetic. Both Logistic Regression and LinearSVC achieve similar accuracy (approximately 0.86) with little to no improvement.

As revealed from the exploratory data analysis, the class imbalanced problem may affect the models' reliability. Therefore, more analysis is needed to explore additional models, check class balance with metrics such as precision and recall, examine confusion matrices, and test different data splits or tune hyperparameters to determine if performance is stable across scenarios before drawing strong conclusions.

The similarity in test scores is an unexpected finding. With a clean dataset containing informative and diverse features, we would expect the classification models to perform at least better than the dummy classifier. Additionally, initial hyperparameter tuning for logistic regression did not affect accuracy (data not shown). This finding highlights the importance of understanding the data through EDA to interpret where accuracy scores come from.

This suggests the next step for deeper EDA, including distributions, to see whether features overlap and whether the model can separate them effectively. Other future questions would be determining which features are most important for classifying an individual as diabetic or not, evaluating the probability estimates, and assessing whether all features are truly helpful for drawing conclusions.

References

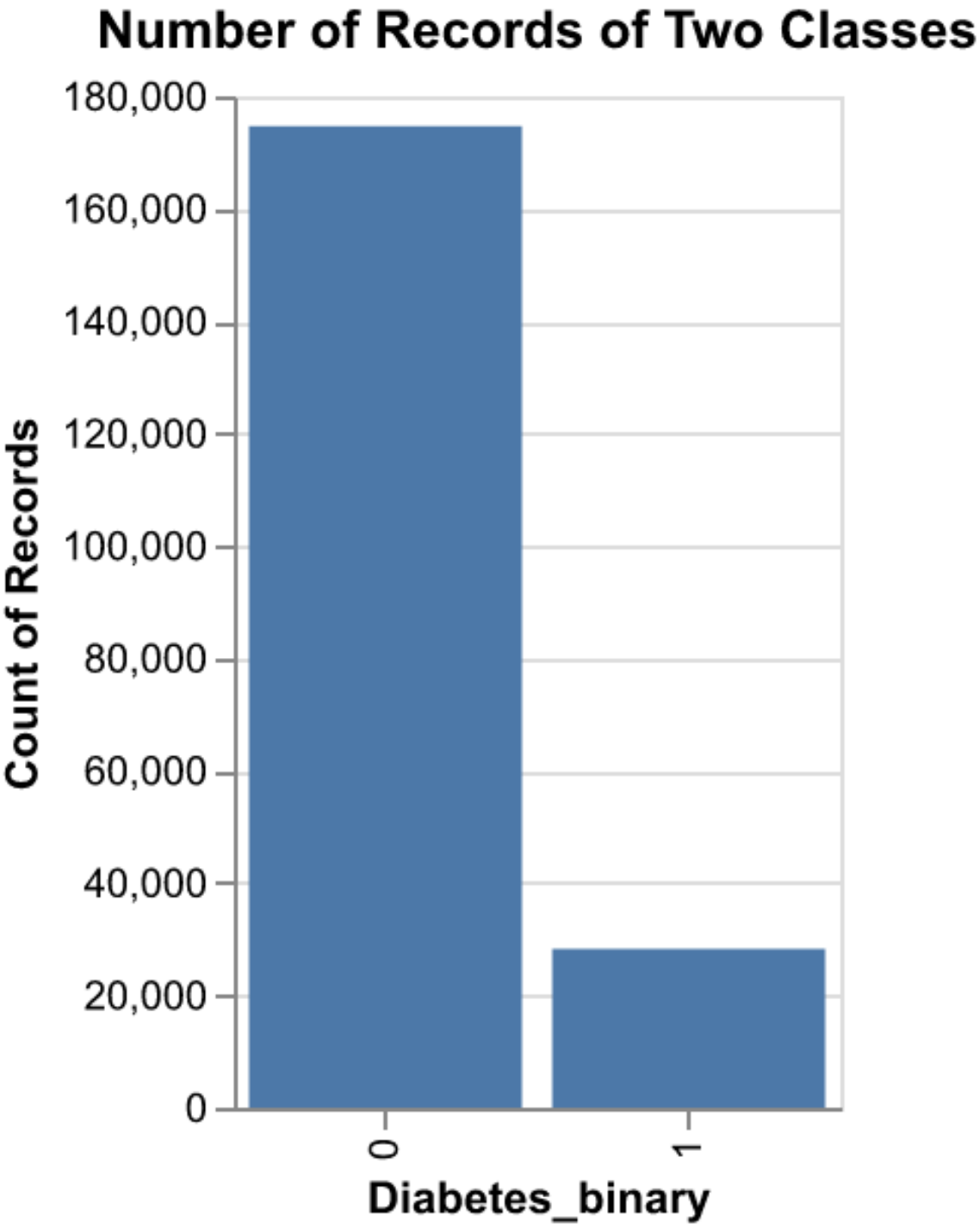

Figure 1: Imbalanced Class Observations

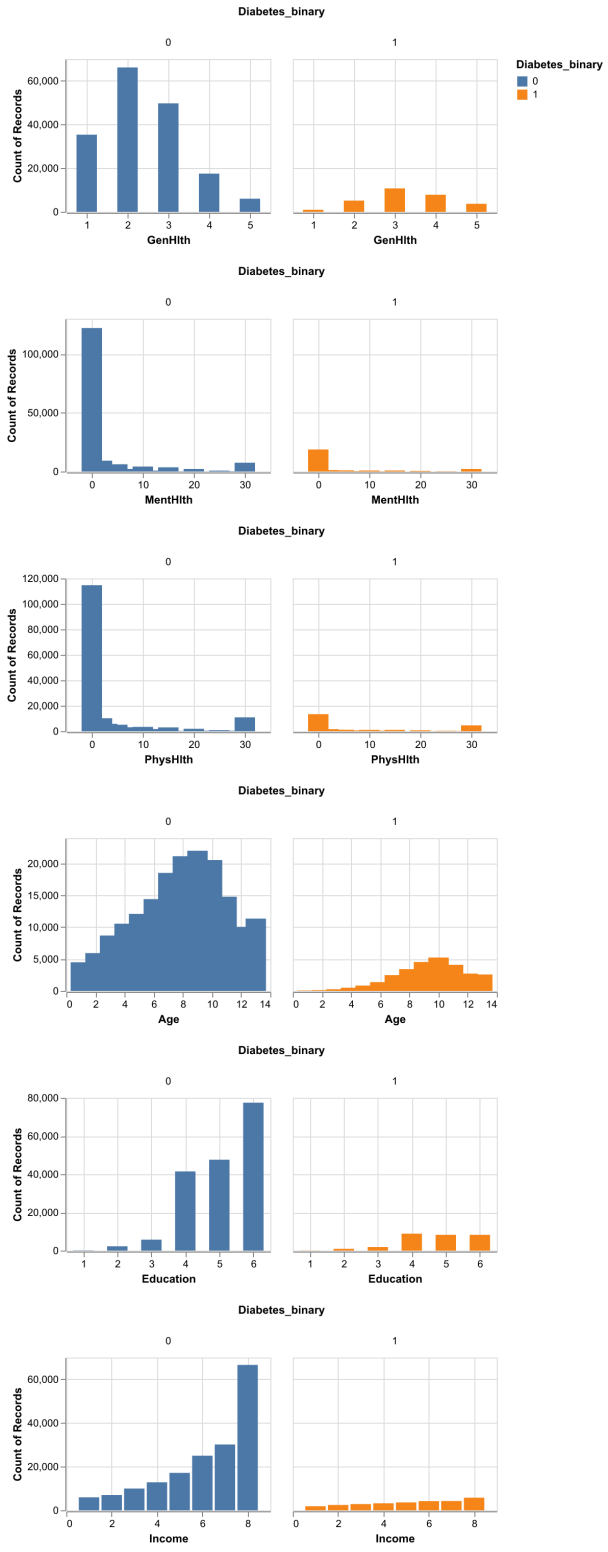

Figure 2: Ordinal Features Observations by Class

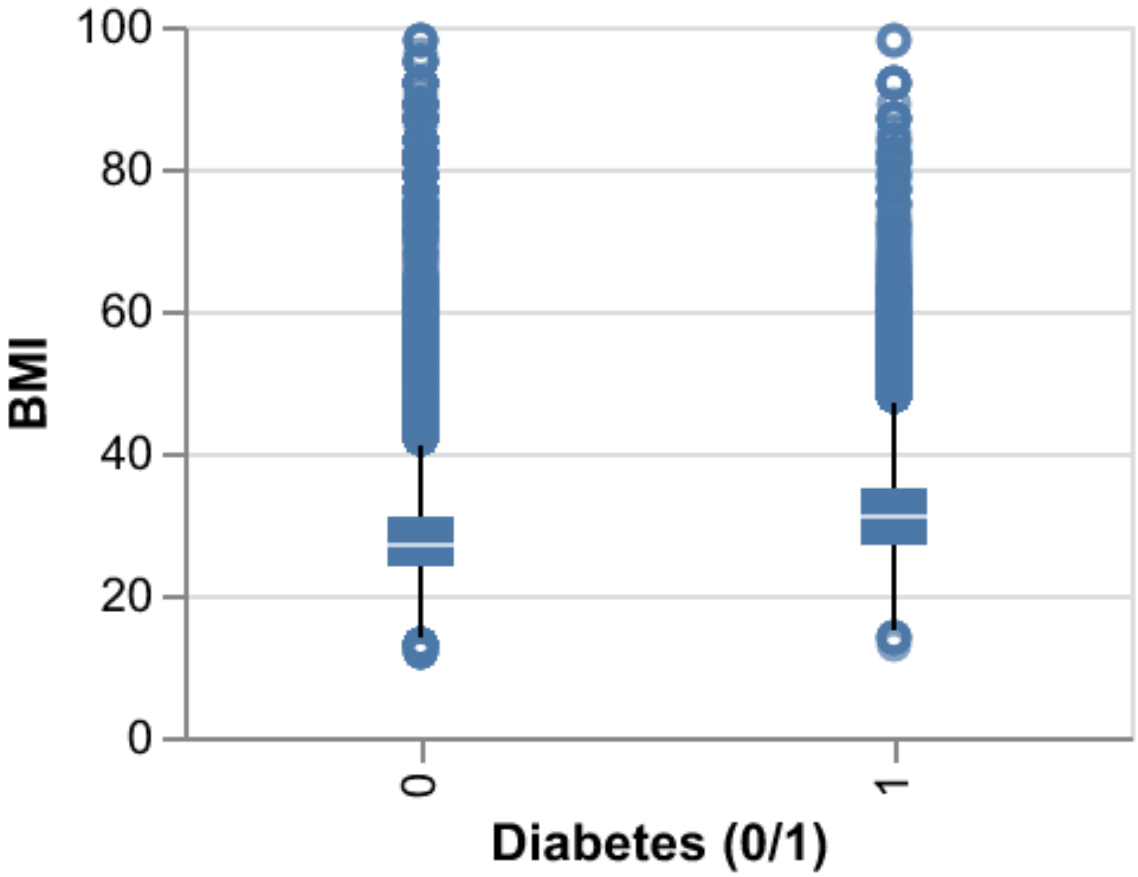

Figure 3: BMI by Class

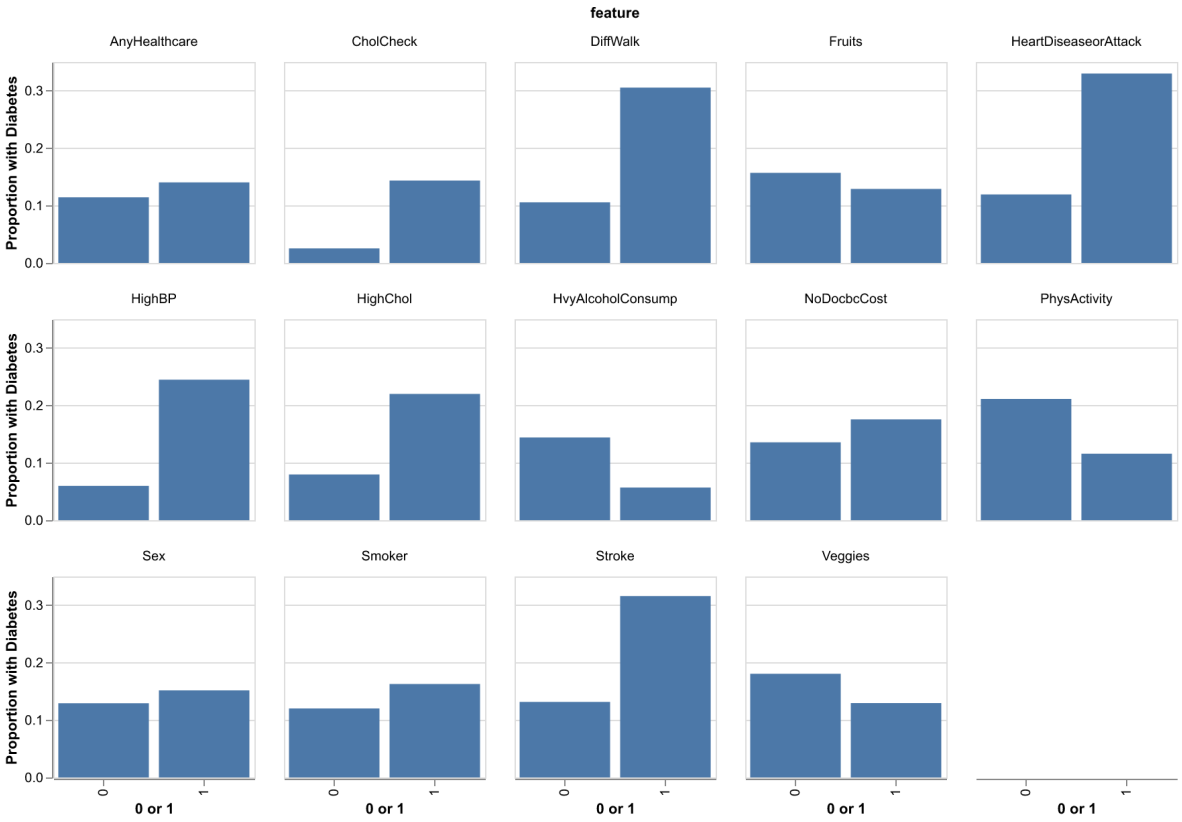

Figure 4: Binary Features by Class