Interpretable and Fair Machine Learning for Diabetes Risk Prediction Using Clinical Data

(Data Science project)

Prepared by:

(CS4-G1)

Murad Beshr	23160018
Abdullah Al-Kabos	23160045
Riyadh Marish	23164062
Abdulsalam Muzafar	23164099
Majed Al-Naayib	21160085

Date:

December 2nd, 2025

1. Abstract

Diabetes is a growing worldwide health challenge where early risk detection enables timely intervention. However, many ML approaches, while producing accurate predictions, remain opaque to the clinician, limiting their adoption in the real world. In this paper, we present an easy-to-reproduce ML pipeline for diabetes risk prediction using the publicly available NHANES 2017-2018 dataset. We combine interpretable and conventional models-calibrated Logistic Regression and XGBoost-and apply homogeneous preprocessing, imputation, and scaling. Making the predictions actionable, we integrated explainable AI, SHAP, for global and per-patient interpretation, evaluated model calibration to obtain reliable risk scores, and assessed model fairness across demographic subgroups: age, gender, and race. Our results show that a calibrated, interpretability-focused pipeline achieves competitive predictive performance while offering clear explanations and clinically useful subgroup differences that will facilitate equitable deployment. We make the code and artifacts publicly available to help ensure reproducibility and foster further validation and clinical translation.

Key words: Diabetes Risk Prediction, Machine Learning, Explainable AI(SHAP), Fairness/Bais Analysis, Healthcare Decision Support

2. Introduction

Diabetes mellitus is one of the most important current public health problems in the world. It inflicts considerable suffering and very high costs on healthcare systems worldwide. Early detection of persons at high risk enables measures that can either prevent or delay the development of the disease. ML has been found to be quite important in diabetes risk prediction due to the fact that it brings to the surface intricate patterns from clinical data of enormous size. In real clinical applications, however, ML models should be interpretable, well-calibrated, and fair among different demographic groups, besides being powerful in predictive ability.

Previous ML studies have often used small or convenience datasets, such as the Pima Indians dataset, and most have employed large, complex models yielding high discrimination scores. However, most of these studies offer limited interpretability, without calibration analysis, and rarely consider fairness in predictions across age, gender, or race. Small differences in pre-processing steps and labeling further limit reproducibility by making the comparison of results, or their translation into clinical practice, non-trivial. Clinicians therefore view black-box models with skepticism due to their uncertain reliability in a diverse range of populations.

These gaps denote persistent challenges: inadequate model transparency, missing calibration assessments, insufficient fairness evaluation, and a lack of reproducible pipelines using large public datasets. Taken together, these issues limit the clinical readiness of ML models for diabetes screening and prevention.

We present here a simple, reproducible ML pipeline using the public NHANES 2017-2018 dataset to address these limitations. We apply calibrated Logistic Regression and XGBoost with standardized preprocessing, evaluate discrimination, calibration, and subgroup performance, and use SHAP to provide both global and per-patient explanations. We release all code and artifacts to ensure reproducibility.

Our contributions are as follows: a complete, reproducible pipeline for NHANES data; a comparison of calibrated Logistic Regression and XGBoost; SHAP-based explainability; extensive calibration and equity analysis; and practical artefacts and advice towards clinical decision support applications.