Progress Report: Predictive Modeling of Diabetes Risk Using Health Behavior Indicators

Frank Ofosu

Clinical Decision Support Modeling

Introduction

Diabetes remains one of the most pressing public health challenges in the United States, affecting more than 34 million individuals and contributing to significant morbidity, mortality, and healthcare costs. Early detection and targeted prevention of diabetes and prediabetes are essential for improving outcomes and reducing economic burden. Traditional screening methods based solely on laboratory measures (e.g., HbA1c, fasting glucose) often fail to capture population-level risk patterns linked to behavioral and lifestyle factors.

This research aims to develop an AI-driven Clinical Decision Support (CDS) model that uses health behavior indicators to predict diabetes risk among adults. By leveraging the *Behavioral Risk Factor Surveillance System (BRFSS 2015)* dataset, the project seeks to build interpretable machine learning models that identify individuals at risk and generate explainable, actionable recommendations for preventive care. The study emphasizes model interpretability to promote clinician trust and to ensure that AI insights are transparent and ethically aligned with patient care goals.

Methodology

Data preparation and model framework:

The project will employ a supervised machine learning approach for binary classification (0 = no diabetes, 1 = prediabetes/diabetes). The initial modeling phase will use the balanced binary dataset from BRFSS 2015, allowing equal representation of both classes for fair baseline training.

The following models will be implemented and compared:

- Logistic Regression: Baseline interpretable linear model to identify significant predictors.
- Random Forest: Ensemble decision tree method capturing nonlinear feature interactions.
- XGBoost: Gradient boosting algorithm for improved accuracy and feature ranking.

Evaluation and Explainability

Models will be evaluated using AUROC, precision-recall (AUPRC), accuracy, F1-score, and calibration plots. The interpretability phase will apply SHAP (SHapley Additive Explanations) to visualize feature contributions to individual predictions. The final CDS prototype will stratify risk into Low, Medium, and High categories and provide clinical action prompts (e.g., "Order HbA1c" or "Recommend lifestyle modification").

Dataset Description

The dataset originates from the CDC Behavioral Risk Factor Surveillance System (BRFSS 2015), a national health-related survey comprising over 400,000 responses on lifestyle, chronic conditions, and preventive service use. It contains three cleaned subsets provided via Kaggle (Alex Teboul, 2021):The dataset contains 3 files and will begin with the balanced binary dataset (Dataset 2), which contains equal proportions of respondents with and without diabetes.

- Dataset 1 – Multi-class (0: none, 1: prediabetes, 2: diabetes)
- Dataset 2 – Balanced binary (50-50 split; 70,692 samples)
- Dataset 3 – Imbalanced binary (real-world distribution; 253,680 samples)

Key Features and Relevance

The features capture both behavioral and clinical risk factors highly correlated with diabetes onset. These indicators are consistent with literature on diabetes determinants and have strong explanatory power for model performance.

- Clinical: BMI, HighBP, HighChol, Stroke, HeartDiseaseorAttack
- Behavioral: Smoking, Physical activity, Fruit/Vegetable intake, Alcohol consumption
- Socioeconomic: Education, Income, Healthcare access
- Perceived health: GenHlth, MentHlth, PhysHlth

Future Work

- Extend models to the imbalanced dataset (Dataset #3) for real-world validation.
- Implement cross-validation and fairness evaluation across subgroups (age, gender, income).
- Enhancing the dataset by including extra data sources, like medication details

References:

Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System (BRFSS) annual data.* U.S. Department of Health & Human Services. https://www.cdc.gov/brfss/annual_data/annual_data.htm

Rahman, S. S., Khatun, F., & Rahman, M. M. (2023). *Diabetes prediction using machine learning and explainable AI. Journal of Healthcare Engineering, 2023,* 1–12. https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388/

Ullah, I., & Lee, Y. (2024). *Robust diabetic prediction using ensemble machine learning. Scientific Reports, 14,* 78519. https://doi.org/10.1038/s41598-024-78519-8

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). *Building risk prediction models for type 2 diabetes using 2014 BRFSS data. Preventing Chronic Disease, 16,* E74. https://doi.org/10.5888/pcd16.190109

Wang, X., Li, H., & Zhang, Y. (2024). A feature optimization study based on a diabetes risk questionnaire. Frontiers in Public Health, 12, 1328353. https://doi.org/10.3389/fpubh.2024.1328353