
PREDICTING DIABETES STATUS FROM SELF-REPORTED HEALTH INDICATORS: PHASE II PROGRESS REPORT

Shon Haskaj

Department of Computer Science
University of Western Ontario
London, ON, Canada
shaskaj@uwo.ca

Tyler Kirk

Department of Computer Science
University of Western Ontario
London, ON, Canada
tkirk4@uwo.ca

Peter Faux

Department of Computer Science
University of Western Ontario
London, ON, Canada
pfaux@uwo.ca

Brendan Karpala

Department of Computer Science
University of Western Ontario
London, ON, Canada
bkarpala@uwo.ca

ABSTRACT

This progress report outlines a proposed framework for predicting diabetes status using self-reported health indicators from the 2015 Behavioral Risk Factor Surveillance System (BRFSS). We frame the task as a binary classification problem distinguishing individuals with no diabetes from those with prediabetes or diabetes. The dataset contains over 253,000 observations with demographic, behavioral, and chronic-condition variables. Our proposed methodology emphasizes interpretable and statistically grounded machine learning techniques aligned with course material, including logistic regression with regularization, decision trees, Random Forests, and gradient-boosted trees (XGBoost). Key considerations include class imbalance, uncertainty quantification, appropriate model selection principles, and evaluation metrics suited to screening contexts. This report describes the planned analytical workflow, modeling approach, and evaluation strategy that will be implemented in the next phase.

Keywords Diabetes prediction · BRFSS · Machine learning · Classification · Regularization · Model selection

1 Introduction

Diabetes poses significant health and economic challenges worldwide. Early identification of individuals at elevated risk enables preventative intervention, but clinical screening can be resource-intensive. Machine learning models trained on population-scale survey data may provide an alternative risk assessment tool by leveraging behavioral and demographic indicators.

The 2015 BRFSS “Diabetes Health Indicators” dataset offers a large, cleaned collection of self-reported health variables suitable for statistical modeling. With over 253,000 respondents and 21 features, the dataset includes information on physical activity, BMI, smoking behavior, mental and physical health status, and chronic conditions. The dataset is highly imbalanced, with the majority of respondents reporting no diabetes.

This progress report proposes a structured and theoretically grounded approach to developing classification models suitable for this dataset. Our direction draws on core DS3000 concepts: uncertainty quantification, probability-based modeling, regularization, decision trees, ensemble methods, maximum likelihood principles, and model selection techniques.

The remainder of the report is organized as follows. Section 2 reviews key literature. Section 3 outlines the proposed methodology. Section 2.1 addresses limitations and risks. Section 4 summarizes next steps.

2 Related Work

Machine learning has been widely applied to diabetes prediction in both clinical and survey-based contexts. Early work using datasets such as the Pima Indians Diabetes dataset explored logistic regression, support vector machines, and decision trees, typically achieving 75–85% accuracy but with limited sample sizes and generalizability.

Large-scale studies using BRFSS data demonstrate the feasibility of developing risk prediction models from behavioral indicators. Xie et al. (2019) evaluated logistic regression, decision trees, Random Forests, SVMs, and boosting methods on the 2014 BRFSS survey, achieving reasonable AUC values and highlighting the importance of recall for public health screening. Nguyen and Zhang (2025) applied similar models to the 2015 BRFSS dataset and emphasized class imbalance handling and model interpretability.

Additional work shows that oversampling techniques such as SMOTE or ADASYN can improve minority-class performance for tree-based ensembles such as XGBoost. The literature converges on several themes relevant to this project: the importance of calibration, the unreliability of accuracy under imbalance, and the advantages of ensemble methods for tabular survey data.

2.1 Limitations and Risks

Key anticipated challenges include:

- **Self-reported measurements:** Potential misreporting and non-random error.
- **Label structure:** Merging prediabetes and diabetes may obscure meaningful distinctions.
- **Class imbalance:** May distort accuracy-based evaluation and complicate model selection.
- **Limited feature diversity:** Lack of clinical biomarkers constrains predictive ceilings.
- **Interpretability trade-offs:** Some ensemble methods offer improved performance at the cost of interpretability.

3 Methodology / Proposed Method

3.1 Dataset Description

The BRFSS dataset contains 253,680 respondents and 21 features derived from the 2015 survey. The target variable `Diabetes_012` includes three categories: no diabetes, prediabetes, and diabetes. For binary classification, we propose merging prediabetes and diabetes into a single “at-risk” class. A secondary balanced version of the dataset (approximately 70,000 rows) is available and will be used only for controlled comparisons, not model selection.

3.2 Data Preprocessing

Proposed preprocessing steps include:

- validation of variable encoding
- one-hot encoding of nominal variables
- preserving ordinal structure for ordered variables
- standardization for scale-sensitive models
- performing all preprocessing *after* train/validation/test splitting to avoid leakage

3.3 Handling Class Imbalance

To address the imbalance, we propose:

- **Class weighting** in logistic regression and SVM
- **SMOTE** applied only within training folds
- **Comparison with the balanced dataset** for sensitivity analysis

3.4 Modeling Approach

In alignment with course material, we propose evaluating the following models:

- Majority-class baseline
- Logistic Regression (L1, L2, Elastic Net regularization)
- Support Vector Classifier (linear and RBF)
- Decision Tree Classifier
- Random Forest (bagging)
- XGBoost (gradient boosting)

Hyperparameter selection will follow DS3000 principles:

- grid search over regularization strengths, tree depth, kernel width, learning rate
- 5-fold stratified cross-validation
- selection guided by validation performance and penalized model complexity

3.5 Evaluation Metrics

Consistent with the course and the screening context, we propose:

- Precision–Recall AUC
- Recall (sensitivity) of the positive class
- F1-score
- ROC-AUC

Threshold tuning will be used to align model decisions with screening priorities.

4 Conclusion

This report presents a comprehensive proposal for modeling diabetes risk using BRFSS self-reported health indicators. The outlined methodology draws directly from DS3000 principles, incorporating probability-based modeling, regularization, decision tree methods, ensemble techniques, and model selection strategies. In the next phase, we will implement this framework, compare models under consistent metrics, examine the effects of class imbalance handling, and prepare well-calibrated, interpretable results for presentation.

5 AI Disclosure

This report was prepared with limited assistance from AI tools (OpenAI ChatGPT 5 Pro). The model was used for the following purposes:

- Converting written project content into LaTeX format to ensure proper structure and formatting.
- Locating and summarizing publicly available related work to support the literature review section.
- Reviewing and skimming the course lecture material to help ensure that the proposed modeling pipeline, evaluation choices, and methodology align with concepts covered in the DS3000 course.
- Refining the clarity and readability of the write up.

All decisions regarding methodology, model selection, interpretation, and writing were made by the authors. The authors verified the accuracy, relevance, and appropriateness of all AI-generated text and sources. No AI tools were used to generate results, perform data analysis, or produce any part of the project's empirical work

References

- [1] Centers for Disease Control and Prevention (CDC). *BRFSS 2015 – Diabetes Health Indicators Dataset*. Kaggle, 2017. Available at: <https://www.kaggle.com/datasets/alextreboul/diabetes-health-indicators-dataset>.
- [2] Xie, Z., Nikolayeva, O., & Lee, J. M. (2019). “Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques.” *Preventing Chronic Disease*, 16:190109. Available at: https://www.cdc.gov/pcd/issues/2019/19_0109.htm.
- [3] Nguyen, B. & Zhang, Y. (2025). “A Comparative Study of Diabetes Prediction Based on Lifestyle Factors Using Machine Learning.” *arXiv:2503.04137*. Available at: <https://arxiv.org/abs/2503.04137>.
- [4] Tasin, M., Rahman, M., & Hossain, M. (2020). “Diabetes Prediction on Pima Indian Dataset Using Machine Learning Techniques.” Referenced in Nguyen & Zhang (2025). Available via arXiv mirror: <https://arxiv.org/>.