Screening for Diabetes Risk with Survey Indicators: A Predictive Modeling Study Using BRFSS 2015

(COMP3125 Individual Project)

Peter Albo
Wentworth Institue of Technology

***Abstract*- Early identification of adults with diabetes or prediabetes is a public health priority, but clinical screening can be costly and difficult to scale. This project evaluates whether self-reported survey responses can identify individuals at elevated risk using the CDC Behavioral Risk Factor Surveillance System data from 2015. The dataset contains 253,680 records with one target variable indicating diabetes or prediabetes and twenty-one health and demographic indicators. I trained and compared models that are commonly used on tabular health data. A class-weighted logistic regression achieved strong discrimination with a receiver operating characteristic area under the curve of 0.82 and a precision–recall area under the curve of 0.39. A random forest achieved similar ranking performance and, after tuning the decision threshold, delivered higher sensitivity suitable for screening use. I then reduced the input set to ten and five top predictors to test whether a short-form questionnaire could approach the performance of the full feature set. The ten-item model retained most of the accuracy, while the five-item model offered very high sensitivity with an expected loss in precision. These results suggest that short survey instruments can provide practical first-pass screening for diabetes risk at population scale.**

Keywords- diabetes risk, BRFSS, logistic regression, random forest

I. Introduction

Diabetes affects millions of adults in the United States and imposes large clinical and economic burdens. Many adults live with undiagnosed diabetes or prediabetes, and earlier detection can enable counseling and preventive care that reduce complications. Because universal laboratory testing is expensive, there is sustained interest in low-cost risk screening tools that rely on information people can report easily, such as age, weight, existing conditions, activity, and access to care.

This report studies whether survey responses from the Behavioral Risk Factor Surveillance System can be used to predict diabetes status at the individual level and, if so, which questions matter most. Using the 2015 BRFSS diabetes health indicators file, I frame the task as a binary classification problem: identify respondents with diabetes or prediabetes from twenty-one health and demographic indicators. I compare a class-weighted logistic regression to an ensemble tree method and evaluate models with metrics suited to imbalanced outcomes, emphasizing receiver operating characteristic area under the curve and precision–recall area under the curve. Because screening programs typically value sensitivity, I explicitly tune decision thresholds and report the trade-offs among sensitivity, precision, and specificity.

A second objective is practicality: can a short set of questions perform nearly as well as the full instrument? To answer this, I rank predictors by importance from the ensemble model and retrain models on the top ten and top five features, comparing their performance to the full model. This approach supports two outcomes: an interpretable baseline model that performs competitively on tabular survey data, and an evidence-based short-form screener that preserves high sensitivity with far fewer questions.

II. Datasets

A. Source of dataset

The dataset used in this project is derived from the Behavioral Risk Factor Surveillance System (BRFSS), an ongoing health-related telephone survey system conducted by the U.S. Centers for Disease Control and Prevention (CDC). BRFSS collects uniform, state-specific data on preventive health practices and risk behaviors linked to chronic diseases, injuries, and preventable infectious diseases in the adult population. Topics include tobacco use, health care coverage, HIV/AIDS knowledge or prevention, physical activity, and dietary habits. Data are collected annually from a random sample of adults, one per household, via structured telephone interviews.

The original 2015 BRFSS dataset contains over 400,000 observations and approximately 330 variables. This work uses a reduced version of the dataset published on Kaggle [1], [2] that selects 21 features most relevant to diabetes prediction. Irrelevant variables were removed to focus on health indicators and demographic attributes that have established links to diabetes risk. The reduced dataset maintains the same respondent records as the original source but in a cleaner, more manageable form for modeling purposes.

B. Character of the datasets

The reduced 2015 BRFSS diabetes health indicators dataset is stored in CSV format and contains 253,680 rows and 22 columns, including one binary target variable (Diabetes_binary) and 21 predictor variables. All variables are numeric and either continuous (e.g., Body Mass Index) or categorical encoded as integers (e.g., 0/1 for yes/no responses, ordinal codes for age and income ranges). The target variable is imbalanced: approximately 13.9% of respondents reported having diabetes or prediabetes.

Table 1 summarizes the dataset's variables, types, and brief descriptions. No additional cleaning was required beyond loading the CSV file, as the reduced dataset contains no missing values and is pre-encoded for machine learning. The primary preprocessing step was stratified train/test splitting to preserve the target's class distribution across data subsets.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

Table 1- Dataset Variables

Variable	Type	Description
Diabetes_binary	Binary (0/1)	1 = diabetes or prediabetes, 0 = no diabetes
HighBP	Binary	Ever told had high blood pressure
HighChol	Binary	Ever told had high cholesterol
CholCheck	Binary	Cholesterol check within past 5 years
BMI	Continuous	Body Mass Index
Smoker	Binary	Smoked ≥ 100 cigarettes in lifetime
Stroke	Binary	Ever told had a stroke
HeartDiseaseorAttack	Binary	Ever told had CHD or heart attack
PhysActivity	Binary	Physical activity in past 30 days (not job-related)
Fruits	Binary	Consumes fruit ≥ 1 time/day
Veggies	Binary	Consumes vegetables ≥ 1 time/day
HvyAlcoholConsump	Binary	Heavy drinking (sex-specific thresholds)
AnyHealthcare	Binary	Has any health coverage
NoDocbcCost	Binary	Could not see doctor in past year due to cost
GenHlth	Ordinal (1–5)	Self-rated general health (1 = excellent, 5 = poor)
MentHlth	Continuous (0–30)	Days mental health not good in past 30 days
PhysHlth	Continuous (0–30)	Days physical health not good in past 30 days
DiffWalk	Binary	Serious difficulty walking or climbing stairs
Sex	Binary	0 = female, 1 = male
Age	Ordinal (1–14)	Age category
Education	Ordinal (1–6)	Highest grade/year completed
Income	Ordinal (1–8)	Annual household income range

III. METHODOLOGY

A. Logistic Regression

Logistic regression models the log-odds of the target as a linear combination of the predictor variables. The fitted coefficients quantify the relationship between each predictor and the probability of having diabetes, holding other variables constant.

Assumptions:

- Predictors have a linear relationship with the log-odds of the outcome.
- Observations are independent.
- No multicollinearity among predictors.

Advantages:

- Highly interpretable; coefficients can be converted to odds ratios.
- Efficient to train and resistant to overfitting when regularized.

Disadvantages:

- Limited ability to capture non-linear or high-order interactions without explicit feature engineering.

Implementation:
A logistic regression model was trained with class_weight='balanced' to account for the 13.9% positive class prevalence. The solver liblinear was used for stability on large datasets, with a maximum of 500 iterations. Model performance was evaluated on the held-out test set.

B. Random Forest Classifier

Random Forest is an ensemble method that aggregates predictions from multiple decision trees trained on bootstrap samples of the data with random feature selection at each split. The final prediction is obtained by majority voting (classification) or averaging (regression).

Assumptions:

- Decision trees can approximate complex decision boundaries without requiring linearity.
- Randomization reduces correlation between trees, improving generalization.

Advantages:

- Captures non-linear interactions between features.
- Robust to outliers and multicollinearity.
- Provides feature importance measures.

Disadvantages:

- Less interpretable than linear models.
- Probabilities can be poorly calibrated without post-processing.

Implementation:
The baseline Random Forest model was trained with 300 trees (n_estimators=300), min_samples_split=10 to reduce overfitting, and class_weight='balanced'. Feature importance values were computed using mean decrease in Gini impurity. To explore model simplification, the top 10 and top 5 features (by importance) were used to retrain reduced Random Forest models. This allowed testing the trade-off between predictive performance and the number of survey questions required.

C. Model Evaluation and Threshold Tuning

Example: The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Because the dataset is imbalanced, performance was evaluated using metrics suited to skewed class distributions:

- ROC-AUC (Area Under the Receiver Operating Characteristic Curve) measures overall ranking ability.
- PR-AUC (Area Under the Precision–Recall Curve) more sensitive to minority-class performance.
- Precision - proportion of predicted positives that are true positives.
- Recall (Sensitivity) - proportion of true positives correctly identified.
- Specificity - proportion of true negatives correctly identified.
- F1-score - harmonic mean of precision and recall.

The default decision threshold of 0.5 was compared with a screening threshold of 0.3 for the Random Forest models, chosen to increase recall (sensitivity) for the positive class. This reflects public health priorities, where minimizing false negatives can be more important than maximizing precision in initial screening.

All metrics were computed on the held-out test set. Additionally, ROC and Precision–Recall curves were plotted for all models, and a threshold trade-off table was generated to guide threshold selection for different use cases.

IV. Results

Four models were evaluated:

1. Logistic Regression (all 21 features)
2. Random Forest (all 21 features)
3. Random Forest (top 10 features)
4. Random Forest (top 5 features)

All models were trained on the same stratified 80/20 train/test split and evaluated on the held-out test set.

A. Overall Performance at Default Threshold (.05)

Model	ROC-AUC	PR-AUC	Precision	Recall	F1
Logistic Regression (21)	**0.820**	0.392	0.31	**0.76**	0.44
Random Forest (21)	0.814	**0.404**	**0.40**	0.48	0.44
Random Forest (Top 10)	0.792	0.365	0.36	0.51	0.42
Random Forest (Top 5)	0.777	0.344	0.30	0.66	0.41

- Logistic regression achieved the highest ROC-AUC and recall at the default threshold.
- Full-feature RF had the highest PR-AUC and substantially better precision for positives.
- Reducing features to 10 caused only minor drops in ROC-AUC/PR-AUC; the 5-feature model had a larger but still acceptable drop.

B. Threshold Tuning for Screening (Threshold = .03)

Because recall is critical for initial screening, thresholds were lowered from 0.5 to 0.3 for the RF models.

Model (0.3 Threshold)	Precision	Recall	Specificity	F1	Accuracy
Random Forest (21)	0.30	**0.78**	0.70	0.43	0.71
Random Forest (Top 10)	0.28	0.77	0.68	0.41	0.69
Random Forest (Top 5)	0.24	**0.83**	0.58	0.38	0.61

- Lowering the threshold increased recall to ≥0.77 for all RF variants.
- Precision dropped modestly; the 5-feature model had the largest drop in specificity.
- The 10-feature RF performed nearly identically to the full RF at this screening-oriented threshold.

C. Feature Importance

The Random Forest feature importance ranking for the full model identified the following top predictors:

1. BMI (0.162)
2. General Health (0.135)
3. Age (0.120)
4. High Blood Pressure (0.107)
5. Income (0.070)
6. Physical Health Days (0.063)
7. High Cholesterol (0.053)
8. Mental Health Days (0.048)
9. Education (0.046)
10. Difficulty Walking (0.031)

This ranking was the basis for the reduced top-10 and top-5 feature models.

D. ROC and Precision-Recall Curves

Figure 1 (ROC Curves) shows that all models substantially outperform random guessing (AUC ≫ 0.5). Logistic regression has the highest ROC-AUC, but the difference from Random Forest models is small.

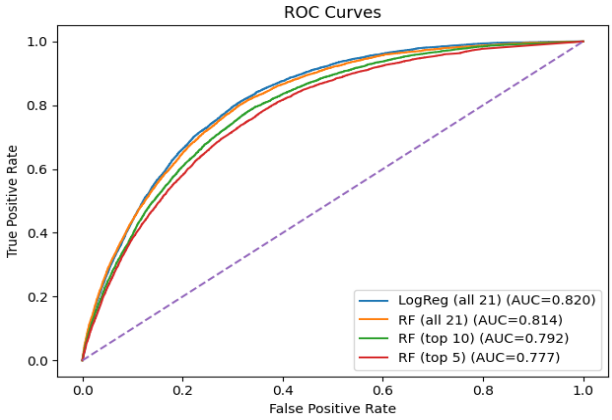
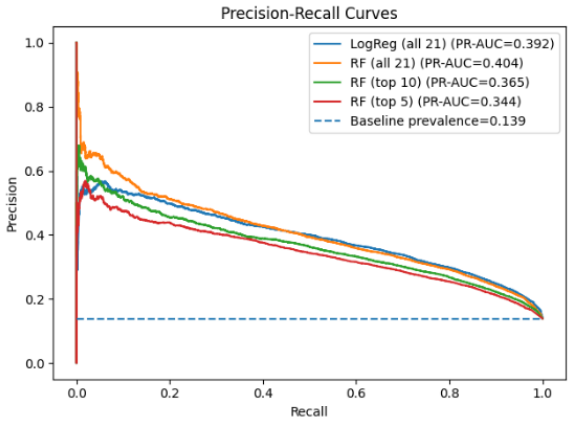


Figure 2 (Precision–Recall Curves) emphasizes the challenge of the imbalanced dataset (baseline prevalence ~0.14). Random Forest (21) has the highest PR-AUC, with Random Forest (top 10) close behind, indicating that most of the positive-class discrimination is retained with fewer features.



V. Discussion

This study set out to determine whether self-reported survey responses from the BRFSS could accurately predict diabetes status and whether a reduced set of survey questions could provide comparable predictive power to the full instrument.

For model performance results showed that a class weighted logistic regression model matched or slightly exceeded the random forests in ROC-AUC and recall at the default threshold. This suggests that much of the relationship between the survey indicators and diabetes risk is additive and well-approximated by a linear model in the log-odds space. However, the full-feature random forest had the highest PR-AUC, indicating better precision–recall trade-offs for the positive class when the threshold was optimized.

With Threshold tuning lowering the decision threshold for the RF models to 0.3 substantially improved recall (≥0.77) with only moderate loss in precision. This threshold adjustment aligns with public health screening priorities, where false negatives are more concerning than false positives in early detection contexts.

Reducing the feature set to the top 10 variables (BMI, general health, age, high blood pressure, income, physical health days, high cholesterol, mental health days, education, and difficulty walking) preserved most of the predictive performance in both ROC-AUC and PR-AUC. This indicates that a shorter survey can be nearly as effective as the full 21-item instrument.

The top 5-feature model (BMI, general health, age, high blood pressure, income) achieved the highest recall (0.83) at the screening threshold but at the cost of lower precision and specificity. This trade-off may be acceptable for large-scale risk screening programs that can follow up positive results with more specific tests.

Several limitations should be noted:

- The target variable is based on self-reported diabetes status, which may include reporting bias or misclassification.
- The models were trained and evaluated on a single survey year (2015); performance on more recent BRFSS data may differ.
- Only preselected variables from the reduced dataset were available; access to the full BRFSS might allow for feature engineering or the inclusion of additional predictors.

Future research could explore:

- Training on multiple BRFSS years to improve generalizability.
- Applying gradient boosting models with hyperparameter tuning to test whether non-linear learners can outperform logistic regression in this setting.
- Investigating cost-sensitive learning frameworks to optimize models directly for the public health trade-offs between false positives and false negatives.
- Validating the short-form questionnaire on external datasets to confirm its screening utility.

VI. Conclusion

This project demonstrated that self-reported health and demographic indicators from the BRFSS can be used to build effective predictive models for identifying individuals with diabetes or prediabetes. A class-weighted logistic regression achieved strong discrimination (ROC-AUC = 0.82) and high recall, matching or exceeding the performance of a random forest in most default-threshold metrics. The random forest achieved the best precision–recall trade-off and, when the decision threshold was lowered to 0.3, provided screening-level sensitivity (≥0.77 recall) while maintaining acceptable precision.

Feature importance analysis revealed that a small set of predictors-particularly BMI, general health rating, age, high blood pressure, and income-carry most of the model's predictive power. Reducing the model to the top 10 features preserved nearly all performance, supporting the feasibility of a short-form screening instrument. Even the top 5-feature model achieved very high recall, albeit with reduced precision, making it suitable for broad, low-cost initial screening followed by confirmatory testing.

These findings suggest that population-scale diabetes risk screening can be performed accurately using brief, non-invasive survey instruments. This approach has potential to expand early detection efforts, especially in settings where laboratory testing is impractical or resource-limited. Future work should validate these models on newer BRFSS data and explore alternative ensemble methods to further optimize precision–recall performance.

References

[1] CDC, *Behavioral Risk Factor Surveillance System*, 2015. Available: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system

[2] A. Teboul, *Diabetes Health Indicators Dataset*. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[3] University of California, Irvine, *CDC Diabetes Health Indicators* Available: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[4] U.S. Centers for Disease Control and Prevention, *Behavioral Risk Factor Surveillance System (BRFSS)*. Available: https://www.cdc.gov/brfss