Can we predict whether an individual is likely to have diabetes based on simple, non-invasive health and demographic information?

Ashley Nguyen, Ahmed Ahmed, Sanket Doddabendigere

Abstract

We developed and evaluated two machine-learning classifiers—logistic regression and random forest—to predict diabetes status using health-indicator data from the CDC Diabetes Health Indicators dataset. Our primary goal was to gauge overall predictive performance on a large, real-world survey dataset and identify which variables most strongly drive model predictions, thereby offering insight into key risk factors. This type of model could be used in real-world healthcare settings, like a free clinic, to quickly identify individuals at higher risk for diabetes using inexpensive, easily collected health indicators and demographic measures. Ultimately, this could help triage necessary follow-up testing and start intervention as early as possible.

Both models were trained on 80 % of the data and tested on the remaining 20 %. The logistic regression baseline achieved an accuracy of 86.2 %, precision of 51.6 %, recall of 15.8 %, F1-score of 24.2 %, and ROC-AUC of 0.819. The random forest showed comparable accuracy (85.9 %), slightly lower precision (48.8 %) but a modest gain in recall (17.9 %) and F1-score (26.2 %), with ROC-AUC of 0.795. Although both models perform well at distinguishing non-diabetic respondents, their low recall highlights a challenge—any true diabetes cases remain undetected, underscoring the need for further refinement before clinical or public-health deployment.

Feature-importance analysis revealed strong agreement on the top predictors. In the logistic regression, the most influential variables were self-rated general health, body mass index (BMI), age, high blood pressure, and high cholesterol. The random forest similarly prioritized BMI and age, while also elevating factors such as income, physical health days, education level, and mental health days. Collectively, these findings reinforce the central role of weight, aging, and overall perceived health in diabetes risk and suggest that socioeconomic and mental-health dimensions also carry predictive value.

In summary, our study demonstrates that straightforward classification algorithms can capture significant signals in survey data but require attention to class imbalance and sensitivity. Future work will attempt to boost recall without sacrificing specificity. These refinements aim to produce a more reliable screening tool and to deepen our understanding of the most actionable risk factors for diabetes prevention and management.

Introduction

Diabetes is a chronic metabolic disease that alters how the body converts food into energy, primarily due to problems with insulin production or effectiveness. When the body fails to produce enough insulin or use it efficiently, blood sugar levels rise, leading to serious long term health complications such as heart disease, kidney failure, vision loss, and nerve damage. Diabetes is classified into several types, the most common being type 2 diabetes, which accounts for 90 to 95 percent of all diagnosed cases. Type 2 diabetes typically develops gradually and can go undetected for years. It is strongly linked to lifestyle factors such as diet, physical activity, and body weight.

The prevalence of diabetes has reached alarming levels both globally and in the United States. Worldwide, around 800 million people are affected, a figure that has doubled since 1990. In the United States alone, over 38 million individuals have diabetes, with one in five unaware of their condition. Diabetes is the eighth leading cause of death in the country and is responsible for more than 400 billion dollars annually in medical costs and lost productivity. Its rapid rise is driven by increasing obesity rates, sedentary lifestyles, and aging populations, disproportionately affecting low income and middle-income communities where access to diagnosis and treatment is limited.

Given the magnitude of the problem and the fact that many diabetics go undetected until difficulties emerge, early detection of at-risk persons is a public health priority. However, widespread laboratory screening can be expensive and logistically difficult, especially in resource-constrained environments. This background inspires the search for simple, non-invasive predictors, such as demographic information, self-reported health status, and basic clinical indicators, to assist identify individuals who might benefit the most from additional testing and preventative actions.

Recent advancements in data science and machine learning have enabled the development of predictive models for diabetes risk based on large-scale health surveys. The CDC Diabetes Health Indicators dataset, for example, contains a wealth of information on U.S. individuals, including a wide range of behavioral, clinical, and socioeconomic variables. By examining these characteristics, we may investigate whether simple, easily obtained indicators like age, body mass index (BMI), blood pressure, cholesterol levels, and self-rated health can reliably identify those at increased risk for diabetes.

By focusing on simple, noninvasive health indicators, the aim is to develop tools that can be implemented in real world settings such as community clinics or public health screenings to efficiently triage individuals for follow up testing. This approach not only holds promise for improving early detection and prevention but also offers practical insights into which health factors are most predictive of diabetes risk in the general population.

Data

We use the CDC's Diabetes Health Indicators dataset, originally published via UC Irvine Machine Learning Repository. This cross-sectional survey captures behavioral and clinical information on U.S. adults, making it ideally suited to explore diabetes risk factors at the population level. Each observation in the dataset represents one respondent and their corresponding answer to the survey.

Variables

Our analysis focuses on 22 predictor variables alongside the binary diabetes outcome (0 or 1).

- Demographics & Socioeconomics: Sex; Age; Education; Income
- Clinical Indicators: High blood pressure; High cholesterol; Cholesterol check; Body-mass index (BMI); Heart disease or attack; Stroke
- Lifestyle & Health Behaviors: Smoker status; Heavy alcohol consumption; Physical activity; Fruit and vegetable intake; Access to any healthcare; Reported cost barriers to seeing a doctor
- Self-Rated Health & Well-Being: General health status; Number of poor physical-health days; Number of poor mental-health days; Difficulty walking

Relevance to Diabetes Prediction

Each variable reflects a known contributor to diabetes onset or progression. For example, elevated BMI, hypertension, and high cholesterol are well-documented clinical precursors to Type 2 diabetes. Fruit/vegetable consumption and physical activity track modifiable lifestyle factors. Self-rated health captures patients' own perception of their overall wellness. By combining these dimensions, our models can learn which features, individually and in combination, strongly predict current diabetes status.

Data Preparation & Challenges

The dataset arrives largely clean as missing-value rates are low, and categorical fields are consistently coded. Our main challenge, therefore, was not in data wrangling but in modeling—specifically, designing an approach that achieves high sensitivity (recall) for the minority "diabetes" class while maintaining overall accuracy. Because Random Forests can naturally handle mixed data types but require careful tuning, a key project hurdle is mastering and applying this method effectively. We address this by:

1. Stratifying our train/test split to preserve class balance
2. Standardizing continuous inputs for regression models
3. Iteratively tuning hyperparameters (e.g., number of trees, maximum depth) to improve true-case detection without inflating false positives.

With the dataset's breadth of health indicators and minimal preprocessing obstacles, our focus remains on building and refining models that can deliver both accurate classification and clear insights into diabetes risk factors.

Methods

To assess whether simple, non-invasive health and demographic information can be used to predict diabetes status, we developed and evaluated two supervised machine learning models: logistic regression and random forest.

After removing rows with missing values, we split the dataset into training and testing subsets using an 80/20 stratified split to maintain the class balance in both sets. Prior to training, we standardized continuous variables such as BMI and age using z-score normalization. For logistic regression, standardized features were used, while random forest models used unscaled inputs since they are not sensitive to feature scaling.

We trained both models using the training data and evaluated them on the held-out test set. Logistic regression was chosen as a baseline due to its interpretability and ability to highlight linear relationships between features and diabetes risk. The random forest model was selected to capture potential non-linear relationships and interactions between predictors. To optimize performance, we conducted iterative hyperparameter tuning for the random forest model, adjusting the number of estimators and tree depth to balance recall and precision.

Given the inherent class imbalance in the dataset, where the majority of respondents did not have diabetes, we emphasized evaluation metrics that provide insight beyond accuracy. Specifically, we assessed each model using precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics help evaluate how well the models identify true cases of diabetes while minimizing false positives. We also generated ROC curves and confusion matrices to visually inspect performance and misclassification patterns.

To interpret model behavior and identify influential features, we extracted and visualized feature importance scores. For logistic regression, we examined standardized coefficient magnitudes, while for the random forest, we used impurity-based feature importances. This allowed us to determine which health indicators most strongly contributed to the models' predictions and assess their consistency with known diabetes risk factors.

Throughout the analysis, we used Python libraries including scikit-learn for model implementation, pandas and NumPy for data manipulation, and seaborn and matplotlib for visualization. The overall goal of the methods was to construct interpretable and generalizable models that could eventually serve as practical tools for early diabetes risk screening in low-resource or community-based healthcare settings.

Results

When we compared two common approaches, Logistic Regression and Random Forest, both achieved approximately 86% overall accuracy. However, accuracy does not account for the fact that the majority of people in our data do not have diabetes, so a model that constantly says "no" would also perform well. To gain a better perspective, we looked at precision (how often a "yes" prediction is true) and recall (how many actual diabetes cases are identified). Logistic Regression properly detects a diabetic case 52% of the time when it predicts "yes," but only 16% of all genuine cases. Random Forest performs similarly, with 49% precision and 18% recall. Both models have modest F1-scores (0.24 and 0.26) and high-ranking ability (ROC-AUC 0.82 for Logistic Regression, 0.80 for Random Forest). These findings indicate that, while our algorithms are good at predicting who is at a higher risk, they miss many real diabetes cases and require further tuning or new data before being utilized for screening.

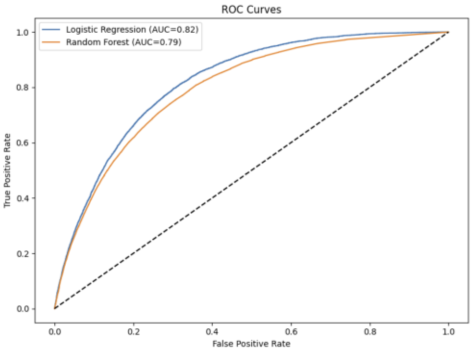
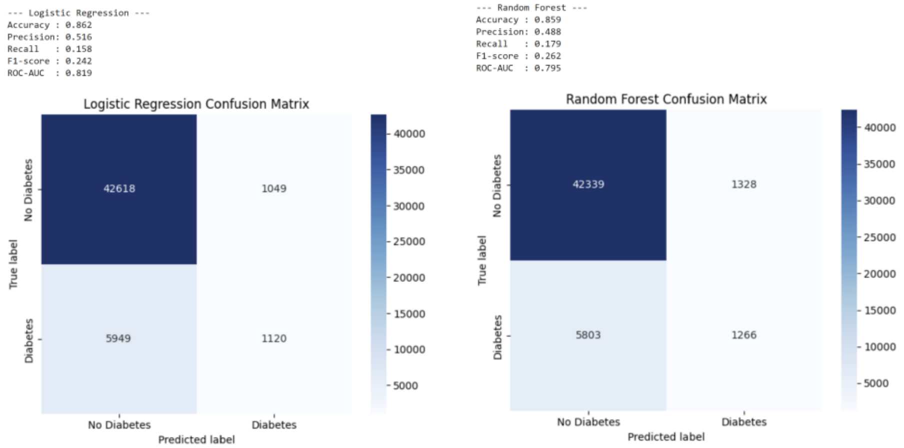


Fig 1: Logistic Regression (blue) and Random Forest (orange) ROC Curves

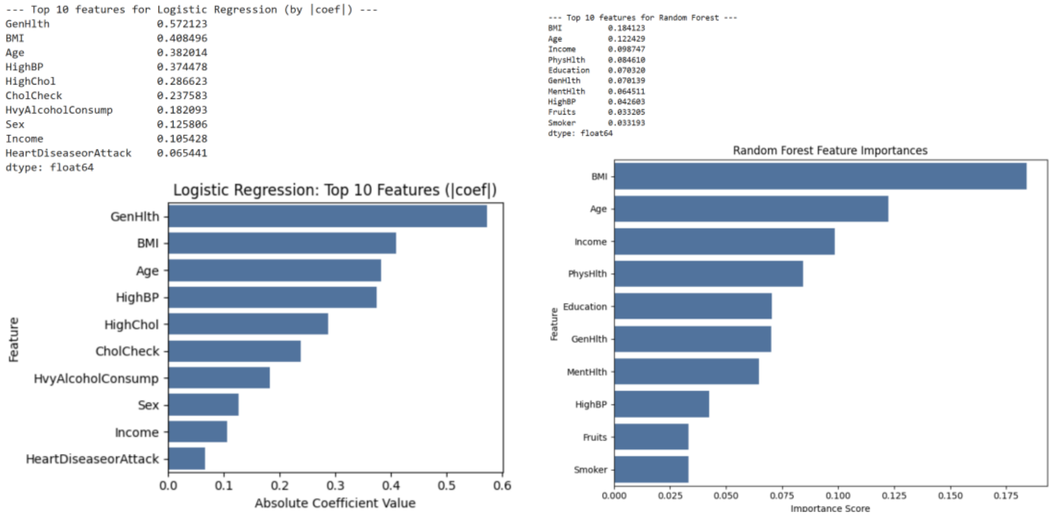


```
--- Logistic Regression ---
Accuracy : 0.862
Precision: 0.516
Recall   : 0.158
F1-score : 0.242
ROC-AUC  : 0.819
```

```
--- Random Forest ---
Accuracy : 0.859
Precision: 0.488
Recall   : 0.179
F1-score : 0.262
ROC-AUC  : 0.795
```

Fig 2: Logistic Regression (left) and Random Forest (right) Confusion Matrix

To better understand what influences each model's decisions, we selected the top 10 predictors. The main signals in both models are age and body mass index (BMI). In Logistic Regression, self-rated general health, high blood pressure, and high cholesterol are tightly related. Physical health days, academic level, and mental health all rank highly in Random Forest. These similarities indicate that simple measures—such as routine blood pressure checks, weight, and how people judge their own health—contain the majority of the information about diabetes risk in the survey.



```
--- Top 10 features for Logistic Regression (by |coef|) ---
GenHlth                 0.572123
BMI                     0.408496
Age                     0.382014
HighBP                  0.374478
HighChol                0.286623
CholCheck               0.237583
HvyAlcoholConsump       0.182093
Sex                     0.125806
Income                  0.105428
HeartDiseaseorAttack    0.065441
dtype: float64
```

```
--- Top 10 features for Random Forest ---
BMI          0.184123
Age          0.122429
Income       0.098747
PhysHlth     0.084610
Education    0.070320
GenHlth      0.070139
MentHlth     0.064511
HighBP       0.042603
Fruits       0.033205
Smoker       0.033193
dtype: float64
```

Fig 3: Top 10 features for Logistic Regression (left) and Random Forest (right)

In general, our models' top predictors are very similar to well-known diabetes risk variables. Body Mass Index (BMI) is one of the strongest indicators, which makes sense given that obesity is a major contributor to insulin resistance in Type 2 diabetes. Age is also an important issue, as the risk of diabetes rises with time due to metabolic changes. High blood pressure and high cholesterol are prevalent in diabetics, and they are part of a group of diseases known as metabolic syndrome. Surprisingly, self-rated overall health and the frequency of poor mental health days both ranked highly, implying that how people perceive their physical and mental well-being may represent deeper underlying health concerns. Our findings support the notion that basic, low-cost health markers can provide useful information about diabetes risk.

Conclusion

Predicting diabetes risk using simple, non-invasive health and demographic indicators shows real promise for early identification of individuals who may benefit from further testing or preventive care. The models developed, logistic regression and random forest, demonstrated that much of the predictive signal for diabetes in a large, representative survey comes from a handful of well-established risk factors, such as body mass index (BMI), age, self-rated health, blood pressure, and cholesterol. Both models achieved high overall accuracy, around 86%, but this figure is somewhat misleading given the class imbalance in the dataset. Many of the respondents do not have diabetes, so a model that always predicts “no” would also appear highly accurate. To get a more meaningful assessment, precision and recall were examined. Precision for logistic regression was 52%, and recall was 16%, while random forest achieved 49% and 18% recall. These modest recall rates highlight a key limitation—many true diabetes cases go undetected, underscoring the classic tradeoff between catching more cases (recall) and minimizing false alarms (precision). While our algorithms are good at predicting who is at a higher risk, they miss many real diabetes cases and require further tuning or new data before being utilized for screening.

The primary challenge encountered was the disparity between diabetic and non-diabetic cases. With considerably fewer positive cases, models automatically bias toward predicting the majority class, increasing accuracy while reducing the capacity to recognize true cases. Addressing this discrepancy with strategies such as class weighting or resampling could be a promising area for future research. Another disadvantage is the reliance on self-reported and easily quantified health variables. While this implies broad application and low-cost screening, it reduces the model's sensitivity when compared to alternatives that use laboratory or genetic data. While the dataset is big and representative of US adults, it may not capture the complete variation of diabetes risk among populations, thereby limiting generalizability.

Despite these challenges, the models' top predictors closely match established clinical knowledge. BMI emerged as the strongest indication, demonstrating obesity's prominent role in Type 2 diabetes risk. Age, high blood pressure, and high cholesterol all ranked high, consistent with their roles in metabolic syndrome. Interestingly, self-rated health and the frequency of poor mental health days were also significant, implying that people's judgments of their own well-being may include other, less obvious risk factors.

These findings indicate various possibilities for future research. Improving recall to identify more actual cases of diabetes should be a goal, even if it means accepting a greater proportion of false positives, because missing true cases can have major health effects. This could be accomplished by experimenting with advanced resampling strategies, altering class weights, or investigating algorithms for imbalanced data. Incorporating richer data sources, such as laboratory findings or longitudinal health records, may also improve predictive performance and enable the identification of people at risk of acquiring diabetes, rather than merely those who already have it. Expanding the analysis to include more varied populations would help to guarantee that the models are fair and widely applicable.

Interpretability remains important, especially for real world deployment in clinical or community settings. Logistic regression offers clear, actionable insights, while more complex models like random forests can capture non-linear patterns but require additional tools to explain their predictions. Employing explainable AI techniques could help bridge this gap, making it easier for healthcare providers to trust and act on model outputs.

The limitations encountered in this project are not simply obstacles but opportunities for deeper investigation and refinement. Each challenge, whether class imbalance, limited feature scope, or questions about generalizability, suggests concrete steps for advancing predictive modeling in public health. Building on these insights can bring future research closer to developing robust, equitable tools that support early detection and prevention of diabetes, helping to reduce the burden of this widespread disease.

References

Bessesen, Daniel. “What Is Diabetes?” National Institute of Diabetes and Digestive and Kidney Diseases, Apr. 2023, www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes.

CDC. “Diabetes Basics.” Www.cdc.gov, 15 May 2024, www.cdc.gov/diabetes/about/index.html.

IHME. “Global Diabetes Cases to Soar from 529 Million to 1.3 Billion by 2050 | the Institute for Health Metrics and Evaluation.” Www.healthdata.org, 22 June 2023, www.healthdata.org/news-events/newsroom/news-releases/global-diabetes-cases-soar-529-million-13-billion-2050.

World Health Organization. “Diabetes.” World Health Organization, 14 Nov. 2024, www.who.int/news-room/fact-sheets/detail/diabetes.