Diabetes Risk Prediction Using Machine Learning:

Fairness Analysis and Public Health Policy Simulation

Salman Shareef
Master of Science in Applied Data Science
University of St. Thomas
Practicum II
Instructor: Dr. Follis
Date: December 2025

Abstract

Diabetes is one of the most prevalent chronic diseases in the United States and contributes significantly to cardiovascular disease, disability, and healthcare costs. This study applies machine learning techniques to predict diabetes risk using the Behavioral Risk Factor Surveillance System (BRFSS) health indicators dataset. Three predictive models—Logistic Regression, Random Forest, and Gradient Boosting—were developed and compared. In addition to predictive performance, this study evaluated algorithmic fairness across sex and income groups and conducted a public health policy simulation based on body mass index (BMI) reduction. Results indicate that Gradient Boosting achieved the highest predictive performance, with income and BMI emerging as key predictors. Fairness analysis revealed persistent socioeconomic disparities in predicted diabetes risk. A simulated BMI reduction intervention demonstrated measurable population-level risk reduction, particularly among lower-income groups. These findings highlight the potential of machine learning to support equitable public health policy and targeted intervention strategies.

1. Introduction

Diabetes affects more than 37 million individuals in the United States and remains a leading contributor to cardiovascular disease, kidney failure, blindness, and premature mortality. Despite advances in medical treatment, the distribution of diabetes risk is not uniform across the population. Socioeconomic status, physical activity, access to healthcare, and lifestyle behaviors play critical roles in shaping diabetes prevalence.

With the rapid growth of data science and artificial intelligence, machine learning offers powerful tools for identifying high-risk individuals, detecting patterns of inequity, and evaluating potential intervention strategies before implementation. However, predictive accuracy alone is insufficient for responsible deployment. Modern public health analytics must also address algorithmic fairness, demographic disparities, and policy relevance.

This practicum project applies supervised machine learning models to predict diabetes risk using national survey data and integrates fairness analysis with a simulated public health intervention. The objectives of this study are to:

1. Develop high-performance diabetes prediction models.
2. Evaluate demographic fairness across sex and income groups.
3. Identify high-risk populations for targeted intervention.
4. Simulate a policy intervention to assess potential population-level impact.

2. Data and Methods

2.1 Dataset

The dataset used in this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS) Diabetes Health Indicators (2015), publicly available through Kaggle. The dataset contains approximately 250,000 observations and 22 variables capturing behavioral, clinical, and socioeconomic characteristics.

Key variables include:

- Body Mass Index (BMI)
- High blood pressure
- High cholesterol

- Physical activity
- Smoking behavior
- Alcohol consumption
- Mental and physical health status
- Healthcare access
- Income level
- Education level
- Age and sex

A binary target variable, **Diabetes_binary**, was constructed such that:

- 1 = Diagnosed diabetes
- 0 = No diabetes or prediabetes

2.2 Data Cleaning and Feature Selection

The dataset was fully numeric and contained no missing values. Features were selected based on clinical relevance, policy significance, and prior epidemiological research. All predictor variables were standardized for Logistic Regression, while tree-based models used raw feature values.

The final modeling dataset included 18 features capturing behavioral risk, chronic conditions, access to care, and socioeconomic status.

2.3 Machine Learning Models

Three supervised classification models were trained:

1. **Logistic Regression** — baseline linear classifier.
2. **Random Forest Classifier** — ensemble tree-based model.
3. **Gradient Boosting Classifier** — high-performance boosting ensemble.

The dataset was partitioned into training and testing subsets using a stratified 75/25 split to preserve outcome prevalence.

Model performance was evaluated using:

- Accuracy
- Receiver Operating Characteristic – Area Under the Curve (ROC–AUC)
- Precision, Recall, and F1-score

2.4 Fairness Analysis

Predicted diabetes rates were disaggregated across:

- Sex
- Income category

This analysis assessed whether model predictions amplified or reflected existing structural inequalities.

2.5 Policy Simulation

A hypothetical public health intervention was simulated by reducing BMI by one unit across the population. The trained Gradient Boosting model was used to re-estimate diabetes risk under this modified scenario. The resulting predicted risk reduction represents a population-level estimate of policy impact.

3. Results

3.1 Model Performance

Among the three models evaluated, Gradient Boosting achieved the highest predictive performance in terms of ROC–AUC and classification stability. Logistic Regression provided interpretable baseline results, while Random Forest achieved strong nonlinear modeling performance.

Gradient Boosting was therefore selected for fairness analysis and policy simulation.

3.2 Feature Importance

The most influential predictors across ensemble models were:

- BMI
- High blood pressure
- Poor general health
- Physical inactivity
- Income level

These variables align with established clinical and social determinants of diabetes.

3.3 Income-Based Health Disparities

A strong inverse relationship between income and diabetes prevalence was observed. Individuals in the lowest income categories exhibited diabetes prevalence exceeding 24%, while those in the highest income categories exhibited rates below 8%.

This gradient demonstrates a pronounced socioeconomic burden of disease and reinforces the role of structural determinants in shaping health outcomes.

3.4 Sex-Based Differences

Male respondents consistently exhibited higher predicted diabetes risk than females, a pattern consistent with national epidemiological trends.

3.5 Healthcare Access and Detection Bias

Individuals with healthcare access exhibited higher observed diabetes rates than those without access. This reflects **diagnosis detection bias**, as uninsured individuals may remain undiagnosed despite underlying disease presence.

3.6 Fairness Analysis

Predicted diabetes probabilities maintained income-based gradients and sex-based differences. The model did not introduce artificial disparities beyond those already present in the underlying data. Instead, it exposed existing structural inequities.

4. Policy Simulation Results

Under the simulated intervention of a one-unit BMI reduction:

- The average predicted diabetes risk across the population decreased measurably.
- The most substantial risk reductions occurred within low-income groups.
- The policy simulation suggests that modest behavioral improvements at the population level could significantly reduce national diabetes burden.

These findings demonstrate how machine learning can support **prospective intervention evaluation** before costly real-world implementation.

5. Discussion

This study demonstrates the value of machine learning as a decision-support tool in public health. By integrating predictive modeling with fairness evaluation and policy simulation, this project extends beyond traditional risk prediction into actionable policy analytics.

Key contributions include:

- Identification of high-risk populations for targeted interventions.
- Quantification of income-based health inequities.
- Demonstration of how small behavioral changes can generate large public health benefits.

However, this study is subject to limitations. The BRFSS dataset is cross-sectional and self-reported, which introduces reporting bias. The policy simulation is hypothetical and assumes uniform BMI reduction across the population, which may not reflect real-world feasibility.

6. Conclusion

This practicum project illustrates how applied machine learning can be used responsibly to inform public health policy. Predictive accuracy, fairness auditing, and policy simulation together create a powerful framework for addressing chronic disease at both the individual and population levels.

As artificial intelligence becomes increasingly integrated into healthcare and public policy, studies such as this underscore the importance of ethical deployment, equity-driven analysis, and transparent decision-support systems.

7. Future Work

Future project extensions may include:

- Longitudinal diabetes risk forecasting
- Causal inference modeling for real policy evaluation
- Integration of geographic and environmental variables
- Deep learning approaches for nonlinear interaction modeling

8. References (Sample APA Format)

Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System (BRFSS)*.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.