🩺 Predicting Diabetes Risk with Machine Learning: A Public Health & Policy Perspective

By Salman Shareef, MS Applied Data Science

⭐ Introduction

Diabetes continues to be one of the most pressing chronic health challenges in the United States. Each year, millions of Americans face increased risk of cardiovascular complications, disability, and reduced quality of life — often driven not only by individual health behaviors but also by broader social and economic factors.

As part of my Practicum II project in the Master of Science in Applied Data Science program, I set out to answer a fundamental question:

Can machine learning help identify high-risk populations and guide public health interventions more effectively?

Using the Behavioral Risk Factor Surveillance System (BRFSS) dataset, I developed predictive models, evaluated fairness across demographic groups, and simulated policy scenarios to estimate their potential impact.

This post summarizes the insights from that work.

📊 Data & Methods

The project used the **BRFSS 2015 Diabetes Health Indicators dataset**, consisting of:

- 📈 **250,000+ respondents**
- 🧩 **22 behavioral & clinical features** such as:

BMI
Blood pressure
Physical activity
Alcohol use
Mental & physical health days
Income level
Health care access

A binary outcome variable was constructed:

1 = Diabetes
0 = No diabetes

I trained three predictive models:

Logistic Regression (baseline)
Random Forest Classifier
Gradient Boosting Classifier (best overall performer)

The pipeline included:

Feature scaling
Train/test split (stratified sampling)
ROC–AUC evaluation
Feature importance analysis

🔍 Key Findings

1️⃣ Gradient Boosting Outperformed Other Models

The Gradient Boosting model achieved the highest **ROC-AUC**, indicating strong predictive ability and stability across subgroups.

It consistently identified well-known risk factors:

High BMI
High blood pressure
Poor general health score

Low physical activity
Low income

2️⃣ Income Was One of the Strongest Predictors

A clear gradient emerged:

Income Level	Diabetes Rate
1 (lowest)	~24–26%
8 (highest)	~8%

This demonstrates a **striking health inequity**, reinforcing findings from public health research:

📉 Lower-income communities bear a disproportionate burden of chronic disease.

3️⃣ Fairness Analysis Revealed Important Patterns

The model's predicted risk rates were compared across demographic groups:

Sex

Males showed higher predicted risk than females — consistent with epidemiological evidence.

Income

Predicted risk decreased steadily as income increased.
This suggests the model reflects true disparities rather than introducing bias.

Health Care Access

Interestingly, individuals with health care access appeared to have higher diabetes rates — but this reflects **detection bias**:

People without health insurance are less likely to be diagnosed, even if they have diabetes.

This is important when considering fairness; sometimes **the data itself reflects structural inequities**.

🚀 Policy Simulation: What If We Reduced BMI?

To demonstrate how machine learning can support policy decisions, I simulated a hypothetical intervention:

What if we could reduce everyone's BMI by 1 unit through a national health promotion program?

The model showed:

- A measurable reduction in predicted diabetes risk
- Greater benefits among low-income groups
- Evidence that **behavioral interventions can narrow inequity gaps**

This type of simulation is exactly how data science can support public health planning.

🧠 Why This Matters

Machine learning is not just about prediction — it's about **understanding populations**, identifying **systemic disparities**, and helping decision-makers evaluate **the cost and benefits of interventions** before they are implemented.

This project demonstrates:

- How ML models can reveal hidden health inequities
- How bias and fairness checks are essential for responsible AI
- How policy simulations can translate model outputs into actionable insights

📌 Conclusion

Diabetes remains a major public health challenge, but machine learning offers tools to better understand risk patterns and guide interventions. By integrating predictive modeling with fairness analysis and policy simulation, we can:

Identify high-risk populations
Highlight structural inequities
Support targeted, effective public health planning

This project represents one step toward using data science not only to model health outcomes, but to **inform policies that promote health equity and improve population well-being**.