7/6/2025

# Diabetes Risk Prediction

## Business Understanding

**Kungulio, Seif H.**
**DATA 650: CAPSTONE PROJECT**

# Improving Diabetes Risk Prediction Among Adults in Texas.

## Introduction

A well-defined business problem sets the direction for a successful analytics project. It ensures the analysis is technically sound and aligned with public health priorities. For this project, the challenge centers on enhancing early detection efforts for diabetes—a critical objective in states like Texas, where the condition's prevalence remains alarmingly high. A focused problem statement helps identify relevant health indicators, guide the selection of predictive modeling techniques, and deliver insights that can support targeted interventions, reduce long-term healthcare costs, and improve patient outcomes.

## Initial Business Problem Statement

How can public health agencies in Texas use self-reported health behavior and demographic data to predict which adults are at high risk of developing diabetes?

## I Used AI (You.com) to Refine Business Problem Statement

I entered my initial business problem into an AI tool (You.com) and asked the following: *How can I make this more measurable or time-bound? What's a better way to define what "high risk" means in this context? Can you suggest business impact language suited for public health?*

The AI tool recommended adding a defined timeline and specifying what success looks like—such as identifying individuals with a 30% or higher risk of developing diabetes within five years, using clinical markers like HbA1c and BMI. It also suggested reframing the question into a more targeted public health objective and incorporating language about potential impacts, like reducing healthcare costs and improving health equity, to make the problem more actionable and aligned with decision-making goals.

## AI Review Summary

Based on the feedback, I revised the statement to include a defined timeline (by 2026) and a measurable risk threshold (30% or higher) using clinical markers like HbA1c and BMI. I shifted the focus from a broad exploratory question to a targeted public health objective aimed at predicting diabetes risk among Texas adults. The AI also recommended emphasizing the expected outcomes—such as lowering healthcare costs and improving health equity—which helped align the language with decision-making priorities and clarify the intended impact.

## Final Revised Problem Statement with Impact Language

Texas public health agencies aim to use self-reported health behavior and demographic data to create a predictive model by 2026 that identifies adults at a 30% or higher risk of developing diabetes within five years, based on clinical markers like HbA1c (>5.7%) and BMI (>30). This targeted approach could lower healthcare costs, enhance health equity, and prevent thousands of diabetes cases, supporting broader public health goals.

## Primary Persona

**Name:** Dr. Angela Torres

**Role:** Director of Chronic Disease Prevention, Texas Department of State Health Services

**Responsibilities:** Designs and managing state-wide health programs for chronic disease prevention. Allocates resources for screenings, awareness, and targeted outreach. Dr. Torres uses data to monitor disease trends and evaluate public health interventions.

**Why This Project Matters:** As Director of Chronic Disease Prevention for the Texas Department of State Health Services, Dr. Torres is responsible for guiding statewide efforts to reduce the burden of diabetes through early detection and prevention. She makes critical decisions such as:

- Determining which populations or regions to prioritize for diabetes screenings
- Setting thresholds for predictive risk scores to direct outreach efforts
- Coordinating with clinics and community organizations to implement preventive care initiatives

However, Dr. Torres faces significant challenges, including limited resources, high statewide demand, and persistent disparities in healthcare access. This project provides a data-driven solution by offering predictive insights that help her target the most vulnerable populations with precision. By identifying high-risk individuals proactively, she can optimize program effectiveness, reduce the incidence of late-stage diagnoses, and allocate public health resources more efficiently—ultimately advancing health equity and improving population-level outcomes across Texas.

## Potential Datasets

To address this problem the dataset should contain the following:

- Demographics: age, education, income, sex, race
- Healthy behavior: smoking, alcohol use, exercise, vitamin/fat/carbohydrate intake
- Clinical indicators: BMI, cholesterol, blood pressure, general health status
- Outcome variable: Diabetes classification (binary/ categorical)

I explored the following sources for data:

- Kaggle – Found Behavioral Risk Factor Surveillance System
- UCI Machine Learning Repository – Found CDC Diabetes Health Indicators
- Kaggle – Found Diabetes Health Indicators Dataset

## Business Case Evidence

Diabetes is a major public health issue in Texas, affecting over 2.8 million diagnosed adults, with an estimated 600,000 more undiagnosed. The state ranks among the top 10 in diabetes prevalence, and over 35% of adults are believed to have prediabetes without knowing it. This leads to preventable complications and strains the healthcare system.

Economically, diabetes costs Texas over $25 billion annually in medical expenses and loses productivity. Predictive analytics, using data like the BRFSS, offers a strategic solution by identifying high-risk individuals for targeted interventions. Programs like the NDPP show that early detection can reduce diabetes onset by up to 71% in older adults.

Equity remains critical, as rural, Hispanic, and African American communities face higher risk due to limited access to care. Predictive models can help close these gaps by enabling more efficient and equitable outreach.

## Conclusion

This project aligns with my career goal of applying predictive analytics to real-world healthcare challenges. Defining the business problem using SMART criteria—and refining it with AI—helped transform a broad question into a measurable and time-bound objective. Identifying Dr. Angela Torres, Director of Chronic Disease Prevention at the Texas Department of State Health Services, as the primary persona rooted the analysis in real-world public health decision-making, highlighting how predictive models can guide early intervention strategies and improve health equity.

The business case clearly underscores the urgency of the issue, with over 2.8 million adults diagnosed with diabetes in Texas and hundreds of thousands more likely undiagnosed. The financial and social toll exceeds \$25 billion annually, making a compelling case for using data-driven approaches to target high-risk populations and allocating limited public health resources more effectively.

I explored multiple datasets from Kaggle and the UCI Machine Learning Repository, selecting those that include behavioral, demographic, and clinical indicators relevant to diabetes risk. Although some datasets lack longitudinal or behavioral depth, they offer a strong foundation for early-stage modeling and risk stratification.

This assignment has laid a strong foundation for my capstone by combining a critical public health problem, actionable data, and a clearly defined stakeholder. It has reinforced the role of analytics in supporting equitable, preventive care and shaped a strategic direction for my continued work in healthcare data science.

## References


- Texas Department of State Health Services (2023). Diabetes in Texas. [https://www.dshs.texas.gov]
- Centers for Disease Control and Prevention (CDC, 2024). National Diabetes Statistics Report.
- American Diabetes Association (2023). The Economic Costs of Diabetes in the U.S.
- National Institute of Diabetes and Digestive and Kidney Diseases (2024). Preventing Type 2 Diabetes.
- Behavioral Risk Factor Surveillance System (BRFSS, 2015). Centers for Disease Control and Prevention.
- National Diabetes Prevention Program (NDPP). [https://www.cdc.gov/diabetes/prevention/]