Project Idea: Diabetes Prediction from Clinical Data

Paula Lozano Gonzalo

April 1, 2025

1 Overview

This supervised learning project aims to predict **diabetes status** (the label) based on various demographic and clinical features. The dataset contains comprehensive health information about 100,000 individuals, making it valuable for developing predictive models and analyzing risk factors for diabetes.

2 Data Source

The dataset is sourced from Kaggle:
`https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows/data`

2.1 Features Available

Table 1: Description of Features

Feature	Description	Possible Values
Patient_ID	Unique identifier	Numerical
Gender	Gender of patient	Male/Female/Other
Age	Age of patient	Numerical
Location	Geographic location	Various city names
Race	Ethnic background	White/Black/Asian/Hispanic/Other
Hypertension	Hypertension status	0 (No)/1 (Yes)
Heart_Disease	Heart disease status	0 (No)/1 (Yes)
Smoking_History	Smoking habits	Never/Former/Current/No Info
BMI	Body Mass Index	Numerical
HbA1c_level	Glycated hemoglobin level	Numerical (4.0-9.0%)
Blood_Glucose_Level	Fasting blood glucose	Numerical (mg/dL)

2.2 Key Feature Explanations

For some features I had to do some further research because I didn't fully understand.

HbA1c_level: This measures average blood sugar levels over the past 2-3 months. Higher values indicate poorer blood sugar control:

- Normal: Below 5.7%
- Prediabetes: 5.7% to 6.4%
- Diabetes: 6.5% or higher

Blood_Glucose_Level: Fasting blood sugar level measurement:

- Normal: Less than 100 mg/dL
- Prediabetes: 100-125 mg/dL
- Diabetes: 126 mg/dL or higher

2.3 Label (Target Variable)

- **Diabetes** (0 = No diabetes, 1 = Diabetes)

3 Problem Type

This is a **binary classification** problem since the label has two possible classes (0 for non-diabetic, 1 for diabetic).

4 Number of Samples

The dataset contains 100,000 samples, making it substantial for training robust machine learning models.

5 Potential Applications

- **Clinical Decision Support**: Help healthcare providers identify high-risk patients for early intervention
- **Public Health Screening**: Target screening programs to populations with higher predicted risk
- **Personal Health Awareness**: Could be integrated into health apps to alert users about diabetes risk
- **Research**: Identify which factors contribute most to diabetes risk in different populations

6 Has This Problem Been Solved Before?

- Diabetes prediction is a well-studied problem in medical machine learning
- Kaggle hosts some solutions in the code tab of the website:
  - `https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows/code`
- However, this specific dataset (with 100k samples and these particular features) appears to be relatively new
- Existing solutions typically use logistic regression, random forests, or neural networks with accuracy around 80-90%