Predictive Analytics and Diabetes Diagnosis

BTMA 531 Group 11

Jan Petallo

Table of Contents

01

Problem Formulation

02

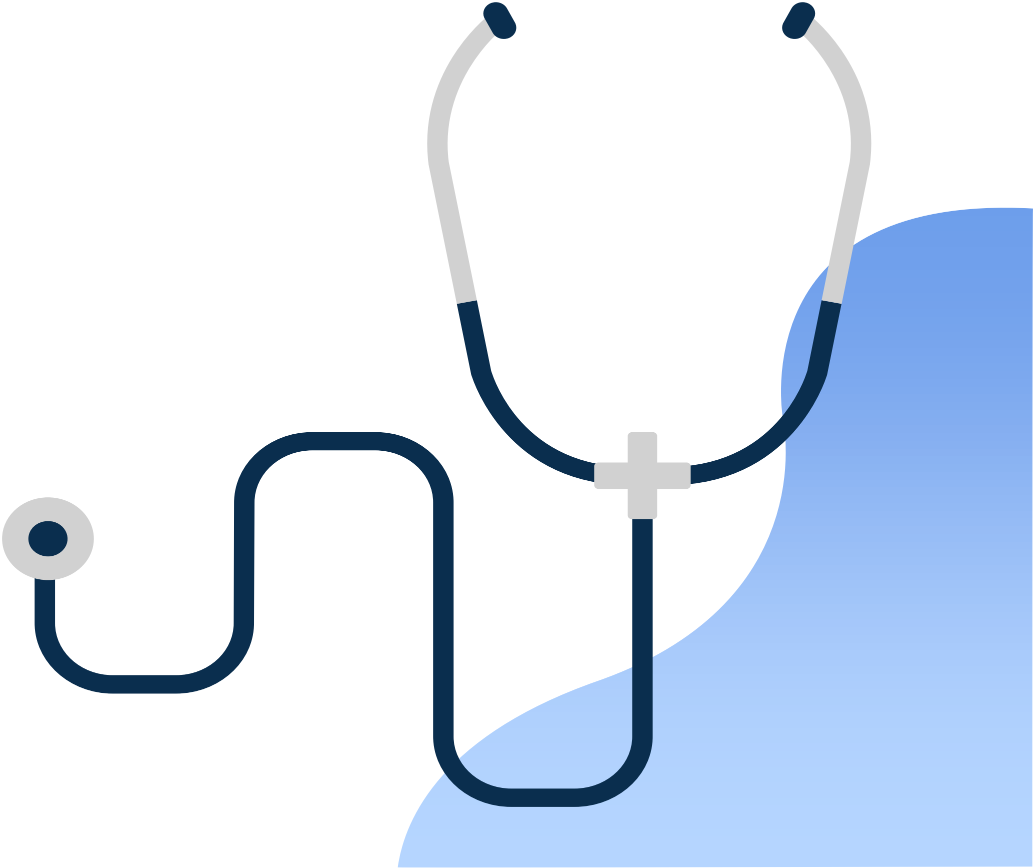
Data Discussion

03

Analysis

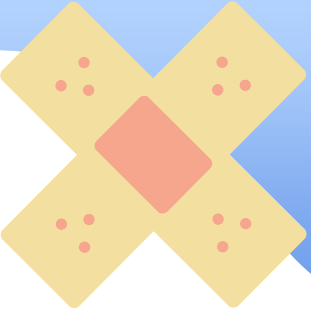
04

Results and Discussion



01

Problem Formulation



The Problem

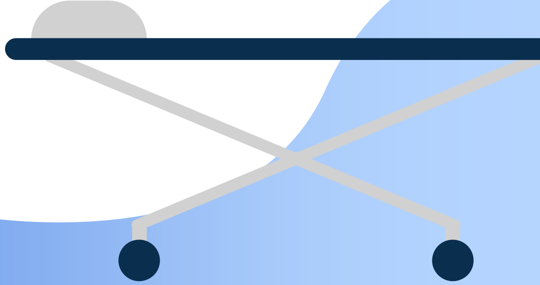
- ❑ Diabetes affects around 3.8 million people in Canada
- ❑ It contributes to various complications:
  - ❑ Stroke
  - ❑ Heart Attacks
  - ❑ Kidney Failures
  - ❑ Reduced lifespan of 5-15 years.
- ❑ Ambiguity regarding dominant factors influencing diabetes risk: Genetics? Lifestyle?

The Goal

- ❑ Use predictive analytics to estimate the likelihood of having prediabetes or diabetes based on health and lifestyle factors
- ❑ Empower the following with actionable insights
  - ❑ Individuals
  - ❑ Healthcare providers
  - ❑ Researchers

02

Data Discussion



About the Dataset

- ❑ Dataset is derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS)
- ❑ Annual survey conducted by Centers for Disease Control and Prevention
- ❑ Ongoing telephone survey collecting data such as health-related risk behaviors and chronic health conditions

236,378 records

~ 203k people without diabetes

~ 33k people with diabetes

Target variable: Diabetes_binary

0 = no diabetes

1 = prediabetes or diabetes

21 feature variables

Represent health indicators and lifestyle factors

Variable	Description
Diabetes_binary	0 = no diabetes, 1 = prediabetes and diabetes
HighBP	0 = no high BP, 1 = high BP
HighChol	0 = no high cholesterol, 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
BMI	Body Mass Index (numerical)
Smoker	0 = no, 1 = yes (Have you smoked at least 100 cigarettes in your entire life?)
Stroke	0 = no, 1 = yes (Ever told you had a stroke)
HeartDiseaseorAttack	0 = no, 1 = yes (Coronary Heart Disease or Myocardial Infarction)
PhysActivity	0 = no, 1 = yes (Physical activity in past 30 days - not including job)
Fruits	0 = no, 1 = yes (Consume Fruit 1 or more per day)
Veggies	0 = no, 1 = yes (Consume Vegetables 1 or more per day)
HvyAlcoholConsump	0 = no, 1 = yes (Heavy drinkers, based on gender-specific criteria)
AnyHealthcare	0 = no, 1 = yes (Have any kind of health care coverage)
NoDocbcCost	0 = no, 1 = yes (Could not see a doctor in past 12 months due to cost)
GenHlth	1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor (General health)
MentHlth	Number of days (0-30) mental health was not good in the past 30 days
PhysHlth	Number of days (0-30) physical health was not good in the past 30 days
DiffWalk	0 = no, 1 = yes (Serious difficulty walking or climbing stairs)
Sex	0 = female, 1 = male
Age	Value between 1 to 13 (Age group)
Education	Value between 1 to 6 (Highest grade or year of school completed)
Income	Value between 1 to 11 (Income group)

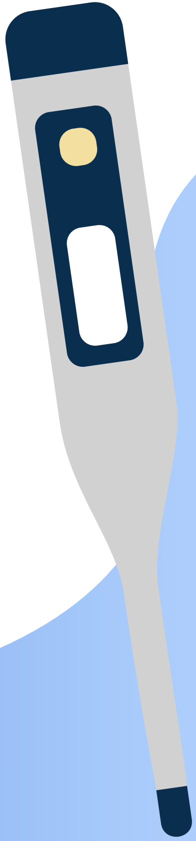
Value	Age Group
1	Age 18 to 24
2	Age 25 to 29
3	Age 30 to 34
4	Age 35 to 39
5	Age 40 to 44
6	Age 45 to 49
7	Age 50 to 54
8	Age 55 to 59
9	Age 60 to 64
10	Age 65 to 69
11	Age 70 to 74
12	Age 75 to 79
13	Age 80 or older

Value	Income Range
1	Less than $10,000
2	$10,000 to < $15,000
3	$15,000 to < $20,000
4	$20,000 to < $25,000
5	$25,000 to < $35,000
6	$35,000 to < $50,000
7	$50,000 to < $75,000
8	$75,000 to < $100,000
9	$100,000 to < $150,000
10	$150,000 to < $200,000
11	$200,000 or more

Value	Education Level
1	Never attended school or only kindergarten
2	Grades 1 through 8 (Elementary)
3	Grades 9 through 11 (Some high school)
4	Grade 12 or GED (High school graduate)
5	College 1 year to 3 years (Some college or technical school)
6	College 4 years or more (College graduate)

03

Analysis



Exploratory Data Analysis (EDA)

01

Class Imbalance

- ❑ 86% no diabetes
- ❑ 14% prediabetes or diabetes

02

Missing Values

Dataset contains no missing values

03

Correlation

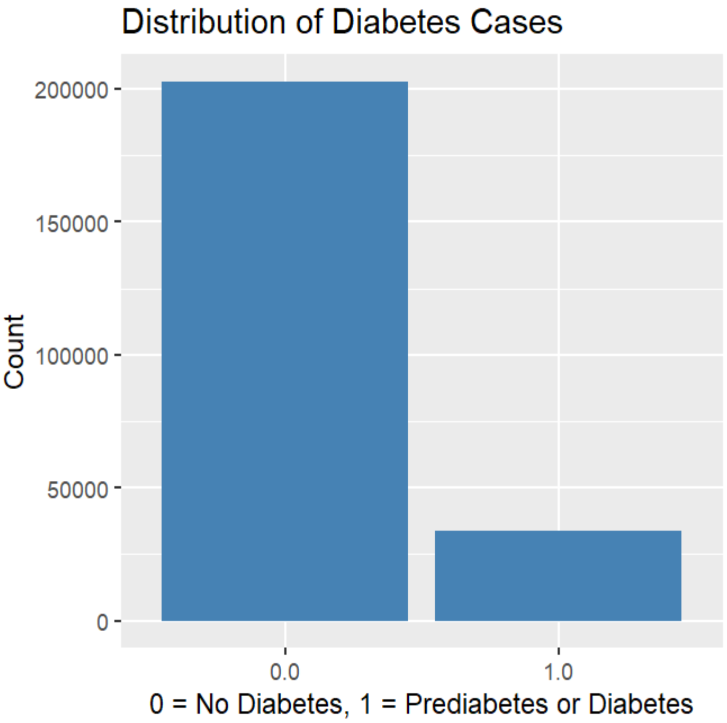
Generally low among numeric features

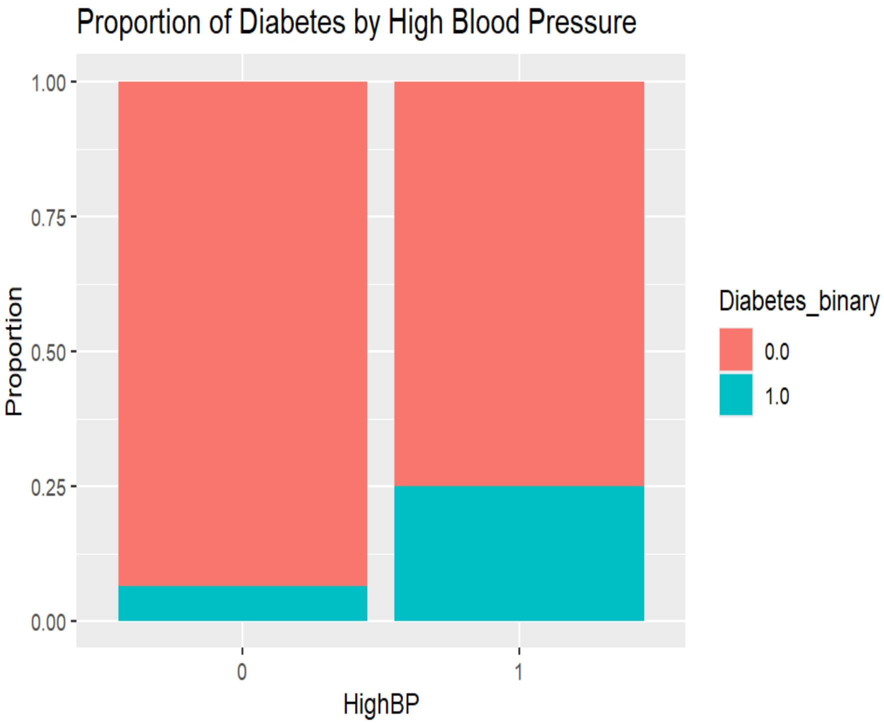
04

Patterns

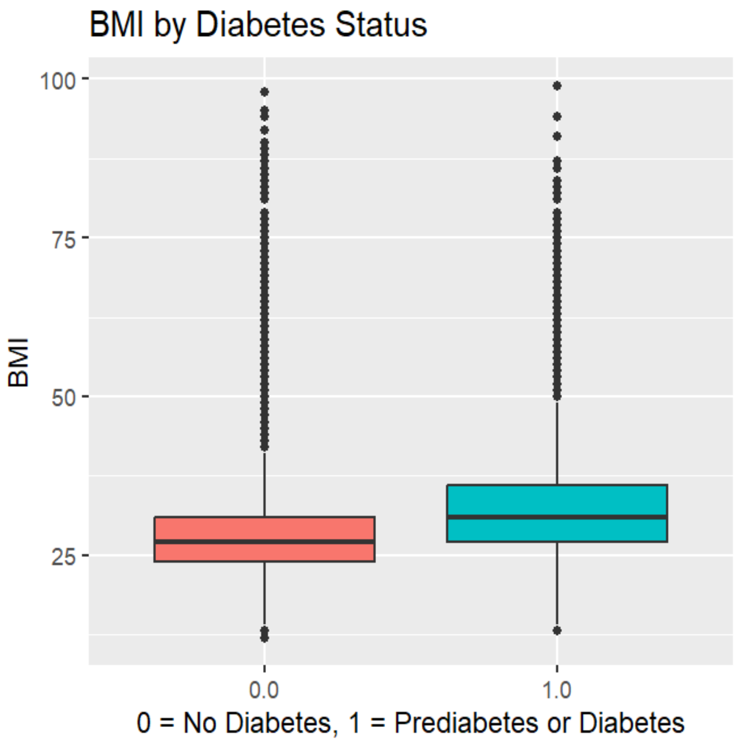
- ❑ High Blood Pressure
- ❑ High BMI
- ❑ Unhealthy physical days

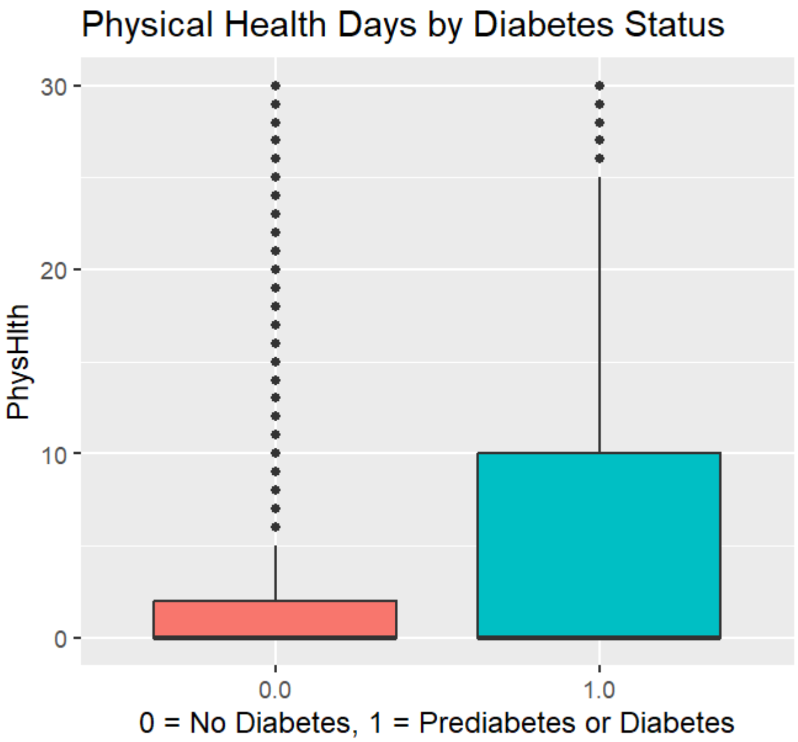
Exploratory Data Analysis (EDA)





Exploratory Data Analysis (EDA)

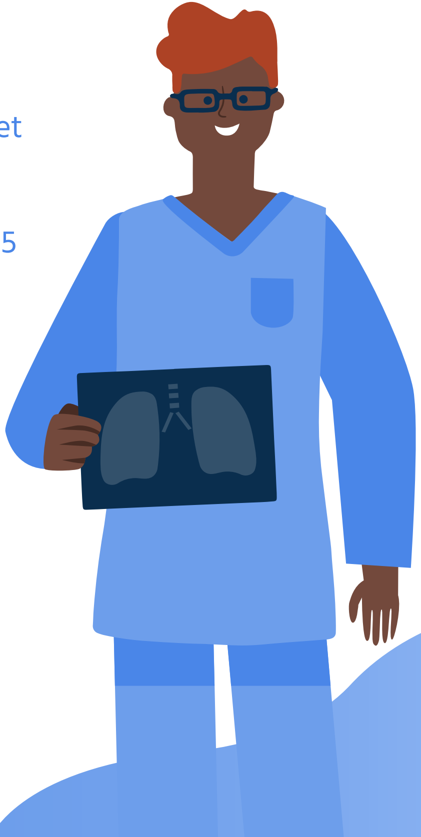




Modeling Approach

Data Split

- ❑ 70% training set
- ❑ 30% test set
- ❑ Seet seed: 2025



Models Used

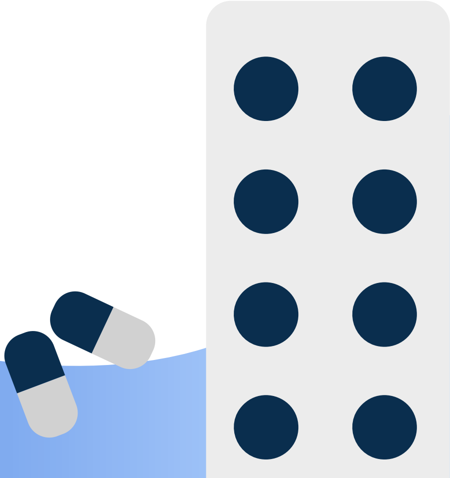
- ❑ **Logistic Regression**:
- ❑ simple, interpretable, standard classifier
- ❑ **Gradient Boosting Machine (GBM):**
- ❑ Captures complex, non-linear patterns

Threshold Selection

- ❑ **0.15**
- ❑ To balance sensitivity and specificity

04

Results and Discussion



Model Comparison

	Accuracy	Sensitivity	Specificity
Logistic Regression	73.14%	75.38%	72.76%
GBM	73.81%	74.83%	73.64%

Relevant Predictors

BMI

The person's Body Mass Index, derived from their height and weight measurements

High BP

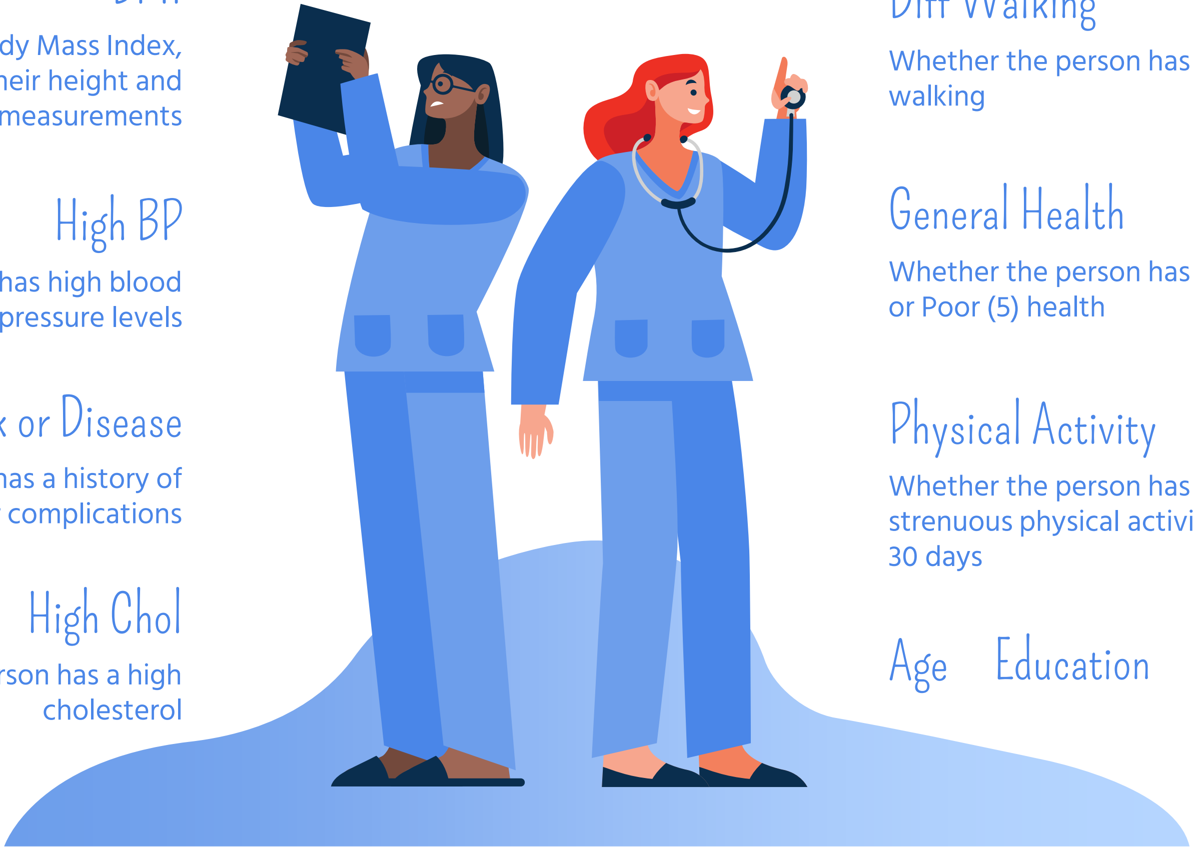
Whether the person has high blood pressure levels

Heart Attack or Disease

Whether the person has a history of cardiovascular complications

High Chol

Whether the person has a high cholesterol



Diff Walking

Whether the person has difficulty walking

General Health

Whether the person has Excellent (1) or Poor (5) health

Physical Activity

Whether the person has done any strenuous physical activity in the last 30 days

Age Education Income

Insights

❑ Poor **general health** and **high BMI** were strongly associated with higher diabetes risk.

❑ **Older adults** face a significantly higher likelihood of diabetes, especially those over 65.

❑ **Lower income and education levels** were linked to higher diabetes prevalence

❑ **Physical inactivity** and **mobility difficulties** also contributed to diabetes risk

Challenges and Limitations

Class imbalance required threshold tuning

No clinical and genetic data

Based on U.S. population, may limit generalizability

GBM required significant computational time

Future Opportunities

01 Include clinical and genetic data

02 Adapt and test the model to non-U.S. populations

03 Explore more advanced models

Recommendations

For Individuals: Prioritize managing BMI and staying physically active; seek regular checkups especially if older or at risk.

For Healthcare Providers: Screen patients with high BMI, high blood pressure, or poor general health more proactively.

For Researchers: Expand predictive models by including more clinical and socioeconomic variables; validate on non-U.S. populations.

Thank You

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

References

Diabetes Canada. (2023, July). *Diabetes in Canada 2023 Backgrounder.* *https://www.diabetes.ca/DiabetesCanadaWebsite/media/Advocacy-and-Policy/Backgrounder/2023_Backgrounder_Canada_English.pdf*

Ginsberg, H. N., Goldberg, R. B., Haffner, S. M., Rivera, G. V., Klein, E. J., Ryan, E. A., ... & ADOPT Study Group. (2000). *Detection and management of prediabetes in the primary prevention of cardiovascular disease and type 2 diabetes.* Circulation, 102(suppl_1), I-377–I-384. *https://doi.org/10.1161/circ.102.suppl_1.I-377*

Hidaji, H. (2025, February 6). *Module 4: Classification* [PowerPoint slides]. University of Calgary D2L site. https://d2l.ucalgary.ca

Hidaji, H. (2025, March 13). *Module 8: Advanced Trees* [PowerPoint slides]. University of Calgary D2L site. https://d2l.ucalgary.ca

Johns Hopkins Medicine. (n.d.). *Diabetes and high blood pressure*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

National Council on Aging. (2022, April 26). *What are 10 warning signs of diabetes in older adults?* https://www.ncoa.org/article/what-are-10-warning-signs-of-diabetes-in-older-adults/

Nazreen, J. (2023, November 27). *Diabetes health indicators dataset*. Kaggle. https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset/data

Public Health Agency of Canada. (2024, October). *Diabetes: Overview.* https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes.html