Diabetes Prediction

Jan Petallo

Haskayne School of Business, University of Calgary

BTMA 531: Data Analytics Tools for Business

Hooman Hidaji

April 14, 2025

Table of Contents

Problem Formulation

Diabetes is a chronic disease affecting around 3.8 million people in Canada over a year old (Public Health Agency of Canada, 2024). Diabetes contributes to various complications including 30% of strokes, 40% of heart attacks, and 50% of kidney failures (Diabetes Canada, 2023). Despite extensive research, there remains ambiguity regarding the dominant factors influencing diabetes risk. This leads to uncertainty among individuals when assessing their risks and taking preventative actions.

This project aims to predict the likelihood of diabetes or prediabetes occurrence based on various health indicators and lifestyle factors. Using predictive analytics, I will identify significant predictors of diabetes, helping to prioritize prevention and early detection strategies. By exploring these factors, this project aims to provide actionable insights for the following:

- At-risk individuals: empower them to understand their health risks and encourage them to take preventive actions
- Healthcare providers: enable them to identify at-risk individuals early
- Researchers: refine existing models and conduct studies

The goal of this study is to develop a predictive model that estimates an individual’s likelihood of having diabetes or prediabetes based on health and lifestyle factors. This includes variables such as BMI, blood pressure status, cholesterol levels, smoking history, physical activity, dietary habits, and others, which will be further described in the next section.

The target variable, *Diabetes_binary*, is a binary classification where 0 (false) represents no diabetes and 1 (true) represents prediabetes or diabetes.

Data

The dataset used for this study is the diabetes_binary_health_indicators_BRFSS2021.csv on Kaggle, derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS). This dataset provides a rich source of health-related data collected from an annual survey conducted by the Centers for Disease Control and Prevention (CDC). The BRFSS dataset has been pre-cleaned and consolidated for machine-learning purposes by a Kaggle contributor. This particular dataset combines individuals with diabetes and prediabetes into one classification (1), contrasting against individuals without diabetes (0). The selected file includes 236,378 records and 21 feature variables, which represent a variety of health indicators and lifestyle factors.

Table 1 List of Variables

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)

The tables describing the Age, Education, and Income groups:

Table 2 Age groups

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

Table 3 Education Groups

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)

Table 4 Income Groups

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

Analysis

Exploratory Data Analysis (EDA)

Before building the models, the dataset was explored to understand the distribution of the target variable and its relationship with other predictors. This step helps identify trends, potential issues, and variables that may influence the model.

The dataset includes over 236,000 observations. No missing values were found, so no cleaning was necessary. The target variable, *Diabetes_binary*, is imbalanced: approximately 86% of the respondents do not have diabetes, while 14% do. This imbalance is important to note when evaluating model performance.

The numeric variables were examined:

- BMI is fairly right skewed with most values between 20 and 40. The Q-Q plot also confirms that BMI is not normally distributed. A boxplot shows that BMI tends to be higher among individuals with diabetes or prediabetes.
- Mental Health Days (MentHlth) and Physical Health Days (PhysHlth) are highly skewed with many zeros, indicating that most respondents report no such issues in the past month. The boxplot for MentHlth shows a slightly higher median for those with prediabetes or diabetes, although the overlap suggests that it may not be a strong independent predictor. In contrast, the boxplot for PhysHlth exhibits a more distinct difference: individuals with diabetes or prediabetes report more unhealthy physical days.

A barplot of HighBP shows a higher proportion of diabetes among those who report high blood pressure, which could mean that this is an important predictor.

Basic correlations among numeric variables were relatively low, with the highest being 0.31 between MentHlth and PhysHlth. This means that there is low information overlap among these numeric predictors.

Data Partitioning

The dataset was randomly split into a 70% training set and a 30% test set to ensure robust model evaluation. This split allows us to train our models on one portion of the data and test them on a separate portion to evaluate their performance. A seed value of 2025 was used for reproducibility, and this same seed was applied throughout the project. Variable type conversions were done as needed prior to model training, particularly for models like GBM that require numeric inputs.

Model 1: Logistic Regression

Logistic regression was selected as the first model because it is a simple, standard and interpretable method for binary classification tasks, such as predicting diabetes status (Hidaji, 2025). In this problem, this model estimates the probability that an observation belongs to a particular class (1 = prediabetes or diabetes), based on the logistic function.

The model was trained on the full set of predictors using the training set. The model summary output indicated that many predictors were statistically significant (p-value < 0.05), including high blood pressure (HighBP), cholesterol level (HighCol), BMI, age group, and physical activity (PhysActivity).

To improve predictive performance given that the target variable is imbalanced, a series of thresholds (from 0 to 1 in increments of 0.05) were tested to evaluate their effect on classification accuracy, sensitivity, and specificity of the model.

```
   Threshold  Accuracy  Sensitivity Specificity
1       0.00 0.1427786 1.0000000000   0.0000000
2       0.05 0.4957977 0.9541728395   0.4194509
3       0.10 0.6480667 0.8588641975   0.6129563
4       0.15 0.7313507 0.7537777778   0.7276152
5       0.20 0.7821023 0.6464197531   0.8047015
6       0.25 0.8139154 0.5422222222   0.8591686
7       0.30 0.8347717 0.4450370370   0.8996858
8       0.35 0.8482669 0.3551604938   0.9303986
9       0.40 0.8562061 0.2749629630   0.9530178
10      0.45 0.8603097 0.2060246914   0.9692872
11      0.50 0.8617198 0.1488395062   0.9804570
12      0.55 0.8619314 0.1013333333   0.9886164
13      0.60 0.8614096 0.0647901235   0.9940943
14      0.65 0.8600558 0.0377283951   0.9970225
15      0.70 0.8585470 0.0186666667   0.9984372
16      0.75 0.8576445 0.0078024691   0.9991939
17      0.80 0.8573906 0.0035555556   0.9996052
18      0.85 0.8572637 0.0011851852   0.9998519
19      0.90 0.8572214 0.0002962963   0.9999506
20      0.95 0.8572214 0.0000000000   1.0000000
21      1.00 0.8572214 0.0000000000   1.0000000
```

In this project, accuracy was not the primary criterion for selecting the cutoff because it can be misleading in imbalanced datasets. A model can have high accuracy by mostly predicting the majority class (0 = without diabetes), even if it fails to identify the minority class (1 = prediabetes or diabetes).

The threshold of 0.15 as it provided more trade-off between sensitivity and specificity. In healthcare, particularly for detecting diseases like diabetes, prioritizing sensitivity helps minimize cases where a person with the condition is misclassified as not having it (false negatives). This is critical because undiagnosed diabetes can lead to severe health complications if not addressed early. Research suggests that false negatives represent missed opportunities for timely treatment, which can be more harmful than misclassifying a person without the condition as having it (false positive), especially when follow-up tests or treatments are low risk (Ginsberg et al., 2000).

Setting the threshold to 0.15 captures more true cases of diabetes or prediabetes, while still maintaining a decent level of accuracy.

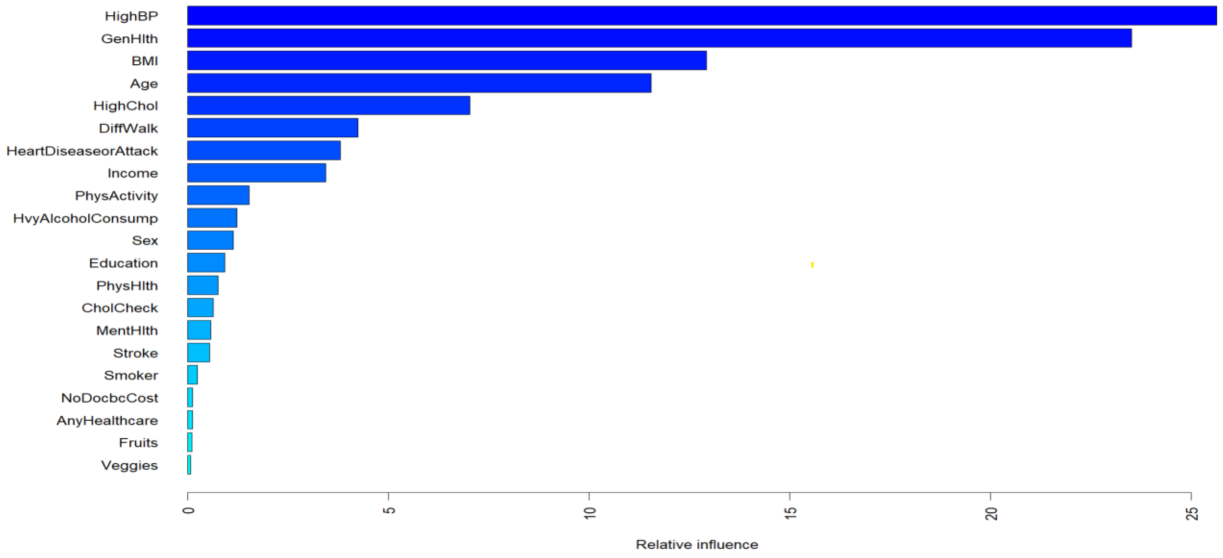
Model 2: Gradient Boosting Machine (GBM)

GBM was chosen as the second model because it is known to offer strong predictive performance through the use of sequential trees that correct errors made by the previous ones. Unlike logistic regression, GBM can capture complex, non-linear interactions between variables (Hidaji, 2025)

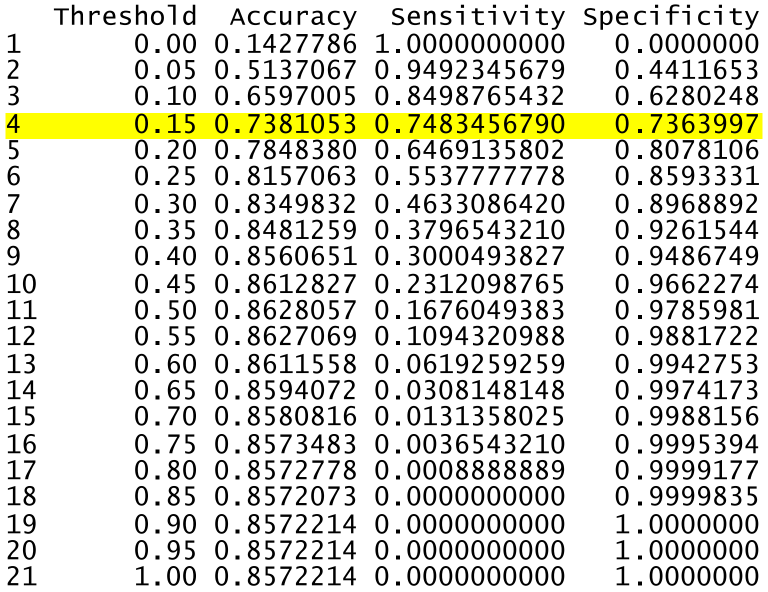
A grid search approach, like seen in the class notes, was used to tune the model across various shrinkage rates (learning rates: 0.01, 0.05, 0.1) and interaction depths (tree complexity: 1, 3, 5). Each combination was evaluated using 5-fold cross-validation based on AUC (Area under the Curve).

	shrinkage	interaction.depth	optimal_trees	max_auc
1	0.05	5	587	0.8256935
2	0.10	5	298	0.8253184
3	0.05	3	1000	0.8245348
4	0.10	3	465	0.8240428
5	0.01	5	1000	0.8221401
6	0.01	3	1000	0.8201834
7	0.10	1	847	0.8189938
8	0.05	1	962	0.8188932
9	0.01	1	1000	0.8135349

As seen in the table above, the best-performing model used a shrinkage of 0.05 and an interaction depth of 5, with 587 trees determined as the optimal number. The GBM model was then trained on the full training set using these selected parameters. Feature importance analysis revealed that High Blood Pressure, General Health, BMI, and Age were the most influential predictors.



As with the logistic regression model, thresholds between 0 and 1 were tested to determine the best cutoff for classifying observations.



```
   Threshold  Accuracy  Sensitivity Specificity
1       0.00 0.1427786 1.0000000000   0.0000000
2       0.05 0.5137067 0.9492345679   0.4411653
3       0.10 0.6597005 0.8498765432   0.6280248
4       0.15 0.7381053 0.7483456790   0.7363997
5       0.20 0.7848380 0.6469135802   0.8078106
6       0.25 0.8157063 0.5537777778   0.8593331
7       0.30 0.8349832 0.4633086420   0.8968892
8       0.35 0.8481259 0.3796543210   0.9261544
9       0.40 0.8560651 0.3000493827   0.9486749
10      0.45 0.8612827 0.2312098765   0.9662274
11      0.50 0.8628057 0.1676049383   0.9785981
12      0.55 0.8627069 0.1094320988   0.9881722
13      0.60 0.8611558 0.0619259259   0.9942753
14      0.65 0.8594072 0.0308148148   0.9974173
15      0.70 0.8580816 0.0131358025   0.9988156
16      0.75 0.8573483 0.0036543210   0.9995394
17      0.80 0.8572778 0.0008888889   0.9999177
18      0.85 0.8572073 0.0000000000   0.9999835
19      0.90 0.8572214 0.0000000000   1.0000000
20      0.95 0.8572214 0.0000000000   1.0000000
21      1.00 0.8572214 0.0000000000   1.0000000
```

The threshold of 0.15 was again selected for consistency and for prioritizing sensitivity, which remains important when the cost of missing cases is high.

Results and Discussion

Findings and Model Performance Comparison

Both models were evaluated using the same test set and the same threshold of 0.15, as discussed in the previous section. This allows for a fair comparison.

Table 5 Model Performance

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.7314	0.7538	0.7276	0.8173
GBM	0.7381	0.7483	0.7364	0.8208

The table above shows that the performance differences between the two models are quite small These results suggest that both models are similarly effective for predicting diabetes or prediabetes in the dataset. Ultimately, the choice between the models depends on the operational context. For applications prioritizing slightly stronger predictive performance, GBM may be preferred. However, if interpretability and simplicity of the model are more critical, logistic regression remains a strong and competitive choice.

Insights

Both the logistic regression coefficients and GBM feature rankings provide useful insights into diabetes prediction. Across both models, several features consistently emerged as strong predictors, including BMI, high blood pressure (HighBP), general health (GenHlth), and Age.

Studies consistently show that individuals with a BMI of 30 or higher have a five to ten times higher risk of developing type 2 diabetes compared to those within a healthy BMI range (British Heart Foundation, 2020). High blood pressure is also ranked the highest in the feature importance. This aligns with the fact that "high blood pressure is twice as likely to strike a person with diabetes than a person without diabetes" (Johns Hopkins Medicine, n.d.).

Individuals who rated their general health poorer was consistently associated with higher diabetes risk. The importance of age is also not surprising given that diabetes prevalence increases with age. In fact, an estimated of 33% of people aged 65 and older are affected by diabetes (National Council on Aging, 2022).

Factors such as physical activity, income and education levels showed associations with diabetes prediction. This reinforces the importance of social factors in identifying at-risk individuals. Individuals with higher income and education levels were significantly less likely to have diabetes, according to the logistic regression model. This is consistent with what was found by Diabetes Canada:

- "The prevalence of diabetes among adults in the lowest income groups is 2.1 times that of adults in the highest income group."
- "Adults who have not completed high school have a diabetes prevalence 1.9 times that of adults with a university education."

This highlights that higher income groups generally have better access to healthcare, healthier food options, and safer environments for physical activity. Similarly, higher educational attainment can be linked to greater health literacy.

These insights help guide efforts by showing which areas to focus on. For example, managing BMI, encouraging physical activity, and controlling blood pressure are important steps to help prevent diabetes

Challenges and Limitations

- Class imbalance made model training and evaluation more difficult and required extra steps like threshold tuning. Relying on accuracy metrics would have led to misleading conclusions.
- The dataset lacked clinical and genetic data, which play a great part in a person's risk of developing diabetes.
- The dataset is based on U.S. population data, which may limit how well our findings apply to other countries. This is because people in different regions may have different diets, lifestyles, access to healthcare, and health systems overall.
- Logistic regression model is simpler and more interpretable but may miss non-linear relationships.
- GBM model required significant computational time. I had to limit the number of parameter combinations tested during hyperparameter tuning to maintain a reasonable runtime.

Future Opportunities

- Incorporate clinical and genetic data to improve model precision and capture additional risk factors.
- Adapt and test the model to non-U.S. populations, especially in regions that differ significantly from the U.S. population
- Explore more advanced models to enhance prediction performance

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.

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