Diabetes Report

By Taylen Inthisane

Executive summary

Primary care providers often struggle to identify high-risk patients early because lifestyle behaviors are inconsistently reported and routine visits occur infrequently. These gaps lead to missed opportunities for early intervention. To address this challenge, I analyzed a large diabetes dataset and developed a machine learning model designed to estimate a patient's diabetes risk at the time of a routine check-up.

My workflow included exploratory data analysis, data cleaning, feature preparation, and model development using three algorithms: Logistic Regression, Random Forest, and XGBoost. Model interpretability was supported through SHAP values, which highlighted the features most strongly associated with diabetes risk. I also created a risk-stratification algorithm that categorizes patients into "Low," "Medium," and "High" risk groups to make the model's outputs more actionable for clinical decision-making.

One of the most striking findings was the prevalence of diabetes in the dataset: out of 100,000 patients, nearly 60,000 had some form of diabetes (Type 1, Type 2, or Gestational). Across all models, metabolic indicators emerged as the strongest predictors of diabetes, particularly HbA1c, fasting glucose, and postprandial glucose. Hereditary factors (family history of diabetes) and lifestyle behaviors (weekly physical activity minutes) also contributed meaningfully to risk.

By identifying the variables most strongly associated with diabetes and translating model outputs into clear risk categories, this project demonstrates how machine learning can support clinicians

in allocating resources, increasing patient awareness, and improving early detection and preventive care.

Introduction

Primary care providers face significant challenges in identifying high-risk patients early, largely due to infrequent clinical visits and inconsistent self-reported lifestyle information. As a result, many patients are not diagnosed with diabetes or do not receive appropriate treatment until complications arise. Without awareness of their risk, patients may continue habits that increase the likelihood of developing diabetes, even when some of those risks could be mitigated through early education and behavioral changes.

Risk is amplified when lifestyle reporting is inaccurate. For example, a patient may claim to follow a healthy diet while regularly consuming fast food, masking an important contributor to metabolic health. Infrequent visits create additional blind spots: clinicians may lock the longitudinal data needed to monitor trends in blood pressure, glucose levels or other indicators that could signal emerging risks.

The goal of this project is to predict a patient's diabetes risk during routine check-ups using only pre-diagnosis data. This approach is highly feasible due to the widespread use of electronic medical records, which provide historical baselines that can be combined with newly collected visit data to reassess risk over time.

Implementing a diabetes risk algorithm can help clinicians flag high-risk patients earlier and provide targeted resources that support preventive care. This is especially important because diabetes does not affect all populations equally. Patients from lower socioeconomic backgrounds often face disproportionate barriers to care, including the high cost of insulin and limited access

to preventive services. By increasing awareness of individual risk, patients may be empowered to take proactive steps that reduce the likelihood of developing diabetes and its associated long-term health and financial burdens.

Data Understanding

The dataset that is used in this project is *Diabetes Health Indicators Dataset by Moha Krishna Thalla* and was accessed from Kaggle. It contains 100,000 patient records and 31 variables spanning demographics, socioeconomic factors, lifestyle behaviors, medical history, vital signs and metabolic indicators. These variables collectively provide a comprehensive view of the patient's health and potential diabetes risk.

Below is an organized overview of the variables included in this dataset:

Demographic Variables

- age – Patient age; diabetes risk generally increases with age.
- gender – Biological sex, which can influence hormonal and metabolic responses.
- ethnicity – Ethnic background; some groups have higher genetic susceptibility to diabetes.

Socioeconomic Variables

- education_level – Highest level of education completed; may correlate with health literacy and access to resources.
- income_level – Income category; higher income can increase access to nutritious food and healthcare services.

- employment_status – Employment may reflect financial stability and access to healthcare.

Lifestyle and Behavioral Variables

- smoking_status – Smoking can influence metabolic health and is associated with insulin resistance.
- alcohol_consumption_per_week – Excessive alcohol intake contributes to weight gain and poor metabolic regulation.
- physical_activity_minutes_per_week – Physical activity improves insulin sensitivity and reduces obesity risk.
- diet_score – Overall diet quality; poor nutrition is a known contributor to diabetes risk.
- sleep_hours_per_day – Adequate sleep supports hormonal regulation and metabolic health.
- screen_time_hours_per_day – High screen time is associated with sedentary behavior and reduced sleep quality.

Medical History Variables

- family_history_diabetes – Genetic predisposition to diabetes.
- hypertension_history – High blood pressure is commonly associated with metabolic disorders.
- cardiovascular_history – Cardiovascular issues often co-occur with diabetes risk factors.

Anthropometric and Vital Sign Variables

- bmi – Body mass index; higher BMI is strongly associated with type 2 diabetes.

- waist_to_hip_ratio – Indicator of central obesity, a key risk factor for metabolic disease.
- systolic_bp – Systolic blood pressure; elevated levels are linked to metabolic dysfunction.
- diastolic_bp – Diastolic blood pressure; similar implications as systolic BP.
- heart_rate – Elevated resting heart rate can reflect underlying metabolic stress.

Lipid Profile Variables

- cholesterol_total – Total cholesterol; shares common risk pathways with diabetes.
- hdl_cholesterol – "Good" cholesterol; low HDL is associated with higher diabetes risk.
- ldl_cholesterol – "Bad" cholesterol; elevated LDL is associated with metabolic disorders.
- triglycerides – High triglycerides often accompany insulin resistance and prediabetes.

Metabolic Indicators

- glucose_fasting – Fasting blood glucose level; a standard measure for assessing diabetes risk.
- glucose_postprandial – Blood glucose after meals; reflects how well the body processes sugar.
- insulin_level – Elevated insulin may indicate insulin resistance.
- hba1c – Average blood glucose over the past 2–3 months; a key diagnostic marker.

Derived and Target Variables

- diabetes_risk_score – A composite score estimating likelihood of diabetes.
- diabetes_stage – Categorical indicator of diabetes status (e.g., No Diabetes, Pre-Diabetes, Type 1, Type 2, Gestational).

- diagnosed_diabetes – Binary target variable indicating whether the patient has diabetes (1 = Yes, 0 = No).

Findings

Most variables in the dataset followed approximately normal distributions. A few such as insulin levels, weekly physical activity and alcohol consumption are right-skewed. The dataset arrived, pre-cleaned, with no missing values, though outlier were present across nearly all variables. I chose not to remove outliers because they reflect the natural variability in patient lifestyle and clinical measurements.

The dataset is imbalance, with 60% of patients diagnosed with diabetes and 40% without, which aligns with correlations observed during EDA.

Several notable trends emerged:

- Age: Older patients were more likely to be diagnosed with diabetes.
- Glucose levels: Patients with elevated fasting or postprandial glucose almost always had diabetes
- Family history: A family history of diabetes significantly increased the likelihood of diagnosis.

One surprising finding was the strength of genetic influence relative to lifestyle behaviors. Because the dataset isolates lifestyle variables from hereditary ones, the impact of family history appears more pronounced than expected. Habitual factors such as diet score or physical activity had weaker associations with diabetes than anticipated.

Another unexpected result was the limited effect of the socioeconomic variables. I initially expected education and income to play a larger role, given their influence on health literacy and access to nutritious food. However, diabetes prevalence appeared relatively consistent across education and income levels.

4. Data Preparation

The dataset came pre-cleaned, so I did not have to address any missing values. Almost every variable contained outliers; however, because the data represents a wide range of lifestyle and health background, I chose not to remove them. These values likely reflect real variation rather than data entry errors, and removing them could distort the underlying population.

There were two variables that required modification. The first was age, which I binned into five groups: 18-44, 45-64, 65-74, 75-84, and 85+. Binning age is consistent with standard medical and epidemiological practice,where patients are grouped into clinically meaningful life-stage categories. These categories reflect non-linear changes in metabolic risks and allow analyses to compare groups that differ in physiologically relevant ways. In other words, binning age does not simply “group numbers together”; it aligns the model with how hospitals, public health agencies, and clinical researchers stratify patients to understand risk patterns across age cohorts.

The second variable that required adjustment was gender. I removed the “Other” category due to ambiguity in the dataset documentation. The term “gender” could refer to gender identity or biological sex, and the dataset did not specify which definition was intended. Because diabetes risk is tied to biological sex rather than gender identity, and because the “Other” category is heterogeneous and cannot be interpreted consistently, I excluded it from the analysis. This decision unfortunately leaves intersex and transitioning patients underrepresented, but I chose

this approach to avoid producing recommendations that could be misleading or potentially harmful.

After addressing these two variables, I encoded all categorical features. This increased the total number of variables in the dataset to 42.

For training and testing data, I used a standard 80/20 train-test split, allocating 80% of the data for model training and 20% for testing. After splitting, I scaled the numeric variables and then examined the distribution of the target variable. The resulting split contained 59.95% diabetic patients and 40.05% non-diabetic patients. Although this is not a 50/50 balance, I chose not to adjust the class distribution because it accurately reflects the underlying population in the dataset, which has a similar 60/40 split. Maintaining this natural distribution avoids artificially altering the prevalence of diabetes and preserves the real-world context of the data.

5. Modeling

To model the data, I evaluated three baseline classifiers: Logistic Regression, Random Forest, and XGBoost. Testing multiple models allowed me to compare performance across linear and non-linear approaches and to assess how well each model captured the underlying relationships in the dataset. Model performance was evaluated using ROC-AUC, precision, recall, F1-score, and the confusion matrix, providing a comprehensive view of both discrimination and classification behavior.

Logistic Regression performed the weakest of the three models, achieving an overall accuracy of 86%. Its precision scores were 0.84 (non-diabetic) and 0.88 (diabetic), with recall values of 0.81 and 0.89, and F1-scores of 0.82 and 0.89, respectively. Despite these limitations, Logistic

Regression still produced a strong ROC-AUC of 0.9345, indicating good ability to distinguish between diabetic and non-diabetic patients across thresholds. As a baseline model, it performed reasonably well, but it did not match the performance of the tree-based models.

Both Random Forest and XGBoost performed similarly across most metrics. Each achieved an overall accuracy of 92%, with precision scores of 0.84 (non-diabetic) and 1.00 (diabetic), recall values of 1.00 and 0.87, and F1-scores of 0.91 and 0.93, respectively. Their ROC-AUC values were nearly identical: 0.9451 for Random Forest and 0.9453 for XGBoost.

Because the models were so close in performance, the confusion matrix became the deciding factor. While both models performed similarly, the Random Forest classifier produced one fewer false positive than XGBoost. Specifically, Random Forest achieved 7,846 true negatives and 2 false positives, compared to XGBoost's 7,845 true negatives and 3 false positives. Although the difference is small, reducing false positives means fewer patients are incorrectly flagged as diabetic and subjected to unnecessary follow-up or resource allocation.

Given the near-identical performance across all other metrics, this reduction in false positives combined with Random Forest's strong interpretability and stability led to selecting Random Forest as the final model for this analysis

6. Model Interpretability

To understand which features most strongly influenced a patient's diabetes diagnosis, I used two complementary interpretability methods: a feature importance chart from the Random Forest model and a SHAP summary plot. Using both approaches provides a more complete picture of

how the model makes decisions, since feature importance captures global influence while SHAP values show both magnitude and direction of impact at the individual-prediction level.

The feature importance chart showed that metabolic variables dominated the model's decision-making. HbA1c, fasting glucose, and postprandial glucose were the three most influential predictors. HbA1c contributed the most at 43%, followed by postprandial glucose at 19%, and fasting glucose at 9%. After these core metabolic markers, the next most influential variable was weekly physical activity, though its contribution was much smaller at around 2%.

The SHAP summary plot confirmed the same three leading predictors HbA1c, fasting glucose, and postprandial glucose but revealed a slightly different ordering. While HbA1c still had the strongest overall impact, fasting glucose appeared more influential in the SHAP analysis than in the feature importance chart. This difference is expected because SHAP values capture how much each feature pushes individual predictions toward or away from diabetes, rather than just measuring how often a feature is used in tree splits.

One notable discrepancy between the two methods was the ranking of secondary variables. The SHAP plot highlighted family history of diabetes as a more influential contributor than the feature importance chart suggested, and it also placed weekly physical activity within the top five predictors. In contrast, the feature importance chart ranked BMI higher than family history. These differences reflect the fact that SHAP values account for interaction effects and non-linear relationships, while feature importance measures average contribution across all trees.

Despite these minor differences, both interpretability methods clearly showed that metabolic indicators were by far the strongest predictors of diabetes, with lifestyle and demographic

variables contributing meaningfully but to a much lesser degree. This alignment across methods increases confidence that the model is capturing clinically consistent patterns in the data.

7. Risk Stratification Tool

To translate model predictions into actionable clinical categories, I developed a risk stratification tool that maps each patient's predicted probability of diabetes into one of three tiers: Low risk, Medium Risk, or High Risk. The thresholds were chosen to balance interpretability and clinical relevance:

- Low Risk: Probability < 0.33
- Medium Risk : 0.33 <= Probability <0.66
- High Risk : Probability >= 0.66

These thresholds reflect intuitive thirds of probability scale and align with common stratification frameworks used in preventive medicine. This model approach prioritizes simplicity and transparency for clinical development.

In practice, these categories allow providers to quickly identify patients who may benefit from additional lab testing, lifestyle counseling, or enrollment in preventive health programs. For example, a patient flagged as high risk could be scheduled for follow-up HbA1c testing, while a medium risk patient might receive targeted lifestyle guidance.

After running the risk stratification algorithm on 19,598 patients, the population was divided into three categories: 10,227 High Risk, 8,945 Low Risk, and 426 Medium Risk. This distribution was striking over 52% of patients were classified as High Risk, indicating that a majority may require additional clinical resources, follow‑up testing, or preventive interventions. Equally

notable was the scarcity of Medium Risk patients, who accounted for just 2% of the population. This suggests that the model tends to make confident predictions, pushing patients toward either end of the risk spectrum rather than clustering them near the decision boundary. The bimodal distribution may reflect strong signal separation in the data, particularly among metabolic variables, and has direct implications for how clinics might allocate resources and prioritize care.

9. Discussion

After running both the feature importance analysis and the SHAP interpretability framework, it became clear that metabolic variables specifically HbA1c, fasting glucose, and postprandial glucose were the most influential predictors of diabetes diagnosis. This aligns with established clinical understanding, as these biomarkers directly reflect glycemic control and are central to diabetes screening and diagnosis. SHAP values further confirmed that these variables consistently pushed predictions toward the diabetic class across individual patients, reinforcing their dominant role in the model's decision-making process. Lifestyle and hereditary variables such as weekly physical activity and family history of diabetes also contributed meaningfully, highlighting the multifactorial nature of diabetes risk.

Among the three models evaluated, the Random Forest classifier demonstrated the strongest overall performance. Although Random Forest and XGBoost produced nearly identical precision, recall, F1-scores, and accuracy, the deciding factor came from the confusion matrix. Random Forest produced one fewer false positive than XGBoost. While this difference may appear small numerically, its clinical implications are significant. A false positive in this context means incorrectly labeling a non-diabetic patient as diabetic, which could lead to unnecessary

follow-up testing, emotional stress, and potential financial burden. Given the rising costs associated with diabetes care, even a single avoided misclassification represents a meaningful improvement in patient experience and resource allocation.

The broader population-level findings further emphasize the importance of early detection. With nearly 60% of the dataset consisting of diabetic patients, the burden of diabetes is substantial. A model that can accurately identify high-risk individuals at routine check-ups has the potential to support earlier intervention, lifestyle modification, and preventive care. Even small improvements in early detection can translate into long-term benefits for patients, especially in a healthcare environment where the cost of diabetes management continues to rise. While diabetes is a manageable condition, access to treatment is not equitable, and early identification remains one of the most effective tools for reducing complications and improving outcomes.

Although the dataset was generally high quality, several limitations should be acknowledged. The most notable issue was the ambiguity of the "gender" variable. The dataset did not specify whether "gender" referred to biological sex or gender identity, which introduces uncertainty into the interpretation of this feature. Because diabetes risk is more closely tied to biological sex, the lack of clarity reduces the reliability of this variable. Additionally, the "Other" category was too vague to interpret meaningfully it could represent intersex individuals, patients undergoing gender transition, or simply data entry inconsistencies. These groups may have unique physiological profiles, particularly related to hormone levels, that were not captured or modeled appropriately.

Another limitation involved computational constraints. Running more computationally intensive models or hyperparameter tuning was challenging due to hardware limitations, which may have

restricted the full exploration of model performance. While Random Forest and XGBoost performed well, more extensive tuning or the use of additional ensemble methods could potentially yield further improvements.

Finally, several lifestyle variables such as diet score, physical activity, and alcohol consumption were self-reported, which introduces subjectivity and potential measurement error. These variables are important because they represent modifiable behaviors, but their reliability is inherently limited.

A valuable direction for future work would be to focus specifically on habitual or behavioral variables to better understand their isolated impact on diabetes risk. Many metabolic and hereditary factors are outside a patient's control, but lifestyle behaviors are modifiable and represent an opportunity for targeted intervention. Exploring a model built primarily on behavioral features could help identify which habits have the strongest influence on diabetes risk and which interventions may be most effective.

However, this approach requires more reliable data. Many lifestyle variables in the current dataset are self-reported and may be subjective or inconsistently measured. An ideal future dataset would include objective, longitudinal behavioral tracking such as wearable device data, continuous glucose monitoring, or structured dietary logs for a smaller cohort. This would allow for a deeper understanding of how daily habits influence diabetes development over time.

Additionally, future work could explore:

- more advanced hyperparameter tuning
- model calibration to improve probability estimates

- fairness analysis across demographic groups
- integration of clinical guidelines to refine risk thresholds

These enhancements would strengthen the model's clinical utility and improve its ability to support early detection and preventive care.

10. Conclusion

This analysis demonstrates that machine learning can play a meaningful role in supporting early detection of diabetes in primary care settings. By leveraging routinely collected demographic, lifestyle, and metabolic indicators, the model effectively identifies patients who are at elevated risk. The value of such a system extends far beyond its implementation cost: early identification gives patients the opportunity to modify behaviors, seek follow-up testing, and receive preventive care before their condition progresses to more severe stages.

The ability to intervene early is critical. Preventive care is fundamentally different from treatment after diagnosis. Early action can reduce long-term complications, improve quality of life, and lessen the financial burden associated with diabetes management. A risk-based approach also helps clinicians allocate resources more efficiently by prioritizing patients who would benefit most from additional monitoring or diagnostic testing.

To deploy this model in a real clinical environment, data engineers and software developers would need to integrate the algorithm into the primary care system's existing electronic health record (EHR) or ERP infrastructure. This would allow the model to automatically pull patient data, generate risk scores, and present actionable insights directly within the clinician's workflow. With proper integration, a diabetes risk detection system could enhance patient care

by enabling earlier intervention, supporting preventive health programs, and ultimately reducing the number of patients who progress to full diabetes without warning.