# Early Detection of Diabetes in Women

12/09/2024

# 1. Abstract

Diabetes is a serious health condition, and early detection is essential to prevent complications and improve outcomes. This study aims to develop predictive tools to identify the risk of diabetes onset in women based on health-related factors such as glucose levels, BMI, age, and family history. Using the Diabetes Dataset from Kaggle, which includes data like blood pressure, insulin levels, and genetic history, two machine learning models, K-Nearest Neighbors (K-NN) and Logistic Regression (LR), were evaluated for their effectiveness in predicting diabetes. Our results show that K-NN achieved an accuracy of 70.13 %, while LR achieved 72%, highlighting their potential as effective tools for early diabetes detection. This research provides valuable insights for improving healthcare interventions and personalized patient care.

# 2. Introduction

## 2.1 Background

In 2021, approximately 537 million adults worldwide were living with diabetes, representing one in ten adults globally. Projections from the International Diabetes Federation (IDF) estimate that this number will rise to 643 million by 2030 and 784 million by 2045 [1], highlighting the growing concern of this chronic condition. The Western Pacific region currently holds the largest share of diabetic patients, reflecting global disparities in prevalence. Middle-income countries are disproportionately affected, with 81% of diabetic adults residing in these regions, where underdiagnosis is common, delaying treatment and increasing complications. The rising prevalence places immense pressure on global healthcare systems, with spending on diabetes management reaching USD 966 billion in 2021, a 316% increase compared to a decade ago. Driven by lifestyle factors such as poor diet, physical inactivity, and obesity, particularly for type 2 diabetes, this trend underscores the urgent need for widespread preventive measures, early detection, and sustainable healthcare strategies.

## 2.2 Motivation & Application

To fully address this challenge, it is crucial to understand the fundamental differences between Type 1 and Type 2 diabetes, as their distinct causes, onset, and treatment approaches require tailored interventions and management strategies. In general terms, Type 1 is caused by genetic and immune system issues, while Type 2 is mostly related to lifestyle. Type 1 diabetes is an autoimmune disease where the body attacks insulin-producing cells in the pancreas, so little or no insulin is made. It usually starts suddenly in childhood or teen years and requires lifelong insulin treatment. Type 2 diabetes happens when the body doesn't use insulin properly or doesn't make enough of it. It develops slowly, mostly in adults, but is becoming more common in children. Type 2 is often linked to being overweight, poor diet, and not being active. Treatment focuses on healthy eating, exercise, medications, and sometimes insulin. Type 2 diabetes is the most prevalent form and has seen a significant rise globally.

In this project, we focus on addressing the growing issue of Type 2 diabetes by helping identify those at risk early. Using a dataset that includes important health information like glucose levels, BMI, age, and family history, we use tools such as **K-Nearest Neighbors (K-NN)** and **Logistic Regression (LR)** to predict diabetes. Our goal is to find practical ways to support early diagnosis, allowing for timely treatment and better management. This work aims to provide useful insights to improve healthcare outcomes and reduce the impact of diabetes on individuals and communities.

# 3. Methodology

## 3.1 Dataset

The Diabetes Dataset is comprised of 768 records of female diabetes patients at the age of 21 years and older. The dataset provides insights into the risks that contribute to diabetes in females and aims to predict the likelihood of the disease occurring. From this data, predictive models can be built to detect and prevent diabetes in individuals.

The dataset is comprised of 768 observations that include eight features and one target variable that outputs whether or not diabetes is present in the patient.

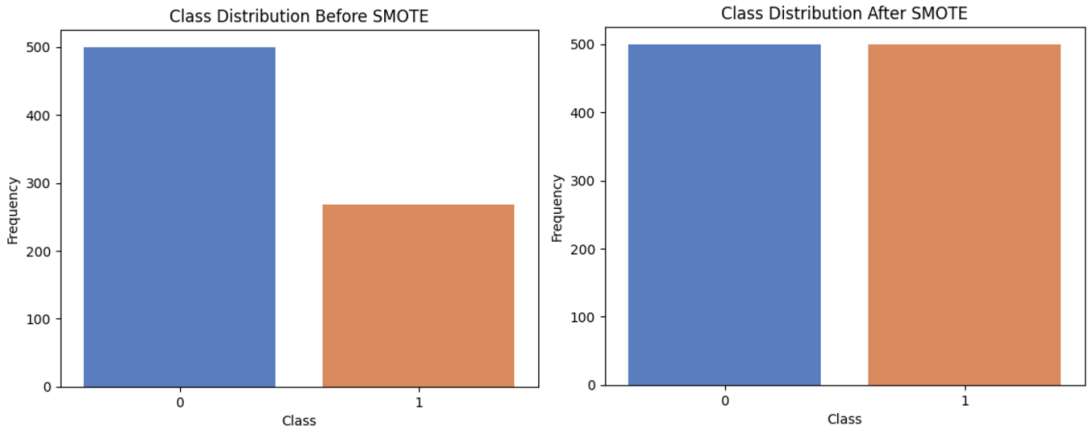
Below is a breakdown of the eight features:

- Pregnancies (Integer): Number of times the patient has been pregnant.
- Glucose (Integer, Continuous): Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.
- Blood Pressure (Integer, Continuous): Diastolic blood pressure (mm Hg).
- Skin Thickness (Integer, Continuous): Triceps skinfold thickness (mm).
- Insulin (Integer, Continuous): 2-hour serum insulin (mu U/ml).
- BMI (Float, Continuous): Body mass index, defined as weight in kg/(height in m)^2.

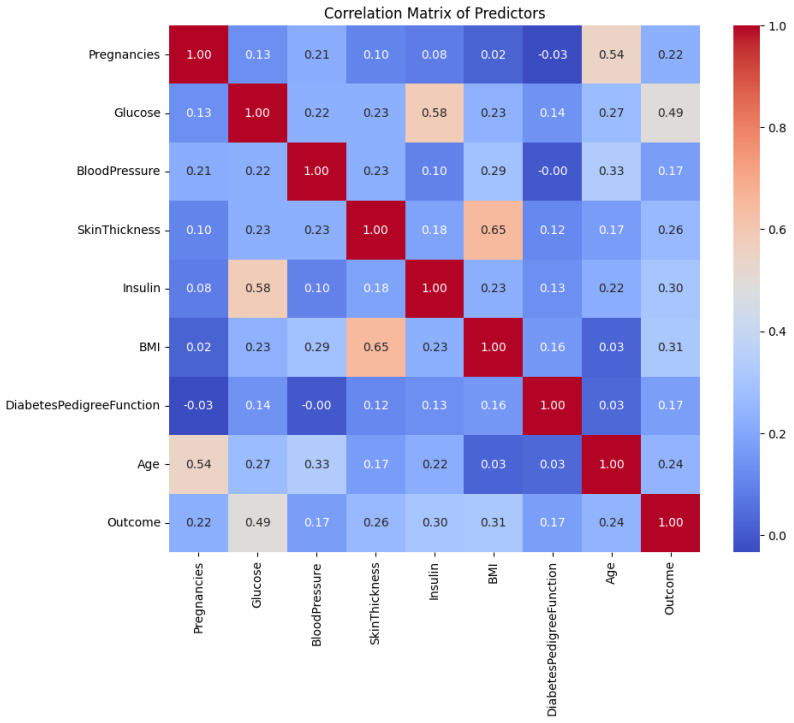
- Diabetes Pedigree Function (Float, Continuous): A score indicating genetic predisposition to diabetes based on family history.
- Age (Integer): Age of the patient (in years).
- Outcome (Binary, Categorical): Target variable where 1 indicates diabetes, and 0 indicates no diabetes.

## 3.2 Preprocessing

We noticed that the dataset was imbalanced, with non-diabetics (500) making up about twice the number of diabetics (268). We realized that this imbalance could pose a significant challenge for training and testing our models. When one class dominates, the model tends to favor it, and might disproportionately predict the non-diabetics, at the expense of accurately identifying diabetic cases. It can be concerning because failing to identify at-risk individuals can delay critical interventions.




To address this imbalance in the dataset, we used a technique called SMOTE (Synthetic Minority Oversampling Technique). This is because the dataset is not large in volume. Therefore, the entire dataset was used as input, but only a small percentage of cases belong to the minority group. It works by identifying each minority class sample's nearest neighbors using the Euclidean distance. From these neighbors, SMOTE selects one at random and generates a new synthetic sample by interpolating between the original sample and its neighbor. The new sample is placed along the line segment connecting these two points, ensuring it lies within the space of the minority class. This process is repeated until there are 50% of people with diabetes and 50% without diabetes. (500 each)

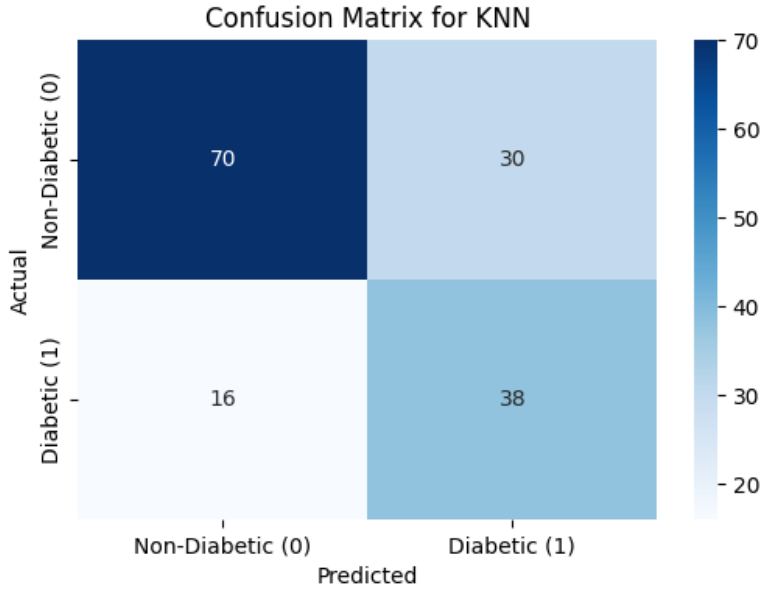



Another key step that we took with preprocessing was examining the relationships between variables in the dataset. Correlation values range from −1 to 1: a value of −1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 suggests no linear relationship. We created the correlation matrix above and found that most variables exhibit low correlations, with values below 0.5. However, there are a couple exceptions for this dataset. For example, there is a clear trend between age and pregnancies, where the number of pregnancies tends to increase with age before stabilizing later in life. Also, glucose and insulin levels show a significant correlation—higher glucose levels are linked to a greater likelihood of a diabetes diagnosis and a higher demand for insulin regulation. Similarly, BMI is strongly correlated with body fat percentage, as individuals with higher BMIs generally have a greater proportion of body fat. Glucose, BMI, skin thickness, and insulin show strong correlations with diabetes outcomes because they reflect key biological factors: glucose indicates blood sugar levels, BMI and skin thickness represent body fat and obesity, which contribute to insulin resistance, and insulin reflects the body's ability to regulate glucose, making them crucial indicators of metabolic health and diabetes risk. These four variables have a correlation above 0.26 when predicting whether a person has diabetes or not.

# 4. Results

## 4.1 KNN Model

As our first method, we used a KNN model. As part of the diabetes classification process, the model calculates how far away a sample, from all other samples in the dataset, is from the sample that needs to be classified. KNN is particularly effective at handling non-linear data, making it a strong candidate for predicting diabetes, as the relationships between health-related attributes such as BMI, insulin levels, and glucose levels are often complex and non-linear. Additionally, KNN works well with multiple features where relationships may not be immediately obvious, which aligns well with our dataset that combines various predictors such as age, BMI, glucose levels, and insulin levels. Using the K nearest samples to the sample to be classified, the class or the numerical value of the sample to be classified is determined.




The confusion matrix above provides an overview of the performance of our KNN model in predicting diabetes on the test dataset. The results show that the model correctly predicted 38 individuals as diabetic (true positives) and 70 individuals as non-diabetic (true negatives). However, it incorrectly classified 30 non-diabetic individuals as diabetic (false positives) and failed to identify 16 diabetic individuals, predicting them as non-diabetic (false negatives). This breakdown allows us to calculate several key performance metrics.

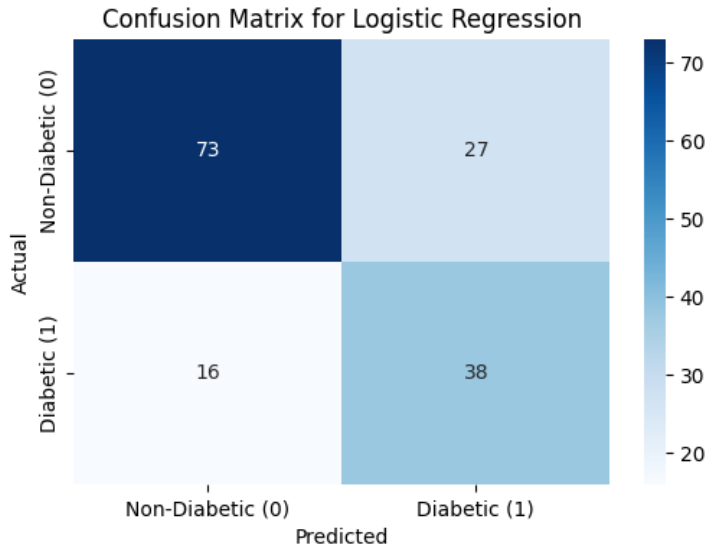
The overall accuracy of the model, which represents the proportion of correctly predicted instances, is approximately 70.13%. This means the model is correct in about seven out of

ten cases. For the non-diabetic class (0), the precision was 81%, meaning the model is highly accurate when predicting non-diabetic cases. However, the recall for this class was 70%, indicating that 30% of actual non-diabetic cases were misclassified. For the diabetic class (1), the precision was lower at 56%, showing a higher rate of false positives, but the recall was 70%, meaning the model correctly identified 70% of actual diabetic cases. The F1-scores were 0.75 for the non-diabetic class and 0.62 for the diabetic class, indicating better balanced performance for non-diabetic predictions.

## 4.2 Logistic Regression

Logistic Regression was chosen as the second method of modeling due to its interpretability and suitability for binary classification problems like predicting diabetes, especially in our dataset. One of the key advantages of Logistic Regression is that it provides clear coefficients for each feature, offering insights into the direction and magnitude of their influence on the target variable (diabetes outcome). This characteristic is particularly useful for understanding the importance of predictors such as BMI, glucose levels, and insulin levels. Additionally, Logistic Regression is specifically designed for binary classification tasks, making it a natural fit for distinguishing between diabetic and non-diabetic individuals in our dataset.

To optimize the model, specific hyperparameters were utilized. The regularization parameter C = 0.1 was applied to prevent overfitting by discouraging reliance on any single feature. Regularization is a technique used to prevent overfitting by penalizing large coefficients, ensuring the model generalizes well to unseen data. Smaller values of C increase the regularization strength, heavily penalizing large coefficients. We set max_iter to 100, which was set to find optimal coefficients efficiently while ensuring convergence. The model works by repeatedly improving its predictions by adjusting the model's coefficients. Each step moves closer to the best fit for the data. However, if the algorithm takes too many steps, it might become inefficient or stuck, so max_iter acts like a time limit. These parameters were chosen to balance computational efficiency and model performance, given the small-to-medium size of our dataset.




Our Logistic Regression model achieved an accuracy of 72%, slightly higher than the KNN model. The confusion matrix shows that the model correctly identified 73 individuals as non-diabetic (true negatives) and 38 individuals as diabetic (true positives). However, 27 non-diabetic individuals were incorrectly predicted as diabetic (false positives), and 16 diabetic individuals were missed, being classified as non-diabetic (false negatives). These results highlight the model's strengths in identifying non-diabetic individuals while maintaining a reasonably good recall for diabetic cases. For the non-diabetic class (0), the model achieved a precision of 82%, meaning it is highly accurate in predicting non-diabetic cases, and a recall of 73%, suggesting that 27% of actual non-diabetic cases were misclassified. For the diabetic class (1), the precision was 58%, showing a relatively higher rate of false positives, while the recall was 70%, indicating that the model correctly identified 70% of actual diabetic cases. The F1-scores were 0.77 for the non-diabetic class and 0.64 for the diabetic class, highlighting better performance for non-diabetic predictions.

# 5. Discussion

In this study, we explored how machine learning models such as K-Nearest Neighbors (K-NN) and Logistic Regression (LR) can help predict diabetes in women using health-related factors like glucose levels, BMI, and insulin. Our results showed that Logistic Regression performed slightly better, achieving an accuracy of 72% compared to 70.13% for K-NN. Logistic Regression's strength lies in its simplicity and interpretability, making it easier to understand which factors contribute most to diabetes risk. Meanwhile, K-NN, though effective at identifying non-linear patterns, struggled with precision, leading to more false positives.

One of the most important takeaways was the role of glucose levels, BMI, and insulin as key predictors of diabetes. This makes sense because elevated glucose levels and insulin imbalances are direct indicators of diabetes, while a higher BMI reflects obesity, which often leads to insulin resistance. These findings reinforce what we already know about diabetes risk and highlight areas where healthcare providers can focus on such as weight management and glucose control, for early detection.

While the models performed reasonably well, the study had some limitations. The dataset was relatively small and initially imbalanced, with twice as many non-diabetic cases as diabetic ones. To address this, we used SMOTE to balance the data, but challenges with false positives persisted. Including additional factors like diet, lifestyle habits, and physical activity could improve predictions in future studies. Exploring more advanced models, such as Random Forests or ensemble methods, may also help achieve better accuracy and precision.

Overall, our findings show that machine learning has real potential to support early diabetes detection. Logistic Regression stands out as a reliable tool due to its accuracy and ability to explain which health factors matter most. By identifying at-risk individuals early, we can improve healthcare interventions and help reduce the growing impact of diabetes worldwide.

# 6. References


1. "IDF Diabetes Atlas." *IDF Diabetes Atlas*, diabetesatlas.org/. Accessed 3 Dec. 2024.
2. Rahman, Md. Hasibur. "Diabetes Dataset." *Kaggle*, 22 Oct. 2024, www.kaggle.com/datasets/hasibur013/diabetes-dataset.
3. *1.3 Billion People Worldwide Projected to Have Diabetes by 2050 - The Washington Post*, www.washingtonpost.com/wellness/2023/07/10/diabetes-worldwide-billion-people/. Accessed 10 Dec. 2024.