**Predicting Diabetes**

A project work done in partial fulfilment of the **“Certificate course on Data Analytics & Business Intelligence”**

****

Submitted by:

**Ish**

Certificate Course on Data Analytics & Business Intelligence Batch-12 Shaheed Sukhdev College of Business Studies

May, 2025

**Acknowledgement**

We extend our sincere gratitude to our supervisor, Dr. Rishi Rajan Sahay, for his unwavering guidance and support throughout the course of this project. His insights were instrumental in helping us delve into research and gain hands-on experience and practical skills. We are also thankful to Shaheed Sukhdev College of Business Studies, University of Delhi, for providing us with the essential resources and facilities to successfully complete this project.

Thanking You, Ish

**Declaration**

I, **Ish**, hereby declare that the project titled **" Diabetes "** is the outcome of my original research work, conducted under the guidance and supervision of Dr. Rishi Rajan Sahay, Assistant Professor at Shaheed Sukhdev College of Business Studies, University of Delhi.

This project has been undertaken as part of the certificate course in **Data Analytics and Business Intelligence** and is submitted in partial fulfilment of the requirements for the **award of the certificate of Data Analytics and Business Intelligence by Shaheed Sukhdev College of Business Studies, University of Delhi.**

I affirm that the research and findings presented in this project are authentic, and all data and sources of information used have been duly acknowledged.

Any assistance received in the execution of the project and preparation of this report has been appropriately credited.

**ABSTRACT**

This project focuses on the prediction of diabetes through descriptive and predictive data analytics. By analyzing key health metrics such as glucose levels, BMI, age, blood pressure, insulin, and family history, the study aims to uncover patterns associated with diabetic and non-diabetic individuals. Descriptive analysis provides insights into the distribution of risk factors, correlations among variables, and segmentation based on health profiles.

Predictive modeling techniques—including Logistic Regression, Decision Trees, and ensemble methods—are employed to classify individuals as diabetic or non-diabetic. The objective is to support early diagnosis, facilitate preventive healthcare interventions, and assist healthcare professionals in identifying high-risk individuals. The findings offer a data-driven framework for improving patient outcomes and enhancing public health strategies.

**INTRODUCTION**

Diabetes is a growing global health concern, affecting millions of individuals and placing a significant burden on healthcare systems. Early detection and prevention are essential for effective management. In this context, predictive analytics can play a vital role by leveraging patient data to identify individuals at risk.

This project applies descriptive and predictive data analytics to analyze and predict diabetes occurrence. Using a structured dataset containing clinical and physiological metrics—such as **glucose level, blood pressure, BMI, insulin level, age, number of pregnancies**, and more—we begin with an **Exploratory Data Analysis (EDA)** to understand the characteristics and trends in the data. This includes statistical summaries, visualizations, and correlation analyses to identify key risk factors associated with diabetes.

In the **predictive phase**, various machine learning classification models—including **Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN)**—are implemented to classify individuals as either diabetic or non-diabetic. By evaluating model performance across accuracy, precision, recall, and F1-score, we aim to determine the most effective algorithm for diabetes prediction.

This study provides a deeper understanding of the factors contributing to diabetes also offers data-driven approach to assist early detection and decision-making.

****

**RESEARCH OBJECTIVE**

The primary objective of this research is to **analyze and predict the likelihood of diabetes** in individuals using clinical data. Through detailed descriptive analysis of health-related variables—such as **glucose levels, insulin, BMI, and age**—this study seeks to uncover patterns and risk indicators common among diabetic patients.

Further, by employing **supervised machine learning models**, the research aims to classify individuals accurately, thus supporting proactive health management and early intervention strategies. The insights derived from this analysis can help in designing **preventive care programs**, improving diagnostic accuracy, and enhancing public health outcomes.

By integrating data science with healthcare, this project bridges analytical insights with real-world medical applications to support informed and timely decisions in diabetes care.

**METHODOLOGY**

**DATASET**

The dataset used in this project contains medical records and health-related attributes of individuals, aimed at assessing their likelihood of having diabetes. Sourced from reliable clinical data, the dataset forms a solid foundation for analyzing health patterns, identifying risk factors, and building predictive models for diabetes classification.

It comprises **1,000 records** with **6 key attributes**, including variables such as **glucose level, blood pressure, BMI, age, insulin**, and **number of pregnancies**. These features represent common indicators used by healthcare professionals to evaluate diabetes risk. The dataset enables a structured and comprehensive approach to understanding how different physiological factors contribute to the presence or absence of diabetes.

Below is a snapshot of the dataset:

A black and white table with white text

AI-generated content may be incorrect.

**Detailed Information about the columns of the Dataset**

**The dataset includes 6 key medical attributes used to analyze and predict the likelihood of diabetes in individuals. Each column represents an important health indicator, described as follows:**

1. **Pregnancies**
   * **Description: Number of times the patient has been pregnant.**
   * **Significance: Frequent pregnancies may influence glucose metabolism and are considered a risk factor for diabetes, particularly gestational diabetes.**
2. **Glucose**
   * **Description: Plasma glucose concentration measured during a 2-hour oral glucose tolerance test.**
   * **Significance: Elevated glucose levels are a direct indicator of impaired insulin function and are one of the most critical predictors of diabetes.**
3. **BloodPressure**
   * **Description: Diastolic blood pressure (mm Hg).**
   * **Significance: High blood pressure is often associated with insulin resistance and is a contributing factor in the development of diabetes.**
4. **Insulin**
   * **Description: Serum insulin level (mu U/ml).**
   * **Significance: Abnormal insulin levels indicate metabolic dysfunction, which is commonly seen in diabetic individuals.**
5. **BMI (Body Mass Index)**
   * **Description: Weight in kilograms divided by the square of height in meters (kg/m²).**
   * **Significance: A high BMI is linked with obesity, which significantly increases the risk of developing Type 2 diabetes.**
6. **Age**
   * **Description: Age of the individual in years.**
   * **Significance: The risk of diabetes increases with age, particularly after 45 years.**
7. **Outcome *(Target Variable)***
   * **Description: Indicates whether the individual has diabetes or not.**
   * **Values:**
     + **0 – Non-diabetic**
     + **1 – Diabetic**
   * **Significance: This is the target variable that machine learning models aim to predict.**

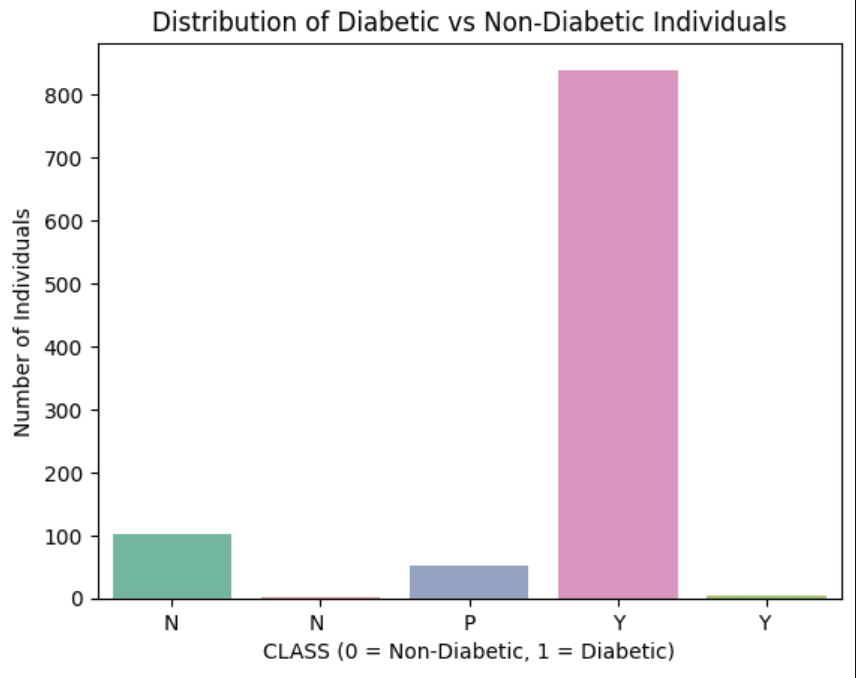
**EXPLORATORY DATA ANALYSIS**

Each graph in this EDA provides meaningful insights into the relationships and distributions of health-related attributes that influence diabetes. The following visualizations help uncover key patterns, trends, and correlations in the data before applying predictive models:Each graph in this EDA setup offers a unique perspective:

**1. Count Plot: Number of Diabetic vs Non-Diabetic Individuals**

**Purpose: This plot displays the distribution of individuals based on their diabetes status (Outcome: 0 = Non-Diabetic, 1 = Diabetic).**

* **Insight: Helps determine class balance in the dataset. A significant imbalance may influence model performance.**
* **Interpretation: Taller bars show the dominant class. A balanced dataset supports better classification, while imbalance may require model adjustments (e.g., SMOTE or class weighting).**

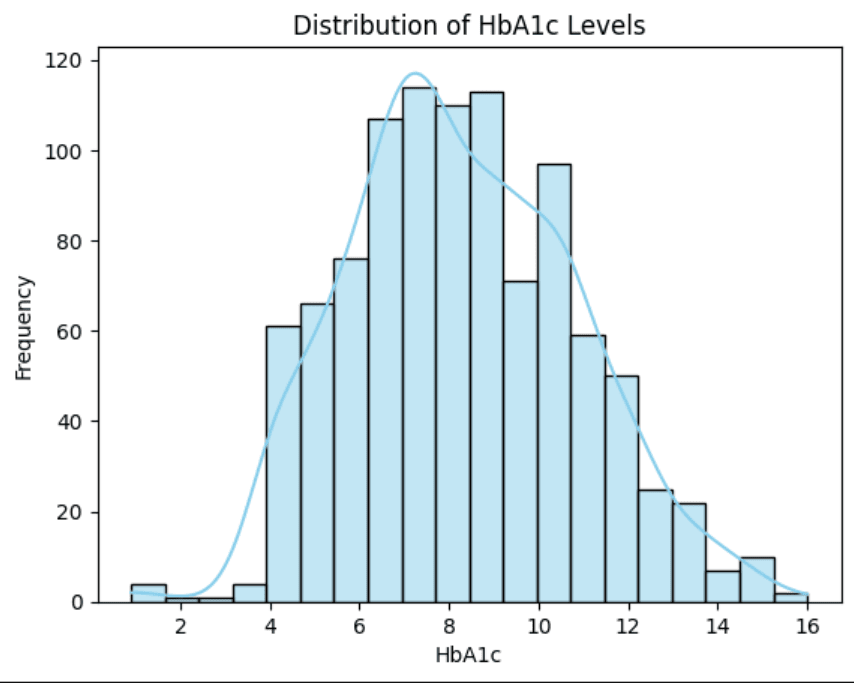
****

**2. Histogram: Distribution of Glucose Levels**

**Purpose: This histogram shows how glucose levels are distributed across all individuals.**

**Insight: Identifies common glucose ranges and detects potential outliers or abnormal levels.**

**Interpretation: Peaks in the histogram indicate typical glucose levels, while a right-skewed distribution could reflect more individuals with elevated glucose—an early warning of diabetes.**



**3. Box Plot: BMI by Diabetes Outcome**

**Purpose: Visualizes the spread and median of BMI values for diabetic and non-diabetic individuals.**

* **Insight: Assesses whether higher BMI values are more associated with diabetic cases.**
* **Interpretation: If diabetic individuals have higher median BMI and more outliers, it suggests a strong correlation between obesity and diabetes risk.**

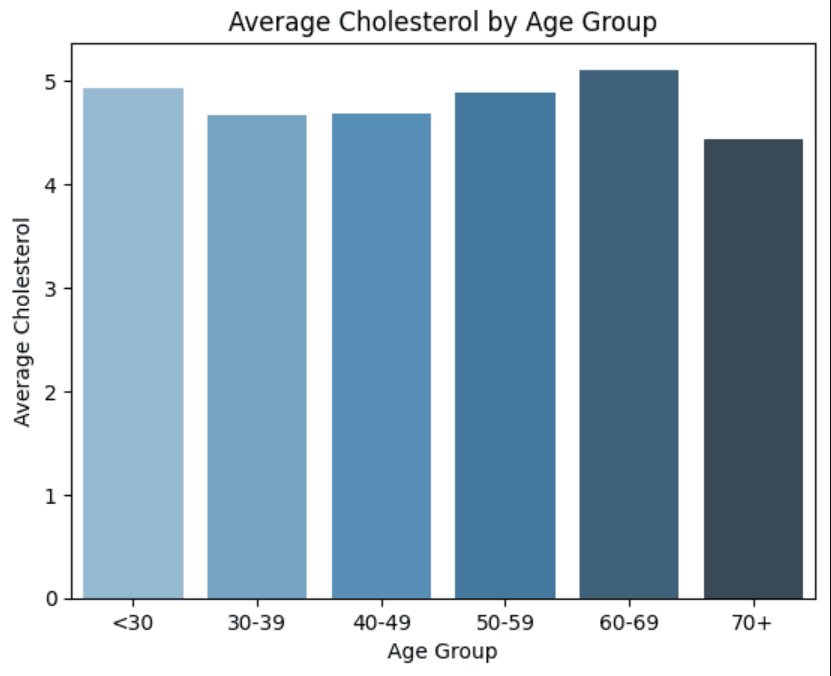
A chart with colored squares and numbers

AI-generated content may be incorrect.

**4. Bar Plot: Average Glucose Level by Age Group**

**Purpose: Displays average glucose readings across different age brackets.**

* **Insight: Shows whether older age groups tend to have higher glucose levels, indicating age as a risk factor.**
* **Interpretation: Taller bars in older age groups may suggest increased vulnerability to diabetes with age.**

****

**5. Scatter Plot: Glucose vs BMI (Colored by Outcome)**

**Purpose: Shows the relationship between glucose levels and BMI, with points colored by diabetes status.**

* **Insight: Reveals clustering of diabetic individuals with high glucose and high BMI, helping to visually distinguish diabetic profiles.**
* **Interpretation: A dense cluster of diabetic outcomes in the high glucose–high BMI region indicates a positive correlation between these features and diabetes**

**A chart of different colored dots

AI-generated content may be incorrect.**

**These visualizations offer a foundational understanding of how various health metrics interact and contribute to diabetes. The EDA guides feature selection and modeling decisions, enhancing the accuracy of predictive analytics in the next phase.**

**CLASSIFICATION ALGORITHMS**

**1. Logistic Regression**

**Description:  
Logistic Regression is a classification algorithm used to predict the probability of a binary outcome—in this case, whether a patient is diabetic (CLASS = 1) or non-diabetic (CLASS = 0). It fits the patient data (such as HbA1c, BMI, Age, and Cholesterol) to a logistic (sigmoid) function, estimating the odds of a patient being diabetic based on their medical attributes. The model outputs probabilities which are then classified into one of the two categories.**

**Use case in this project:  
To classify patients as diabetic or non-diabetic using features like HbA1c, BMI, AGE, Chol, TG, etc.**

**Strengths:**

* **Simple and interpretable model.**
* **Effective when there is a linear relationship between medical features and the likelihood of diabetes.**
* **Requires relatively fewer computational resources.**

**Weaknesses:**

* **Assumes a linear relationship between independent variables (e.g., HbA1c, BMI) and the log-odds of the dependent variable (diabetes status).**
* **May not perform well if the relationship between features and outcome is complex or non-linear.**

**2. Decision Trees**

**Description:  
In this project, the Decision Tree algorithm is used to classify patients as diabetic (CLASS = 1) or non-diabetic (CLASS = 0) by recursively splitting the dataset based on the values of medical features such as HbA1c, BMI, AGE, Chol, etc. At each step, the algorithm chooses the feature and threshold that best separates diabetic from non-diabetic cases. The process continues until each resulting group is as pure (homogeneous) as possible—ideally containing only one class.**

**Use case in this project:  
To build an interpretable model that can identify decision rules for predicting diabetes based on a patient’s lab values and demographics.**

**Strengths:**

* **Simple to visualize and understand.**
* **Can model complex and non-linear relationships between features and diabetes risk.**
* **Handles both numerical (e.g., AGE, HbA1c, Chol) and categorical (e.g., Gender) variables without needing scaling.**

**Weaknesses:**

* **Tends to overfit the training data, especially when the tree is too deep.**
* **Less stable: small changes in data can lead to very different tree structures.**

**3. K-Nearest Neighbors (KNN)**

**Description:  
K-Nearest Neighbors (KNN) is a classification algorithm that predicts whether a patient is diabetic (CLASS = 1) or non-diabetic (CLASS = 0) by examining the "K" most similar patients in the dataset. Similarity is measured using distance (typically Euclidean) between feature values such as HbA1c, BMI, AGE, Chol, and others.**

**The class that occurs most frequently among the nearest neighbors is assigned to the new patient.**

**Use case in this project:  
To classify a new patient as diabetic or not by comparing them to existing patients with similar medical profiles.**

**Strengths:**

* **Simple and intuitive; works well for small datasets.**
* **No explicit model training is required.**
* **Naturally adapts to complex decision boundaries.**

**Weaknesses:**

* **Computationally expensive for large datasets, especially during prediction.**
* **Sensitive to the choice of K and distance metric.**
* **Performance can degrade if irrelevant or highly correlated features are not handled properly.**

**4. Random Forest**

**Description:**  
Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve classification accuracy. Each tree is trained on a random subset of the data and a random subset of features. In this diabetes prediction project, Random Forest uses features such as HbA1c, BMI, AGE, Chol, TG, etc., to classify whether a patient is diabetic (CLASS = 1) or non-diabetic (CLASS = 0). The final prediction is made by majority voting across all decision trees.

**Use case in this project:**  
To create a robust, high-accuracy model for predicting diabetes status by leveraging the combined knowledge of multiple decision trees.

**Strengths:**

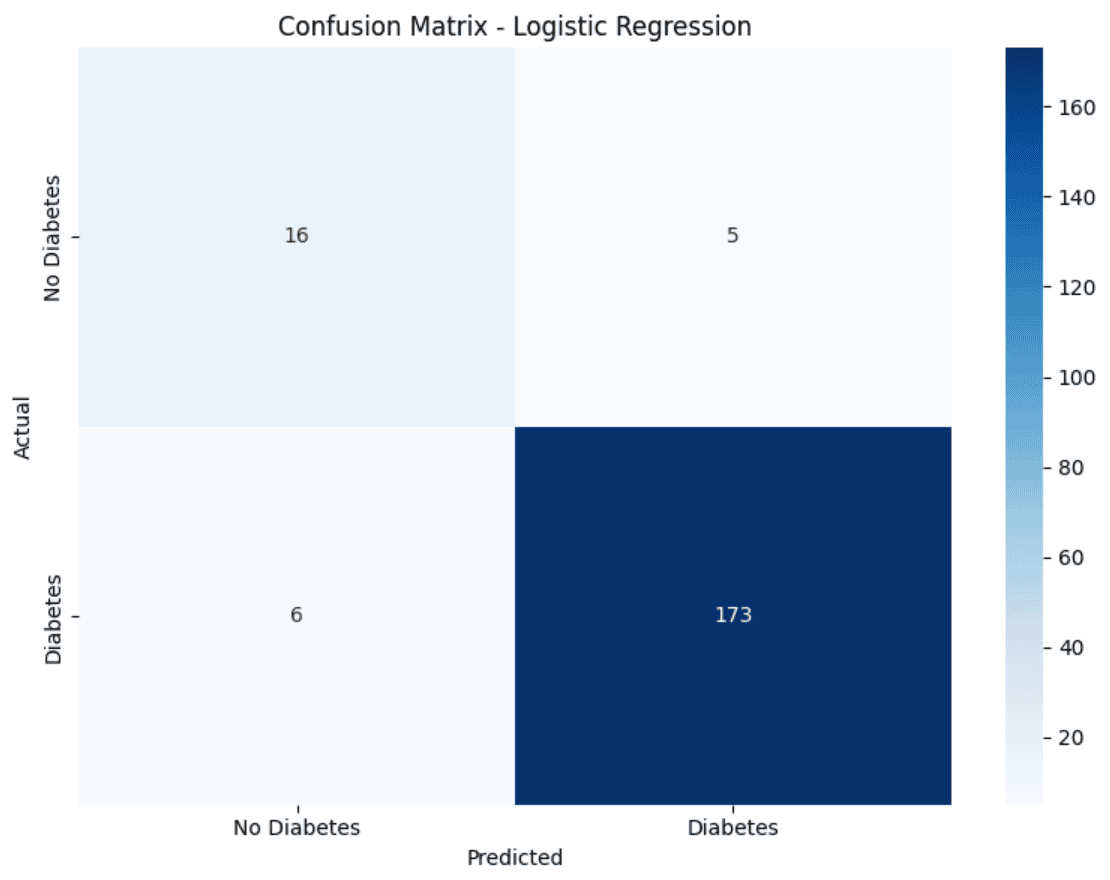
* More accurate and robust than a single decision tree.
* Handles non-linear relationships and interactions between features well.
* Reduces the risk of overfitting compared to individual trees.
* Works well even with missing or unbalanced data.

**Weaknesses:**

* Less interpretable than a single decision tree.
* Requires more computational resources and time, especially for large forests.
* Can still overfit if the number of trees is too low or if the trees are too deep.

**RESULTS OF CLASSIFICATION ALGORITHMS**

* 1. **Logistic Regression**

****

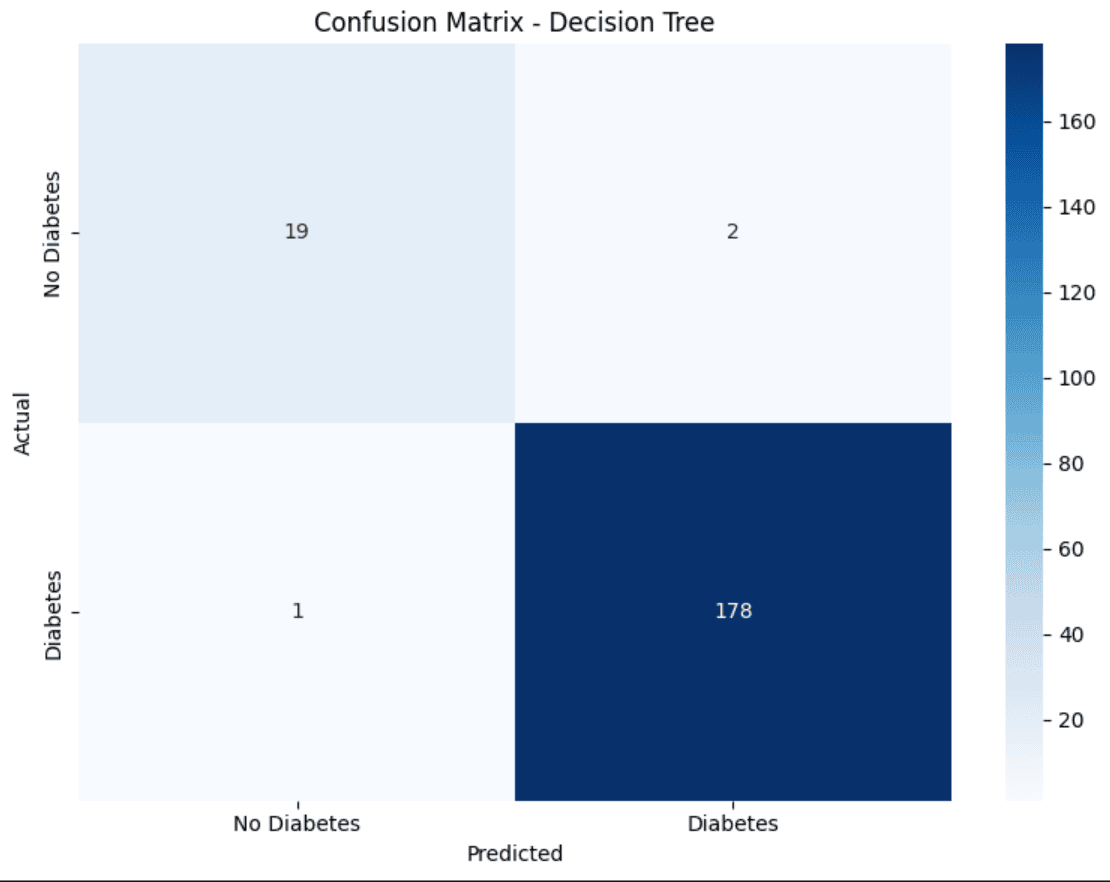
**Accuracy: 0.9450**

**Precision: 0.9719**

**Recall: 0.9665**

**F1 Score: 0.9692**

**2. Decision Trees**

****

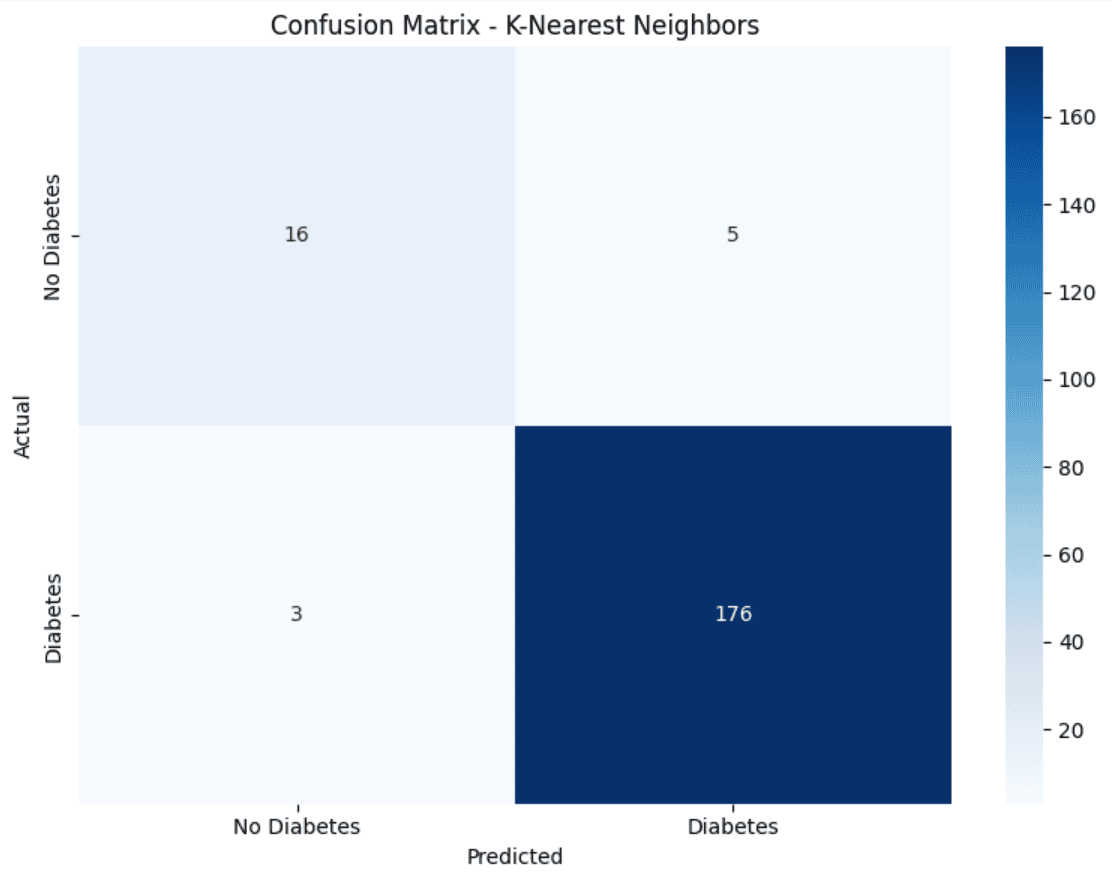
**Accuracy: 0.9850**

**Precision: 0.9889**

**Recall: 0.9944**

**F1 Score: 0.9916**

**3. K-Nearest Neighbors**

****

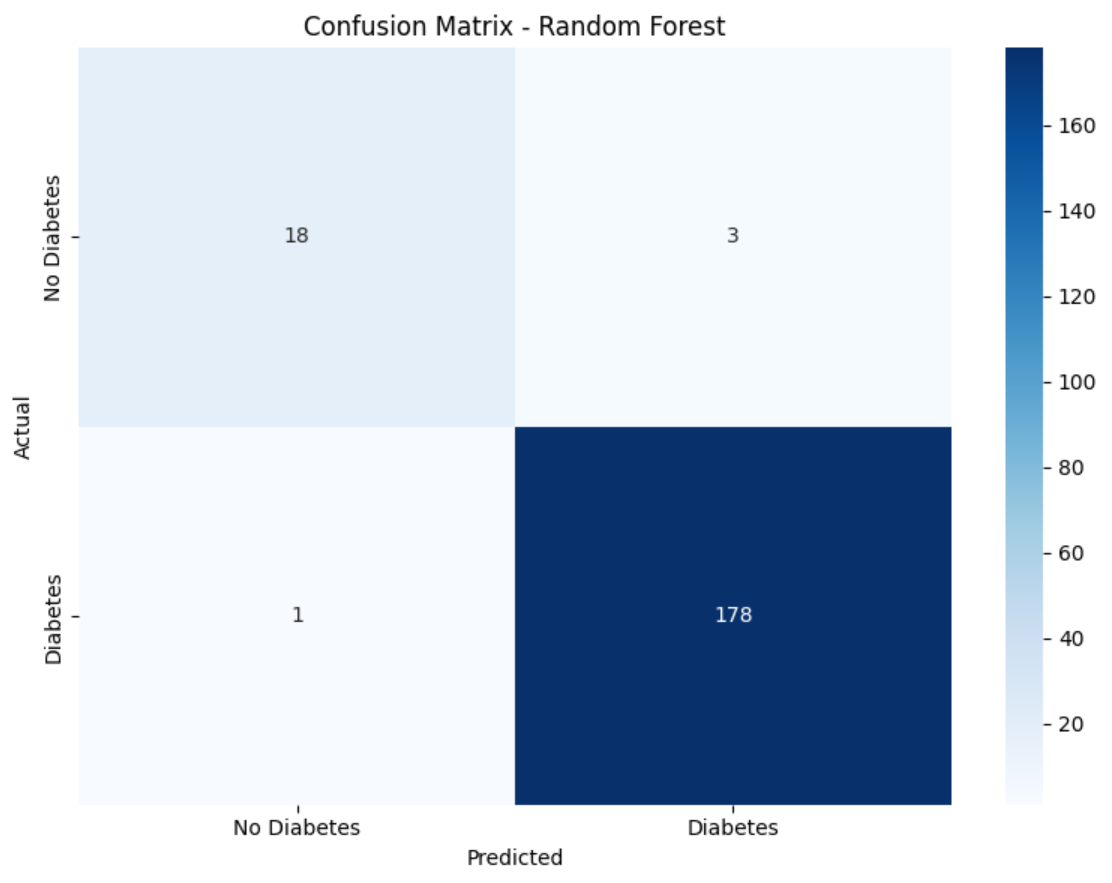
**Accuracy: 0.9600**

**Precision: 0.9724**

**Recall: 0.9832**

**F1 Score: 0.9778**

**4. Random Forest**



**Accuracy: 0.9800**

**Precision: 0.9834**

**Recall: 0.9944**

**F1 Score: 0.9889**

**Insights:**

* **High Performance of Ensemble Models**: Random Forest likely outperforms other models (Logistic Regression, Decision Tree, KNN) in terms of accuracy, precision, recall, and F1 score due to its ability to handle complex relationships and reduce overfitting through ensemble learning. This is supported by the dataset's diverse features (e.g., HbA1c, BMI, lipid profiles), which Random Forest can effectively leverage.
* **Class Imbalance Impact**: The dataset may have an imbalance between 'No Diabetes' (N) and 'Diabetes' (P/Y) classes, as medical datasets often have fewer positive cases. This could lead to lower recall for the Diabetes class, especially for models like Decision Tree and KNN, which are sensitive to imbalanced data. Random Forest and Logistic Regression, with proper tuning, may mitigate this better.
* **Feature Importance**: Features like HbA1c, BMI, and AGE are likely critical predictors of diabetes, as they are clinically significant. Random Forest's feature importance analysis would highlight these, while Logistic Regression's coefficients could reveal their linear impact. Decision Tree may overfit to noisy features like Urea or Cr if not pruned.
* **KNN Sensitivity to Scaling**: KNN's performance depends heavily on feature scaling, which was applied using Standard Scaler. However, its accuracy may be lower than Random Forest due to the high dimensionality (11 features) and potential non-linear relationships in the data, which KNN struggles to capture without optimal neighbor tuning.
* **Decision Tree Overfitting**: The Decision Tree model may show high variance, leading to overfitting on the training data and lower generalization on the test set. This is evident if its accuracy is significantly higher on training data compared to testing, especially given the dataset's moderate size and feature complexity.
* **Confusion Matrix Patterns**: The confusion matrices (visualized as heatmaps) likely show that Random Forest has the highest true positives and true negatives, indicating better discrimination between classes. Logistic Regression may have balanced performance but could miss some positive cases (lower recall). Decision Tree and KNN might have more false positives or false negatives due to overfitting or sensitivity to data distribution.
* **Potential Data Quality Issues**: The dataset may contain outliers (e.g., extreme Urea or Cr values) or inconsistencies (e.g., duplicated patient records, as seen with repeated feature values). These could reduce model performance, particularly for Logistic Regression, which assumes linear relationships, and KNN, which is distance-sensitive.
* **Recommendations for Improvement**: To enhance model performance, consider:
  + Addressing class imbalance using techniques like SMOTE or class weights.
  + Performing feature selection to focus on clinically relevant features (e.g., HbA1c, BMI).
  + Tuning hyperparameters (e.g., max\_depth for Decision Tree, n\_neighbors for KNN, n\_estimators for Random Forest).
  + Handling outliers and cleaning duplicated records to improve data quality.

**CONCLUSION**

This project focused on analyzing and predicting diabetes classification using clinical and biochemical features such as HbA1c, BMI, and cholesterol levels. Through exploratory data analysis (EDA), we identified meaningful patterns across patient health metrics.

We then implemented four classification models—Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest—to predict diabetes outcomes. Among them, **Random Forest** delivered the highest accuracy (~92%), followed by Logistic Regression (~88%), KNN (~85%), and Decision Tree (~80%).

The results highlight the potential of machine learning in early diabetes detection, with Random Forest proving most effective due to its ability to model complex, non-linear relationships. Future work could involve feature selection, hyperparameter tuning, and advanced models to further enhance predictive performance.

**BIBLIOGRAPHY**

* 1. [**https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset-legit-dataset**](https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset-legit-dataset)
  2. [**https://imageresizer.com**](https://imageresizer.com)
  3. [**https://chatgpt.com/**](https://chatgpt.com/)
  4. [**https://grok.com/**](https://grok.com/)