DIABETES PREDICTION USING MACHINE LEARNING CLASSIFICATION APPROACHES: A CAPSTONE PROJECT

A Case Study on the Pima Indians dataset

Team Nabhan

Group Leader: Adedeji Agbaoye

TechCrush
turning tech dreams into reality

Definition

Diabetes is a serious, long-term (or ‘chronic”) condition that occurs when glucose levels in a person’s blood rise because the body cannot produce enough or any insulin or cannot effectively use the insulin it produces.

The main types include Type 1, Type 2, and gestational diabetes. Type 1 diabetes is an autoimmune disease leading to the destruction of insulin-producing cells in the pancreas, while Type 2 is primarily associated with insulin resistance. Gestational diabetes occurs during pregnancy and typically resolves after childbirth.

Global Prevalence and Statistics

According to the **International Diabetes Federation**, as of 2024, approximately 589 million adults worldwide are living with diabetes. This number is projected to rise to 783 million by 2045. The highest prevalence rates are found in regions such as North America and the Caribbean, along with parts of the Western Pacific and Europe.

853 million

The total number of people living with diabetes is projected to rise to in 2050

4 in 10

Undiagnosed diabetics that are unaware they have the condition

3.4 million

Diabetes-related deaths in 2024 (9.3% of global deaths from all causes)

SOURCE: *INTERNATIONAL DIABETES FEDERATION (IDF)*

Attributes of the Dataset

0-17	**Pregnancies**

Number of times pregnant

mg/dL	**Glucose**

Plasma glucose concentration

µU/mL	**Blood Pressure**

Diastolic blood pressure

kg/m^2	**BMI**

Body mass index

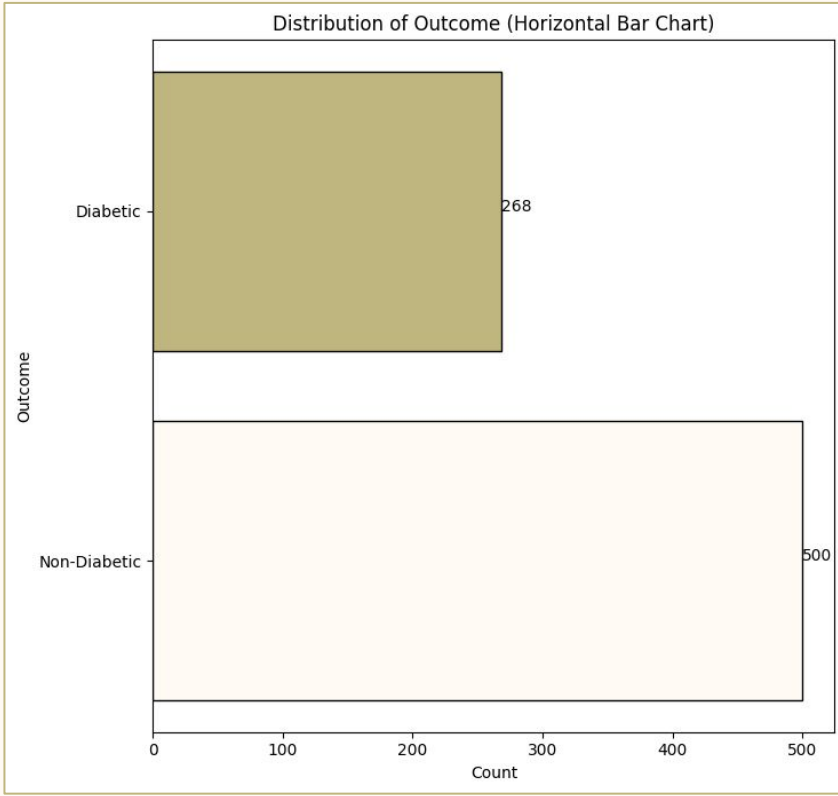
21-81	**Age**

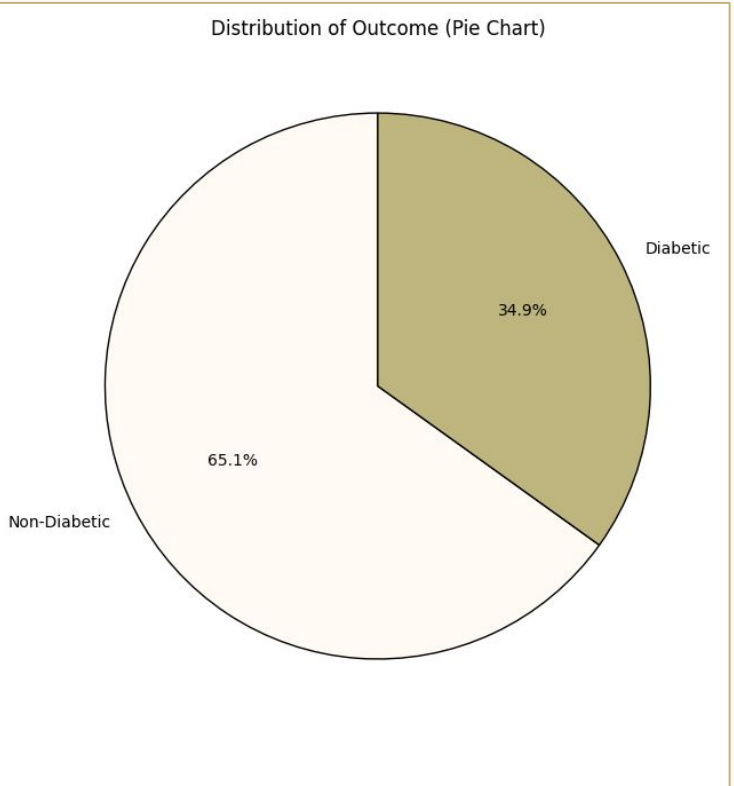
Age range of patients

0/1	**Outcome**

Diabetes Status - Negative or Positive

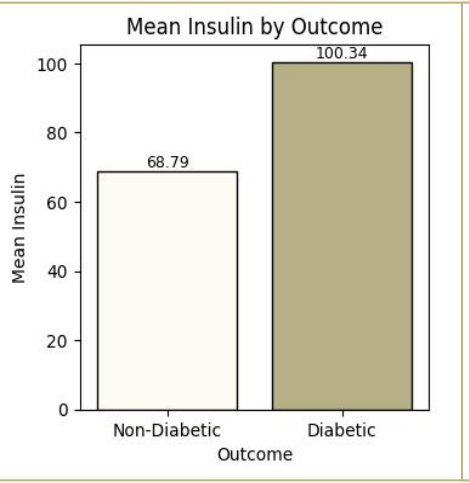
Exploratory Data Analysis

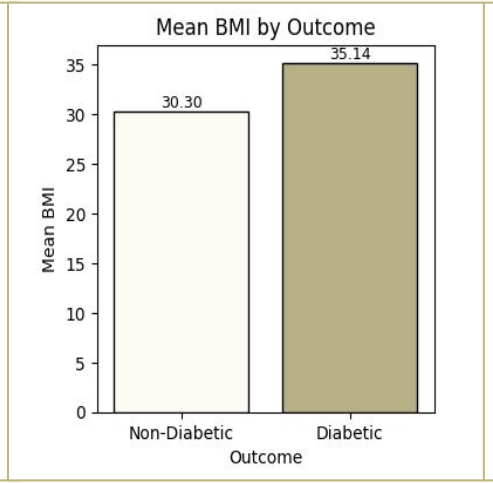


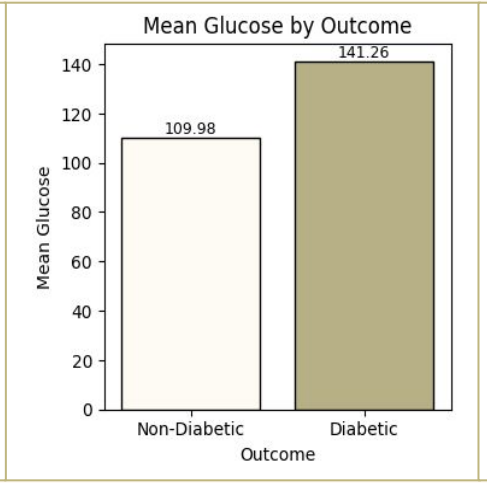


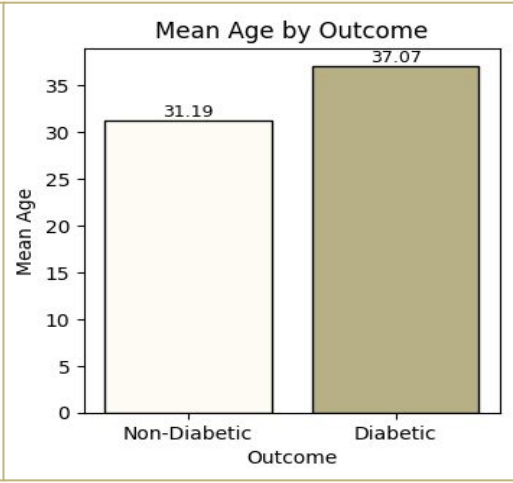
Slightly unbalanced dataset

Average distribution per Outcome

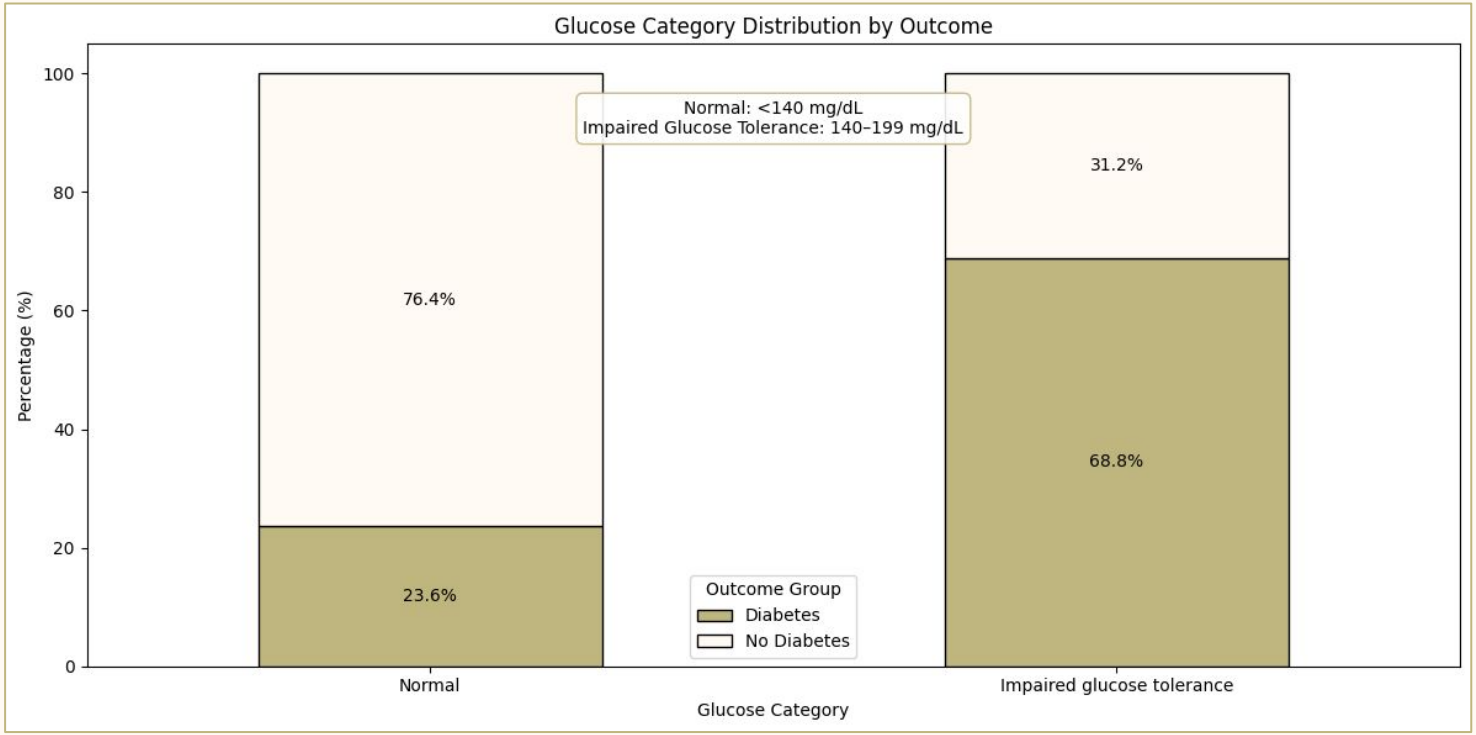






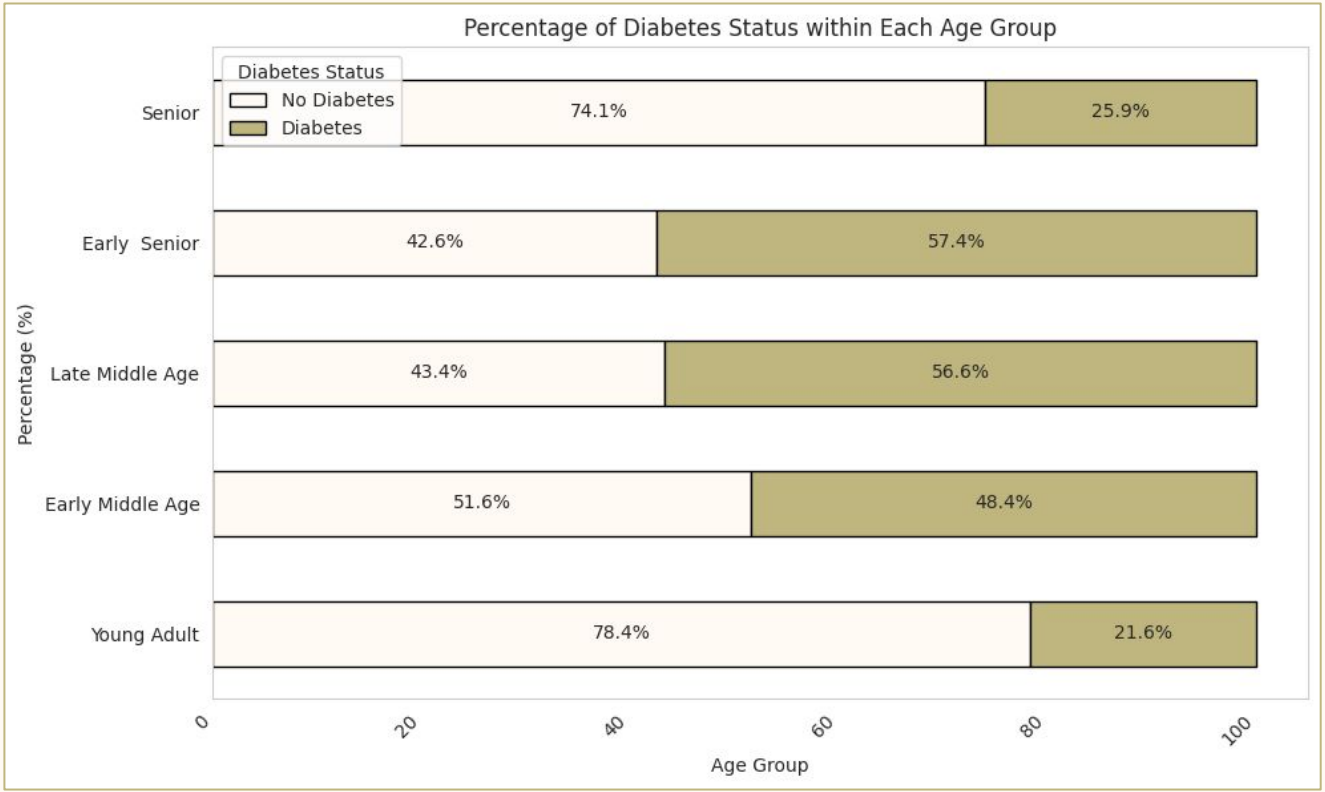


Clear establishment between mean features in both outcomes



Elevated blood glucose strongly relates to the presence of diabetes.

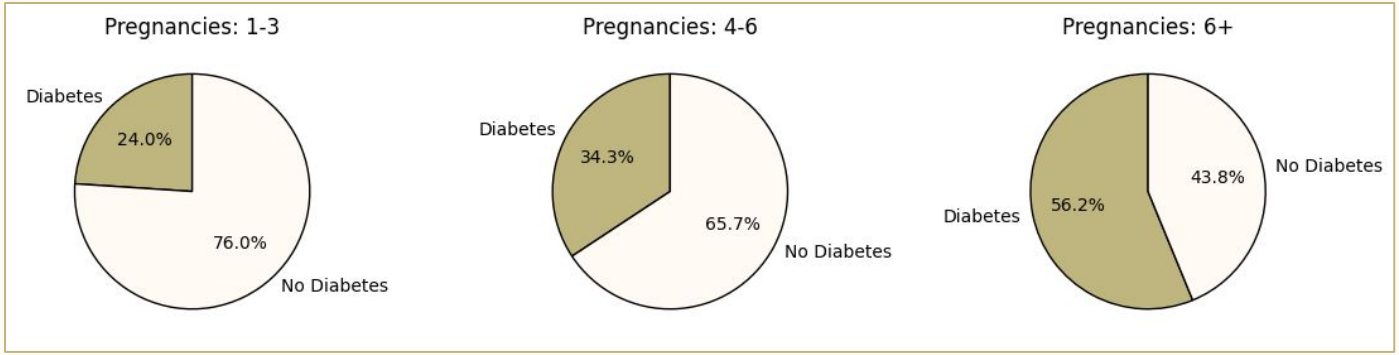
Outcome distribution



*Young Adult - 21-30
Early Middle Age - 31-40
Late Middle Age - 41-50
Early Senior- 51-60
Senior - 61+

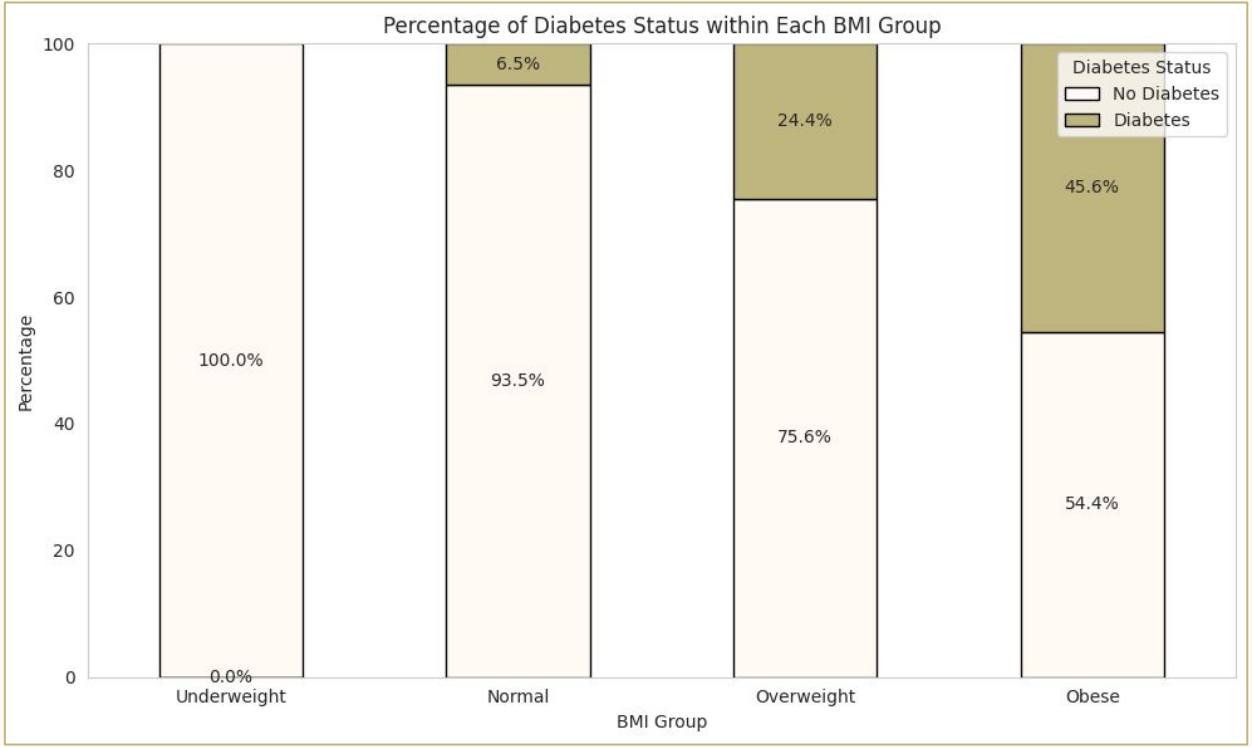
Greater risk of diabetes among older adults and the elderly

Outcome distribution



Greater risk of gestational diabetes among patients with more pregnancies

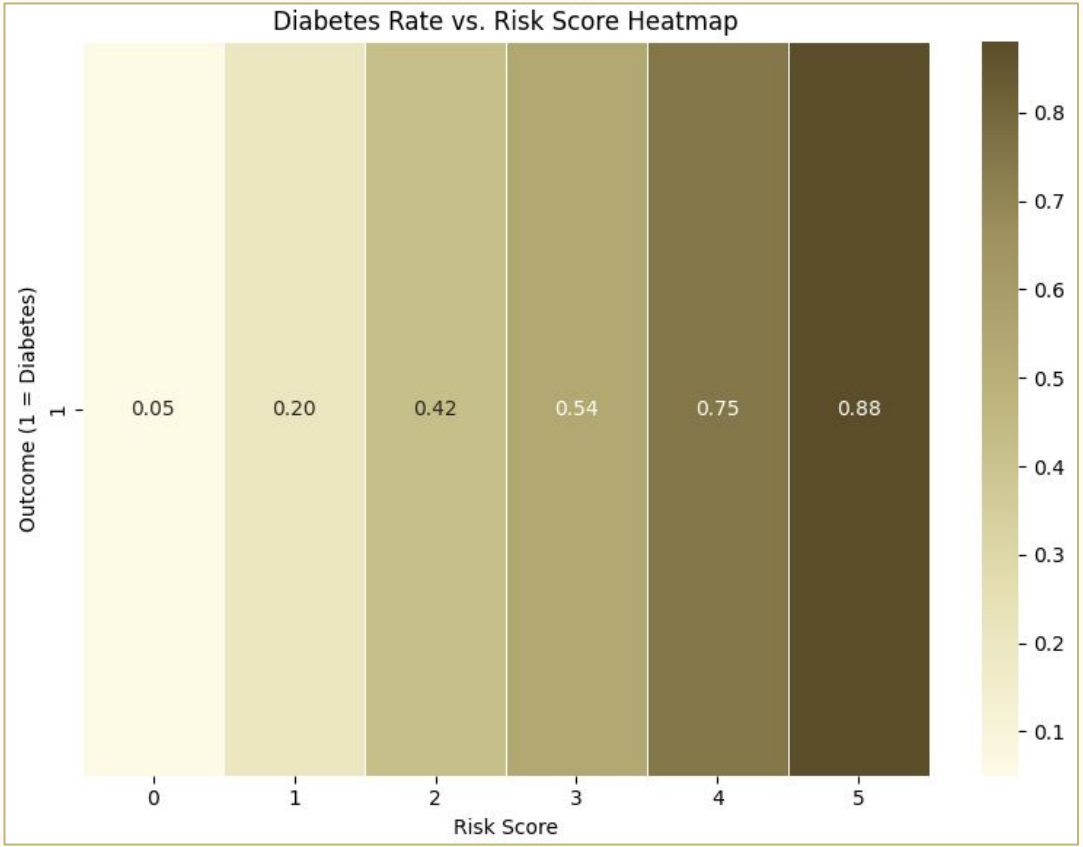
Outcome distribution



*Underweight: <18.5
Normal: 18.5-24.9
Overweight: 25-29.9
Obese: >30

Greater risk of diabetes in overweight and obese individuals

Risk Analysis



*+1 for Glucose ≥140
+1 for BMI ≥30
+1 for Age ≥40
+1 for Insulin ≥150
+1 for pregnancies ≥3

Risk score corresponds to attribute thresholds exceeded

Modelling

Objective

Predict diabetes outcomes using ML models

Motivation

Early diagnosis → better outcomes

Tools

Python, Pandas, Matplotlib, Seaborn, Scikit-learn

Models

Logistic Regression, Random Forest, Decision Tree, RandomizedSearchCV, xgb, etc.

Model Performance (Best results)

Logistic Regression

	Predicted: No Diabetes	Predicted: Diabetes
Actual: No Diabetes	82 (TN)	17 (FP)
Actual: Diabetes	21 (FN)	34 (TP)

Hypertuned Random Forest

4455	Predicted: No Diabetes	Predicted: Diabetes
Actual: No Diabetes	79 (TN)	20 (FP)
Actual: Diabetes	17 (FN)	38 (TP)

*TP – True Positives
FP – False Positives
TN – True Negatives
FN – False Negatives

Balanced performance but room for improvement

Model performance (Best results)

66%

Patients predicted to have diabetes that actually did

*For hypertuned random forest classification model

69%

Actual diabetic patients were correctly identified

Patient medical history

Anaya

0.83

Chances of Anaya to have diabetes

0.17

Chances of Anaya to not have diabetes

Pregnancies:	2	**Insulin**	128 µU/ml
Glucose Level	130 mg/dL	**BMI**	35 kg/m^2
Blood Pressure	70 mmHg	**Pedigree Function**	0.82
Skin Thickness	30mm	**Age**	43

Anaya is most likely diabetic

Recommendations

- Regular monitoring of blood glucose is key to managing diabetes
- Early screening should target individuals who are obese, over 40, or show high glucose and insulin levels.
- Since glucose is the strongest predictor of diabetes, and insulin resistance often precedes it, individuals with abnormal readings should be flagged for early intervention, even if not yet diabetic.
- Obese adults in early middle age should receive more aggressive weight management support.
- Screening for gestational diabetes for all women between the 24th and 28th week of pregnancy, but should be conducted earlier in pregnancy for women at high-risk

THANK YOU!

DO YOU HAVE ANY QUESTIONS?