Predictive Model Building Process

Problem Statement

The objective is to build a predictive model using patient health indicators to determine whether or not a person is diabetic. The dataset used is the **Pima Indians Diabetes Database**, which includes various diagnostic measurements.

Exploratory data analysis

Population

The dataset used in this project consists of **768 medical records** of female patients of Pima Indian heritage, all aged 21 and above. These records include **eight features** related to health diagnostics:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI (Body Mass Index)
- Diabetes Pedigree Function
- Age

Alongside these features, the dataset includes an **Outcome** variable, indicating whether the patient is diabetic (1) or non-diabetic (0)

Descriptive statistics

FEATURE	MEAN	STD DEV	MIN	25%	50%	75%	MAX
PREGNANCIES	3.85	3.37	0	1.00	3.00	6.00	17.00
GLUCOSE	120.89	31.97	0	99.00	117.0	140.2	199.0

BLOOD PRESSURE	69.11	19.36	0	62.00	72.00	80.00	122.0
SKIN THICKNESS	20.54	15.95	0	0.00	23.00	32.00	99.00
INSULIN	79.80	115.24	0	0.00	30.50	127.2	846.0
BMI	31.99	7.88	0	27.3	32.00	36.60	67.10
DIABETES PEDIGREE FUNCTION	0.47	0.33	0.08	0.24	0.37	0.63	2.42
AGE	33.24	11.76	21.0	24.00	29.00	41.00	81.00

Class imbalance

In initial analysis of the **Outcome** variable revealed a class imbalance:

- **500 non-diabetic** patients
- **268 diabetic** patients

SMOTE (Synthetic Minority Over-sampling Technique), which equalized both classes to 500 records each was used.

Outlier treatment

Several numerical features were found to have **implausible values**, particularly zeros in columns where a zero would not be biologically valid (e.g., zero Glucose or BMI). These were interpreted as missing or erroneous data points.

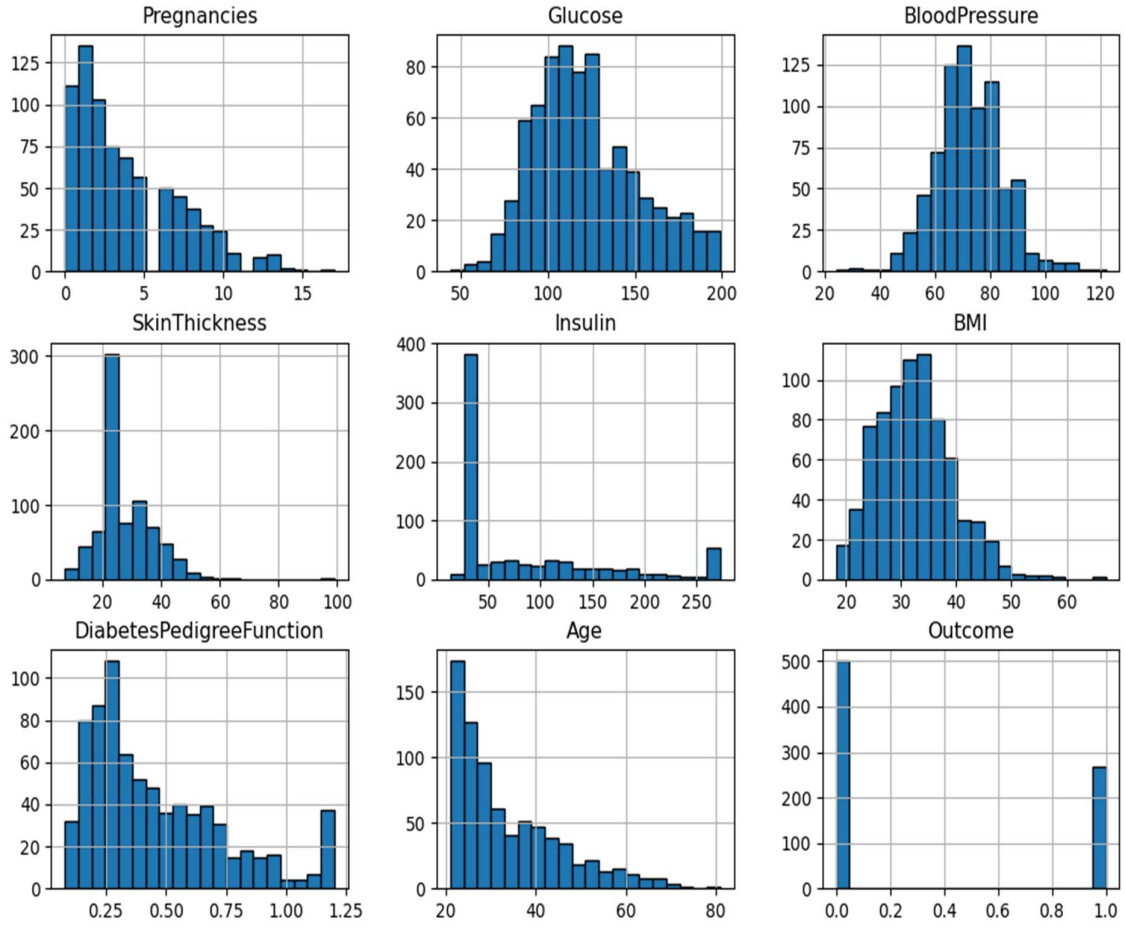
To handle outliers, the **Interquartile Range (IQR) method** was used to detect and cap extreme values. The following variables were treated for outliers:

- **Insulin**
- **Glucose**
- **BMI**
- **Skin Thickness**

Data visualization

Various visual tools were used to understand the data better:

- **Histograms** and **boxplots** were created for all features to inspect distributions and detect outliers.



Findings:

Pregnancy: The majority of the samples have fewer pregnancies.

Glucose: There is higher frequency in low to moderate glucose levels.

Skin Thickness: The majority has a skin thickness of 20-30mm and a sharp drop after 40mm.

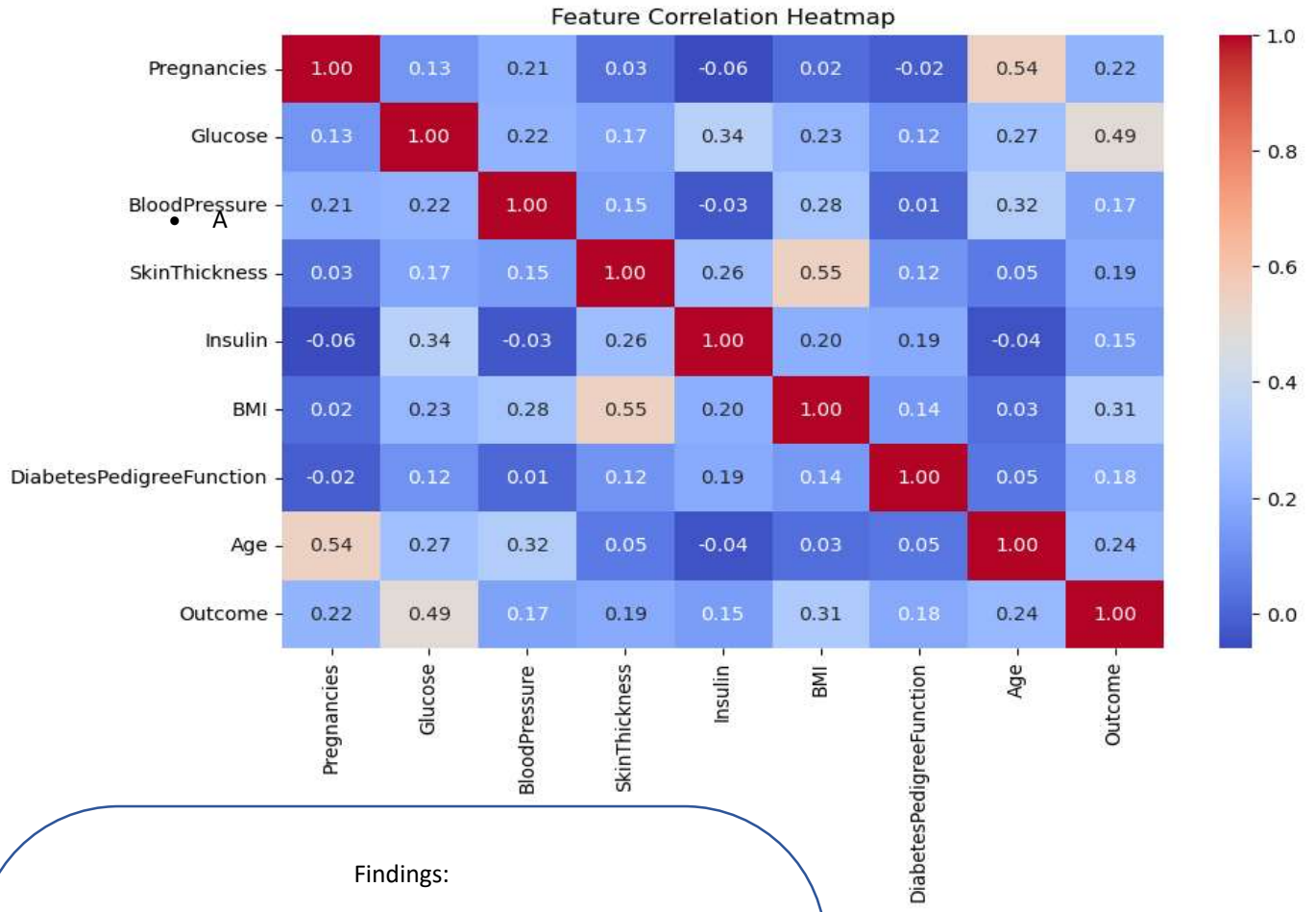
Insulin: Most individuals have insulin levels below 100.

BMI: A balanced spread of BMI values.

Diabetes Pedigree Function: Most scores are concentrated towards lower values.

Age: Samples are more of the young have diabetes.

- A **heatmap of Pearson correlation** coefficients was generated to study relationships between features and the target variable.



- A

Findings:

Glucose and Outcome (0.49): A strong positive correlation, indicating higher glucose levels are associated with a higher likelihood of diabetes.

BMI and Skin Thickness (0.55): This correlation suggests that individuals with a higher BMI tend to have greater skin thickness, which is a potential indicator of insulin resistance.

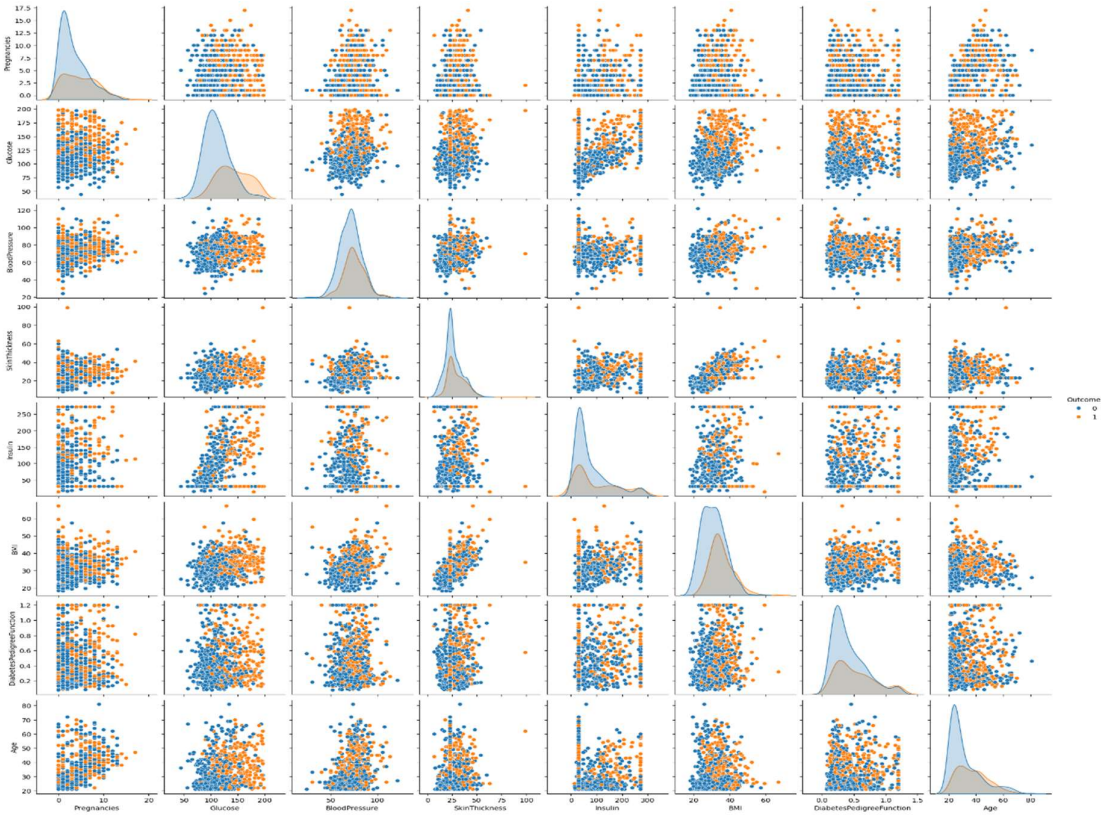
Insulin (and BMI) Correlations: Insulin significantly correlates with BMI (0.55), implying that higher insulin levels tend to be associated with higher BMI.

Outcome and Age (0.24): A weak positive correlation, indicating that older age is somewhat related to a higher likelihood of the diabetes outcome.

Insulin and Glucose (0.34): Indicates a positive relationship, suggesting that higher insulin levels may be related to elevated glucose levels.

Diabetes Pedigree Function and Outcome (0.18): A weak positive correlation, suggesting that a higher diabetes pedigree score may relate slightly to a higher likelihood of diabetes.

Feature	Correlation	Strength
Glucose	0.49	Moderate
BMI	0.31	Moderate
AGE	0.24	Weak
Pregnancies	0.22	Weak
Skin Thickness	0.19	Weak
Diabetes Pedigree function	0.18	Weak
Blood Pressure	0.16	Weak
Insulin	0.14	Weak



FINDINGS

Pregnancy: A higher number of diabetes patients tend to have a higher number of pregnancies.

Glucose: An increase in glucose levels in diabetes patients suggests correct data.

Blood Pressure: BP also tends to differentiate well from the outcome, indicating a higher value with a positive outcome, which is already proven.

Skin Thickness: Higher skin thickness also tends to have a higher likelihood of positive outcomes (diabetes).

Insulin: Diabetes individuals have insulin levels above 100.

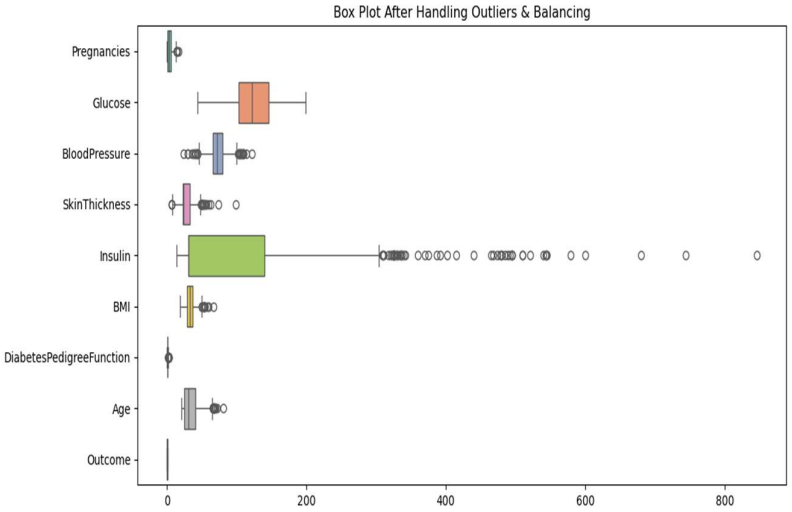
BMI: A higher BMI is an indication of the presence of diabetes.

Diabetes Pedigree Function: A positive correlation is found with diabetes.

Age: Old individuals tend to show a higher incidence of diabetes.

Feature engineering

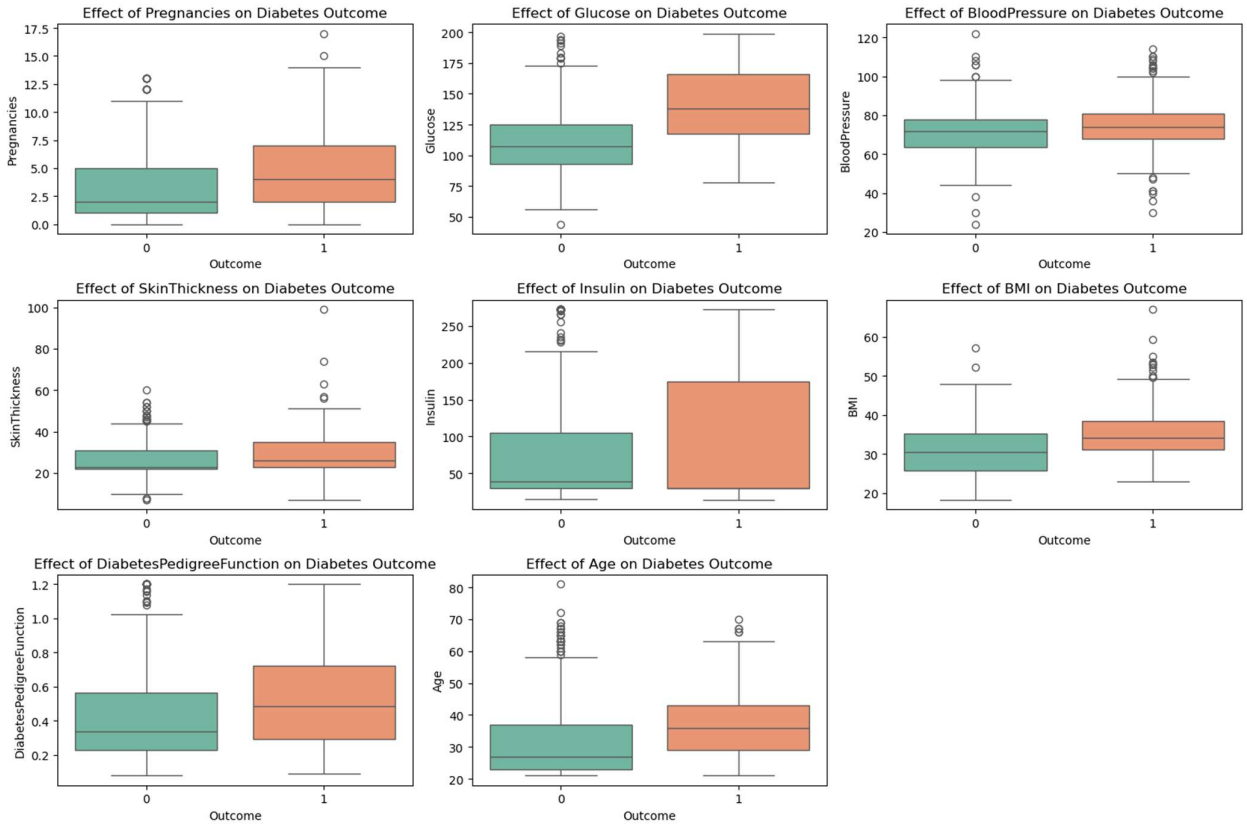
Class imbalance correction



PARAMETER	KS TEST P-VALUE
PREGNANCIES	0.95
GLUCOSE	0.06
BLOOD PRESSURE	0.71
SKIN THICKNESS	0.91
INSULIN	0.64
BMI	0.12
DIABETESPEDIGREEFUNCTION	0.25
AGE	0.06

Handled via **SMOTE**, which successfully improved model performance on minority class with all p value > 0.05.

Feature Distributions vs Diabetes Outcome



Findings

Pregnancies: The median number of pregnancies for non-diabetic individuals (0) is lower than for diabetic individuals.

Glucose: Diabetic individuals exhibit significantly higher glucose levels than non-diabetic individuals, evidenced by the higher median and a wider range.

BP: Similar to glucose, the median blood pressure significantly differs, with diabetic individuals presenting higher values.

Skin Thickness: The skin thickness is notably lower in non-diabetic individuals than in diabetic individuals, with a higher median for the latter.

Insulin: Insulin levels are substantially higher for diabetic individuals. The median is significantly shifted upwards compared to non-diabetic individuals.

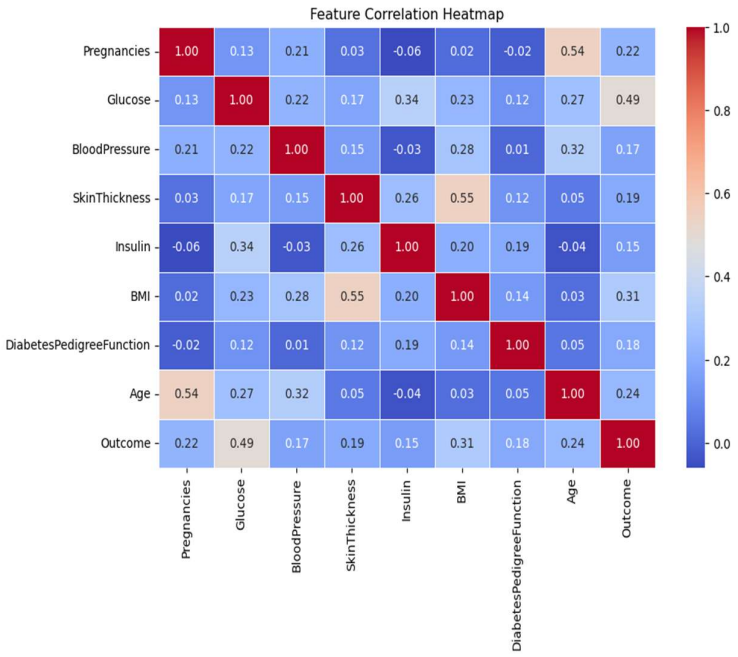
BMI: Diabetic individuals have higher BMI values than their non-diabetic counterparts, as shown by the higher median.

Diabetes Pedigree Function: The values are similar for both groups, but the median for diabetic individuals is slightly higher, indicating that family history may play a role.

Age: Diabetic individuals tend to be older than non-diabetic individuals, as indicated by the higher median age in the diabetic group. There are some outliers, especially among the diabetic group, suggesting that older individuals may contribute to variability.

The consistent pattern across multiple features shows that diabetic individuals generally exhibit higher values in critical health metrics such as glucose, insulin, BMI, and age.

Feature Correlation Matrix



Findings

Glucose (0.49) has the strongest correlation with diabetes, though still in the moderate range.

BMI (0.31) is also moderately correlated, indicating that higher BMI is linked to diabetes risk.

Age (0.24), Pregnancies (0.22), and others show weak correlations, meaning they contribute but are not primary predictors.

Blood Pressure, Insulin, Skin Thickness, and Pedigree Function have relatively lower correlations.

Feature Scaling

MinMaxScaler + Logistic Regression gave the best accuracy overall (0.7727).

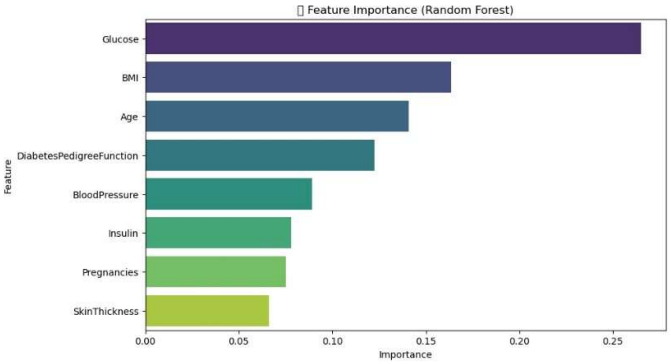
KNN improved significantly from 0.6558 (no scaling) to 0.7468 with scaling.

Random Forest remains relatively stable across scalers, as expected (since it's not distance-based).

Model building

Model Performance Comparison

Model	*Accuracy (Before scaling*	*Accuracy (After scaling)*	
		MinMaxScaler	StandardScaler
Logistic Regression	0.7597	0.7727	0.7662
KNN	0.6558	0.7468	0.7468
Random Forest	0.7532	0.7532	0.7597



Feature Importance – Random Forest Classifier

After applying Standard Scaler and fitting the Random Forest model, the following features were found to be most influential in predicting diabetes:

RANK	FEATURE	IMPORTANCE SCORE	INTERPRETATION
1	**Glucose**	0.2649	Most important
2	**BMI**	0.1633	Strong contributor
3	**Age**	0.1410	Moderate contribution
4	**DiabetesPedigreeFunction**	0.1224	Moderate
5	**Blood Pressure**	0.0892	Weak
6	**Insulin**	0.0778	Weak
7	**Pregnancies**	0.0750	Weak
8	**Skin Thickness**	0.0664	Least important

Glucose is the strongest indicator of diabetes, followed by BMI and Age.

Conclusion

Through systematic analysis, preprocessing, and evaluation:

- **Glucose, BMI, and Age** were identified as the top predictors.
- **Logistic Regression with MinMaxScaler** gave the best overall performance.
- **Outlier treatment**, **class balancing**, and **scaling** significantly improved model accuracy.

Files in repo:

Diabetes prediction model.ipynb: Code notebook

README.md: Project description

requirements.txt: Python dependencies

LICENSE: MIT License

diabetes: Contains dataset used