**Feature Understanding**

 **PatientID**: A unique identifier for each patient. It helps differentiate between individuals in the dataset.

 **Pregnancies**: Number of times the patient has been pregnant. It may influence the risk of diabetes in women.

 **PlasmaGlucose**: Plasma glucose concentration after a glucose tolerance test (mg/dL). Higher levels can indicate diabetes.

 **DiastolicBloodPressure**: Diastolic blood pressure (mm Hg). Measures the pressure in blood vessels between heartbeats.

 **TricepsThickness**: Skinfold thickness of the triceps (mm), often used to estimate body fat.

 **SerumInsulin**: Serum insulin level (μU/mL). High levels may indicate insulin resistance or diabetes.

 **BMI**: Body Mass Index (kg/m²), calculated as weight divided by the square of height. It is a measure of body fat.

 **DiabetesPedigree**: A score indicating the likelihood of diabetes based on family history. Higher values suggest a stronger genetic predisposition.

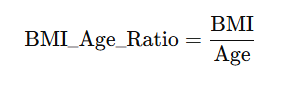
 **Age**: Age of the patient (years).

 **Diabetic**: Target variable (0 = Non-diabetic, 1 = Diabetic). Indicates whether the patient is diagnosed with diabetes.

**Feature extraction**

**1. BMI-to-Age Ratio**

Normalizes BMI by age, accounting for weight changes with age.

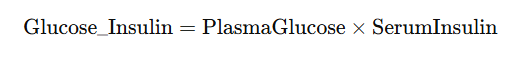


**Why It Helps**:

* **Normalizes BMI for age**: Adjusts BMI values relative to the patient’s age, capturing whether the weight is unusually high or low for their life stage.

**2. Glucose-Insulin Product**

Highlights insulin resistance, a key diabetes indicator. A high value suggests impaired glucose metabolism.

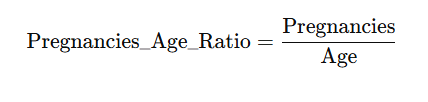


**Why It Helps**:

* **Highlights insulin resistance**: A high product suggests the pancreas is producing insulin, but glucose remains elevated, a hallmark of Type 2 diabetes.

**3. Pregnancies-to-Age Ratio**

Captures how frequent pregnancies are relative to the patient's age, which may indicate a correlation with diabetes risk.

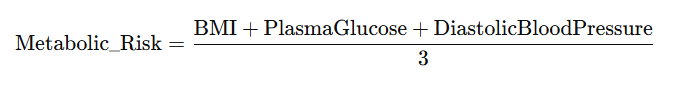


**Why It Helps**:

* **Accounts for reproductive patterns**: Frequent pregnancies at a young age may increase metabolic stress, potentially raising diabetes risk.

**4. Metabolic Risk Score**

A composite score combining BMI, Plasma Glucose, and Diastolic Blood Pressure to measure overall metabolic health.

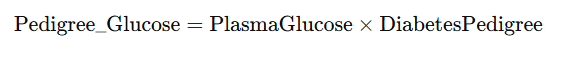


**Why It Helps**:

* **Aggregates critical risk factors**: Combines BMI (obesity), glucose (blood sugar), and blood pressure (vascular health) into a single score.

**5. Pedigree-Adjusted Glucose**

Weights the glucose levels by the genetic predisposition to diabetes, emphasizing family history in glucose-related risks.

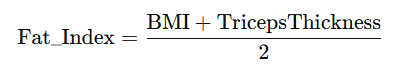


**Why It Helps**:

* **Weights glucose levels by family history**: Captures how much family history influences glucose levels, emphasizing genetic predisposition.

**6. Fat Distribution Index**

Combines BMI and Triceps Thickness to represent overall body fat distribution, as both are independent fat indicators.

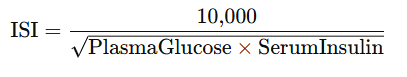


**Why It Helps**:

* **Improves body fat representation**: BMI captures overall body fat, while triceps thickness measures subcutaneous fat. Combining these gives a better picture of fat distribution.

**7. Insulin Sensitivity Index (ISI)**

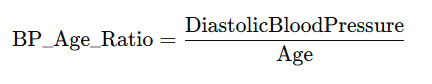
This is a commonly used formula to estimate insulin sensitivity, which helps determine the body's ability to utilize insulin effectively.



**Why It Helps**:  
ISI provides a non-linear interaction between glucose and insulin, which is crucial for detecting Type 2 diabetes.

**8. Blood Pressure Normalized by Age (BP-Age Ratio)**

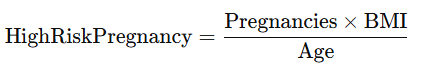
Adjusts blood pressure to account for the natural increase in blood pressure with age.



**Why It Helps**:  
Identifies patients with disproportionately high blood pressure for their age, which can indicate early-stage metabolic issues.

**9. High-Risk Pregnancy Index**

Combines pregnancies with BMI and age to capture whether multiple pregnancies in younger patients with high BMI increase diabetes risk.



**Why It Helps**:  
This feature flags women who might have higher risks of gestational diabetes or its long-term effects.

### After Feature extraction

### 1. Exploratory Data Analysis (EDA)

Before modeling, thoroughly analyze the data to uncover patterns and validate assumptions.  
**Steps**:

* **Distribution Analysis**: Plot histograms for continuous features (e.g., PlasmaGlucose, BMI) to check for skewness and outliers.
* **Correlation Heatmap**: Identify relationships between features and the target variable.
* **Feature Importance**: Use statistical tests (e.g., t-tests, ANOVA) or feature importance from simple models to rank features.

**Why It’s Important**:  
Reveals hidden insights and validates whether your engineered features improve separability.

**2. Data Balancing (If Necessary)**

If the target variable (e.g., Diabetic) is imbalanced, address it using techniques such as:

* **Oversampling**: SMOTE (Synthetic Minority Oversampling Technique).
* **Undersampling**: Downsample the majority class to match the minority.
* **Class Weighting**: Adjust loss functions to penalize misclassifications of the minority class.

**Why It’s Important**:  
Prevents bias in the model towards the majority class, ensuring better predictions for underrepresented groups.