Project: Diabetes Prediction

**1.Data Understanding and Acquisition:**

* ABOUT DATASET:

The "Pima Indians Diabetes Dataset" refers to a publicly available collection of medical data on female Pima Indians aged 21 and older, gathered by the National Institute of Diabetes and Digestive and Kidney Diseases. Its purpose is to be used in machine learning and statistical models to predict whether an individual has diabetes based on features like pregnancies, glucose levels, blood pressure, skin thickness, insulin, body mass index (BMI), age and a diabetes pedigree function.

The Data Type is Structured data type.

**2. DATA PREPARATION AND FEATURE ENGINEERING:**

* EXPLORATORY DATA ANALYSIS:

**2.1. Data Overview and Structure**

The dataset contains diagnostic measurements for female patients of Pima Indian heritage, with the goal of predicting the onset of diabetes.

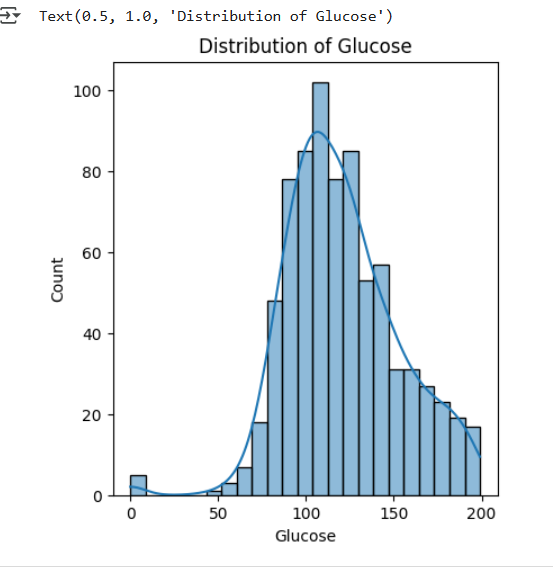
**2.2 Target Variable Analysis**

The first critical step in a classification problem is to check the balance of the target variable.

So, the target variable is outcome.

**2.3 Data Visualization**

**2.3.1 Histogram: Distribution of Glucose**

****

Shape: The distribution is approximately normal (bell-shaped) with a slight right skew

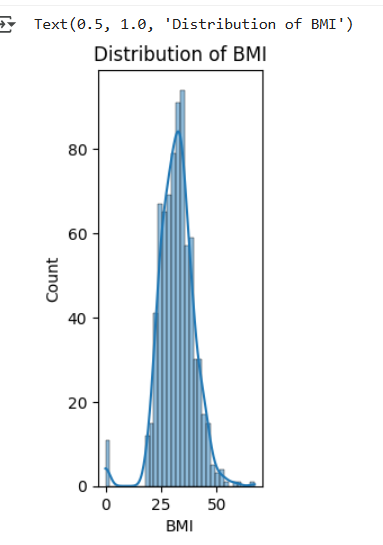
Central tendency: Most values cluster around 100-120, which appears to be the peak

Range: Glucose levels span from near 0 to about 200

Pattern: The KDE curve (smooth line) confirms the normal distribution pattern

Outliers: There are a few cases with very low glucose values (near 0) and some higher values extending to 200

**2.3.2. Distribution of BMI**

****

Shape: The distribution is approximately normal (bell-shaped) and symmetric

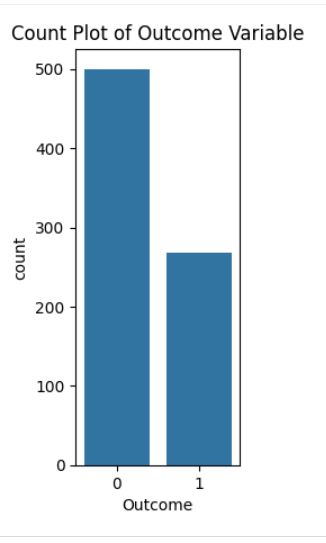
Central tendency: Most values are concentrated around 30-35 BMI, which is the peak

Range: BMI values span from near 0 to about 70

Pattern: The KDE curve shows a smooth, unimodal distribution

Outlier: There's one notable case with very low BMI (near 0), which may be a data entry error or missing value.

**2.3.3. Count plot of a binary outcome variable:**

****

Two categories: The outcome has only two values (0 and 1)

Class imbalance: Category 0 has approximately 500 cases, while category 1 has about 270 cases

Ratio: Roughly a 65:35 split between the two classes

Implication: There's a moderate class imbalance, with the negative class (0) being more prevalent

In a medical context (like diabetes prediction), this typically means:

0 = No condition (majority class - ~500 patients)

1 = Has condition (minority class - ~270 patients)

This imbalance is important for modelling, as it may require techniques like resampling or adjusting class weights to prevent bias toward predicting the majority class.

**2.4 Data Transforming and preprocessing:**

**1. Data Cleaning**

Zero Value Handling:

* Identified suspicious zero values in medical features (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI)
* Imputation Method: Replaced zeros with median values to maintain data integrity
* Most affected: Insulin (374 zeros) and skin Thickness (227 zeros)

**2. Dataset Statistics (After Cleaning)**

* Total samples: 768 patients
* Target distribution: 34.9% diabetic (Outcome=1), 65.1% non-diabetic (Outcome=0)
* Key feature ranges:
  + Glucose: 44-199 mg/dL (mean: 121.7)
  + BMI: 18.2-67.1 (mean: 32.5)
  + Age: 21-81 years (mean: 33.2)

**3. Train-Test Split**

* Training set: 537 samples (70%)
* Testing set: 231 samples (30%)
* Random state: 42 (ensures reproducibility)

**4. Feature Scaling (Standardization)**

* Method: StandardScaler (z-score normalization)
* Process:
  + Scaler fitted only on training data to prevent data leakage
  + Both train and test sets transformed using the same scaler
* Purpose: Ensures all features have mean=0 and std=1, improving model performance and convergence

Status: Data is now ready for machine learning model training.

**3.MODEL DEVELOPMENT AND TRANING:**

**Logistic Regression Model** - Performance Report

**Model Training**

* Algorithm: Logistic Regression
* Training data: 537 samples (scaled)
* Status: Model trained successfully

**Model Performance Summary**

Overall Accuracy: 74.46%

The model correctly predicts diabetes outcome in approximately 3 out of 4 cases.

**Confusion Matrix Analysis**

|  | Predicted: No Diabetes (0) | Predicted: Diabetes (1) |
| --- | --- | --- |
| Actual: No Diabetes (0) | 124 (True Negative) | 27 (False Positive) |
| Actual: Diabetes (1) | 32 (False Negative) | 48 (True Positive) |

**Interpretation:**

* True Negatives (124): Correctly identified non-diabetic patients
* True Positives (48): Correctly identified diabetic patients
* False Positives (27): Incorrectly predicted diabetes (Type I error)
* False Negatives (32): Missed actual diabetes cases (Type II error) - More concerning in medical context

**Classification Report**

Class 0 (No Diabetes):

* Precision: 79% - When model predicts "no diabetes," it's correct 79% of the time
* Recall: 82% - Model identifies 82% of all actual non-diabetic cases
* F1-Score: 81% - Balanced measure of precision and recall

Class 1 (Diabetes):

* Precision: 64% - When model predicts "diabetes," it's correct 64% of the time
* Recall: 60% - Model identifies only 60% of all actual diabetic cases
* F1-Score: 62% - Lower performance on positive class

**OVERALL SUMMARY:**

A Logistic Regression baseline model was successfully developed to predict diabetes with 74.46% accuracy. While the model shows reasonable performance, the 60% recall for diabetic patients is insufficient for medicaldeployment. The model correctly identifies 6 out of 10 diabetic patients but misses 4, which poses significant health risks.