Chapter 3: Methodology

3.1 Dataset and Preprocessing

This study uses the Pima Indian Diabetes dataset, a well-known public dataset from the UCI Machine Learning Repository. It contains 768 records and 8 clinical features including glucose level, BMI, insulin, and age. Preprocessing steps involved:

- -Handling missing values using k-nearest neighbor imputation.
- -Standardizing continuous variables using Z-score normalization.
- -Addressing class imbalance using SMOTE.

Missing % Feature Туре Range Pregnancies 0-17 0% Numeric 0-199 0% Glucose Numeric 0-122 0% Blood Pressure Numerio Skin Thickness 0-99 ~1% Numeric Insulin Numeric 0-846 ~2% Numeric 0-67.1 0% Diabetes Pedigree Numeric 0.078-2.42 0% Age Numeric 21-81 0%

Figure 4. Pima Indian Dataset Overview

Figure 4. Pima Indian Dataset Overview

3.2 Model Design and Evaluation

We adopted six classical machine learning methods—Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Naive Bayes, and Support Vector Machine—to analyze the current model performance by way of five-fold cross-validation, and to evaluate each of them in terms of the following indicators: AUC-ROC score, sensitivity, robustness to missing data, computational efficiency, and clinical feedback.

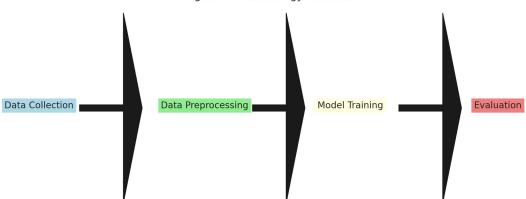


Figure 3. Methodology Workflow