

In this experiment, classification models were built on the Pima Indians Diabetes dataset to predict whether a patient has diabetes. Data preprocessing involved replacing invalid zero values in medical features (Glucose, BMI, Insulin, etc.) with median values. The dataset was split into training (80%) and testing (20%) sets.

A Random Forest classifier was first trained using all features. Feature importance scores showed that **Glucose, BMI, Age, and Insulin** were the most influential predictors. Retraining the model using only the top features produced similar performance, demonstrating that redundant features can be removed without major accuracy loss.

Principal Component Analysis (PCA) was then applied to reduce dimensionality. Although PCA reduced feature space, performance slightly decreased due to loss of interpretability and some information.

For comparison, Logistic Regression and Naïve Bayes models were trained. Logistic Regression performed reasonably well but struggled with nonlinear relationships. Naïve Bayes was the fastest but had the lowest performance due to its independence assumption.

Conclusion:

The **Random Forest with selected top features** is the best model, as it provides high accuracy and F1-score while reducing complexity. It balances performance, interpretability, and efficiency better than PCA, Logistic Regression, and Naïve Bayes.