

## **Report on Feature Selection and Model Comparison:**

This work investigates the impact of feature selection and model choice on the performance of machine learning classifiers using the Diabetes dataset. The dataset was divided into training and testing sets to ensure fair evaluation. A Random Forest classifier was first trained using all available features to establish a strong baseline. Feature selection was then applied by identifying the most important features, allowing the model to achieve similar performance with reduced complexity. Principal Component Analysis was also used to reduce dimensionality, though this led to a slight reduction in predictive performance due to loss of information.

For comparison, Logistic Regression and Naïve Bayes models were trained and evaluated using accuracy, precision, recall, and F1-score metrics. Random Forest consistently delivered the best results, demonstrating its strength in handling non-linear relationships. Logistic Regression showed reasonable performance, while Naïve Bayes performed less effectively due to its simplifying assumptions. Overall, the study highlights that appropriate feature selection improves efficiency and that Random Forest is the most reliable model for this dataset.

.