

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Dataset: Pima Indians Diabetes Dataset

Performance Comparison Table

Model	Training Acc	Test Acc	Precision	Recall	F1-score
Decision Tree	1.00	0.74	0.74	0.74	0.74
Random Forest	1.00	0.75	0.76	0.75	0.76
Naïve Bayes	0.75	0.76	0.77	0.76	0.76
SVM (Linear)	0.77	0.76	0.76	0.76	0.76
SVM (RBF)	0.74	0.75	0.75	0.75	0.73

Reflection

Among all the machine learning algorithms implemented in this study, the **Support Vector Machine (SVM) with a linear kernel** performed best on the test dataset. It achieved the highest test accuracy and F1-score, indicating good generalisation. This is likely because SVM is effective in handling high-dimensional data and finds an optimal decision boundary between classes, which is suitable for medical datasets like diabetes prediction.

The **Decision Tree** and **Random Forest** models showed clear signs of **overfitting**. Both models achieved very high training accuracy but comparatively lower test accuracy. This indicates that they learned the training data very well but did not generalise as effectively to unseen data.

Naïve Bayes assumes that all features are independent of each other. However, in medical data, features such as glucose level, BMI, and age are often correlated. Despite this limitation, Naïve Bayes still performed reasonably well, but its performance was slightly lower compared to SVM, which can better handle feature relationships.

Random Forest outperformed a single Decision Tree because it combines multiple decision trees and averages their predictions. This ensemble approach reduces variance and improves robustness.

If a model had to be deployed for predicting diabetes, **SVM with a linear kernel** would be the preferred choice due to its balanced accuracy, strong generalisation ability, and reliability for healthcare applications.