

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Compare Results

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1-score
DecisionTree	1.00	0.69	0.56	0.54	0.55
Random Forest	1.00	0.74	0.64	0.59	0.61
NaïveBayes	0.77	0.71	0.59	0.63	0.61
SVM(Linear)	0.79	0.71	0.61	0.50	0.55
SVM(RBF)	0.74	0.67	0.55	0.30	0.39

Reflection

In this activity, multiple machine learning algorithms were applied to the Pima Indians Diabetes dataset to evaluate their performance in predicting diabetes. According to the results in the comparison table, **Random Forest achieved the best overall performance on the test data**, with the highest test accuracy (0.74) and the strongest F1-score (0.61). This indicates that Random Forest was able to balance precision and recall more effectively than the other models. Its ensemble structure, which combines multiple decision trees, helps reduce variance and improves generalization on unseen data.

The **Decision Tree model showed clear signs of over fitting**, as it achieved perfect training accuracy (1.00) but a noticeably lower test accuracy (0.69). This suggests that the model memorized the training data rather than learning generalizable patterns. Although Random Forest also reached perfect training accuracy, it performed better than a single Decision Tree on the test set due to averaging predictions from many trees.

The **Naïve Bayes model performed reasonably well**, achieving moderate test accuracy and recall, which indicates it was effective at identifying diabetic cases despite its assumption that features are independent. The **SVM models**, particularly SVM with an RBF kernel, showed weaker performance with low recall and F1-score, suggesting difficulty in correctly identifying diabetic patients.

Overall, **Random Forest would be the preferred model for deployment** in a diabetes prediction system because of its strong test performance, robustness, and balanced evaluation metrics.