

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Model	Train accuracy	Test Accuracy	Precision	Recall	F1-score
Decision Tree	1.00	0.55	0.55	0.53	0.54
Random Forest	1.00	0.74	0.64	0.59	0.61
Naïve Bayes	0.77	0.71	0.58	0.62	0.60
SVM (Linear)	0.77	0.71	0.61	0.50	0.55
SVM (RBF)	0.77	0.66	0.55	0.29	0.38

Reflection Questions:

Q1. Which algorithm performed best on the test data? Why might that be the case?

→ Random Forest performed the best on the test data with the accuracy (0.74) and F1-score (0.61). As Random Forest uses various decision trees for predictions. Then it combines all predictions, which avoids overfitting issues and gives more stability, higher accuracy on test data.

Q2. Did any algorithm show signs of **overfitting** (high training accuracy, low test accuracy)?

→ Decision Trees shows clear sign of over-fitting, as it got 1.00 accuracy on training data. But it got only 0.55 accuracy on test data, which is much lower than the training accuracy. It shows that it has memorized training data pattern.

Q3. How does Naive Bayes handle correlated features compared to Decision Tree or SVM?

→ Decision Tree and SVMs can handle correlated features by learning relationships between variables. Where else, Naive Bayes thinks all features are independent, which can cause performance issues when variables are correlated.

Q4. Why might Random Forest outperform a single Decision Tree?

→ Random Forest can outperform a single Decision Tree as it combines multiple Decision Trees and these trees are trained parallel, which makes the training faster. It majorly counters overfitting issue of Decision Tree. Random Forest gives more reliable prediction and also improves the accuracy on test data.

Q.5 If you had to deploy a model for predicting diabetes, which one would you choose, and why?

→ I will choose Random Forest algorithm to deploy a model because it has achieved highest test accuracy (0.74) and F1-score (0.61). It reduces overfitting through training on non-linear relationships and provides interpretable feature importance, making it more reliable and leads to better predictions.