

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Compare Results

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1-score
Decision Tree	1.0000	0.6883	0.5577	0.5370	0.5472
Random Forest	1.0000	0.7403	0.6400	0.5926	0.6154
Naïve Bayes	0.7752	0.7143	0.5862	0.6296	0.6071
SVM (Linear)	0.7948	0.7143	0.6136	0.5000	0.5510
SVM (RBF)	0.7443	0.6688	0.5517	0.2963	0.3855

Reflection

Based on the test accuracy and evaluation metrics, **Random Forest** performed best on the diabetes dataset. It achieved the highest test accuracy (0.7403) along with strong precision, recall, and F1-score compared to other models. This strong performance is likely due to Random Forest's ensemble nature, which combines multiple decision trees to reduce variance and improve generalization on unseen data.

Both **Decision Tree** and **Random Forest** showed signs of **overfitting**, as their training accuracy reached 1.0000 while test accuracy was significantly lower. This indicates that these models learned the training data too well, including noise, which reduced their performance on test data. However, Random Forest mitigated overfitting better than a single Decision Tree through averaging multiple trees.

Naïve Bayes handled the dataset reasonably well despite its assumption that features are independent. In practice, Naïve Bayes can still perform effectively even when features are correlated, although this assumption may limit its maximum performance compared to **SVM** or tree-based models.

Random Forest outperformed a single Decision Tree because it reduces overfitting and improves robustness by aggregating predictions from many trees trained on random subsets of data and features.

If deploying a diabetes prediction model, **Random Forest** would be the preferred choice due to its balanced accuracy, stability, and reliable performance across multiple evaluation metrics.