

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Compare Results

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1-score
Decision Tree	1.00	0.70	0.77	0.79	0.78
Random Forest	1.00	0.77	0.82	0.83	0.83
Naïve Bayes	0.75	0.75	0.82	0.82	0.82
SVM (Linear)	0.77	0.78	0.82	0.86	0.84
SVM (RBF)	0.72	0.73	0.73	0.95	0.82

Reflection

After completing this activity, the difference between various ML algorithms has become much clearer to me. From my results, Random Forest and Naïve Bayes performed quite well on the test data. The most interesting part was seeing the Decision Tree performance; it reached 100% accuracy during training but dropped significantly during testing. This is a classic case of "overfitting," where the model simply memorized the training data instead of learning general patterns.

In medical datasets like this one, features such as BMI and Glucose levels are often related. Naïve Bayes has a limitation here because it assumes all features are independent, which isn't really true in biological cases. On the other hand, SVM and Random Forest seemed to handle these complex connections between features much more effectively.

If I had to select a model for a real-world medical application, I would go with Random Forest. In healthcare, focusing on Recall is very important. We don't want to tell a patient they are healthy when they actually have diabetes (False Negative). Because Random Forest combines multiple trees, it is more stable and reliable for making such critical predictions.