

Activity: Implementing and Comparing ML Algorithms on the Diabetes Dataset

Compare Results: (0-no Diabetes)

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

Reflection

In this activity, different machine learning algorithms were implemented and compared to predict diabetes using the Pima Indians Diabetes dataset. The models evaluated were Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM) with Linear and RBF kernels. The performance of each model was assessed using accuracy, precision, recall, F1-score, and confusion matrix.

From the results, **SVM with a Linear kernel performed the best on the test data**, achieving the highest test accuracy (0.77) and the highest weighted F1-score (0.77). The training and testing accuracies of this model were very close, which indicates good generalization and no significant over fitting. This suggests that the dataset is nearly linearly separable and well suited for a linear SVM.

The **Decision Tree and Random Forest models showed signs of over fitting**, as both achieved perfect training accuracy but lower test accuracy. This indicates that these models learned the training data too well and did not generalize as effectively to unseen data. Although Random Forest is generally more robust than a single Decision Tree, it did not outperform SVM in this case.

The **Naïve Bayes model performed reasonably well**, but its performance was slightly lower because it assumes feature independence, which is not realistic for medical data where features such as glucose level, BMI, and age are correlated.

Overall, **SVM (Linear)** would be the preferred model for deployment in a diabetes prediction system due to its balanced performance and strong generalization ability.