

Short written on reflection question:

From the machine learning algorithms, I have tested and trained from the diabetes dataset. Random Forest performed best on the test dataset, giving the highest accuracy and more stable results compared to the other ML models. The results I have got mainly due to its ensemble nature, where multiple decision trees work together to reduce prediction errors and improve generalization on unseen data and give better results.

The Decision Tree algorithm given the signs of overfitting. It also achieved very high training accuracy but noticed lower test accuracy, showing that the model learned the training data too specifically and struggled to generalize. This is a common problem with single decision trees when no depth or complexity constraints are applied.

Naive Bayes handled the given data well but was affected by correlated features such as glucose level, BMI, and age. Since Naive Bayes assumes conditional independence between features, correlations can reduce its predictive effectiveness compared to Decision Tree or SVM, which can model complex relationships.

Random Forest performed as a single Decision Tree because it uses multiple trees built on different subsets of data and features, reducing variance and overfitting. If deploying a model for diabetes prediction, Random Forest would be the preferred choice due to its reliable accuracy, robustness, and suitability for real-world healthcare decision-making applications globally.