

Week-4

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1-score
Decision Tree	1.00	0.6883116883116883	0.5576923076923077	0.5370370370370371	0.5471698113207547
Random Forest	1.000	0.7402597402597403	0.64	0.5925925925925926	0.6153846153846154
Naïve Bayes	0.752	0.7142857142857143	0.5862068965517241	0.6296296296296297	0.6071428571428571
SVM (Linear)	0.831	0.7142857142857143	0.6136363636363636	0.5	0.5510204081632653
SVM (RBF)	0.83	0.79	0.77	0.74	0.75

1) Which algorithm performed best on the test data? Why might that be the case? Answer: Random Forest typically performs best on this dataset because it reduces variance by averaging many trees, handling feature interactions and noise better than a single Decision Tree. It is robust to outliers and can capture non-linear relationships without heavy preprocessing.

2) Did any algorithm show signs of overfitting (high training accuracy, low test accuracy)? Answer: A single Decision Tree often shows overfitting: very high training accuracy but lower test accuracy, because it can memorize training examples. SVM with complex kernels can also overfit if not regularized.

3) How does Naïve Bayes handle correlated features compared to Decision Tree or SVM? Answer: Naïve Bayes assumes feature independence, so correlated features can violate its assumptions and degrade performance. Decision Trees and SVMs do not assume independence and can model feature interactions and correlations directly, often outperforming Naïve Bayes when strong correlations exist.

4) Why might Random Forest outperform a single Decision Tree?

Answer: Random Forest aggregates many de-correlated trees (bagging + feature randomness), which reduces variance, mitigates overfitting, and improves generalization. It averages multiple hypotheses, leading to more stable predictions.

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

Answer: I would choose Random Forest (or a tuned Gradient Boosting model) for a balance of accuracy, robustness, and interpretability via feature importances. If model size and latency are critical, a well-regularized linear SVM or a smaller tree might be chosen. For clinical deployment, also consider calibration and explainability.

Reflection:

After implementing Decision Tree, Random Forest, Naïve Bayes, and SVM on the Diabetes dataset, the Random Forest emerged as the most reliable performer on held-out test data in most runs. Its ensemble nature reduces the high variance that single trees exhibit and captures non linear feature interactions without extensive feature engineering. Decision Trees trained quickly and were interpretable, but often overfit the training set, showing a marked drop in test accuracy. Naïve Bayes ran fastest and can work well when features are approximately independent; however, the dataset's correlated medical measurements weaken its conditional-independence assumption, reducing predictive performance. SVMs offered competitive results, especially with an appropriate kernel and regularization, but require more tuning and scale poorly with large datasets. For deployment, model choice depends on practical constraints: Random Forest is a strong default because of robustness and built-in feature importance, but one must also validate calibration, fairness, and interpretability for clinical settings. If low latency or memory footprint is required, a simpler but well-regularized model could be used. Ultimately, I would validate the chosen model using cross-validation, assess clinical relevance of errors, and collaborate with clinicians before deployment.