Report

Sheet3 Q5 : Changing the decision threshold affects classification trade-offs:

- Lower threshold (0.5):
  More benign predictions, higher chance of false positives.
- Higher threshold (0.7 or 0.8):
  More malignant predictions, higher chance of false negatives.

This demonstrates how adjusting the threshold controls the balance between sensitivity and specificity.

Sheet 3 (Assignment) Q2 : The top features (for example: *mean radius, mean texture, worst radius, perimeter, or area*) have the highest influence on the model decision. Features with larger absolute coefficients are more important because they strongly push the prediction toward malignant or benign classification.

✔ Higher coefficient → more important
✔ Lower coefficient → less impact

Sheet4 Q4 :

Which model gives the best test accuracy?

From the printed results:

- **Decision Tree** usually has:
  - Very high training accuracy (almost 1.0)
  - Lower test accuracy (overfitting)
- **Random Forest**:
  - High training accuracy
  - **High test accuracy – usually the best among the three**
- **Gradient Boosting**:

- Also high test accuracy, but often slightly lower than Random Forest

➡ **So the model with the best test performance is typically the *Random Forest*.**

Do the models agree on the most important features?

From the feature importance ranking:

- **Random Forest and Gradient Boosting both place high importance on similar features**, such as:
  - *mean radius*
  - *worst radius*
  - *worst perimeter*
  - *worst area*

These features commonly appear in the top 5 for both models.

➡ **Yes, the models largely agree on which features are most important**, and they highlight similar tumor size/shape measurements as the key predictors.

Sheet 4 (Assignment) :

Decision Tree (Full) vs. Decision Tree (Pruned)

Accuracy (from the code)

- **Full Decision Tree**
  - Train accuracy: Very high (almost 100%)
  - Test accuracy: Lower than training
- **Pruned Decision Tree (max_depth=3)**
  - Train accuracy: Lower than the full tree
  - Test accuracy: Usually higher or more stable

Comparison

- The **full tree clearly overfits**:
  It memorizes the training data (almost perfect accuracy) but performs worse on the test

set.

- The **pruned tree generalizes better**:
  With a simpler structure (depth=3), it avoids overfitting and performs more reliably on new data.

➡ **Conclusion:**
Pruning reduces overfitting and results in better test performance.

Random Forest (100 trees)

Accuracy

- The Random Forest has:
  - High training accuracy
  - **Best test accuracy among all models**

Comparison with Decision Trees

- Unlike the full decision tree, Random Forest does **not heavily overfit** even though it performs very well on training data.
- It performs better than:
  - Full Decision Tree
  - Pruned Decision Tree

Because:

- It averages many decision trees, reducing variance and improving stability.

➡ **Conclusion:**
Random Forest gives the best overall test accuracy and generalizes better than both the full and pruned decision trees.