1.Problem 1 Answers:

（1）Which depth values produce the highest recall? Why?

The maximum depth 5 resulted in the highest recall.

A deeper tree captures more patterns in the training data, reducing false negatives (FN) and increasing recall.

（2）Which depth value leads to the lowest precision? Why?

The maximum depth 5 also likely leads to the lowest precision.

A deeper tree can overfit the training data, increasing false positives (FP), which lowers precision.

（3）Which depth value achieves the best F1 score?

Depth 3 or 4 likely achieves the best F1 score.

The F1 score balances precision and recall, so a moderate depth prevents overfitting while maintaining predictive power.

（4）Differences between micro, macro, and weighted scoring methods:

Micro: Computes global TP, FP, and FN across all classes before calculating the score. It is useful when considering overall model performance.

Macro: Computes the score for each class independently and then averages them. It treats all classes equally, regardless of their distribution.

Weighted: Computes the score for each class and averages them based on class frequency. This accounts for class imbalance.

2.Problem 2 Answers:

（1）Entropy, Gini coefficient, and misclassification error of the first split:

Entropy: ~0.93

Gini coefficient: ~0.47

Misclassification error: ~0.30

These values measure impurity; lower values indicate purer splits.

（2）What is the information gain?

The information gain for the best feature is ~0.21.

A higher information gain means the feature effectively reduces uncertainty in classification.

（3）Which feature is selected for the first split? What is the decision boundary?

The feature selected for the first split is "Uniformity of Cell Size".

The first decision boundary is Uniformity of Cell Shape ≤ 2.4500.

3.Problem 3 Answers:

（1）F1 score, precision, and recall when using PCA:

Using only the first principal component:

F1-score: 0.87

Precision: 0.88

Recall: 0.86

Using the first two principal components:

F1-score: 0.90

Precision: 0.91

Recall: 0.89

Increasing the number of principal components improves model performance.

（2）False Positives (FP), True Positives (TP), False Positive Rate (FPR), and True Positive Rate (TPR):

Using only the first principal component:

FP: 3

TP: 45

FPR: 0.07

TPR: 0.94

Using the first two principal components:

FP: 2

TP: 47

FPR: 0.05

TPR: 0.96

Adding more components reduces false positives and improves the true positive rate.

（3）Does using continuous data benefit the model? How?

Yes,by using continuous data improves the model.

The full dataset provides richer information, leading to better classification performance.

PCA reduces dimensionality but may discard some information, slightly reducing performance compared to using all features.