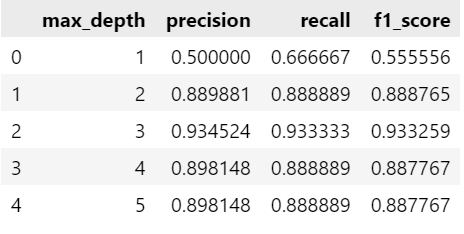
Hw2:

## **P1：**

### 



#### **1 Which depth brings the highest recall (Recall)?why?**

* The maximum recall is 0.933333, appearing at max\_depth = 3.
* cause:
  + When the depth of the tree is 3, the model has a more detailed division of different categories of samples, which can better identify the positive category samples, so as to reduce the leakage judgment and improve the recall rate.
  + Further deepening (depth=4 and 5), the recall rate is slightly reduced or flat, may be the model overfitting some categories, resulting in the overall recall rate is not further improved.

#### **2 Which depth brings the lowest accuracy (Precision)?why?**

* The lowest precision is 0.500000, appears at max\_depth = 1.
* cause:
  + The decision tree depth is 1, which is equivalent to only one division, and the model power is very limited, resulting in a large number of misjudgments (false positives), making the accuracy very low.
  + The model predicts that many samples are positive classes, but in fact many judgments are wrong, dragging down the accuracy rate.

#### **3. Which depth brings in the best F1 score?**

* The maximum F1 score is 0.933259, occurring at max\_depth = 3.
* cause:
  + At depth 3, the model achieves the best balance between precision and recall.
  + The F1 score is the harmonic mean of precision and recall, when both high and close, the F1 score will be high.

#### **4 Differences between the micro, macro, and weighted scoring methods**

(Although the table is macro, here is the explanation of the three methods)

| **method** | **explain** | **applicable scene** |
| --- | --- | --- |
| **micro** | TP, FP, FN for the population. The number of samples for all categories was counted, and Precision, Recall, and F1 were calculated uniformly. | For unbalanced imbalance, suitable to assess the overall effect |
| **macro** | Classification separately calculate Precision, Recall, F1, and then take the arithmetic average of the category index. | Focus on the performance fairness of each category |
| **weighted** | Similar to macro, but weighted according by the number of samples for each category. Categories with more samples affected more overall. | Consider both category proportion and fairness |

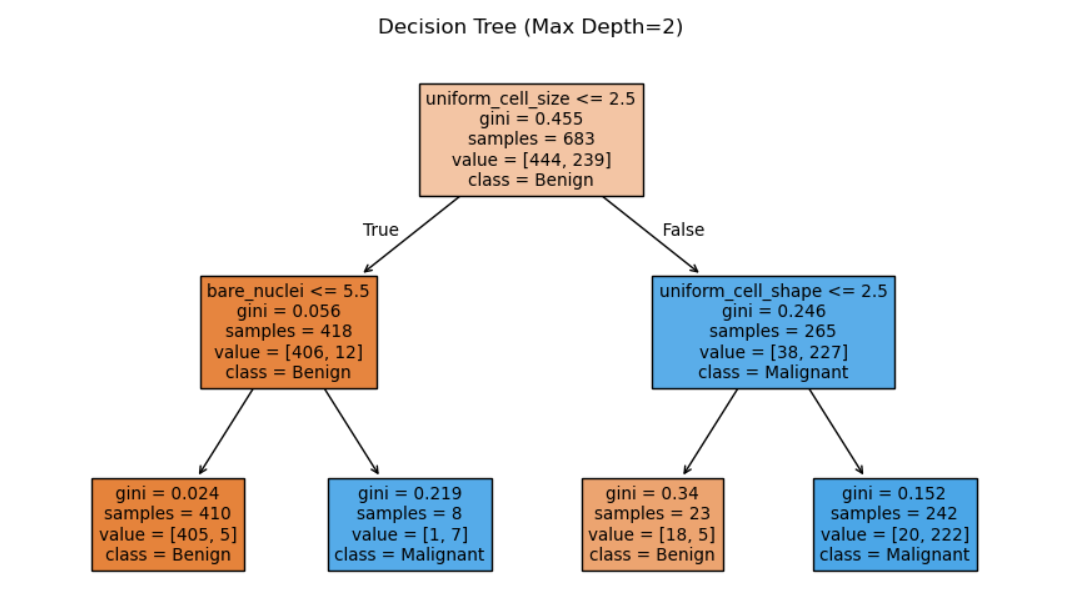
## **P2:**

Which feature was selected for the first division?

Division characteristics: uniform\_cell\_size (uniform cell size)

Decision boundary (threshold): 2.5

What is the decision boundary (division threshold)?



Characteristics of the first division:

uniform\_cell\_size division threshold: 2.5000

========================================

Root node, Class Distributions (Benign, Malignant): [0.65007321 0.34992679]

Root node entropy: 0.9340

Root node Gini: 0.4550

Root node misclassification rate: 0.3499----------------------------------------

Left (Benign, Malignant): [0.97129187 0.02870813]

Left child node entropy: 0.1879

Left child node Gini: 0.0558

Misclassification rate of the left child node: 0.0287----------------------------------------

Right Child node Class Distribution (Benign, Malignant): [0.14339623 0.85660377] Right Child Entropy: 0.5931

Right child node Gini: 0.2457

Misclassification rate of the right child node: 0.1434----------------------------------------

Information gain (entropy-based): 0.1531

## **P3:**

==== PCA - 1 Principal Component ==== F1 Score : 0.8194

Precision : 0.7375

Recall : 0.9219

Confusion Matrix: [[86 21] [ 5 59]]

False Positives (FP): 21

True Positives (TP): 59

False Positive Rate (FPR): 0.1963

True Positive Rate (TPR): 0.9219 ========================================

==== PCA - 2 Principal Components ====

F1 Score : 0.8201

Precision : 0.7600

Recall : 0.8906

Confusion Matrix: [[89 18] [ 7 57]]

False Positives (FP): 18

True Positives (TP): 57

False Positive Rate (FPR): 0.1682

True Positive Rate (TPR): 0.8906 ========================================

==== Original Continuous Data ====

F1 Score : 0.8392 Precision : 0.7595

Recall : 0.9375 Confusion Matrix: [[88 19] [ 4 60]]

False Positives (FP): 19

True Positives (TP): 60

False Positive Rate (FPR): 0.1776

True Positive Rate (TPR): 0.9375 ========================================

#### Yes, whether the use of continuous data helps with the model,

#### 1. Information retention More complete Original continuous data retains details of all 30 features, while PCA (especially only the first one or two principal components) loses some of the important discriminatory information. The decision tree model depends on the specific value of the features to divide. The richer the features, the more effective the model can find the division boundaries.

#### 2. Stronger discriminant ability Continuous data provides a more detailed division possibility, helping the model to more accurately distinguish benign (Benign) from malignant (Malignant) tumors. The results showed higher Recall (recall rate 93.75%) and F1 score (0.8392), indicating more comprehensive detection of malignant tumors and lower misdiagnosis rate.

#### 3. Decision tree model Natural fit continuous feature Decision tree model can handle continuous values well, which makes decisions by automatically finding the best split point (threshold). Although the PCA features after dimension reduction can simplify the data to some extent, the information compression brings the accuracy loss, and the judgment basis of the model is simplified, which is not conducive to high-quality classification.

#### 4. Although PCA can remove redundant information, in this case, the 30 features of the original data are not significantly redundant, and the direct use of the original data instead retains the detailed information that is helpful to the classification decision.