Homework 2

Problem 1

✧ Result

Depth	Recall score	Precision score	F1 score
1	0.6842	0.5332	0.5820
2	0.9737	0.9757	0.9736
3	1.0000	1.0000	1.0000
4	1.0000	1.0000	1.0000
5	1.0000	1.0000	1.0000

✧ Description

The highest recall is 1.0000, and the corresponding depth values are 3, 4, and 5. This is because as the depth increases, the decision tree can fully partition the dataset based on multiple features, increasing the value of True Positives (TP) and reducing the value of False Negatives (FN), thereby improving the recall score.

The lowest precision is 0.5332, and the corresponding depth value is 1. This is because the decision tree with a depth of 1 is too simple, the model can only perform a simple partition of the dataset based on a single feature, which will lead to many misclassification cases, resulting in a low precision rate.

The best F1 score is 1.0000, and the corresponding depth values are 3, 4, and 5. The F1 score comprehensively considers both precision and recall. When both precision and recall reach the maximum, the F1 score also reaches the maximum value.

✧ Difference between the micro, macro, and weighted methods of score calculation

Micro-average: It calculates the evaluation indicators by computing the **global** True Positives (TP), False Positives (FP), and False Negatives (FN).

Macro-average: It calculates the evaluation indicators for **each class** separately and then takes the average value. It treats each class equally and does not consider the differences in the number of samples among various classes.

Weighted-average: It calculates the evaluation indicators for **each class** separately, but it conducts **a weighted average** according to the number of samples in each class.

Problem 2

✧ Result

	Entropy	Gini	Misclassification Error
Before the first split	0.9266	0.4500	0.3419
After the first split	0.0995/0.5783	0.0255/0.2376	0.0129/0.1378
After the first split(weighted)	0.2850	0.1076	0.1076
Information Gain	**0.6417**	**0.3424**	

✧ Description

The feature selected for the first split is uniformity_of_cell_size. The value determined the decision boundary is 2.5 .

Problem 3

✧ Result

	F1 score	Precision score	Recall score
Original data	0.8889	0.8889	0.8889
After the first PCA	0.9320	0.9796	0.8889
After the second PCA	0.9320	0.9796	0.8889

	FP	TP	FPR	TPR
Discrete data	5	55	0.0472	0.8730
Continuous data	6	48	0.0674	0.8889
After the first PCA	1	48	0.0112	0.8889
After the second PCA	1	48	0.0112	0.8889

✧ Description

Continuous data can help the model better capture positive examples, so the value of TPR for continuous data performs better than that for discrete data. However, initially, continuous data resulted in more cases of misclassifying negative samples as positive ones. This situation improved after applying PCA.

Overall, in this case, using continuous data has certain advantages.