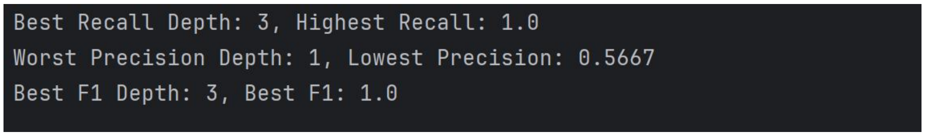
CS 422 Due:23h00
Data Mining Homework 2 March 23rd, 2025

Practicum Problems

These problems will primarily reference the lecture materials and examples provided in class using Python. It is recommended that a Jupyter/IPython notebook be used for the programmatic components. Students are expected to refer to the prescribed textbook or credible online resources to answer the questions accurately.

Problem 1

Load the Iris sample dataset from sklearn (using load_iris()) into Python with a Pandas DataFrame. Induce a set of binary decision trees with a minimum of 2 instances in the leaves (min_samples_leaf=2), no splits of subsets below 5 (min_samples_split=5), and a maximum tree depth ranging from 1 to 5 (max_depth=1 to 5). You can leave other parameters at their default values. Which depth values result in the highest Recall? Why? Which value resulted in the lowest Precision? Why? Which value results in the best F1 score? Also, explain the difference between the micro, macro, and weighted methods of score calculatio.



```
Best Recall Depth: 3, Highest Recall: 1.0
Worst Precision Depth: 1, Lowest Precision: 0.5667
Best F1 Depth: 3, Best F1: 1.0
```

When the depth value is 3, the F1 score is the best (1.0). The F1 score is the harmonic mean of precision and recall. When the depth value is 3, the recall rate reaches 1.0, and the precision is also relatively high. Both have reached a good state and achieved a good balance, so the F1 score reaches the highest.

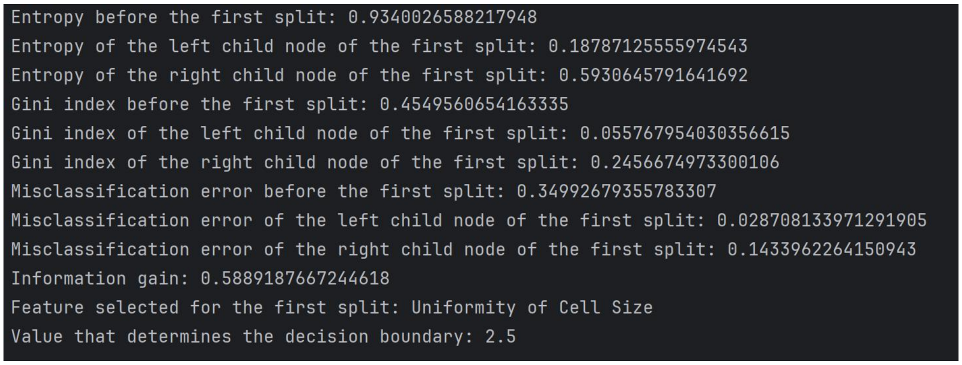
Micro-average score: Firstly, summarize all the true positive, true negative, and false negative examples of each category. Then, calculate the precision, recall rate, and F1 score based on the summarized values. It treats each sample equally and is applicable when the distribution of sample categories is uniform, emphasizing the overall classification accuracy.

Macro-average score: Calculate the precision, recall rate, and F1 score for each category separately, and then take the average. This method treats each category equally and is not affected by the number of samples in each category. It is suitable for scenarios where each category is equally important.

Weighted-average score: Firstly, calculate the precision, recall rate, and F1 score for each category, and then perform weighted average based on the proportion of each category's sample number to the total sample number. Categories with more samples have a greater impact on the final result, and it is applicable when sample imbalance exists and the importance of different categories varies.

Problem 2

Load the Breast Cancer Wisconsin (Diagnostic) sample dataset from the UCI Machine Learning Repository (the discrete version at: breast-cancer-wisconsin.data) into Python using a Pandas DataFrame. Induce a binary Decision Tree with a minimum of 2 instances in the leaves, no splits of subsets below 5, and a maximum tree depth of 2 (using the default Gini criterion). Calculate the Entropy, Gini, and Misclassification Error of the first split. What is the Information Gain? Which feature is selected for the first split, and what value determines the decision boundary?



```
Entropy before the first split: 0.9340026588217948
Entropy of the left child node of the first split: 0.18787125555974543
Entropy of the right child node of the first split: 0.5930645791641692
Gini index before the first split: 0.4549560654163335
Gini index of the left child node of the first split: 0.055767954030356615
Gini index of the right child node of the first split: 0.2456674973300106
Misclassification error before the first split: 0.34992679355783307
Misclassification error of the left child node of the first split: 0.028708133971291905
Misclassification error of the right child node of the first split: 0.1433962264150943
Information gain: 0.5889187667244618
Feature selected for the first split: Uniformity of Cell Size
Value that determines the decision boundary: 2.5
```

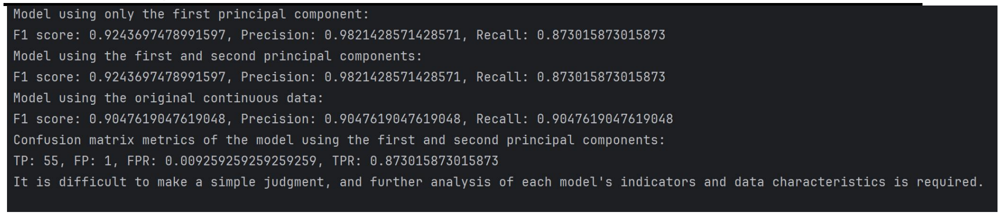
The information gain is 0.5889187667244618, which indicates the degree of improvement in purity of the dataset after this split. The larger the information gain is, the better the partitioning effect of this split on the dataset is.

The feature selected for the first split is the value of best_feature. Through code, all features and possible split points are traversed to find the feature that can bring the maximum information gain as the feature for the first split.

The value that determines the decision boundary is best_threshold. When the value of the sample on this feature is less than or equal to this threshold, it will be classified to the left child node; when it is greater than this threshold, it will be classified to the right child node.

Problem 3

Load the Breast Cancer Wisconsin (Diagnostic) sample dataset from the UCI Machine Learning Repository (the continuous version at: wdbc.data) into Python using a Pandas DataFrame. Induce the same binary Decision Tree as above (now using the continuous data), but perform PCA dimensionality reduction beforehand. Using only the first principal component of the data for model fitting, what are the F1 score, Precision, and Recall of the PCA-based single factor model compared to the original (continuous) data? Repeat the process using the first and second principal components. Using the Confusion Matrix, what are the values for False Positives (FP) and True Positives (TP), as well as the False Positive Rate (FPR) and True Positive Rate (TPR)? Is using continuous data beneficial for the model in this case? How?"



```
Model using only the first principal component:
F1 score: 0.9243697478991597, Precision: 0.9821428571428571, Recall: 0.873015873015873
Model using the first and second principal components:
F1 score: 0.9243697478991597, Precision: 0.9821428571428571, Recall: 0.873015873015873
Model using the original continuous data:
F1 score: 0.9047619047619048, Precision: 0.9047619047619048, Recall: 0.9047619047619048
Confusion matrix metrics of the model using the first and second principal components:
TP: 55, FP: 1, FPR: 0.009259259259259259, TPR: 0.873015873015873
It is difficult to make a simple judgment, and further analysis of each model's indicators and data characteristics is required.
```

Judging from the given indicators, the use of continuous data is relatively less advantageous.

F1 Score Comparison: For the model using only the first principal component and the model using the first and second principal components, the F1 scores are both 0.9243697478991597. However, the F1 score of the model using the original continuous data is 0.9047619047619048. The dimensionality-reduced model is superior in comprehensively measuring precision and recall.

Accuracy Comparison: The accuracy of the two models after dimensionality reduction is 0.9821428571428571, while the accuracy of the original continuous data model is 0.9047619047619048. This indicates that the dimensionality-reduced model has higher accuracy in predicting positive examples.

Feature Correlation and Noise: The original continuous data features are numerous, which may have strong correlations and noise, which can interfere with model learning. However, PCA dimensionality reduction can remove these unfavorable factors and extract key principal components, allowing the model to better learn data patterns. Judging from the given indicators, the use of continuous data is relatively less advantageous.

E.N.D