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 calculation

	depth	recall	precision	F1
0	1	0.666667	0.833333	0.555556
1	2	0.974359	0.976190	0.974321
2	3	1.000000	1.000000	1.000000
3	4	1.000000	1.000000	1.000000
4	5	1.000000	1.000000	1.000000

1. The recall is the highest when the depth is 3, 4, and 5.Reason: When the tree depth is >2, the model has learned enough features to correctly classify most samples.

2. The lowest accuracy is achieved when the depth is 1.Reason: A depth of 1 is too simple to distinguish between multiple classes

3. The highest F1 score is obtained when depth >2.Reason: When depth >2, the model has balanced the classification ability, and the recall and precision are high, so the F1 is high.

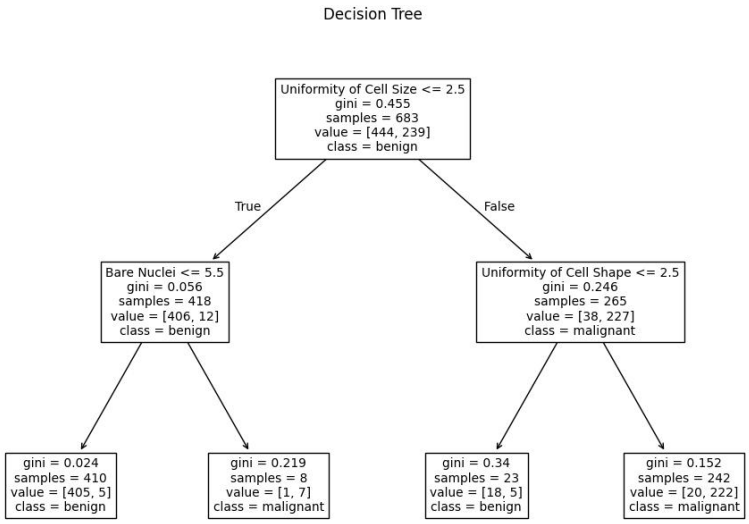
4. Micro average: The index is calculated globally, and the TP, FP, and FN of all categories are summed up and calculated.Macro average: The metrics for each category are calculated independently and then the arithmetic average is taken.Weighted average: The average is weighted according to the number of samples in each class.

Problem 2

Load the Breast Cancer Wisconsin (Diagnostic) sample dataset from the UCI Machine Learning Repository (the discrete version at: breast-cancer-

E.N.D

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：0.934

Gini：0.455

Misclassification Error：0.350

Information Gain：0.589

Selected features:

Feature: Uniformity of Cell Size

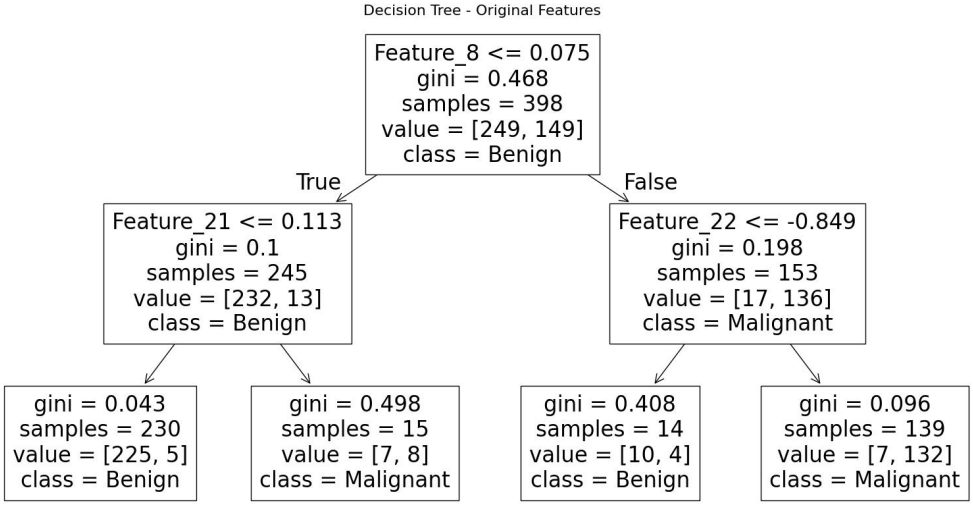
Decision boundary value: 2.5

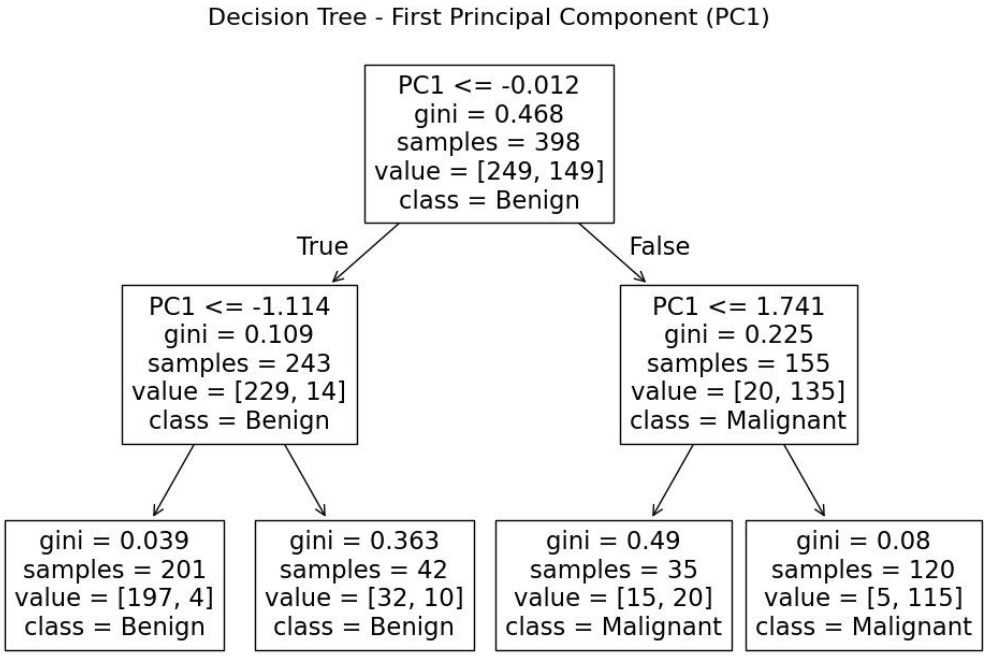
Problem 3

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

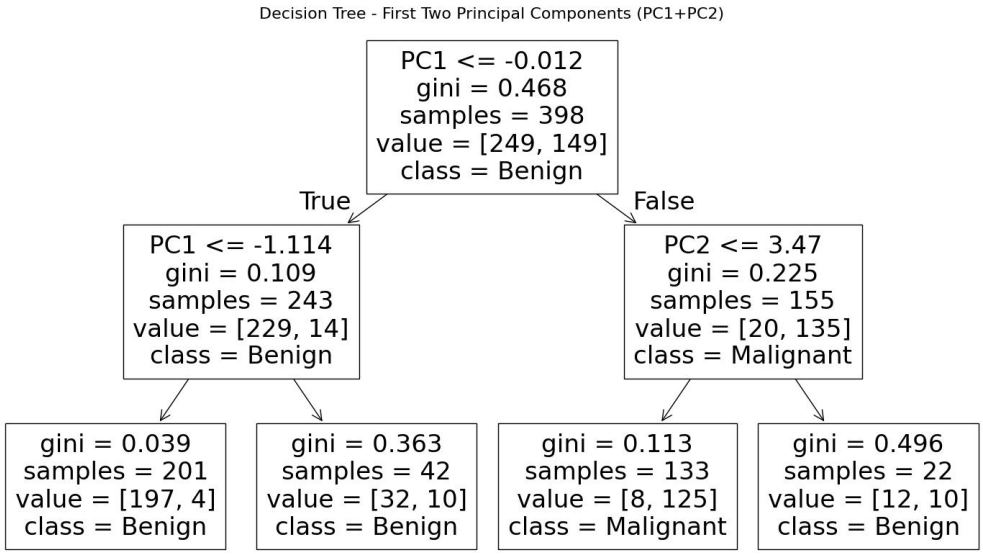
E.N.D

well as the False Positive Rate (FPR) and True Positive Rate (TPR)? Is using continuous data beneficial for the model in this case? How?"





E.N.D



Model performance of Raw data：

F1 Score: 0.9048

Precision: 0.9048

Recall: 0.9048

FP: 6, TP: 57

FPR: 0.0556

TPR/Recall: 0.9048

Model performance of PCA Single principal component：

F1 Score: 0.8992

Precision: 0.8788

Recall: 0.9206

FP: 8, TP: 58

FPR: 0.0741

TPR/Recall: 0.9206

Model performance of PCA two principal components：

F1 Score: 0.8852

Precision: 0.9153

Recall: 0.8571

FP: 5, TP: 54

E.N.D

FPR: 0.0463

TPR/Recall: 0.8571

E.N.D