Machine Learning Course Autumn semester

By Dr. Mahdi Eftekhari Homework #2

Task Description: ID3 Algorithm

In this task, you will implement the ID3 algorithm to create a decision tree classifier using a provided dataset. You will experiment with **feature selection** to determine which features contribute most effectively to the model's accuracy.

Instructions:

1. **Dataset Understanding:**
   Begin by loading the provided dataset. Review its structure, including the features and the target variable. Summarize your findings and note any issues, such as missing values.

2. **Feature Selection and Experimentation:**
   You will try selecting different subsets of features to find the best combination for maximizing accuracy. Follow these steps:

   1. Create multiple subsets of features to test (e.g., using different combinations by hand or through techniques).
   2. For each selected subset, train your ID3 model and calculate its accuracy.

3. **Model Training and Evaluation:**
   Use each feature subset to train your decision tree classifier. Evaluate and compare the accuracy for each combination of features using metrics such as:

   1. **Confusion Matrix:**
      Calculate and present the confusion matrix for each model. This will help you visualize the performance of the classifier, showing the correct and incorrect predictions for each class.

   2. **Precision, Recall, and F1-Score:**

      1. **Precision:** The ratio of true positive predictions to the total predicted positives.
      2. **Recall:** The ratio of true positive predictions to the total actual positives.

3. **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.
4. **Calculate Height of Each Decision Tree**

In your analysis, you are required to calculate the height of each decision tree you create. Here are the steps:

- Build your decision trees using various subsets of features.
- Determine the height by counting the edges in the longest path from the root node to the furthest leaf node.
- Document the height alongside performance metrics like accuracy.
- Analyze how the height of the tree correlates with its complexity and interpretability, and discuss potential overfitting concerns.
- Compare the heights of decision trees from different feature sets and present your findings.

4. **Results Comparison:**
   After evaluating the accuracies for all feature subsets, compare the results. Discuss which features (or combinations of features) resulted in the highest accuracy. Conclude with a statement on which feature subset you found to be the most effective.

5. **Documentation:**
   Document your process, findings, and conclusions within the notebook. Make sure to clearly outline the accuracy results for each feature set and provide reasoning for your selections.

6. **Submission:**
   Ensure that all cells are run and the notebook is functioning as expected. Submit your completed notebook along with any additional notes on your experiment.

Remember, each tree you build is not just a model, but a stepping stone on your journey to becoming a proficient data scientist—let your curiosity lead the way 😎