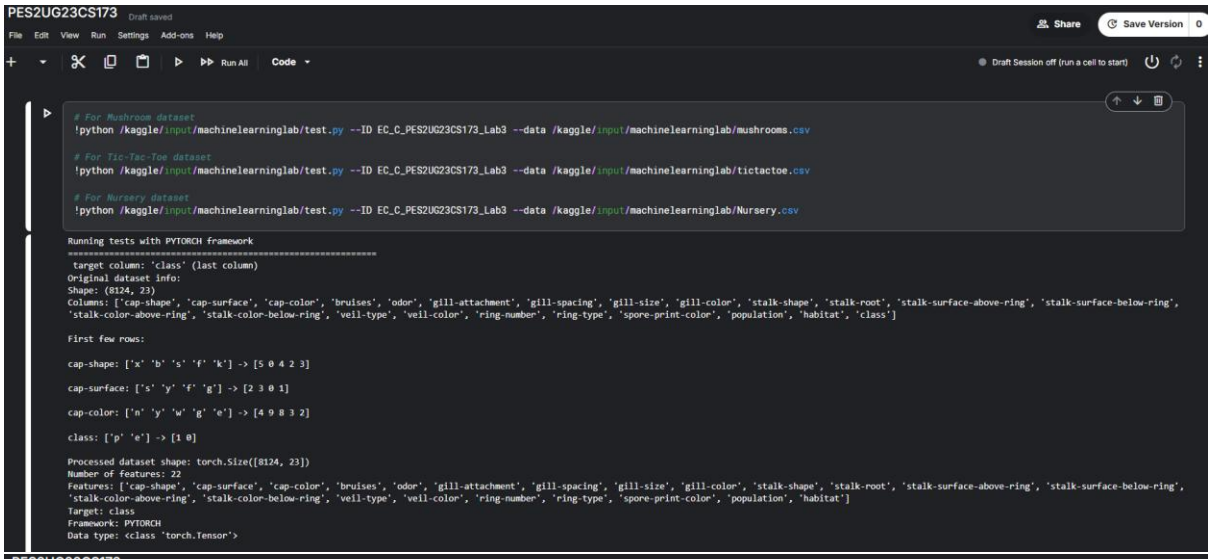
Machine Learning Lab Assignment - 1

Name: Dhruv Maheshwari | Sem 5 – C | Date: 19th Aug 2025

Link to Kaggle Notebook:
https://www.kaggle.com/code/dhruvmaheshwari2004/pes2ug23cs173

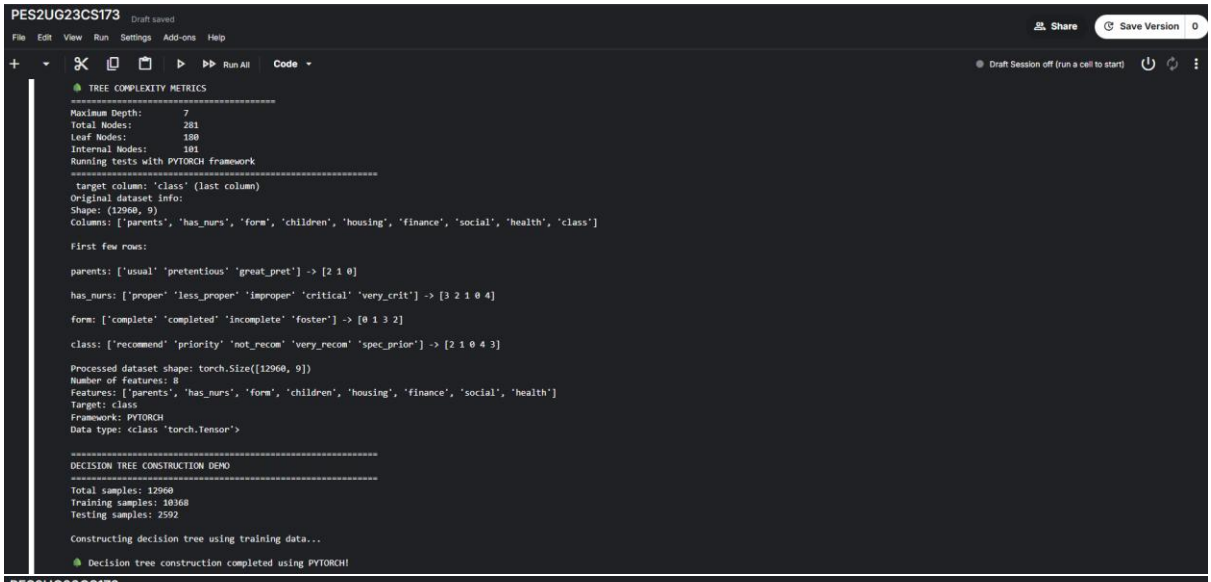


Entropy Calculation: To determine a dataset's entropy, a function is developed. The degree of class mixing in a given set of examples can be ascertained by measuring the impurity or randomness of the data using entropy.

Average Information: The average information of a particular attribute is determined by another function. In essence, this is a weighted average of the entropy of each subset that was produced by dividing the dataset according to that characteristic.

Information Gain: A third function determines the information gain for each attribute using the two functions mentioned above. The decrease in entropy that arises from dividing the data according to a specific feature is measured by information gain. For a split, the best attribute to use is the one with the highest information gain.

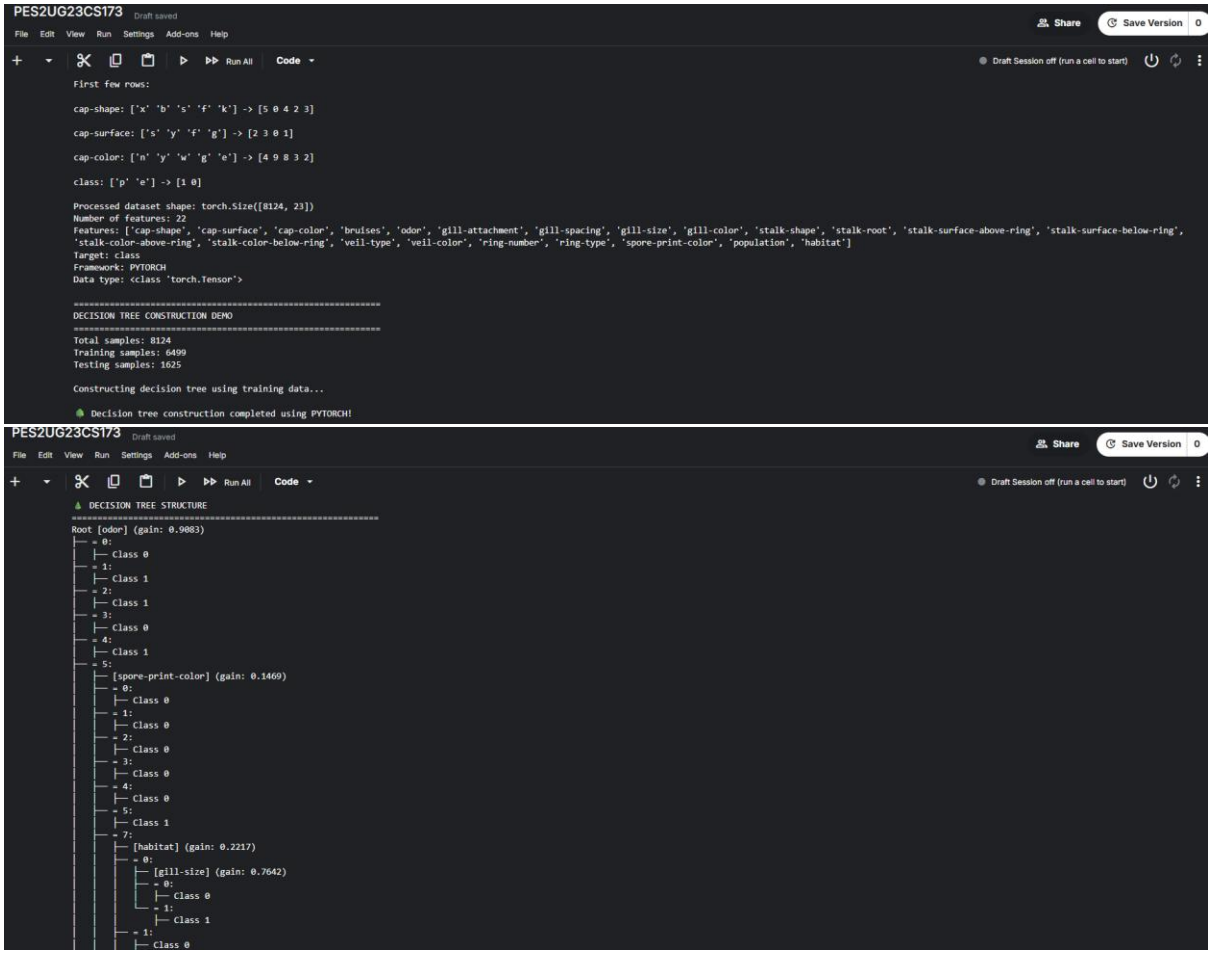
Attribute Selection: The last function iterates through all of the features that are available and chooses the one with the greatest information gain in order to determine which attribute should be used for the subsequent split in the tree.



Execution: The implemented code is run on each dataset using commands to record performance metrics like F1-score, recall, accuracy, and precision.

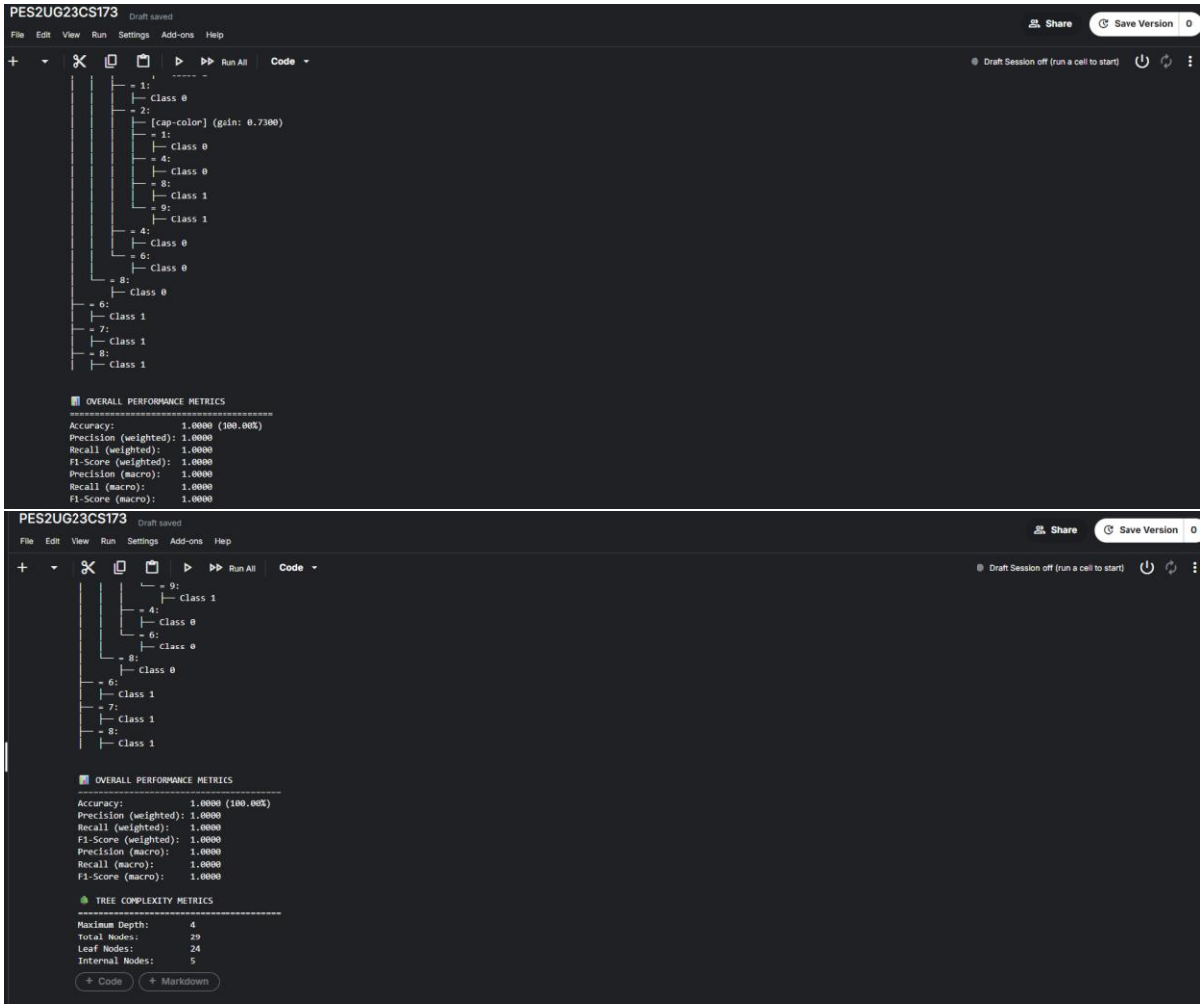
Analysis: A thorough analysis is based on the results of these tests. A thorough comparison of the findings from all three datasets is included in the notebook. This analysis includes:

Performance Metrics: A comparison of each dataset's classification metrics.



Tree Characteristics: An analysis of the depth and total number of nodes in the resultant tree.

Feature Importance: An examination of the features (the tree's root and early splits) that were deemed most crucial for decision-making.



Insights: An end-of-term report explaining how the performance and shape of the decision tree varied in response to dimensions such as dataset size, feature dimension, and class distribution for each particular problem. The report also indicates the benefits of applying an interpretable model such as a decision tree.