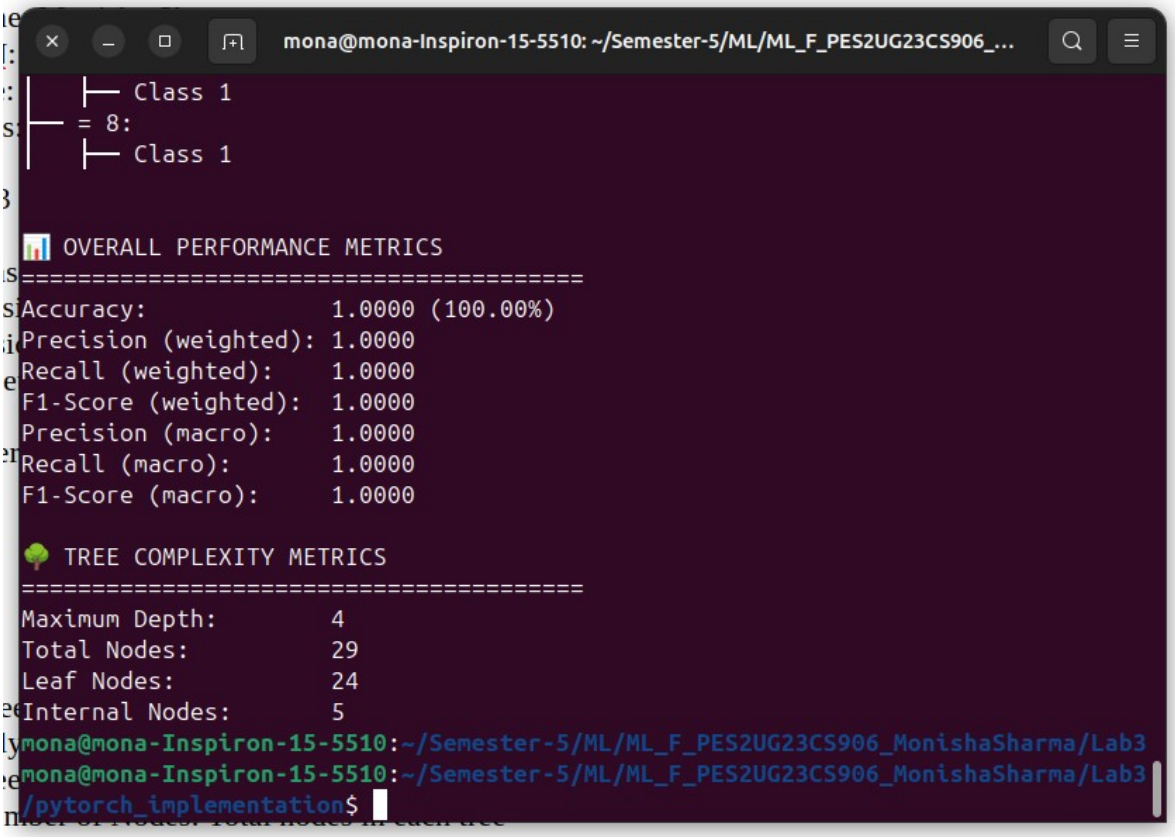
Name: Monisha Sharma
SRN: PES2UG23CS906
Date: 24/08/2025
Class: 5F

Lab3 Decision Tree Classifier – Multi-dataset Analysis

Dataset 1: Mushroom Classification
Classification Task: Predict whether a mushroom is edible or poisonous based on its physical characteristics.
Target Classes: Edible (e) → 0, Poisonous (p) → 1

1. Performance Comparison
2.Tree Characteristics Analysis



3. Dataset-Specific Insights

• Feature Importance: Which attributes contribute most to classification
Most important attribute: odor as it is the root split of the decision tree (gain = 0.9083). Certain odor values (e.g., foul, fishy, musty) almost perfectly identify poisonous mushrooms.

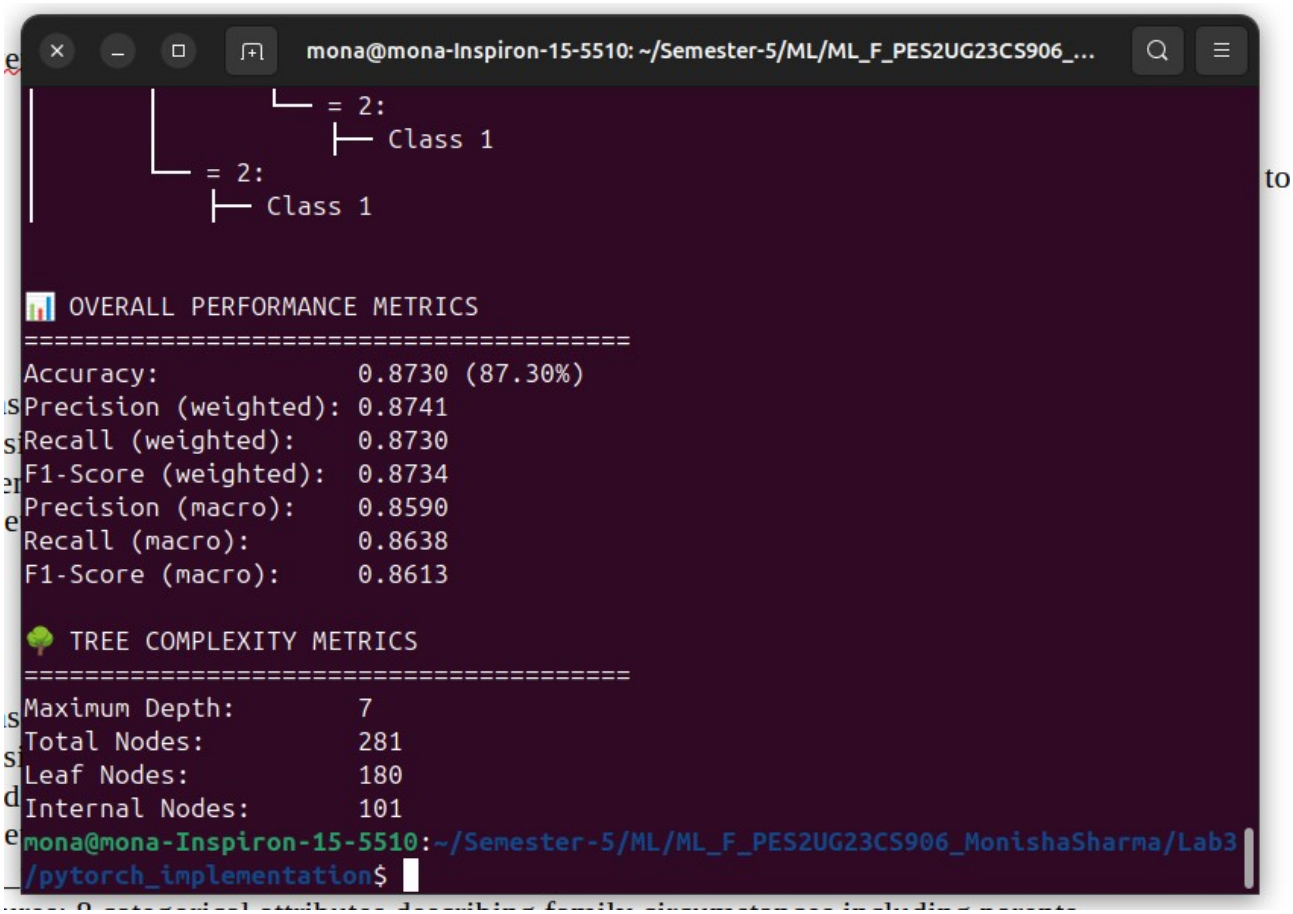
• Class Distribution: How balanced are the target classes
nearly equally balanced indicated by the values obtained

• Decision Patterns: Common decision paths in the tree
The first split (odor) already classifies the majority of mushrooms. The remaining splits refine ambiguous cases where odor = "none."

If odor = foul, creosote, fishy, musty, pungent, spicy → poisonous
If odor = almond or anise → edible
If odor = none → look deeper:

• Overfitting Indicators: Signs of overfitting in tree structure
High training accuracy (100%) and equally high test accuracy (100%) suggests not overfitting because test set confirms perfect generalization.
However, if tree depth is high with many nodes, on smaller/noisy datasets this could lead to overfitting.

Dataset 2: Tic-Tac-Toe Endgame
Classification Task: Predict the outcome of a Tic-Tac-Toe game (win/loss) based on the current board configuration.
Target Classes: Positive (win) → 1, Negative (loss/draw) → 0



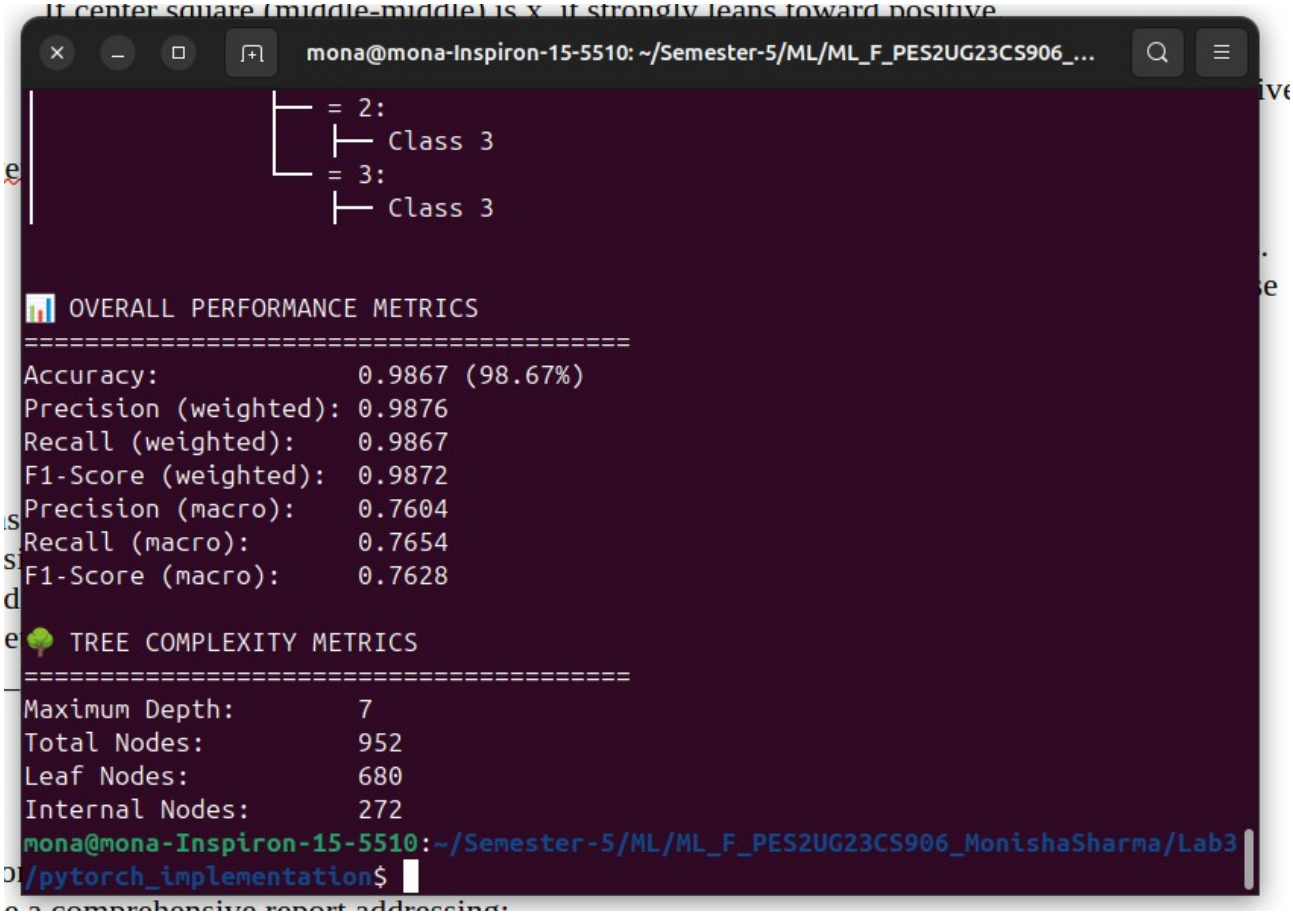
• Feature Importance: Which attributes contribute most to classification
The root node is middle-middle-square with an information gain of 0.0834, meaning it's the most important feature.

• Class Distribution: How balanced are the target classes
the model leans toward a more 75-25 relation in classes

• Decision Patterns: Common decision paths in the tree
    If center square (middle-middle) is x, it strongly leans toward positive.
    If center is o or blank, the decision depends on corners (top-right, top-left, bottom-right).
    Deep branches show combinations like top-left = x + bottom-middle = o → Class = positive.

• Overfitting Indicators: Signs of overfitting in tree structure
    exhibits overfitting tendencies. Structure contains deep, highly specific branches that correspond to rare board states, and several leaf nodes are supported by very few samples. While early splits on the center and corner squares generalize well, subsequent splits chase marginal entropy reductions, producing near-pure leaves. This indicates that the model memorizes fine-grained board patterns instead of learning broader winning strategies, -> overfitting.

Dataset 3: Nursery School
Classification Task: Predict the recommendation level for nursery school admission based on family and social factors.
Target Classes: 5 classes - recommend=2, priority=1, not_recom=0, very_recom=4, spec_prior=3



• Feature Importance: Which attributes contribute most to classification
    The root node is “health” (gain: 0.9595), showing it is by far the most decisive factor in determining nursery recommendations.
• Class Distribution: How balanced are the target classes
    Weighted metrics are very high, because the majority class is predicted well.
    Macro metrics are much lower, indicating weaker performance on minority classes.

• Decision Patterns: Common decision paths in the tree

If health = not_recom (worst value) → immediate class = not_recom.
If health = critical and has_nurs = critical → check parents and finance:
If health = very_crit (best value): finer splits rely on parents and has_nurs:
Finance nearly perfectly decides the class once health, has_nurs, and parents are fixed.

• Overfitting Indicators: Signs of overfitting in tree structure

The model performs extremely well on the dataset overall but shows overfitting risk for minority classes due to deep, complex branches and reliance on low-support splits.

4. Comparative Analysis Report

a) Algorithm Performance:

• Which dataset achieved the highest accuracy and why?
Mushrooms dataset: Achieved the highest accuracy (100%), both on training and test sets. The attribute odor almost perfectly separates edible from poisonous mushrooms.

• How does dataset size affect performance?
Larger datasets help reduce overfitting, but balance matters more than size alone.

• What role does the number of features play?
Fewer, highly informative features (like odor in mushrooms) give cleaner trees, while many multi-valued features (nursery) or combinatorial ones (tic-tac-toe) cause larger, more complex trees.

b)Data Characteristics Impact:

• How does class imbalance affect tree construction?
Mushrooms: Balanced Tic-Tac-Toe: Skewed Nursery: Strongly imbalanced.
Class imbalance leads to biased predictions and deeper trees that attempt to isolate rare classes, increasing overfitting risk.

•Which types of features (binary vs multi-valued) work better?
Binary features work best for decision trees, especially when strongly predictive. Multi-valued features expand complexity and risk overfitting.

c)Practical Applications:

• For which real-world scenarios is each dataset type most relevant?

- Mushrooms:
    - Application: Food safety – quick edible vs. poisonous classification.
    - Strength: Highly interpretable rules like “odor = foul → poisonous.”
- Tic-Tac-Toe:
    - Application: Game strategy modeling and teaching machine learning concepts.

        - Strength: Captures human-like decision paths in games but less useful in real-world decision-making.
    - Nursery:
        - Application: Resource allocation and recommendation systems (e.g., matching children to nurseries based on family/social conditions).
        - Strength: Handles multi-class, multi-criteria decisions, similar to real-world admission or priority assignment systems.

•What are the interpretability advantages for each domain?

- Mushrooms: Extremely interpretable. Rules are short, biologically plausible, and easily communicated to non-experts.
- Tic-Tac-Toe: Partially interpretable. Early splits stategic but deeper branches are too specific to rare cases.
- Nursery: Partially interpretability. Top-level splitsmake intuitive sense, but deeper rules are too complex making the overall tree harder to explain.
