MACHINE LEARNING

Decision Tree Classifier – Multi - Dataset Analysis

Analysis Requirements

1. Performance Comparison

   Compare the following metrics across all three datasets:

   - Accuracy: Overall classification accuracy
   - Precision: True positives / (True positives + False positives)
   - Recall: True positives / (True positives + False negatives
   - F1-Score: Harmonic mean of precision and recall

Dataset	Accuracy	Precision (weighted)	Recall (weighted)	F1-score
Mushroom	1.0000 (100%)	1.0000	1.0000	1.0000
Nursery	0.9867 (98.67%)	0.9876	0.9867	0.9872
Tictactoe	0.8730 (87.30%)	0.8741	0.8730	0.8734

Observations:

- The **Mushroom dataset** achieved **perfect accuracy**. Likely due to strong discriminative features (e.g., odor, gill-size) that almost deterministically separate poisonous vs edible mushrooms.
- The **Nursery dataset** performed extremely well (98.67%), despite more classes and larger size.
- The **TicTacToe dataset** had the lowest accuracy (87.3%), possibly due to more ambiguous patterns in the game outcomes.

2. Tree Characteristics Analysis Analyze and compare:
   • Tree Depth: Maximum depth of the constructed trees
   • Number of Nodes: Total nodes in each tree
   • Most Important Features: Attributes selected as root and early splits
   • Tree Complexity: Relationship between tree size and dataset characteristics

Dataset	Max Depth	Total Nodes	Leaf Nodes	Internal Nodes
Mushroom	4	29	24	5
Nursery	7	952	680	272
Tictactoe	7	281	180	101

Observations:

- **Mushroom tree** is very shallow (depth 4), with only 29 nodes, yet achieves perfect accuracy → dataset is easy to classify with a few strong features.
- **Nursery tree** is deeper (7 levels) with very large node count (952). Complexity is proportional to multi-class nature and many categorical splits.
- **TicTacToe tree** is also depth 7, but smaller (281 nodes) compared to Nursery. It reflects a balanced but more nuanced decision-making process.

3. Dataset-Specific Insights For each dataset, analyze:
   • Feature Importance: Which attributes contribute most to classification
   • Class Distribution: How balanced are the target classes
   • Decision Patterns: Common decision paths in the tree
   • Overfitting Indicators: Signs of overfitting in tree structure

Mushroom Dataset

- **Feature Importance:** Odor and gill-size are typically the most discriminative features in mushroom classification.
- **Class Distribution:** Nearly balanced edible vs poisonous classes.
- **Decision Patterns:** Few features (like odor = foul) directly determine poisonous mushrooms → explains shallow tree.
- **Overfitting Indicators:** None → shallow tree + perfect accuracy suggests very clean separations.

Nursery Dataset

- **Feature Importance:** Parents, finance, social, health strongly influence decisions.
- **Class Distribution:** Multi-class, slightly imbalanced (e.g., "not_recom" is more frequent).
- **Decision Patterns:** Combination of parent status + social/finance lead to majority classifications.
- **Overfitting Indicators:** High node count (952) could indicate overfitting, but strong performance on test set suggests generalization is still good.

TicTacToe Dataset

- **Feature Importance:** Center square ("middle-middle") and rows/columns are most important for win/loss prediction.
- **Class Distribution:** Balanced between positive and negative outcomes.
- **Decision Patterns:** Win/loss patterns require deeper splits (7 levels). Some board states are ambiguous.
- **Overfitting Indicators:** Depth 7 tree with 281 nodes may be close to overfitting, reflected in lower accuracy compared to others.

4. Comparative Analysis Report Write a comprehensive report addressing:

a) Algorithm Performance:

a. • Which dataset achieved the highest accuracy and why?

b. • How does dataset size affect performance?

c. • What role does the number of features play?

b)Data Characteristics Impact:

• How does class imbalance affect tree construction?

•Which types of features (binary vs multi-valued) work better?

c)Practical Applications:

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

•What are the interpretability advantages for each domain?

a) Algorithm Performance

- **Highest Accuracy:** Mushroom dataset (100%) → due to strong categorical features that directly map to class labels.
- **Dataset Size Effect:** Nursery (large dataset) still achieved high accuracy, showing scalability. TicTacToe (small dataset) struggled more.
- **Number of Features:** More features (Nursery: 8) → more splits, larger tree; Mushroom (22 features) → only a few features dominate, leading to small but perfect tree.

b) Data Characteristics Impact

- **Class Imbalance:** Nursery has imbalance across multiple classes → slight drop in macro precision/recall (0.76). Mushroom (binary, balanced) → no issue.
- **Binary vs Multi-valued Features:**
    - Mushroom → many categorical but decisive features (works best).
    - TicTacToe → categorical board positions, but interdependencies are harder to capture with trees.

- Nursery → multi-valued categorical features → deeper splits needed.

c) Practical Applications

- **Mushroom dataset:** Relevant for food safety, agriculture, biology. High interpretability and trustworthiness due to shallow tree.
- **Nursery dataset:** Relevant for childcare resource allocation, social services. Interpretability helps in policy-making but tree is complex.
- **TicTacToe dataset:** Useful for game AI, reinforcement learning examples. Shows limits of decision trees in capturing strategic interactions.

d) Performance Improvement Suggestions

- **Mushroom:** Already perfect → no improvement needed.
- **Nursery:** Apply tree pruning to reduce complexity without hurting accuracy; try ensemble methods (Random Forest).
- **TicTacToe:** Use ensemble methods (Random Forest) or feature engineering (derived features like "two in a row") to improve accuracy beyond 87%.