MACHINE LEARNING WEEK 3 LAB

Comparative Analysis Report:

1. Performance Comparison Across Datasets

Mushroom Dataset

- **Accuracy:** 1.0000 (100%)
- **Precision (weighted):** 1.0000
- **Recall (weighted):** 1.0000
- **F1-score (weighted):** 1.0000
- **Precision (macro):** 1.0000
- **Recall (macro):** 1.0000
- **F1-score (macro):** 1.0000

Reasoning: The mushroom dataset is perfectly separable because attributes like *odor* almost completely determine edibility vs. poison.

Tic-Tac-Toe Dataset

- **Accuracy:** 0.8730 (87.30%)
- **Precision (weighted):** 0.8741
- **Recall (weighted):** 0.8730
- **F1-score (weighted):** 0.8734
- **Precision (macro):** 0.8643
- **Recall (macro):** 0.8638
- **F1-score (macro):** 0.8630

Reasoning: Board configurations require deeper splits, and some positions overlap between winning and non-winning states, lowering performance compared to mushroom dataset.

Nursery Dataset

- **Accuracy:** 0.9867 (98.67%)
- **Precision (weighted):** 0.9867
- **Recall (weighted):** 0.9867
- **F1-score (weighted):** 0.9867
- **Precision (macro):** 0.7624

- **Recall (macro):** 0.7628
- **F1-score (macro):** 0.7628

Reasoning: While weighted metrics are very high, macro precision/recall are lower due to class imbalance—some classes have fewer samples, reducing balanced performance.

2. Tree Characteristics Analysis

Mushroom

- **Maximum Depth:** 4
- **Total Nodes:** 29
- **Leaf Nodes:** 24
- **Internal Nodes:** 5
- **Key Features:** *Odor* is the most important feature at the root, giving almost perfect information gain.
- **Tree complexity:** Very low because strong features provide pure splits early.

Tic-Tac-Toe

- **Maximum Depth:** 7
- **Total Nodes:** 281
- **Leaf Nodes:** 180
- **Internal Nodes:** 101
- **Key Features:** *Middle square* and *corner squares* dominate early splits.
- **Tree complexity:** Higher because no single feature fully classifies; combinations of positions are required.

Nursery

- **Maximum Depth:** 7
- **Total Nodes:** 952
- **Leaf Nodes:** 680
- **Internal Nodes:** 272
- **Key Features:** *Parents*, *finance*, and *social* attributes dominate early splits.
- **Tree complexity:** Very high due to many multi-valued categorical features.

3. Dataset-Specific Insights

- **Mushroom:** Balanced classes, clear decision patterns (odor → class), almost zero overfitting risk.
- **Tic-Tac-Toe:** Balanced classes, but deeper trees are required. Moderate overfitting risk if not pruned.
- **Nursery:** Strong class imbalance in some labels. Larger, deeper tree can overfit minority classes.

4. Comparative Analysis

(a) Algorithm Performance:

- **Highest accuracy:** Mushroom (100%) due to pure splits.
- **Dataset size impact:** Larger datasets (Mushroom, Nursery) produce stable results, but size alone doesn't guarantee perfect accuracy (Tic-Tac-Toe is small but still decent).
- **Number of features impact:** More features (Nursery, Mushroom) → deeper and more complex tree, but performance depends on information gain quality, not just count.

(b) Data Characteristics Impact:

- Class imbalance in Nursery reduces macro precision/recall.
- Binary features (Mushroom) are cleaner and lead to smaller trees. Multi-valued features (Nursery) increase depth and complexity.

(c) Practical Applications:

- Mushroom → Food safety inspection systems.
- Tic-Tac-Toe → Game AI decision-making.
- Nursery → Automated admission recommendation systems.

(d) Improvements:

- Use **tree pruning** to avoid overfitting in Nursery/Tic-Tac-Toe.
- Use **balanced sampling or class weights** for Nursery to improve macro metrics.
- Limit **max depth** or use **feature selection** for large feature sets.