Analysis and Insights

Performance Comparison

Accuracy: Overall classification accuracy

- mushrooms.csv = 1.0000 (100.00%)
- tictactoe.csv = 0.8936 (89.36%)
- Nursery.csv = 0.9867 (98.67%)

Precision: True positives / (True positives + False positives)

Weighted

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8930
- Nursery.csv = 0.9876

Macro

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8846
- Nursery.csv = 0.7604

Recall: True positives / (True positives + False negatives)

Weighted

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8936
- Nursery.csv = 0.9867

Macro

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8846
- Nursery.csv = 0.7654

F1-Score: Harmonic mean of precision and recall

Weighted

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8932
- Nursery.csv = 0.9872

Macro

- mushrooms.csv = 1.0000
- tictactoe.csv = 0.8816
- Nursery.csv = 0.7628

Insights

- Mushrooms dataset is perfectly classified
- Nursery also performs extremely well in weighted metrics but macro metrics drop which implies that some classes are harder to predict for than others.
- Tic-Tac-Toe has the lowest performance which implies the patterns are harder to generalize.

Tree Characteristics Analysis

Tree Depth: Maximum depth of the constructed trees

- mushrooms.csv = 4
- tictactoe.csv = 7
- Nursery.csv = 7

Number of Nodes: Total nodes in each tree

- mushrooms.csv = 59
- tictactoe.csv = 306
- Nursery.csv = 992

Most Important Features: Attributes selected as root and early splits

- mushrooms.csv = 46
- tictactoe.csv = 196
- Nursery.csv = 710

Tree Complexity: Relationship between tree size and dataset characteristics

- mushrooms.csv = 13
- tictactoe.csv = 110
- Nursery.csv = 282

Insights

- Mushrooms: It is a very Simple and Shallow tree.
- Tic-Tac-Toe: It is a Deeper Tree with a larger size needed to capture the game logic.
- Nursery: It is a very large tree and naturally the complexity increases with this dataset.

Dataset-Specific Insights

Mushrooms

- Feature Importance: (Odor, spore-print-color, bruises) dominate in comparison to other columns.
- Class Distribution: Balanced.
- Decision Patterns: Early splits around odor.
- Overfitting Indicators: None since its a simple tree with perfect accuracy

Tic-Tac-Toe

- Feature Importance: Middle squares (middle-middle-square) dominate since central positions are most decisive.
- Class Distribution: Balanced.
- Decision Patterns: Requires multiple splits to capture win conditions.
- Overfitting Indicators: Moderate tree depth; generalizes reasonably.

Nursery

- Feature Importance:(Parents, has_nurs, form) are critical splits.
- Class Distribution: Imbalanced -inferred from differing macro and weighted average values
- Decision Patterns: Complex, with many deep splits.
- Overfitting Indicators: overfitting minority classes.

Comparative Analysis Report

Algorithm Performance

- Highest Accuracy-Mushrooms because certain features directly map to desired output
- Nursery is the Largest dataset and it also gives us the most complex decision tree. So I conclude that dataset size is proportional to complexity(i.e larger the dataset more complex the decision tree).
- More features imply more nodes

Data Characteristics Impact

- We see that Nursery dataset has a high class imbalance as inferred from the differing macro and weighted values of the dataset.It leads to more nodes being needed and an increased complexity.

- Binary features as seen in Mushroom datatset lead to cleaner splits and since it gives us the most accuracy amongst all the dataset we infer that Binary is better than multi-values features.
- Shallow trees are more interpretable than larger trees.

Practical Applications

- Mushrooms datset could be used in medical,food safety applications
- Tic Tac Toe dataset could be used in reinforcement learning examples
- Nursery dataset could be used in social policy and recommendation systems.