Decision Tree Performance Analysis Report

A) Algorithm Performance

Highest Accuracy Datasets

The decision tree model demonstrated exceptional performance on the **mushrooms** dataset, achieving an accuracy of **100%**. This was followed by the **Nursery** dataset at **98.67%** and the **Tic-Tac-Toe** dataset at **87.30%**.

The mushrooms dataset's perfect score can be attributed to its highly discriminative features. The 'odor' attribute, in particular, served as the root node with an impressive information gain of **0.9083**. This single feature was powerful enough to create a very clean decision boundary, effectively separating edible from poisonous mushrooms.

Dataset Size Impact on Performance

An analysis of the three datasets reveals that size alone doesn't guarantee superior performance:

- **Nursery:** 12,960 samples → 98.67% accuracy
- **Mushrooms:** 8,124 samples → 100% accuracy
- **Tic-Tac-Toe:** 958 samples → 87.30% accuracy

While the largest dataset (Nursery) had strong performance, the moderately-sized mushrooms dataset achieved perfect accuracy. The smallest dataset (Tic-Tac-Toe) performed the worst. This suggests that while a sufficiently large dataset is important for robust learning, the **quality and relevance of the features** are more critical to a model's success.

Role of Number of Features

The number of features also did not directly correlate with performance:

- **Mushrooms:** 22 features, Max depth: 4 → 100% accuracy
- **Nursery:** 8 features, Max depth: 7 → 98.67% accuracy
- **Tic-Tac-Toe:** 9 features, Max depth: 7 → 87.30% accuracy

The mushrooms dataset, with the most features, achieved a perfect score with a surprisingly shallow tree (depth of 4), highlighting that a small number of **highly relevant features** can be far more effective than a large number of only moderately relevant ones.

B) Data Characteristics Impact

Class Imbalance Effects

The mushrooms dataset illustrates the benefit of balanced, highly discriminative features, resulting in a minimal tree complexity (**29 total nodes, 24 leaf nodes**).

In contrast, the Nursery dataset's macro-averaged precision of **76.04%** was significantly lower than its weighted precision of **98.76%**. This disparity is a strong indicator of **class imbalance**, where the model struggles to correctly predict less frequent classes.

Binary vs. Multi-valued Features

This analysis shows that multi-valued categorical features can be extremely effective if they have high discriminative power. The 'odor' feature in the mushrooms dataset, with its nine distinct values, provided excellent class separation. The key takeaway is that the **information content** of a feature is what matters, not whether it is binary or multi-valued.

The Tic-Tac-Toe dataset, with its simple three-valued features ('x', 'o', 'blank'), struggled more, resulting in a lower accuracy and a much deeper, more complex tree (**281 total nodes**).

C) Practical Applications

Real-world Relevance

- **Mushrooms Dataset:** Applicable in **food safety** and automated mushroom identification apps. The model's excellent interpretability (e.g., "if odor is fishy/spicy/foul, then poisonous") makes it ideal for life-critical applications where trust and understanding of the rules are paramount.
- **Nursery Dataset:** Relevant for **educational policy** and childcare allocation systems. The decision factors are meaningful to administrators, making the model's output easily interpretable in a real-world administrative context.
- **Tic-Tac-Toe Dataset:** Useful for developing **game AI** and educational tools. While the interpretability is moderate, it provides a great example for demonstrating sequential decision-making.

Performance Improvement Strategies

- **For the Mushrooms Dataset:** Since accuracy is already perfect, the focus would be on **tree simplification** to improve computational efficiency.
- **For the Nursery Dataset:** To address the identified class imbalance, strategies like **oversampling minority classes**, **using weighted loss functions**, or employing **ensemble methods** would be effective.
- **For the Tic-Tac-Toe Dataset:** To improve performance, one could **expand the dataset** with more game states or add **derived features** (e.g., "number of X's in a row") that encode strategic information.

Conclusion

This analysis demonstrates that the quality and information content of features are more influential on decision tree performance than dataset size or feature quantity. The mushrooms dataset's success is a perfect example of how a few highly discriminative features can lead to a simple, accurate, and easily interpretable model. In practical applications, this simplicity is a significant advantage, particularly in domains where decisions are critical and need to be understood by humans.