Comparative Analysis Report

Name: Megha Dhanya	SRN: PES2UG23CS336	Section: F

Write a comprehensive report addressing:

a) Algorithm Performance

1. Which dataset achieved the highest accuracy and why?
The Mushroom dataset achieved the highest accuracy. This was because it contained very strong discriminative features, especially the attribute odor, which alone can almost completely determine whether a mushroom is edible or poisonous. As a result, the decision tree could separate classes perfectly with simple rules.

2. How does dataset size affect performance?
Larger datasets generally improve performance by giving the decision tree more examples to learn from. The Nursery dataset, being the largest, allowed the model to learn diverse patterns and achieved high overall accuracy. Smaller datasets like Tic-Tac-Toe provided fewer examples, which made it harder for the tree to generalize and lowered accuracy.

3. What role does the number of features play?
The number of features influences both tree depth and complexity. In Mushroom, many features were available but only a few were highly useful, leading to a small and accurate tree. In Tic-Tac-Toe, all nine features (board positions) mattered together, so the tree became deeper and more complex. In Nursery, the moderate number of features combined with a large dataset resulted in a very large tree but good performance overall.

b) Data Characteristics Impact

1. How does class imbalance affect tree construction?
Class imbalance makes the decision tree biased toward the majority class. In the Nursery dataset, the most frequent class (“not_recom”) dominated predictions, which increased overall accuracy but reduced performance on minority classes such as “spec_prior” and “very_recom.”

2. Which types of features (binary vs multi-valued) work better?
Binary or nearly-binary features tend to work better for decision trees because they create clean and simple splits, as seen in the Mushroom dataset with odor. Multi-valued features, such as those in the Nursery dataset, lead to more branches per split, making the tree larger and harder to interpret.

c) Practical Applications

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

- Mushroom: Relevant for food safety, where quick and interpretable rules can identify poisonous mushrooms.
- Tic-Tac-Toe: Useful in game strategy analysis or as an educational example for studying pattern recognition.
- Nursery: Applicable in school admission systems, where family and social factors are used to recommend admission categories.

2. What are the interpretability advantages for each domain?

- Mushroom: Very high interpretability, since the rules are short and strongly tied to physical properties like odor.
- Tic-Tac-Toe: Helps explain which board configurations lead to certain outcomes, though the tree can be large.
- Nursery: Provides clear conditions that explain how combinations of family and social factors affect recommendations, supporting transparent decision-making.

3. How would you improve performance for each dataset?

- Mushroom: Already perfect, no major improvements needed.

- Tic-Tac-Toe: Performance could be improved with pruning or ensemble methods (e.g., Random Forests) to better capture feature interactions.
- Nursery: Handling class imbalance through re-sampling, class weighting, or cost-sensitive learning would improve fairness across all recommendation classes.