Comparative Analysis Report

NAME: NEHA HARISH

SRN: PES2UG23CS378

1. Algorithm Performance:
    - Which dataset achieved the highest accuracy and why?

    =>Mushroom dataset achieved the highest accuracy because the features (like Odor) are strongly predictive of the target (edible vs poisonous). Many of these attributes have a direct causal relationship with class labels, leading to near-perfect separability.

    - How does dataset size affect performance?

    =>Larger datasets improve generalization by covering more diverse patterns, but also increase training time and model complexity. Smaller datasets may lead to overfitting, especially when the tree depth is high relative to the sample size.

    - What role does the number of features play?

    =>High feature count improves model accuracy if features are informative, but risks overfitting if irrelevant. Low feature count limits expressiveness of the model, but keeps trees simpler and more interpretable.

2. Data Characteristics Impact:

    - How does class imbalance affect tree construction?

    =>Decision trees tend to favour majority classes when imbalance exists.

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

    =>Binary features (Tic-Tac-Toe) simplify tree construction and reduce splits. Multi-valued features (Mushroom, Nursery) allow richer rules but increase tree branching, leading to more complex structures.

3. Practical Applications:

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

    => Mushroom dataset: Food safety applications (predicting poisonous mushrooms). Tic-Tac-Toe dataset: Game AI (training models to learn optimal

strategies). Nursery dataset: Decision support in childcare and school admission systems.

- What are the interpretability advantages for each domain?

=>Mushroom: Easy to interpret rules. Useful for non-technical users. Tic-Tac-Toe: Rules mirror human strategies, making them educational. Nursery: Helps explain social decisions like admission priority in an interpretable rule-based format.

- How would you improve performance for each dataset?

=>The Mushroom dataset already achieves near-perfect accuracy, so only minimal improvements are needed, such as pruning redundant features to simplify the model. For the Nursery dataset, performance can be improved by applying feature selection or dimensionality reduction to remove less useful features and by balancing the classes through oversampling or using weighted decision splits. In the case of the Tic-Tac-Toe dataset, the dataset can be expanded using board symmetries as a form of augmentation, and pruning can be applied to prevent overfitting given the relatively small amount of data.