Comparative Analysis Report

A. Algorithm Performance

1. *Which dataset achieved the highest accuracy and why?*

Ans: Mushrooms Dataset achieved the highest accuracy with 100%. The mushroom.csv dataset is perfectly separable. There are strong categorical feature- class correlations. ID3 can find pure partitions, so it classifies without any error.

2. *How does dataset size affect performance?*

Ans: Larger datasets like Mushroom and nursery give the decision tree more patterns to learn, hence improving stability and generalization. Smaller datasets like tictactoe, limits how well the model generalizes resulting in a lower accuracy. However, larger dataset increases tree complexity, leading to overfitting if pruning isnt used.

3. *What role does the no. of features play?*

Ans: The no. of features directly impacts a model's accuracy, complexity and generalization ability.

More no. of features improves the accuracy,however it increases the risk of overfitting. This leads to higher training time and computational cost. Fewer no. of features makes the model faster and simpler, the accuracy dips lower,and can cause underfitting - where model fails to capture important relationships.

We need to find an optimal no. of features that provide just enough information for the model to learn efficiently without introducing excessive noise or complexity.

This is achieved through feature selection or dimensionality reduction. E.g. Principle component analysis(PCA)

B. Data Characteristics impact

1. *How does the class imbalance affect tree construction?*

Ans: In decision tree learning, class imbalance can bias the model towards the majority class - leading to poor detection of minority classes. The tree may grow shallow for majority classes but fail to create strong splits for minority classes , reducing recall. A solution for this is to use resampling , class weights or ensemble methods to handle imbalance.

2. *Which type of features (binary vs multi-valued) word better?*

Ans: Binary features simply tree splits often making trees smaller and more interpretable. Multi-valued features provide richer information but may increase tree depth and complexity. The effectiveness depends on the nature of the dataset used - balanced binary splits often work better for classification, while multi-valued features are helpful if categories encode meaningful variation

C. Practical applications

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

Ans: Mushroom dataset is relevant for toxicology and food safety- classifying edible vs poisonous mushrooms in the wild.
Nursery Dataset is relevant for decision support systems - assigning children to appropriate nurseries based on social, financial and family conditions.
Tic-tac-toe dataset is relevant for game AI - learning strategies and predicting winning moves in turn-based games.

2. *What are the interpretability advantages for each domain?*
   Ans:
   - Mushroom: Decision tree rules are highly interpretable (oder = foul ⇒ poisonous), which is useful for safety critical decisions.
   - Nursery: Rules can helps policy makers or institutions justify admission decisions fairly and transparently.
   - Tictactoe: Rules show strategic patterns which are interpretable for understanding gameplay

3. *How would you improve performance for each dataset?*
   Ans:
   - Mushroom is already at 100%, but pruning could simplify the tree for interpretability without hurting accuracy
   - Nursery : add pruning or ensemble methods(random forest) to handle noise and to avoid overfitting
   - Tictactoe: use feature engineering or switch to more expressive models(SVM) since pure ID3 struggles with overlapping binary patterns.