Comparative Analysis Report

PES2UG23CS334
Medha Raj

a) Algorithm Performance

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

Mushrooms dataset has highest accuracy because of strong feature-target correlations.The mushroom dataset achieved the highest accuracy because certain features, especially odor, almost perfectly determine whether a mushroom is edible or poisonous, making the classification task very straightforward compared to nursery and tic-tac-toe.

a.2 How does dataset size affect performance?

Larger datasets typically improve performance consistency, while smaller datasets risk overfitting.Larger datasets like mushrooms and nursery give the model more examples to learn patterns and generalize better, while smaller datasets like tic-tac-toe can lead to overfitting and less consistent accuracy.

a.3 What role does the number of features play?

Having more features increases complexity, but not necessarily accuracy; datasets like mushrooms benefit from fewer but highly predictive features, while nursery suffers from too many overlapping attributes.

b) Data Characteristics Impact

b.1 How does class imbalance affect tree construction?

Class imbalance biases the decision tree toward predicting the majority class, which reduces its ability to correctly classify minority classes and results in uneven performance.

b.2 Which types of features (binary vs multi-valued) work better?

Binary features produce simpler, shallower trees that are easier to interpret, while multi-valued features can be more powerful but risk overfitting if their values don't strongly correlate with the target.Multi-valued features can be powerful if values correlate strongly with class, but binary features give simpler, more interpretable trees.

c) Practical Applications

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

- The mushroom dataset is most relevant to food safety and toxicology, the nursery dataset applies to school admissions and resource allocation, and the tic-tac-toe dataset is useful for developing game AI and testing learning strategies.

c.2 What are the interpretability advantages for each domain?

Mushroom classification is highly interpretable since simple rules like odor determine outcomes, tic-tac-toe is intuitive because rules align with game strategies, while nursery is moderately interpretable due to the complexity of many interacting features.

d) How would you improve performance for each dataset?

Performance can be improved by pruning redundant features in mushrooms, applying feature selection and balancing class distribution in nursery, and augmenting data with board symmetries plus pruning in tic-tac-toe to avoid overfitting.