ID3 Decision Tree Analysis Report

Name: Mohammed Aahil Parson
SRN: PES2UG23CS342

1. **Performance comparison**:

Dataset	Accuracy	Precision	Recall	F1-Score
mushrooms.csv	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Nursery.csv	**98.67%**	**98.76%**	**98.67%**	**98.72%**
tictactoe.csv	**87.30%**	**87.41%**	**87.30%**	**87.34%**

2. **Tree Characteristics Analysis**

Dataset	Tree Depth	Number of Nodes	Important Features	Tree Complexity
mushrooms.csv	**Shallow(4)**	**Low**	**Odor, spore-print-color**	**Simple and has less number of splits**
Nursery.csv	**Medium (7)**	**High**	**Financial, Social and health**	**It has got High Complexity**
tictactoe.csv	**Medium (7)**	**Medium**	**Central and corner positions**	**It has got Medium Complexity**

3. Dataset-Specific Insights

mushrooms.csv Dataset:
• Feature Importance: Odor attribute dominates classification choices
• Class Distribution: Equally distributed between edible and poisonous classes
• Decision Patterns: One factor typically makes the decision
• Overfitting Indicators: Little overfitting owing to good feature predictability

Nursery.csv Dataset
• Feature Importance: Economic, social status, and health elements
• Class Distribution: Unbalanced, with "not_recom" class prevailing
• Decision Patterns: Complex multi-attribute decision paths
• Indications of Overfitting: Deep tree implies potential overfitting

tictactoe.csv Dataset
• Feature Importance: Hence, center and corner board positions.
• Class Distribution: Distributed equally between positive and negative results
• Decision Patterns: Logical game-winning combinations
• Overfitting Indicators: Potential memorization of some board positions

4. Comparative Analysis Report

a) Algorithm Performance

Which dataset achieved the highest accuracy and why?
mushrooms.csv dataset had the best accuracy rate (100%) because of highly discriminative features
that is, the smell descriptor most frequently associated with mushroom poisoning.

How does dataset size affect performance?
Larger datasets such as Nursery.csv create deeper trees with increased branches, which leads to overfitting.
Smaller, well-structured datasets like mushrooms.csv produce simpler, more generalizable models.

What role does the number of features play?
More features add to the complexity of trees. mushrooms.csv (22 features) performed best with simple.
splits, whereas Nursery.csv (8 attributes) needed deeper splitting because of multi-valued categorical attributes.

b) Data Characteristics Impact

How does class imbalance affect tree construction?

Class imbalance in the Nursery.csv dataset causes the tree biased towards the majority classes.Forecasts of "not_recom" outcomes and reduced performance on minority groups.

Which types of features work better?

Binary features create less complex decision boundaries and smaller trees. Multi-valued categorical features in Nursery.csv dataset increases tree complexity and branching factor, so interpretation becomes challenging.

c) Practical Applications

Real-world uses for each type of dataset:
• **mushrooms.csv**:used in classification of flowers and other fruits, used in Food Safety to prevent getting poisoning from packed food.
• **tictactoe.csv**: used for game development, can be used for strategic based classification for model training and model inference.
• **Nursery.csv**: School Admissions, can be used for training the model for better resource allocation.

Interpretability benefits for every field:
• **mushrooms.csv**: Easy availability of safety protocols to fieldworkers
• **tictactoe.csv**: Transparent game's strategy logic
• **Nursery.csv**: Criteria which are comprehensible for fairness assessment

How would you improve performance for each dataset?
• **mushrooms.csv**: Apply post-pruning for size reduction of the tree while maintaining precision.
• **tictactoe.csv**: Use early stopping mechanism to avoid memorization of infrequent board positions.
• **Nursery.csv**: Apply class weighting to handle imbalance and restrict maximum tree depth to minimize overfitting.