Lab Report Template - Decision Tree (ID3) Analysis

Name: Mayuran Ravi Pillai

SRN:PES2UG23CS333

Section: F

Date: 25-08-25

Objective

Implement the ID3 Decision Tree algorithm and perform comparative analysis across three datasets (Mushroom, Tic-Tac-Toe, Nursery).

Dataset Descriptions

Dataset 1: Mushroom Classification

Dataset 2: Tic-Tac-Toe Endgame

Dataset 3: Nursery School

Performance Comparison

Dataset	Accuracy	Precision	Recall	F1-Score
Mushroom	1.000(100%)	1.000	1.000	1.000
Tic-Tac-Toe	0.8836(88.36%	0.8827	0.8836	0.8822
Nursery	0.9887(98.87%	0.9888	0.9887	0.9887

Tree Characteristics

Dataset	Tree Depth	No. of Nodes	Most Important Features

Mushroom	4	29	Very shallow & compact tree
Tic-Tac-Toe	7	260	Medium complexity but still inaccurate due to noise
Nursery	7	983	Huge, complex but accurate

Dataset-Specific Insights

Mushroom Dataset:

Feature Importance: Odor , Spore-print-color, Gill size.

Class Distribution: Balanced (edible vs poisonous).

Decision Patterns: Few strong rules can perfectly separate classes.

Overfitting Indicators: None since it is a shallow tree and is perfectly accurate.

Tic-Tac-Toe Dataset:

Feature Importance: Central cell, Corners, Edges (important to decide outcome of game).

Class Distribution: Likely balanced.

Decision Patterns: Board states with similar moves : harder to separate.

Overfitting Indicators: Tree depth is 7 but not very accurate. The dataset complexity is the limiting factor.

Nursery Dataset:

Feature Importance: Parents, Financial status, Health, Housing.

Class Distribution: Imbalanced (Some class labels have very few samples).

Decision Patterns: Many branch paths due to multivalued features.

Overfitting Indicators: Large tree but accuracy is still high, not much overfitting.

Comparative Analysis

1. Which dataset achieved the highest accuracy and why?
Answer: Mushrooms (100%) due to features being perfectly separate between edible and poisonous.

2. How does dataset size affect performance?
Answer: Larger data sets lead to more complex trees but accuracy stays the same due to more examples.

3. What role does the number of features play?
Answer: Multivalued categorical features in Nursery led to deeper, much larger trees as compared to binary features in tic-tac-toe

4. How does class imbalance affect tree construction?
Answer: It affected Nursery slightly (macro F1 < weighed F1).

5. Which feature types (binary vs multi-valued) work better?
Answer: Binary features lead to more ambiguity and lesser accuracy whereas multivalued categorical features lead to stronger splits and better classification.

Practical Applications

For which real-world scenarios is each dataset type most relevant? What are the interpretability advantages for each domain?

Answer:

Mushrooms -> Food safety, Bioclassification. It's interpretable and reliable since the tree is small.

Nursery -> Decision support. Large trees make interpretation harder but accuracy is high.

Tic-tac-toe -> Game AI or strategy learning. Accuracy < 90% shows decision trees might not capture optimal strategies perfectly.

Improvements

How would you improve performance for each dataset?
Answer: Pruning could be done to Nursery to reduce complexity without losing much accuracy. Feature engineering could be done for tic-tac-toe