LAB 3 - DECISION TREES - REPORT

Name: Meghana Saisri Bisa SRN: PES2UG23CS337 Semester and Section: 4F

- 1. Datasets
- mushrooms.csv: Predict edible (0) vs poisonous (1) mushrooms based on categorical attributes like odour, colour etc.,
- Nursery.csv: Predict nursery school admission recommendation among 5 classes, based on family/social attributes.
- tictactoe.csv: Predict is a board state is a win (1) or not (0). Features are the 9 board positions.

2. Performance Metrics

Dataset	Accuracy	Precision	Recall	F1
		(Weighted)	(Weighted)	(Weighted)
mushrooms.csv	1.0	1.0	1.0	1.0
Nursery.csv	0.98	0.98	0.98	0.98
tictactoe.csv	0.88	0.88	0.88	0.88

Observations:

- mushrooms.csv: Perfect scores (1.0) means some features almost completely determine whether a mushroom is poisonous or edible.
- Nursery.csv: Very high performance because it is a larger dataset with 5 classes making most decisions easy and also demonstrates ID3 algorithm handles multi-class problems well.
- tictactoe.csv: Lower performance compared to the others because of the complex interactions between board positions. A simple decision tree cannot capture all the winning conditions perfectly.

3. Tree Characteristics

Dataset	Maximum Depth	Total Nodes	Leaf Nodes	Internal Nodes
mushrooms.csv	4	29	24	5
Nursery.csv	7	983	703	280
tictactoe.csv	7	260	165	95

Observations:

- Mushroom tree is shallow because of dominant features.
- Nursery tree is largest due to multiple target classes and many categorical attributes with high cardinality.
- Tic-tac-toe tree is large since winning depends on complex interactions across multiple board positions.

- 4. Dataset Specific Insights
- mushrooms.csv: Strong single attributes (odour, colour) almost perfect separate classes and therefore high interpretability.
- Nursery.csv: Many attributes interact; imbalanced classes affect splits.
- tictactoe.csv: Features are symmetric, the root node often focuses on the center position since it most decisive in gameplay.

5. Comparative Analysis

- Best performing dataset: Mushroom achieved near perfect accuracy due to highly discriminative attributes.
- Dataset size affects:
 - Larger datasets (Nursery.csv, mushrooms.csv) provided more training examples which helped the tree generalise better.
 - Tic-tac-toe is smaller but still required maximum depth is 7 to capture positional dependencies resulting in lower accuracy despite the depth.
- Number of features:
 - More categorical levels (Nursery.csv, mushrooms.csv) produced larger trees.
 - Tic-tac-toe has fewer features but interdependencies still require depth.
- Class Imbalance: Nursery is affected by imbalance, Mushroom is fairly balanced and tic-tac-toe is binary so less affected.
- 6. Practical Applications
- Mushroom: Food safety classification highly interpretable and reliable for edible vs poisonous prediction.
- Nursery: Decision support for school admissions interpretable but mist be handled carefully due to bias from imbalanced data.
- Tic-Tac-Toe: Demonstrates game state evaluation useful in teaching rule-based AI and strategy learning.

7. Improvements

- Apply pruning to reduce overfitting and simplify complex trees.
- Use ensemble methods (Random Forest, Gradient Boosted Trees) to improve accuracy on complex datasets.
- Handle class imbalance using re-sampling or weighted loss functions, especially for Nursery.

8. Most Important Features

- Mushroom Dataset
 - Root Node: Odour almost perfectly separates edible vs poisonous mushrooms.
 - Early Splits: Spore-print-colour and Gill-size refine classification further.

Nursery Dataset

- o Root Node: Parents strongly affects admission recommendation.
- Early Splits: Has_nurs and Finance influence splits significantly, reflecting socio-economic factors.

Tic-Tac-Toe Dataset

- Root Node: Middle-middle (center position) the most decisive move in gameplay.
- Early Splits: Top-left and Bottom-right commonly appear, since corner positions are strategic.

9. Overfitting Indicators

- Mushroom Dataset
 - o Tree is shallow (depth 4) with only 29 total nodes.
 - o Clear feature dominance (odour) means no major overfitting risk.

Nursery Dataset

- Very large tree (983 nodes, depth 7) indicates overfitting tendencies.
- Class imbalance and high cardinality categorical attributes contribute to complex splits.

Tic-Tac-Toe Datset

- o Tree has 260 nodes with depth 7.
- Some branches repeat symmetric patterns (redundant splits), showing mid overfitting.

10. Decision Patterns

Mushroom Dataset

- If $odour = foul \rightarrow always poisonous$.
- o If odour = none → tree checks spore-print-color and gill-size to decide class.

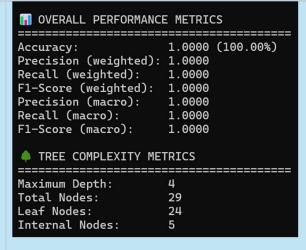
Nursery Dataset

- o If parents = usual and finance = convenient → often classified as "priority".
- o If has_nurs = very convenient → typically classified as "recommend".

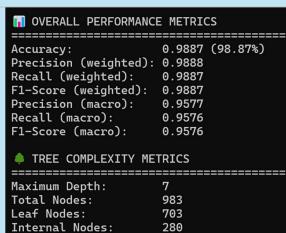
- Tic-Tac-Toe Dataset
 - If middle-middle = x AND top-left = x AND bottom-right = x → predicted win.
 - If $middle-middle = o \rightarrow typically classified as negative (loss/draw).$

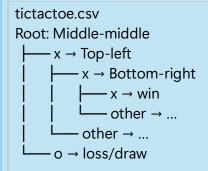
11. Tree Visualisation and Metrics

mushrooms.csv Root: Odour foul → poisonous none → Spore-print-color white → edible other → Gill-size



Nursery.csv Root: Parents — usual → Has_nurs — very convenient → recommend — convenient → Finance — convenient → priority — inconv → not_recom — pretentious → Finance ...





```
OVERALL PERFORMANCE METRICS
Accuracy:
                      0.8836 (88.36%)
Precision (weighted): 0.8827
Recall (weighted):
                      0.8836
F1-Score (weighted):
                      0.8822
Precision (macro):
                      0.8784
Recall (macro):
                      0.8600
F1-Score (macro):
                      0.8680
TREE COMPLEXITY METRICS
Maximum Depth:
Total Nodes:
                      260
Leaf Nodes:
                      165
```

95

Internal Nodes: