ML LAB WEEK 3

1. Performance Comparison:

Dataset	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)	Precision (macro)	Recall (macro)	F1 (macro)
Mushroom	1.0000 (100%)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Tic-Tac-Toe	0.8730 (87.30%)	0.8741	0.8730	0.8734	0.8590	0.8638	0.8613
Nursery	0.9867 (98.67%)	0.9876	0.9867	0.9872	0.7604	0.7654	0.7628

Insights:

- **Mushroom** dataset is perfectly classified (clean, categorical, strong signal from odour).
- **Tic-Tac-Toe** dataset has good accuracy (~87%), but not perfect (game logic is trickier, more overlap).
- **Nursery** dataset has very high weighted metrics (~99%), but macro metrics are much lower (~76%), meaning rare classes are misclassified (class imbalance).

2. Tree Characteristics Analysis:

Dataset	Max Depth	Total Nodes	Leaf Nodes	Internal Nodes	Notes
Mushroom	4	29	24	5	Very compact tree, interpretable
Tic-Tac-Toe	7	281	180	101	Deep tree, many nodes, complex paths
Nursery	7	952	680	272	Very large tree, complex due to multi-valued features

Insights:

- **Mushroom**: Simple, shallow tree → interpretable, efficient.
- **Tic-Tac-Toe**: Deep tree → many splits, complexity reflects game rules.
- **Nursery**: Large tree → multi-valued categorical features blow up node count.

3. Dataset-Specific Insights:

Mushroom Dataset

- **Feature Importance**: Odor dominates, followed by spore-print-colour.
- **Class Distribution**: Balanced (edible vs poisonous).
- **Decision Patterns**: If odour = foul → poisonous; odour = none + spore=black → edible.
- **Overfitting**: None (clean, perfectly separable). **Tic-Tac-Toe Dataset**
- **Feature Importance**: Middle-square and corners dominate splits.
- **Class Distribution**: Positive vs negative fairly balanced.
- **Decision Patterns**: If a row/column/diagonal is filled → positive (win).
- **Overfitting**: Some (tree memorizes many specific board states).

Nursery Dataset

- **Feature Importance**: Parents, finance, social factors dominate.
- **Class Distribution**: Imbalanced (majority = not_recom, minority = spec_prior/very_recom).
- **Decision Patterns**: Favourable finance/social → priority/recommend; otherwise not_recom.
- **Overfitting**: Yes, large branching due to multi-valued categorical features.

4. Comparative Analysis Report:

a) Algorithm Performance

- **Highest Accuracy**: Mushroom (100%) → clear, strong predictors.
- **Dataset Size Effect**: Nursery (largest dataset) → still high accuracy, but tree very large.
- **Number of Features**: More features/multi-valued → deeper, bushier trees (Nursery > Tic-Tac-Toe > Mushroom).

b) Data Characteristics Impact

- **Class Imbalance**: Hurts Nursery's macro metrics — rare classes not well predicted.
- **Feature Types**: Binary features (odour yes/no) → clean splits. Multi-valued (housing, finance) → complexity, weaker generalization.

c) Practical Applications

- **Mushroom**: Food safety, highly interpretable ("odour=foul → poisonous").
- **Tic-Tac-Toe**: Educational dataset, shows how trees can capture game rules.
- **Nursery**: Admission systems, but interpretability is lower due to tree size.

d) Improving Performance

- **Mushroom**: Already perfect.
- **Tic-Tac-Toe**: Apply pruning or max-depth constraints to avoid memorization.
- **Nursery**: Use ensemble methods (Random Forest), or collapse multi-valued features to simplify.

5. Summary:

- **Mushroom dataset**: ID3 achieves perfect classification (best case for decision trees).
- **Tic-Tac-Toe dataset:** good accuracy (~87%), but tree is complex and overfits some board states.
- **Nursery dataset**: excellent weighted accuracy (~99%), but macro metrics reveal class imbalance issues, with complex, less interpretable trees.

Name: Mrigank Jain

Semester: 5

Section: F

SRN: PES2UG23CS352