ML LAB WEEK 3

REPORT: PES2UG23CS387

TICTACTOE DATASET

```
📊 OVERALL PERFORMANCE METRICS
==========================================
Accuracy:                0.8730 (87.30%)
Precision (weighted):    0.8741
Recall (weighted):       0.8730
F1-Score (weighted):     0.8734
Precision (macro):       0.8590
Recall (macro):          0.8638
F1-Score (macro):        0.8613

🌳 TREE COMPLEXITY METRICS
==========================================
Maximum Depth:           7
Total Nodes:             281
Leaf Nodes:              180
Internal Nodes:          101
```

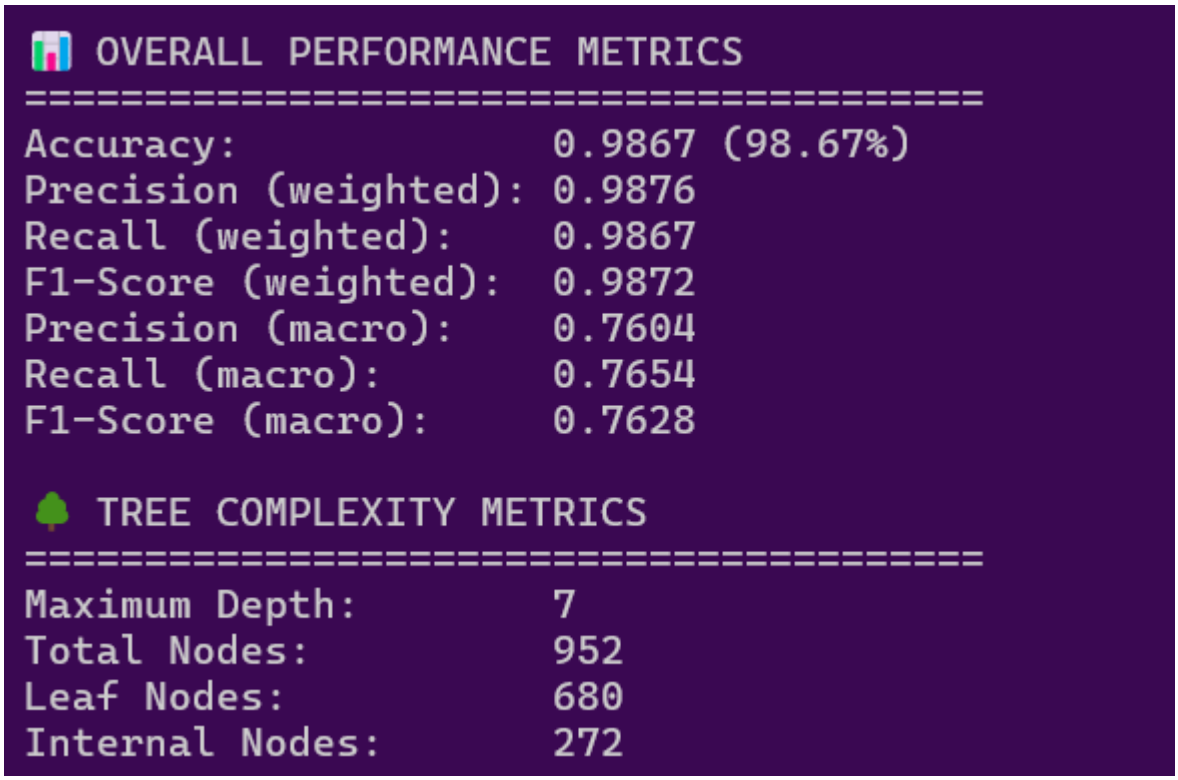
Roots and early splits are usually the middle squares because they decide winning conditions.

Class distribution is balanced between win (positive) and loss(negative)

Decision patterns: Tree first checks the critical winning lines

Overfitting indicators: Highly complex tree but still generalizes (87% accuracy)

NURSERY DATASET



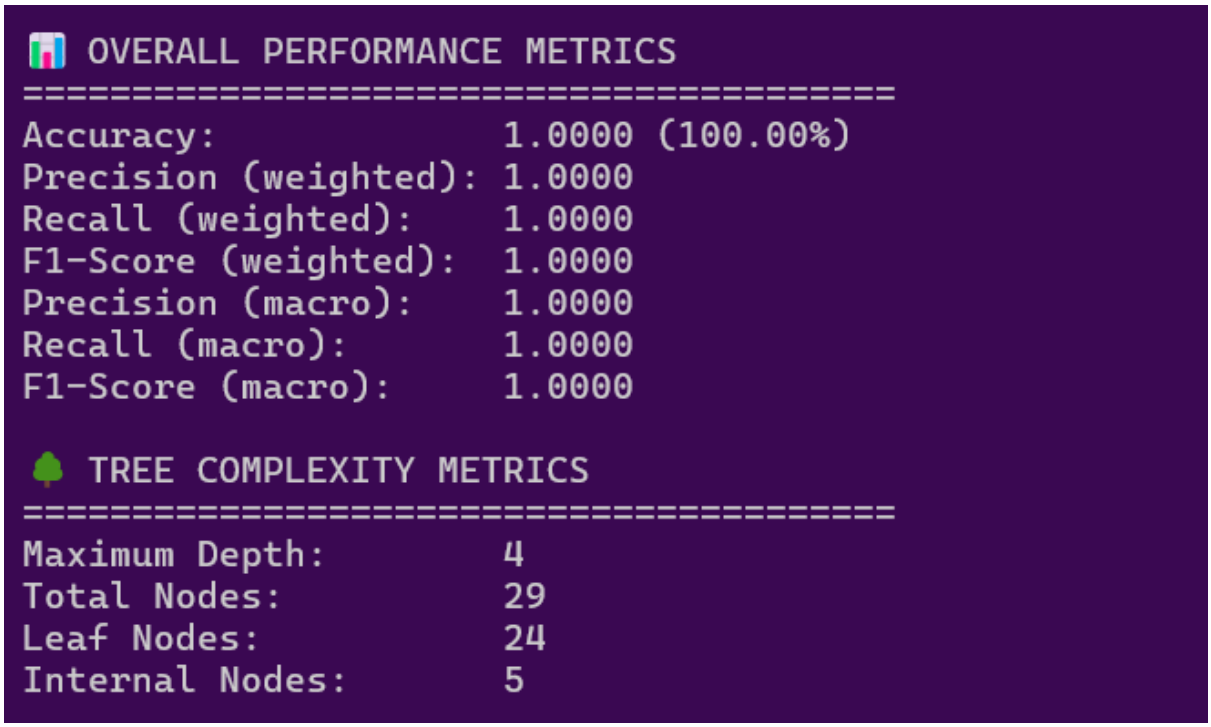
Root splits have attributes like *parents*, *has_nurs*

Class Distribution is very imbalanced because most are **not_recom**, while the other values are rare.

Decision Patterns: Longer and more complex because of many classes and imbalance of values.

Overfitting indicators: Accuracy very high, but recall is lower meaning tree struggles with minority classes. Therefore, it is *overfitting* to the **not_recom** class. (Deeper branches to cover rare cases)

MUSHROOM DATASET



Odor and spore-print-colour dominate the root and early splits.

Class Distribution: Fairly balanced

Decision Patterns: Very simple rules, single feature splits

Overfitting indicators: Achieves 100% accuracy, not overfitting because it is seperable by a few categorical features.

Algorithm Performance:

Highest Accuracy: Mushroom dataset (100%) because features like odor are perfect predictors.

Effect of Dataset Size: Larger datasets (Nursery: 12960, Mushroom: 8124) are more stable and accurate models. Smaller dataset (Tictactoe: 958) leads to lower accuracy, more variance.

Role of Features: More features and multi-valued attributes (Nursery, Mushroom) give richer splits. Tictactoe has only 9 simple attributes, so tree is more complex to capture patterns.

Data Characteristics Impact

Class Imbalance: Nursery has imbalance because tree tends to predict not_recom often; affects macro precision/recall because of bias in predictions. Mushrooms is mostly balanced, so no bias.

Binary vs Multi-valued: Multi-valued categorical features (Mushroom, Nursery) help trees split cleanly and improve accuracy. Binary (Tictactoe outcomes) makes patterns harder to learn.

Practical Applications

Tictactoe: Game AI, strategy modelling, rule-based systems. Advantage: interpretable strategies (which moves lead to wins).

Nursery: Admission or resource allocation systems. Advantage: clear decision rules for fairness/priority.

Mushroom: Food safety and toxicology. Advantage: easily interpretable Eg. If odor = foul, poisonous.

Improving Performance:

Tictactoe: Use pruning or Random Forests to reduce overfitting and capture general patterns.

Nursery: Handle class imbalance (resampling, weighted loss).

Mushroom: Already perfect no improvements needed

CONCLUSION

Mushrooms dataset was **clean** and **separable**

simple, perfect model

Nursery dataset was **large** and **imbalanced**

very accurate but careful handling of minority classes

Tictactoe dataset was **small** and **binary**

very complex tree lower accuracy