MACHINE LEARNING
LAB 3

NAME: DIYA CHANDRASHEKHAR
SRN: PES2UG23CS182
SEM: 4
SECTION: C

Mushrooms.csv

```
📊 OVERALL PERFORMANCE METRICS
========================================
Accuracy:             1.0000 (100.00%)
Precision (weighted): 1.0000
Recall (weighted):    1.0000
F1-Score (weighted):  1.0000
Precision (macro):    1.0000
Recall (macro):       1.0000
F1-Score (macro):     1.0000

🌳 TREE COMPLEXITY METRICS
========================================
Maximum Depth:        4
Total Nodes:          29
Leaf Nodes:           24
Internal Nodes:       5
```

Tictactoe.csv

```
📊 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
```

Nursery.csv

```
📊 OVERALL PERFORMANCE METRICS
========================================
Accuracy:              0.9867 (98.67%)
Precision (weighted): 0.9876
Recall (weighted):    0.9867
F1-Score (weighted):  0.9872
Precision (macro):    0.7604
Recall (macro):       0.7654
F1-Score (macro):     0.7628

🌳 TREE COMPLEXITY METRICS
========================================
Maximum Depth:         7
Total Nodes:           952
Leaf Nodes:            680
Internal Nodes:        272
```

a)

a. Which dataset achieved the highest accuracy and why?

The mushroom.csv dataset has the highest accuracy with 100%. There are categorical features in the mushroom dataset (like cap shape, colour, odour, etc.) that are very good at predicting class (edible vs. poisonous). The decision tree split rules are very effective and clear because many features have a direct correlation with the class.

b. How does dataset size affect performance?

Decision trees become deeper and more complicated with more datasets, but performance relies on how balanced and predictive the features are.

c. What role does the number of features play?

Trees with more features have more split options, which increases their predictive power but also increases complexity.

b) How does class imbalance affect tree construction?

As seen in the nursery dataset, imbalance causes the model to favour the majority classes while ignoring the minority.

Which types of features (binary vs multi-valued) work better?

Multi-valued categorical features work better for decision trees because they capture stronger signals in fewer splits.

c) For which real-world scenarios is each dataset type most relevant?
Mushroom dataset: bioinformatics, agriculture, food safety.
Tic-Tac-Toe dataset: game AI research and strategy learning.
Nursery dataset: admission/placement systems.

What are the interpretability advantages for each domain?
The mushroom dataset is very interpretable and easy to understand.
The tictactoe dataset is hard to interpret since the tree needs multiple splits, hence its less intuitive
The nursery dataset is complex but still provides clear rules helping interpretability.

How would you improve performance of each dataset?
Mushroom.csv already has a 100% accuracy, pruning can be used to make the tree smaller
Ensambles can be used in nursery and tictactoe