# Machine Learning Lab

| **Name** - G S S Surya Prakash | **SRN** - PES2UG23CS192 | **Class** - 5 'C' | **Topic** - Decision Tree Classifier |
|---|---|---|---|

## Lab - 3

## ScreenShots of the code execution for each CSV file:

a. tictactoe.csv

```
PS C:\Users\gonel\OneDrive\Desktop\Sem5\ML\code\pytorch_implementation> python test.py --ID EC_C_PES2UG23CS192_Lab3 --data tictactoe.csv
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!
```

```
📊 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
```

```
🌲 DECISION TREE STRUCTURE
============================================================
Root [middle-middle-square] (gain: 0.0834)
├── = 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: 0.6058)
│   │       │   ├── = 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
│   │           │   ├── = 1:
│   │           │   │   ├── Class 1
│   │           │   └── = 2:
│   │           │       ├── Class 0
│   │           ├── = 1:
│   │           │   ├── [middle-left-square] (gain: 0.9183)
│   │           │   ├── = 0:
│   │           │   │   ├── Class 1
│   │           │   ├── = 1:
│   │           │   │   ├── Class 1
│   │           │   └── = 2:
│   │           │       ├── Class 0
│   │           └── = 2:
│   │               ├── Class 1
```

b. nursery.csv

```
PS C:\Users\gonel\OneDrive\Desktop\Sem5\ML\code\pytorch_implementation> python test.py --ID EC_C_PES2UG23CS192_Lab3 --data nursery.csv
Running tests with PYTORCH framework
============================================================
 target column: 'class' (last column)
Original dataset info:
Shape: (12960, 9)
Columns: ['parents', 'has_nurs', 'form', 'children', 'housing', 'finance', 'social', 'health', 'class']

First few rows:

parents: ['usual' 'pretentious' 'great_pret'] -> [2 1 0]

has_nurs: ['proper' 'less_proper' 'improper' 'critical' 'very_crit'] -> [3 2 1 0 4]

form: ['complete' 'completed' 'incomplete' 'foster'] -> [0 1 3 2]

class: ['recommend' 'priority' 'not_recom' 'very_recom' 'spec_prior'] -> [2 1 0 4 3]

Processed dataset shape: torch.Size([12960, 9])
Number of features: 8
Features: ['parents', 'has_nurs', 'form', 'children', 'housing', 'finance', 'social', 'health']
Target: class
Framework: PYTORCH
Data type: <class 'torch.Tensor'>

============================================================
DECISION TREE CONSTRUCTION DEMO
============================================================
Total samples: 12960
Training samples: 10368
Testing samples: 2592

Constructing decision tree using training data...

🌳 Decision tree construction completed using PYTORCH!
```

```
📊 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
```

```
│   │   │   │   │   │   ├── Class 1
│   │   │   │   │   └── = 1:
│   │   │   │   │       ├── [form] (gain: 0.3555)
│   │   │   │   │       ├── = 0:
│   │   │   │   │       │   ├── Class 3
│   │   │   │   │       ├── = 1:
│   │   │   │   │       │   ├── Class 1
│   │   │   │   │       ├── = 2:
│   │   │   │   │       │   ├── Class 3
│   │   │   │   │       └── = 3:
│   │   │   │   │           ├── Class 3
│   │   │   │   ├── = 1:
│   │   │   │   │   ├── [form] (gain: 0.1011)
│   │   │   │   │   ├── = 0:
│   │   │   │   │   │   ├── [children] (gain: 0.7219)
│   │   │   │   │   │   ├── = 0:
│   │   │   │   │   │   │   ├── Class 1
│   │   │   │   │   │   ├── = 1:
│   │   │   │   │   │   │   ├── Class 3
│   │   │   │   │   │   ├── = 2:
│   │   │   │   │   │   │   ├── Class 3
│   │   │   │   │   │   └── = 3:
│   │   │   │   │   │       ├── Class 3
│   │   │   │   │   ├── = 1:
│   │   │   │   │   │   ├── Class 3
│   │   │   │   │   ├── = 2:
│   │   │   │   │   │   ├── Class 3
│   │   │   │   │   └── = 3:
│   │   │   │   │       ├── Class 3
│   │   │   │   └── = 2:
│   │   │   │       ├── [children] (gain: 0.5044)
│   │   │   │       ├── = 0:
│   │   │   │       │   ├── [form] (gain: 0.8113)
│   │   │   │       │   ├── = 0:
│   │   │   │       │   │   ├── Class 1
│   │   │   │       │   ├── = 1:
│   │   │   │       │   │   ├── Class 1
│   │   │   │       │   └── = 2:
│   │   │   │       │       ├── Class 3
│   │   │   │       ├── = 1:
│   │   │   │       │   ├── [form] (gain: 0.9183)
│   │   │   │       │   ├── = 0:
```

c. mushrooms.csv

```
PS C:\Users\gonel\OneDrive\Desktop\Sem5\ML\code\pytorch_implementation> python test.py --ID EC_C_PES2UG23CS192_Lab3 --data mushrooms.csv
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-typ
e', '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 8 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-ty
pe', '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!
```

```
📊 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
```

```
🌲 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
```

## a) Algorithm Performance

Qn. Which dataset achieved the highest accuracy and why?
Ans: The Mushroom dataset achieves the highest accuracy as it has a well-defined categorical features directly correlated with the binary target. This is because the features are discriminative and less noisy, enabling clear splits that reduce entropy.

Qn. How does dataset size affect performance?
Ans: Larger the datasets, more information for the decision tree to learn robust decision boundaries, improving accuracy and reducing overfitting. Smaller datasets (like nursery or tic-tac-toe) risk underfitting or overfitting depending on feature relevance and data variability.

Qn. What role does the number of features play?
Ans: More features can increase the model's capacity to learn complex patterns if relevant, but can also cause overfitting with irrelevant or redundant features. Features which provide high information gain contribute to better splits and higher accuracy.

## b) Data Characteristics Impact

Qn. How does class imbalance affect tree construction?
Ans: Class imbalance can bias the tree towards majority classes, leading to skewed splits and poor minority class recognition. Metrics like weighted precision/recall and pruning techniques can help counteract imbalance effects.

Qn. Which types of features (binary vs multi-valued) work better?
Ans: Binary features often create simpler, clearer splits, improving interpretability and reducing overfitting. Multi-valued categorical features offer richer information but increase tree complexity and may cause more fragmented splits, needing careful handling.

## c) Practical Applications

Qn. For which real-world scenarios is each dataset type most relevant?
Ans:

- Mushroom: Food safety and biological classification where clear binary categorization is critical.
- Tic-tac-toe: Game outcome prediction or small strategic decision tasks with limited discrete states.

- Nursery: Social and family-based recommendation systems with multiple categories reflecting real-world complexity.

Qn. What are the interpretability advantages for each domain?
Ans: The mushroom dataset's binary classification and features provide straightforward explanations for decisions (safe vs poisonous). The tic-tac-toe dataset's spatial game states map intuitively to tree splits, aiding understanding. The nursery dataset, with multi-class and more complex categorical factors, offers richer but more complex interpretability challenges.

Qn. How to Improve Performance for Each Dataset
Ans:

- Mushroom: Consider feature selection focusing on top information gain attributes to simplify the tree and implement pruning to reduce overfitting.
- Tic-tac-toe: Use cross-validation to tune tree depth to prevent overfitting.
- Nursery: Handle class imbalance via resampling or weighted splitting criteria; consider hierarchical tree structures due to multi-class target and complex features.