# HOMEWORK GINI IMPURITY

## Homework and Lab – Part 1

| Student | Study Hours | Pass (1)/Fail (0) |
|---|---|---|
| 1 | 2 | 0 |
| 2 | 4 | 0 |
| 3 | 6 | 1 |
| 4 | 8 | 1 |
| 5 | 10 | 1 |

You are provided with the following dataset. Predict whether a student will pass (1) or fail (0) a test based on their study hours.

**Part 1: By Hand (Pen and Paper)**

- Step 1: List all possible split points for Study Hours (between consecutive unique values: 3, 5, 7, 9).
- Step 2: For each possible split, compute the Gini impurity for the resulting groups ("Study Hours ≤ split", "Study Hours > split").
- Step 3: Calculate the weighted average Gini impurity for each split.
- Step 4: Choose the split with the lowest Gini impurity. Draw the first-level decision tree.
- Step 5: For any groups with impurity > 0, repeat Steps 1–4.

---

## Homework and Lab – Part 2

| Student | Study Hours | Pass (1)/Fail (0) |
|---|---|---|
| 1 | 2 | 0 |
| 2 | 4 | 0 |
| 3 | 6 | 1 |
| 4 | 8 | 1 |
| 5 | 10 | 1 |

Part 2: Python

- Implement the same process manually using Python (without using any machine learning libraries).
- Output the Gini impurity at each step, show chosen split, and visualize the final tree using text or simple graph.
- Optional: Cross-check your manual solution with sklearn's DecisionTreeClassifier using the same data.

Part A

| Student | Study hours | pass/fail |
|---|---|---|
| 1 | 2 | 0 |
| 2 | 4 | 0 |
| 3 | 6 | 1 |
| 4 | 8 | 1 |
| 5 | 10 | 1 |

3, 5, 7, 9

split at 3

Gini ($\leq 3$) fail = 0 → 1 person

Gini ($> 3$) 4 samples 1 fail, 3 pass

$P(\text{pass}) = \frac{3}{4}$ $P(\text{fail}) = \frac{1}{4}$

$\text{Gini} = 1 - ((3/4)^2 + (1/4)^2)$

$= 1 - (0.5625 + 0.0625)$

$= 1 - 0.625$

$= 0.375$

$\text{weighted gini} = \frac{1}{5} \times 0 + \frac{4}{5} \times 0.375$

$= 0.3$

split at 5

gini ($\leq 5$) 0.0 Both fail so impurity 0

Gini ($> 5$) 1, 1, 1 All pass so impurity 0

$\text{weighted gini} = \frac{2}{5} \times 0 + \frac{3}{5} \times 0$

$= 0$

split at 7

Gini impurity (≤7) 0, 0, 1

2 fail & 1 pass.

Gini fail = $\frac{2}{3}$ gini pass = $\frac{1}{3}$

Gini = $1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)$

= 0.444

Gini (>7) 1, 1 Both pass so zero.

= $\left(\frac{3}{5}\right) \times 0.444 + \left(\frac{2}{5}\right) \times 0$

= 0.2667

split at 9

gini (≤9) 0, 0, 1, 1

2 fail and 2 pass

gini fail = $\frac{2}{9}$ Gini pass = $\frac{2}{4}$

Gini = $1 - \left(\frac{1^2}{2^2} + \frac{1^2}{2^2}\right)$

Gini = 0.5

Gini>9 = 1 one sample so zero

= $\frac{4}{5} \times 0.5 + \frac{1}{5} \times 0$

= 0.4

Lowest split at 5 (Gini = 0)

Study hours ≤ 5

Yes NO

(≤5) : [0,0] (>5) : [1,1,1]

Fail Pass

0 1.

Part B

```python
# Data setup
students = [
    {'hours': 2, 'pass': 0},
    {'hours': 4, 'pass': 0},
    {'hours': 6, 'pass': 1},
    {'hours': 8, 'pass': 1},
    {'hours': 10, 'pass': 1}
]

# Possible split points between unique values
split_points = [3, 5, 7, 9]

def gini(groups):
    total = sum(len(group) for group in groups)
    score = 0.0
    for group in groups:
        size = len(group)
        if size == 0:
            continue
        proportion_pass = sum([row['pass'] for row in group]) / size
        proportion_fail = 1 - proportion_pass
        score += size / total * (1.0 - (proportion_pass ** 2 + proportion_fail ** 2))
    return score

for split in split_points:
    left = [row for row in students if row['hours'] <= split]
    right = [row for row in students if row['hours'] > split]
    impurity = gini([left, right])
    print(f"Split at {split}: Gini = {impurity:.4f}")

# Build the decision tree (first level)
print("\nBest split at Study Hours <= 5:\n")
print("If Study Hours <= 5 -> 'Fail' (0)")
print("If Study Hours > 5 -> 'Pass' (1)")
```

Output

```
Split at 3: Gini = 0.3000
Split at 5: Gini = 0.0000
Split at 7: Gini = 0.2667
Split at 9: Gini = 0.4000

Best split at Study Hours <= 5:

If Study Hours <= 5 -> 'Fail' (0)
If Study Hours > 5 -> 'Pass' (1)
```