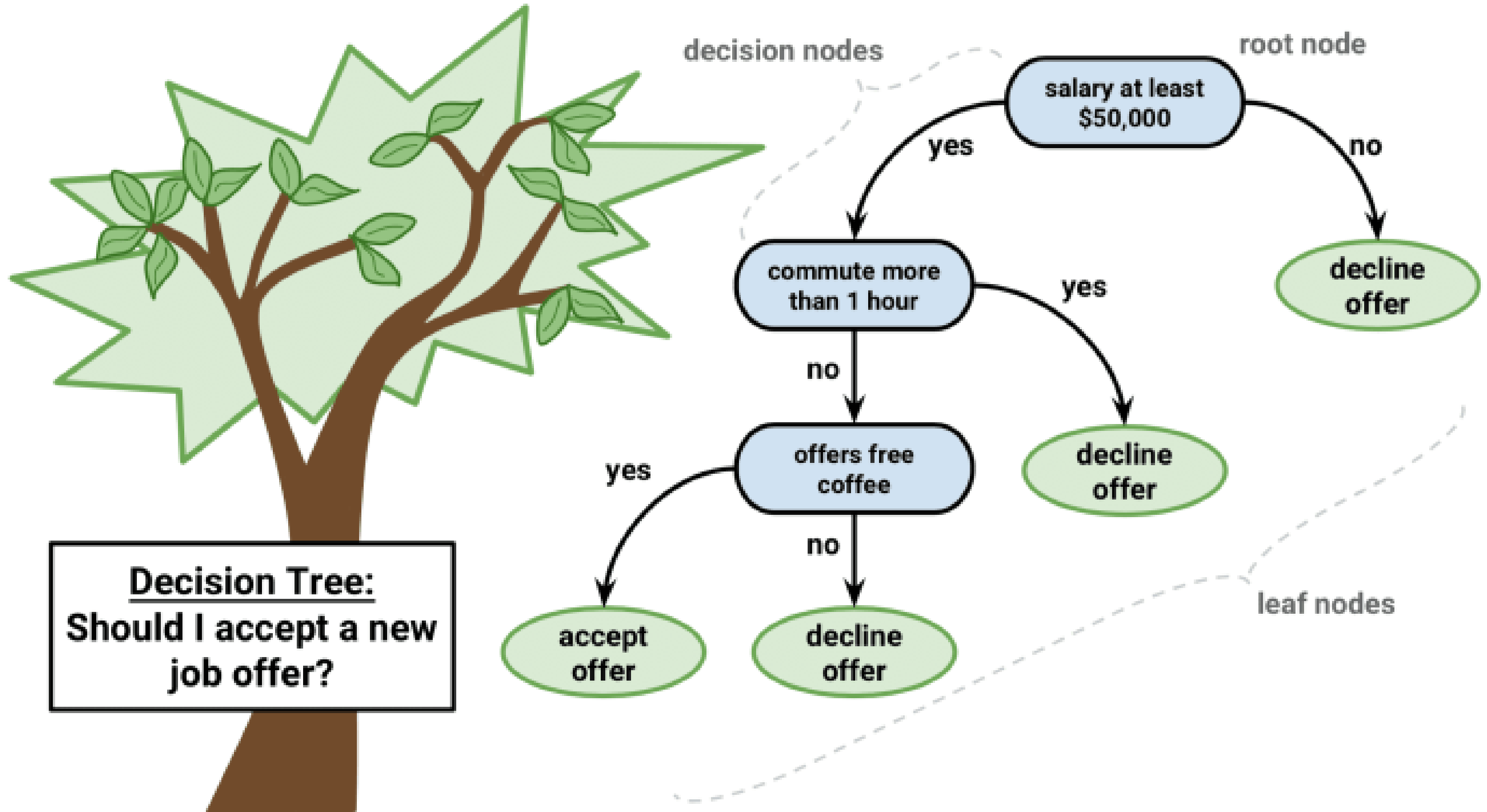# Decision Tree (ID3 Algorithm)
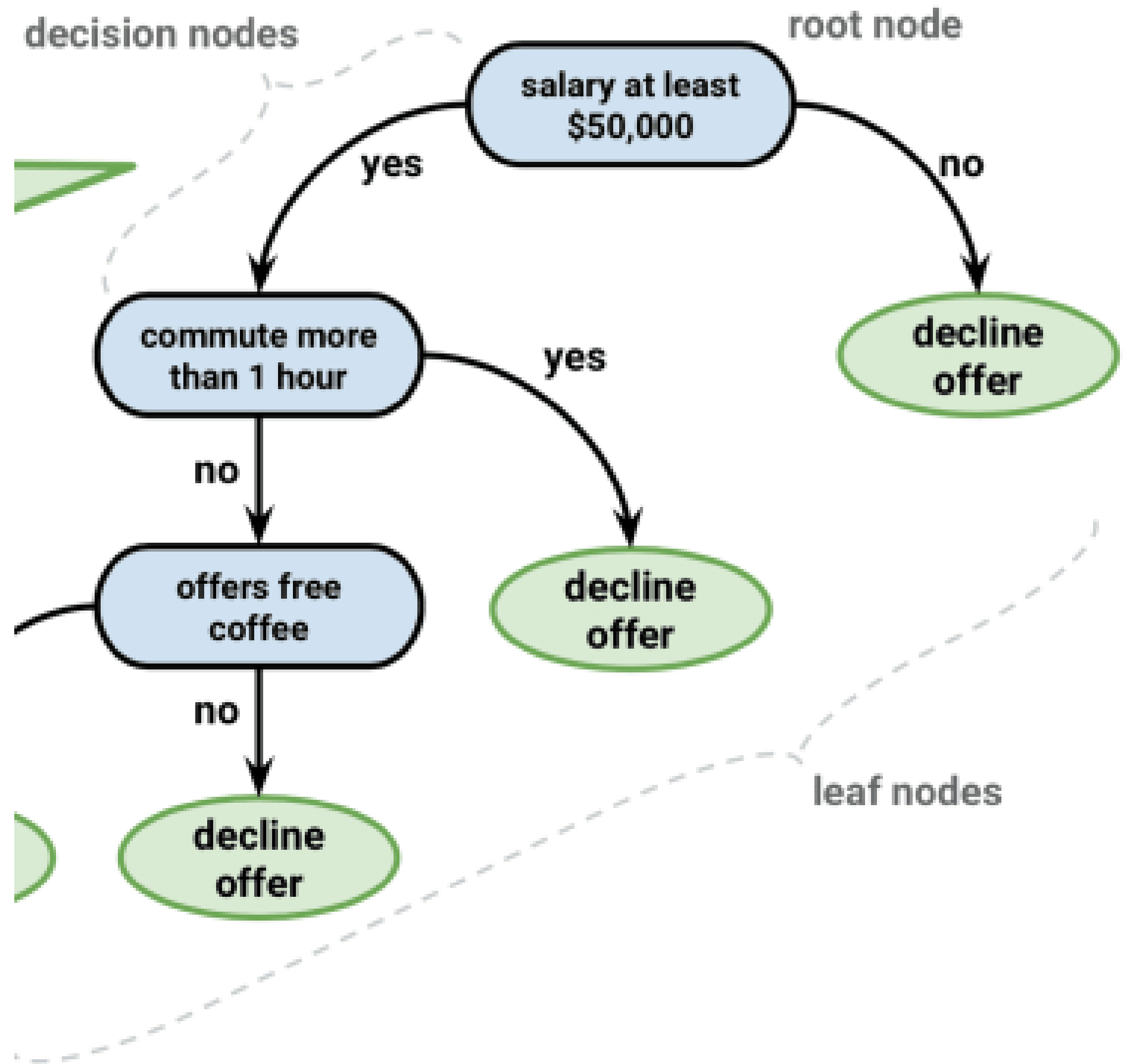
# (Numerical)

# DECISION TREE and ID3 Algorithm

| S. No. | Outlook | Temperature | Humidity | Windy | PlayTennis |
|--------|---------|-------------|----------|-------|------------|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | Yes |
| 6 | Rainy | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rainy | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rainy | Mild | High | Strong | No |

# Make a decision tree that predicts whether tennis will be played on the day?

decision nodes

root node

salary at least $50,000

yes

no

commute more than 1 hour

decline offer

yes

no

decline offer

offers free coffee

yes

no

Decision Tree: Should I accept a new job offer?

accept offer

decline offer

leaf nodes

# WHAT IS DECISON TREE?

A *Decision Tree* is a tree where *each node* represents a **Feature (Attribute)**, each link (branch) represents a **decision (rule)** and each leaf represents an **outcome**.



decision nodes

root node

salary at least $50,000

yes

no

commute more than 1 hour

yes

decline offer

no

decline offer

offers free coffee

no

decline offer

leaf nodes

# Algorithms

## Cart

- Gini Index

## ID3

- Entropy Function
- Information Gain

| S. No. | Outlook | Temperature | Humidity | Windy | PlayTennis |
|--------|---------|-------------|----------|-------|------------|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | Yes |
| 6 | Rainy | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rainy | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rainy | Mild | High | Strong | No |

# MAKE A DECISION TREE THAT PREDICTS WHETHER TENNIS WILL BE PLAYED ON THE DAY?

| S. No. | Outlook | Temperature | Humidity | Windy | PlayTennis |
|--------|---------|-------------|----------|-------|------------|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | Yes |
| 6 | Rainy | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rainy | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rainy | Mild | High | Strong | No |

# STEP 1: CREATE A ROOT NODE

- HOW TO CHOOSE THE ROOT NODE?

  The attribute that best classifies the training data, use this attribute at the root of the tree.

# STEP 1: CREATE A ROOT NODE

- HOW TO CHOOSE THE ROOT NODE?

  The attribute that best classifies the training data, use this attribute at the root of the tree.

- HOW TO CHOOSE THE BEST ATTRIBUTE?

  So from here, *ID3 algorithm* begins

- Calculate **Entropy** (Amount of uncertainity in dataset):

$$Entropy = \frac{-p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2 \left(\frac{n}{p+n}\right)$$

- Calculate **Average Information**:

$$I(Attribute) = \sum \frac{p_i + n_i}{p+n} Entropy(A)$$

- Calculate **Information Gain**: (Difference in Entropy before and after splitting dataset on attribute A)

$$Gain = Entropy(S) - I(Attribute)$$

1. COMPUTE THE **ENTROPY** FOR DATA-SET **ENTROPY(S)**

2. FOR EVERY ATTRIBUTE/FEATURE:

   1. CALCULATE ENTROPY FOR ALL OTHER VALUES **ENTROPY(A)**

   2. TAKE **AVERAGE INFORMATION ENTROPY** FOR THE CURRENT ATTRIBUTE

   3. CALCULATE **GAIN** FOR THE CURRENT ATTRIBUTE

3. PICK THE **HIGHEST GAIN ATTRIBUTE.**

4. **REPEAT** UNTIL WE GET THE TREE WE DESIRED.

# 1.

| S. No. | Outlook | Temperature | Humidity | Windy | PlayTennis |
|--------|---------|-------------|----------|-------|------------|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rainy | Mild | High | Weak | Yes |
| 5 | Rainy | Cool | Normal | Weak | Yes |
| 6 | Rainy | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rainy | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rainy | Mild | High | Strong | No |

$P = 9$

$N = 5$

Total = 14

- Calculate **Entropy(S):**

$$Entropy = \frac{-p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2 \left(\frac{n}{p+n}\right)$$

$$Entropy(S) = \frac{-9}{9+5} log_2 \left(\frac{9}{9+5}\right) - \frac{5}{9+5} log_2 \left(\frac{5}{9+5}\right)$$

$$Entropy(S) = \frac{-9}{14} log_2 \left(\frac{9}{14}\right) - \frac{5}{14} log_2 \left(\frac{5}{14}\right) \quad = 0.940$$

- For each Attribute: (let say **Outlook**)

  - Calculate Entropy for each Values, i.e for 'Sunny', 'Rainy','Overcast'

| Outlook | PlayTennis |
|---------|------------|
| Sunny | No |
| Sunny | No |
| Sunny | No |
| Sunny | Yes |
| Sunny | Yes |

| Outlook | PlayTennis |
|---------|------------|
| Rainy | Yes |
| Rainy | Yes |
| Rainy | No |
| Rainy | Yes |
| Rainy | No |

| Outlook | PlayTennis |
|---------|------------|
| Overcast | Yes |
| Overcast | Yes |
| Overcast | Yes |
| Overcast | Yes |

| Outlook | p | n | Entropy |
|---------|---|---|---------|
| Sunny | 2 | 3 | 0.971 |
| Rainy | 3 | 2 | 0.971 |
| Overcast | 4 | 0 | 0 |

Calculate **Entropy(Outlook='Value'):**

$$Entropy = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

E (Outlook=sunny) = $-\frac{2}{5} \log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$

E (Outlook=overcast) = $-1 \log(1) - 0 \log(0) = 0$

E (Outlook=rainy) = $-\frac{3}{5} \log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$

- Calculate **Average Information Entropy**:

$$I(Outlook) = \frac{p_{sunny} + n_{sunny}}{p + n} Entropy(Outlook = Sunny)+$$

$$\frac{p_{rainy} + n_{rainy}}{p + n} Entropy(Outlook = Rainy)+$$

$$\frac{p_{Overcast} + n_{Overcast}}{p + n} Entropy(Outlook = Overcast)$$

$$I(Outlook) = \frac{3 + 2}{9 + 5} * 0.971 + \frac{2 + 3}{9 + 5} * 0.971 + \frac{4 + 0}{9 + 5} * 0 \quad = 0.693$$

- Calculate **Gain**: attribute is Outlook

$$Gain = Entropy(S) - I(Attribute)$$

$$Entropy(S) = 0.940$$

$$Gain(Outlook) = 0.940 - 0.693 = 0.247$$

- For each Attribute: (let say **Temperature**)
  - Calculate Entropy for each Temp, i.e for 'Hot', 'Mild' and 'Cool'

| Temperature | PlayTennis |
|---|---|
| Hot | No |
| Hot | No |
| Hot | Yes |
| Hot | Yes |

| Temperature | PlayTennis |
|---|---|
| Mild | Yes |
| Mild | No |
| Mild | Yes |
| Mild | Yes |
| Mild | Yes |
| Mild | No |

| Temperature | PlayTennis |
|---|---|
| Cool | Yes |
| Cool | No |
| Cool | Yes |
| Cool | Yes |

| Temperature | p | n | Entropy |
|---|---|---|---|
| Hot | 2 | 2 | 1 |
| Mild | 4 | 2 | 0.918 |
| Cool | 3 | 1 | 0.811 |

- Calculate **Average Information Entropy**:

$$I(Temperature) = \frac{p_{hot} + n_{hot}}{p + n} Entropy(Temperature = Hot) +$$

$$\frac{p_{mild} + n_{mild}}{p + n} Entropy(Temperature = Mild) +$$

$$\frac{p_{Cool} + n_{Cool}}{p + n} Entropy(Temperature = Cool)$$

$$I(Temperature) = \frac{2 + 2}{9 + 5} * 1 + \frac{4 + 2}{9 + 5} * 0.918 + \frac{3 + 1}{9 + 5} * 0.811 => 0.911$$

- Calculate **Gain**: attribute is Temperature

$$Gain = Entropy(S) - I(Attribute)$$

$$Entropy(S) = 0.940$$

$$Gain(Temperature) = 0.940 - 0.911 = 0.029$$

- For each Attribute: (let say **Humidity**)

  - Calculate Entropy for each Humidity, i.e for 'High', 'Normal'

| Humidity | PlayTennis |
|----------|------------|
| Normal   | Yes        |
| Normal   | No         |
| Normal   | Yes        |
| Normal   | Yes        |
| Normal   | Yes        |
| Normal   | Yes        |
| Normal   | Yes        |

| Humidity | PlayTennis |
|----------|------------|
| High     | No         |
| High     | No         |
| High     | Yes        |
| High     | Yes        |
| High     | No         |
| High     | Yes        |
| High     | No         |

| Humidity | p | n | Entropy |
|----------|---|---|---------|
| High     | 3 | 4 | 0.985   |
| Normal   | 6 | 1 | 0.591   |

- Calculate **Average Information Entropy**:

$$I(Humidity) = \frac{p_{High} + n_{High}}{p + n} Entropy(Humidity = High) +$$

$$\frac{p_{Normal} + n_{Normal}}{p + n} Entropy(Humidity = Normal)$$

$$I(Humidity) = \frac{3 + 4}{9 + 5} * 0.985 + \frac{6 + 1}{9 + 5} * 0.591 => 0.788$$

- Calculate **Gain**: attribute is Humidity

$$Gain = Entropy(S) - I(Attribute)$$

$$Entropy(S) = 0.940$$

$$Gain(Humidity) = 0.940 - 0.788 = 0.152$$

- For each Attribute: (let say **Windy**)
  - Calculate Entropy for each Windy, i.e for 'Strong' and 'Weak'

| Windy | PlayTennis |
|-------|-----------|
| Weak | No |
| Weak | Yes |
| Weak | Yes |
| Weak | Yes |
| Weak | No |
| Weak | Yes |
| Weak | Yes |
| Weak | Yes |

| Windy | PlayTennis |
|-------|-----------|
| Strong | No |
| Strong | No |
| Strong | Yes |
| Strong | Yes |
| Strong | Yes |
| Strong | No |

| Windy | p | n | Entropy |
|-------|---|---|---------|
| Strong | 3 | 3 | 1 |
| Weak | 6 | 2 | 0.811 |

- Calculate **Average Information Entropy**:

$$I(Windy) = \frac{p_{Strong} + n_{Strong}}{p + n} Entropy(Windy = Strong) +$$

$$\frac{p_{Weak} + n_{Weak}}{p + n} Entropy(Windy = Weak)$$

$$I(Windy) = \frac{3 + 3}{9 + 5} * 1 + \frac{6 + 2}{9 + 5} * 0.811 => 0.892$$

- Calculate **Gain**: attribute is Windy

$$Gain = Entropy(S) - I(Attribute)$$

$$Entropy(S) = 0.940$$

$$Gain(Windy) = 0.940 - 0.892 = 0.048$$
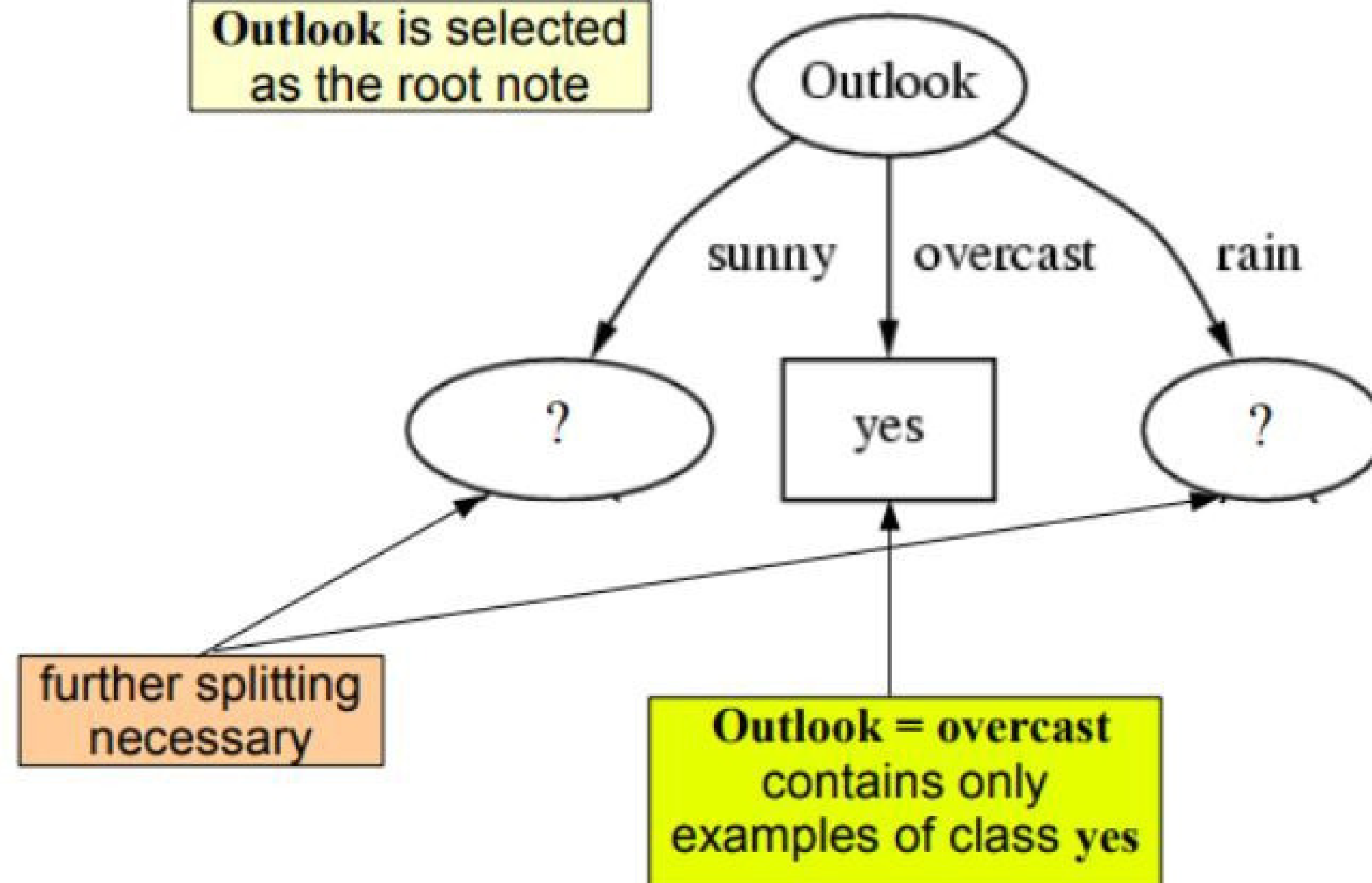
- PICK THE HIGHEST GAIN ATTRIBUTE.

| Attributes | Gain |
|---|---|
| Outlook | 0.247 |
| Temperature | 0.029 |
| Humidity | 0.152 |
| Windy | 0.048 |

ROOT NODE:

**OUTLOOK**

| Outlook | Temperature | Humidity | Windy | PlayTennis |
|---------|-------------|----------|-------|------------|
| Overcast | Hot | High | Weak | Yes |
| Overcast | Cool | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |

**Outlook** is selected as the root note

Outlook

sunny    overcast    rain

?    yes    ?

further splitting necessary

**Outlook = overcast** contains only examples of class **yes**

- Repeat the same thing for sub-trees till we get the tree.

| Outlook | Temperature | Humidity | Windy | PlayTennis |
|---------|-------------|----------|-------|------------|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |

**OUTLOOK = "SUNNY"**

| Outlook | Temperature | Humidity | Windy | PlayTennis |
|---------|-------------|----------|-------|------------|
| Rainy | Mild | High | Weak | Yes |
| Rainy | Cool | Normal | Weak | Yes |
| Rainy | Cool | Normal | Strong | No |
| Rainy | Mild | Normal | Weak | Yes |
| Rainy | Mild | High | Strong | No |

**OUTLOOK = "RAINY"**

| Outlook | Temperature | Humidity | Windy | PlayTennis |
|---------|-------------|----------|-------|------------|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |

P=          N=

2          3

Total=

5

- ENTROPY:

$$Entropy = \frac{-p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2 \left(\frac{n}{p+n}\right)$$

$$Entropy(S_{sunny}) = \frac{-2}{2+3} log_2 \left(\frac{2}{2+3}\right) - \frac{3}{2+3} log_2 \left(\frac{3}{2+3}\right)$$

=>0.971

- For each Attribute: (let say **Humidity**):

  - Calculate Entropy for each Humidity, i.e for 'High' and 'Normal'

| Outlook | Humidity | PlayTennis |
|---------|----------|------------|
| Sunny | High | No |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | Normal | Yes |
| Sunny | Normal | Yes |

| Humidity | p | n | Entropy |
|----------|---|---|---------|
| high | 0 | 3 | 0 |
| normal | 2 | 0 | 0 |

- Calculate **Average Information Entropy**:    I(Humidity) = 0

- Calculate **Gain:**                                      Gain = 0.971

- For each Attribute: (let say **Windy**):

  - Calculate Entropy for each Windy, i.e for 'Strong' and 'Weak'

| Outlook | Windy | PlayTennis |
|---------|-------|------------|
| Sunny | Strong | No |
| Sunny | Strong | Yes |
| Sunny | Weak | No |
| Sunny | Weak | No |
| Sunny | Weak | Yes |

| Windy | p | n | Entropy |
|-------|---|---|---------|
| Strong | 1 | 1 | 1 |
| Weak | 1 | 2 | 0.918 |

- Calculate **Average Information Entropy**:     I(Windy) = 0.951

- Calculate **Gain:**     Gain = 0.020

- For each Attribute: (let say **Temperature**):

  - Calculate Entropy for each Windy, i.e for 'Cool', 'Hot' and 'Mild'

| Outlook | Temperature | PlayTennis |
|---------|-------------|------------|
| Sunny | Cool | Yes |
| Sunny | Hot | No |
| Sunny | Hot | No |
| Sunny | Mild | No |
| Sunny | Mild | Yes |

| Temperature | p | n | Entropy |
|-------------|---|---|---------|
| Cool | 1 | 0 | 0 |
| Hot | 0 | 2 | 0 |
| Mild | 1 | 1 | 1 |

- Calculate **Average Information Entropy**:    I(Temp) = 0.4
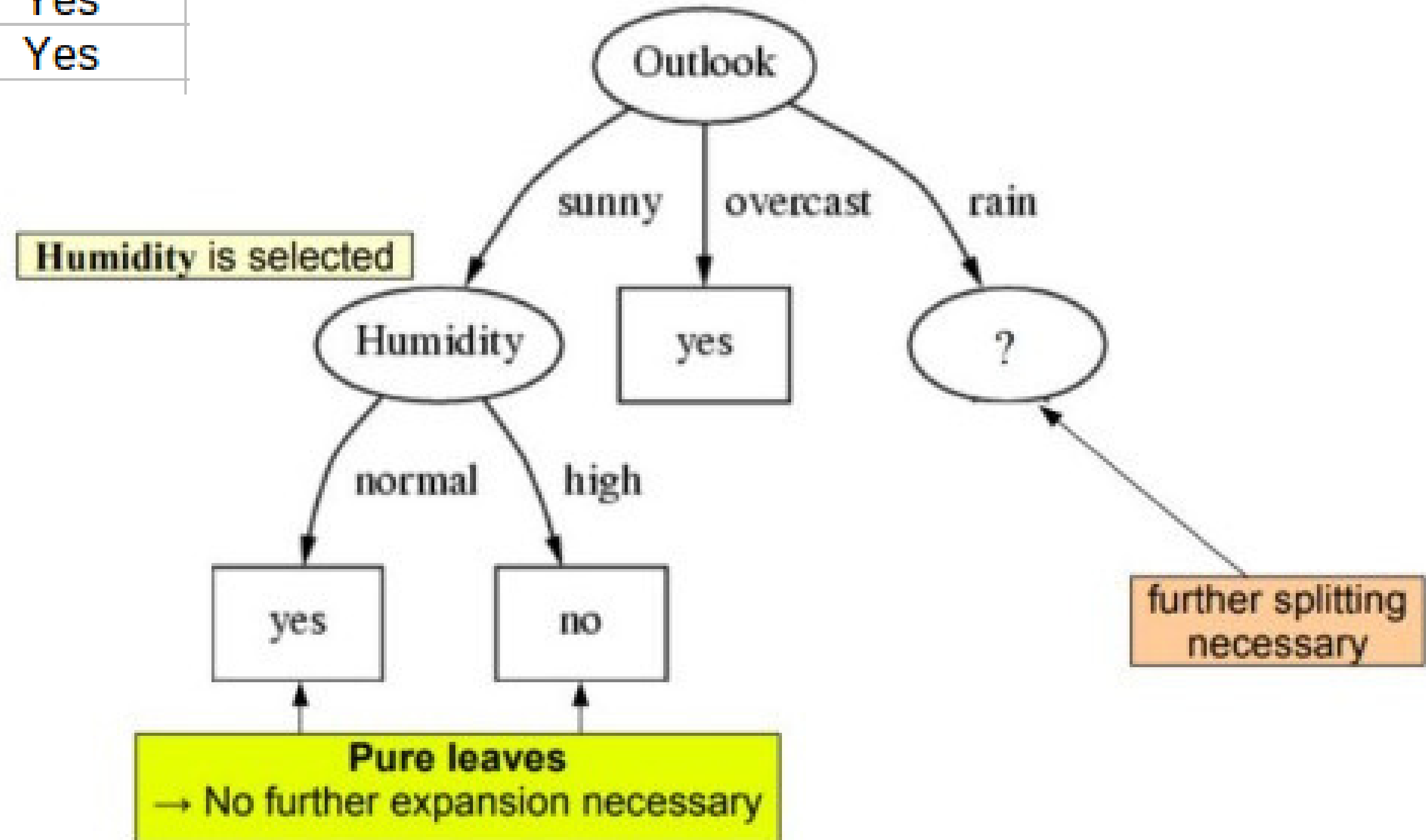
- Calculate **Gain:**    Gain = 0.571

- PICK THE HIGHEST GAIN ATTRIBUTE.

| Attributes | Gain |
|---|---|
| Temperature | 0.571 |
| Humidity | 0.971 |
| Windy | 0.02 |

NEXT NODE IN SUNNY:

**HUMIDITY**

| Outlook | Humidity | PlayTennis |
|---------|----------|------------|
| Sunny | High | No |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | Normal | Yes |
| Sunny | Normal | Yes |

Outlook

sunny    overcast    rain

Humidity is selected

Humidity    yes    ?

normal    high

yes    no

Pure leaves
→ No further expansion necessary

further splitting
necessary

| Outlook | Temperature | Humidity | Windy | PlayTennis |
|---------|-------------|----------|-------|------------|
| Rainy | Mild | High | Weak | Yes |
| Rainy | Cool | Normal | Weak | Yes |
| Rainy | Cool | Normal | Strong | No |
| Rainy | Mild | Normal | Weak | Yes |
| Rainy | Mild | High | Strong | No |

P=          N=

3             2

Total = 5

- ENTROPY:

$$Entropy = \frac{-p}{p+n}log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}log_2\left(\frac{n}{p+n}\right)$$

$$Entropy(S_{Rainy}) = \frac{-3}{3+2}log_2\left(\frac{3}{3+2}\right) - \frac{2}{3+2}log_2\left(\frac{2}{2+3}\right)$$

=>0.971

- For each Attribute: (let say **Humidity**):
  - Calculate Entropy for each Humidity, i.e for 'High' and 'Normal'

| Outlook | Humidity | PlayTennis |
|---|---|---|
| Rainy | High | Yes |
| Rainy | High | No |
| Rainy | Normal | Yes |
| Rainy | Normal | No |
| Rainy | Normal | Yes |

| Attribute | p | n | Entropy |
|---|---|---|---|
| High | 1 | 1 | 1 |
| Normal | 2 | 1 | 0.918 |

- Calculate **Average Information Entropy**:    I(Humidity) = 0.951

- Calculate **Gain:**                                  Gain = 0.020

- For each Attribute: (let say **Windy**):

  - Calculate Entropy for each Windy, i.e for 'Strong' and 'Weak'

| Outlook | Windy | PlayTennis |
|---------|-------|------------|
| Rainy | Strong | No |
| Rainy | Strong | No |
| Rainy | Weak | Yes |
| Rainy | Weak | Yes |
| Rainy | Weak | Yes |

| Attribute | p | n | Entropy |
|-----------|---|---|---------|
| Strong | 0 | 2 | 0 |
| Weak | 3 | 0 | 0 |

- Calculate **Average Information Entropy**: $\quad$ I(Windy) = 0

- Calculate **Gain:** $\quad\quad\quad\quad\quad$ Gain = 0.971

- For each Attribute: (let say **Temperature**):

  - Calculate Entropy for each Windy, i.e for 'Cool', 'Hot' and 'Mild'

| Outlook | Temperature | PlayTennis |
|---------|-------------|------------|
| Rainy | Mild | Yes |
| Rainy | Cool | Yes |
| Rainy | Cool | No |
| Rainy | Mild | Yes |
| Rainy | Mild | No |

| Attribute | p | n | Entropy |
|-----------|---|---|---------|
| Cool | 1 | 1 | 1 |
| Mild | 2 | 1 | 0.918 |

- Calculate **Average Information Entropy**:      I(Temp) = 0.951

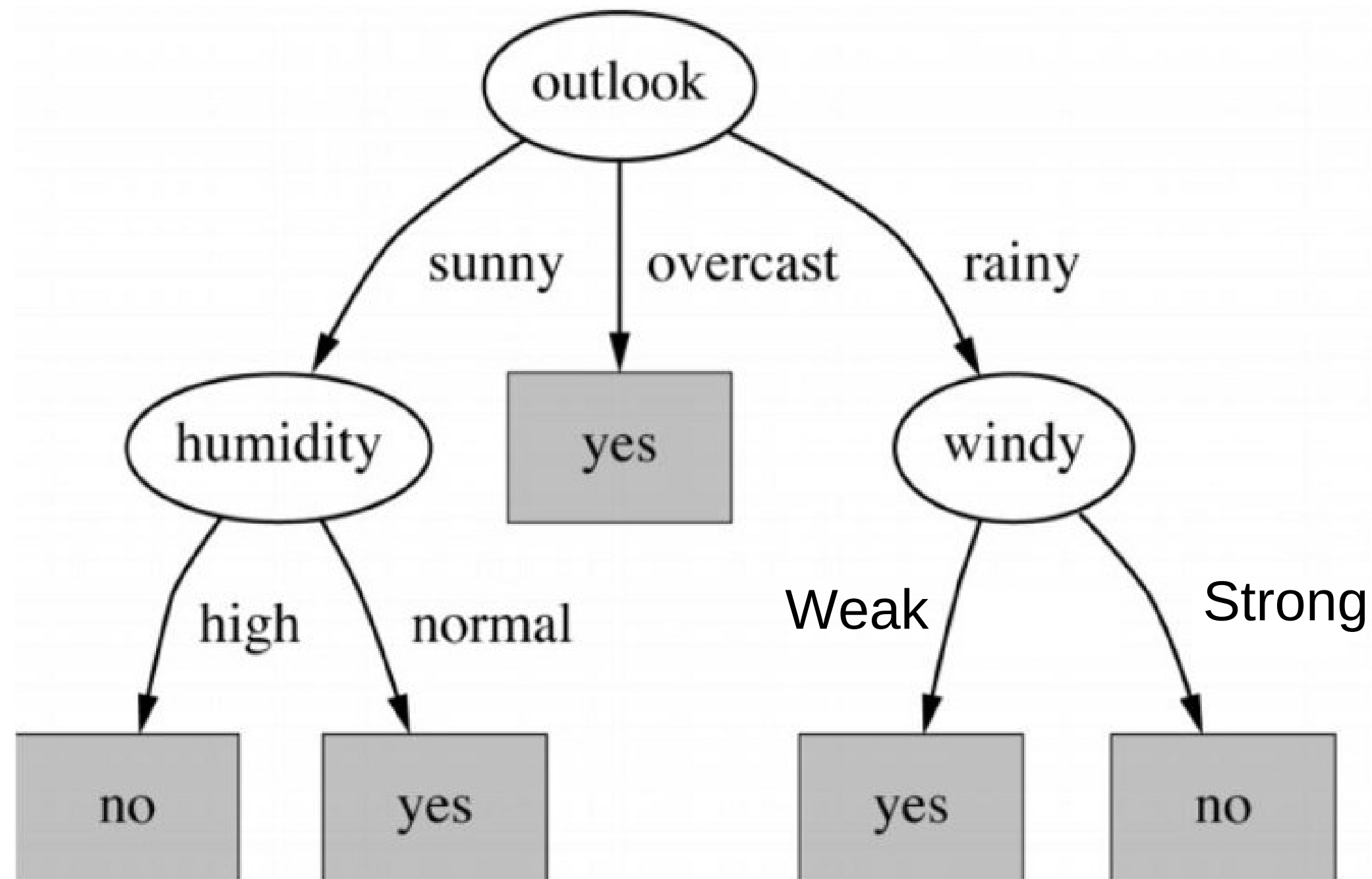- Calculate **Gain:**                Gain = 0.020

- PICK THE HIGHEST GAIN ATTRIBUTE.

| Attributes | Gain |
|:---:|:---:|
| Humidity | 0.02 |
| Windy | 0.971 |
| Temperature | 0.02 |

NEXT NODE IN

RAINY:

WINDY

# Final decision tree

Thank You