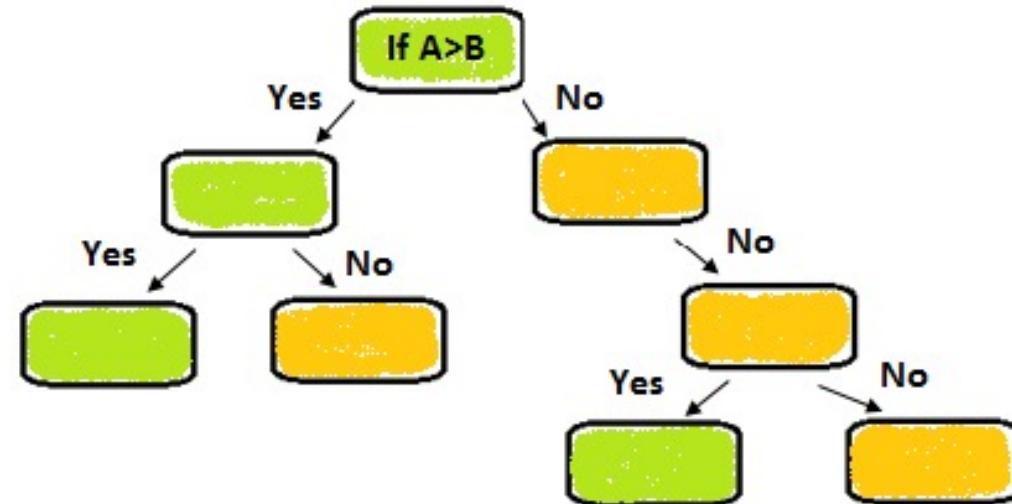
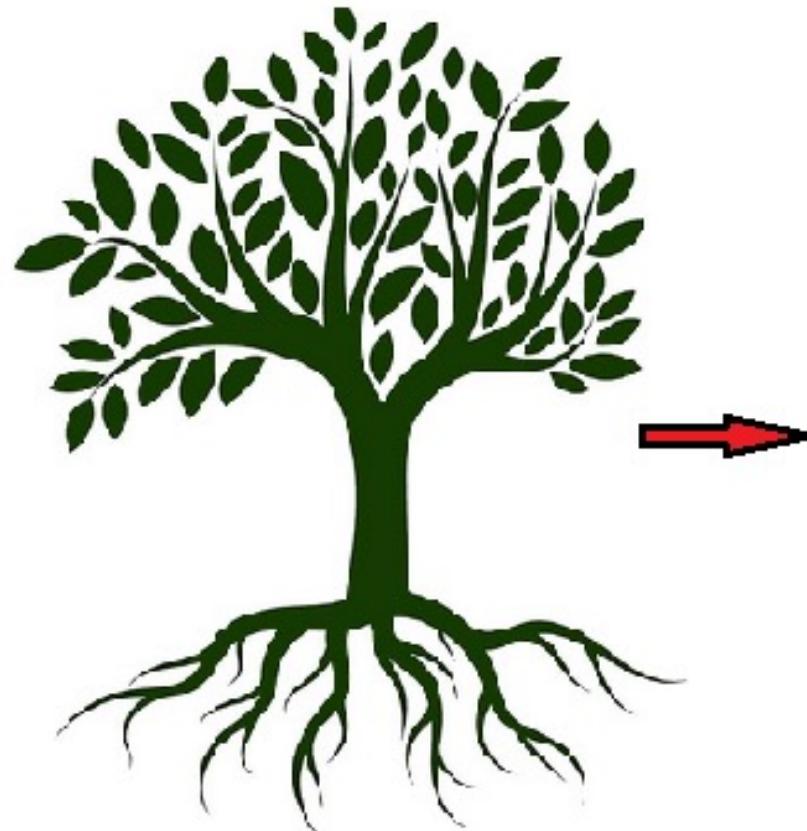


Decision Tree for Classification



Vinh Dinh Nguyen
PhD in Computer Science

Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

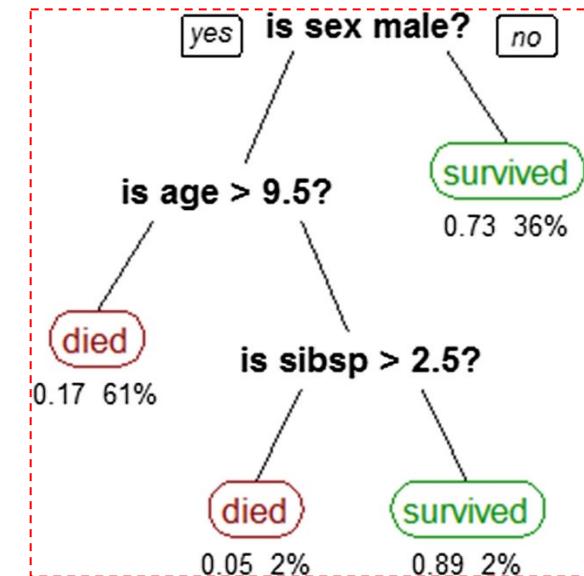
Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

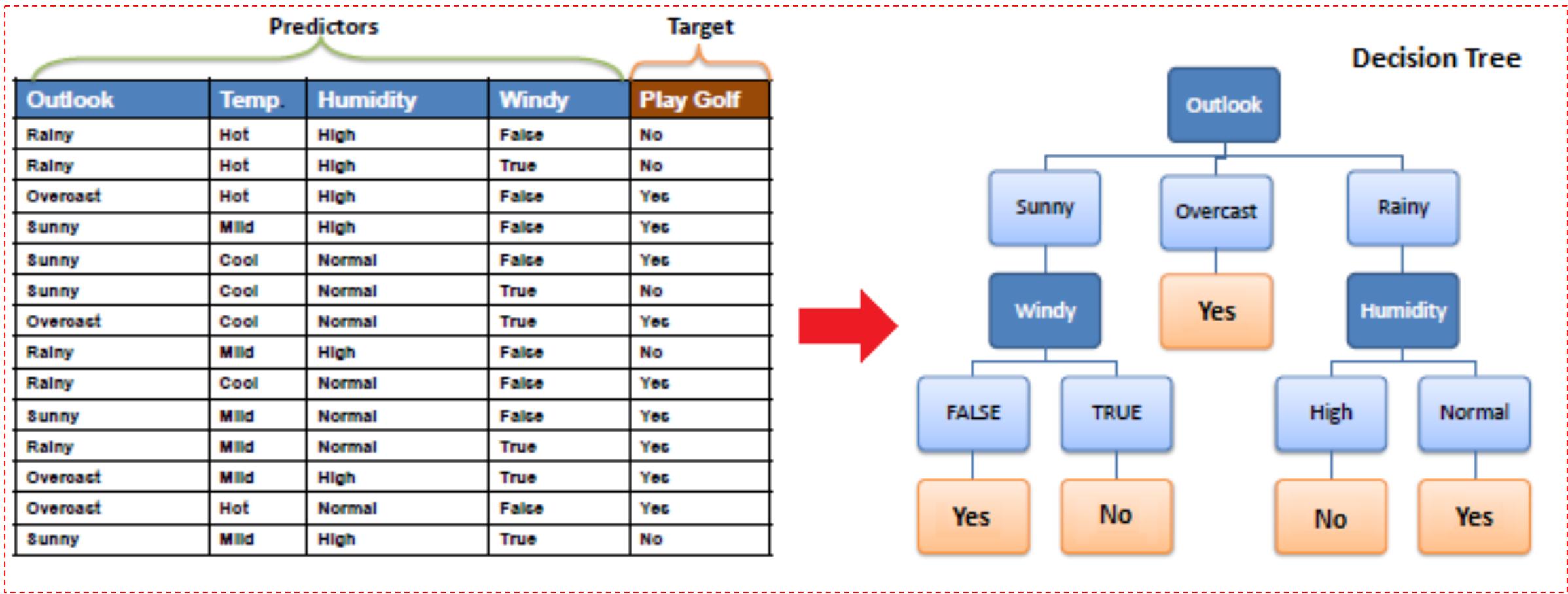
Sample Decision Tree

A decision tree predicting the survival of a passenger on the Titanic using the sex, age, and siblings or spouses onboard attributes

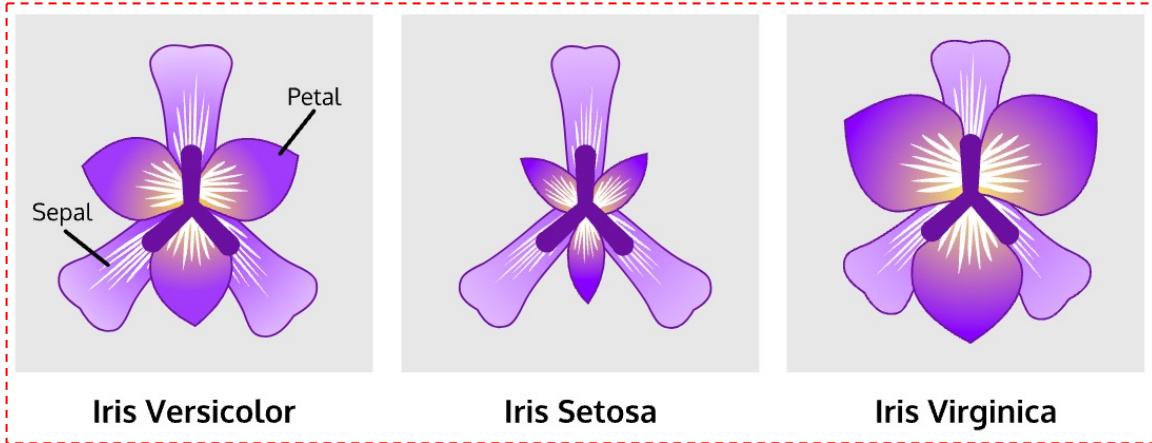
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 0 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |



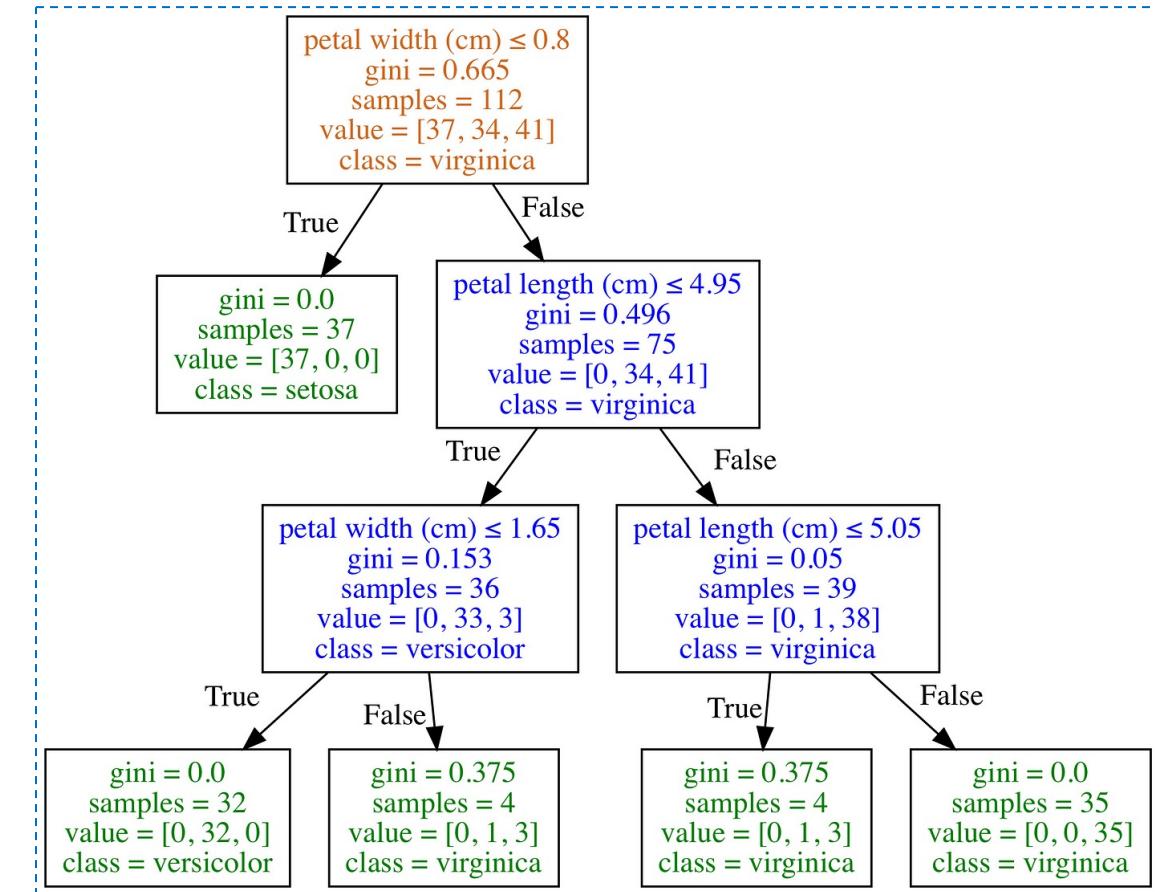
Sample Decision Tree



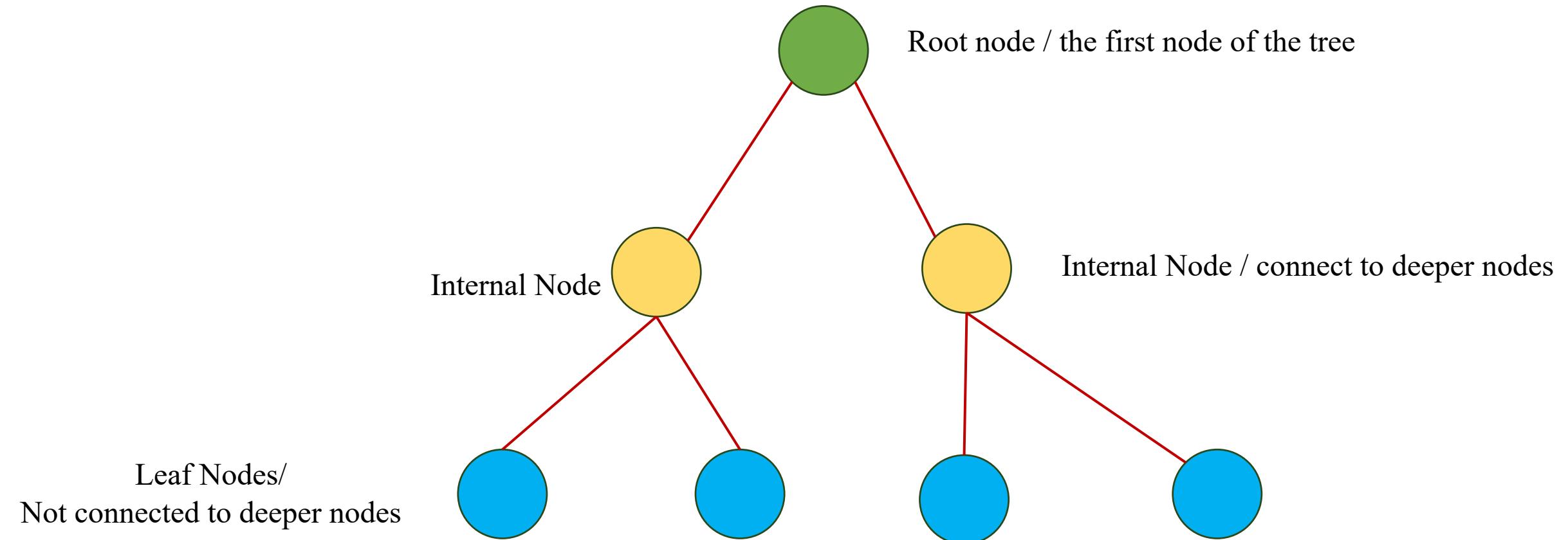
Sample Decision Tree



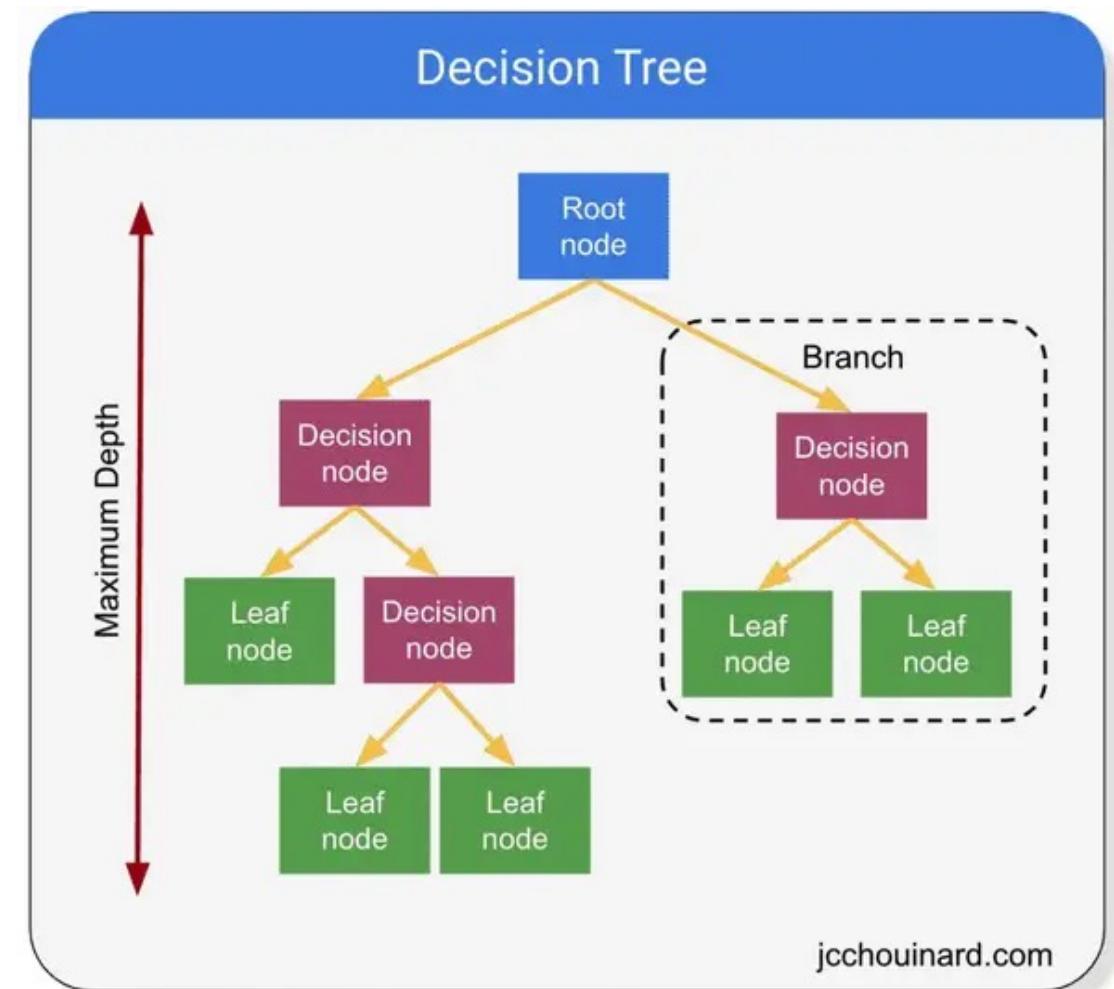
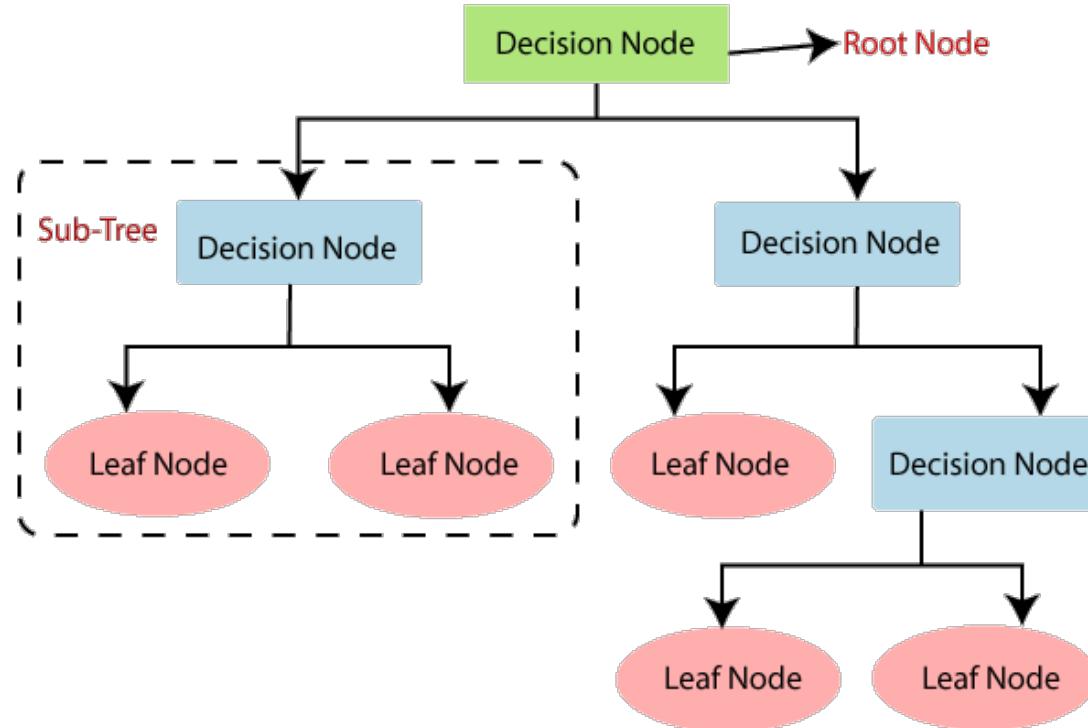
| | sepal_length | sepal_width | petal_length | petal_width | species |
|---|--------------|-------------|--------------|-------------|-------------|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |



Basic Tree Terminology



Basic Tree Terminology



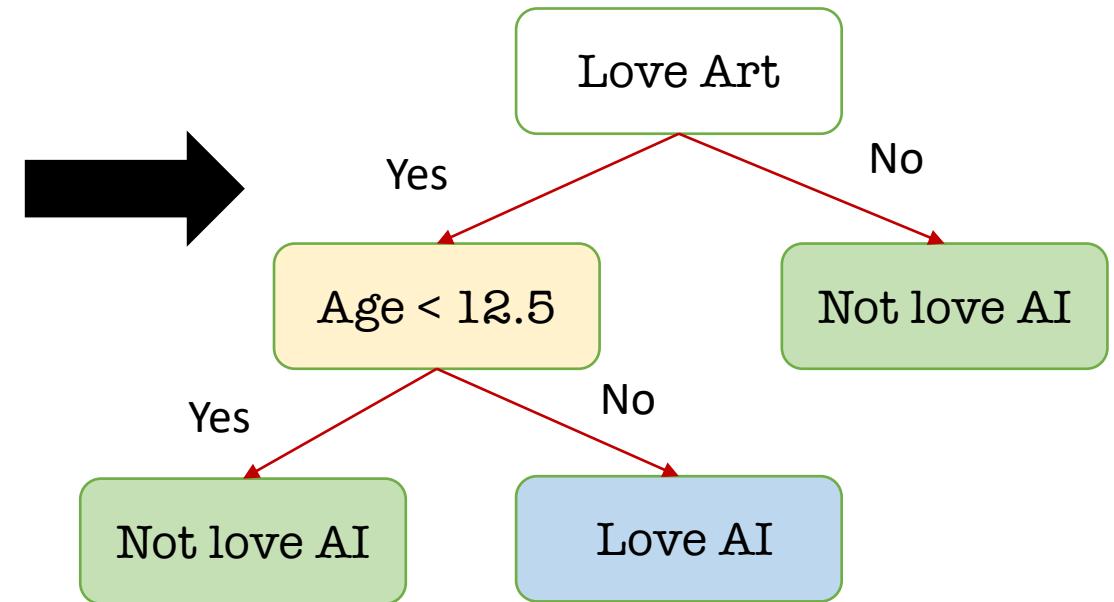
jcchouinard.com

Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

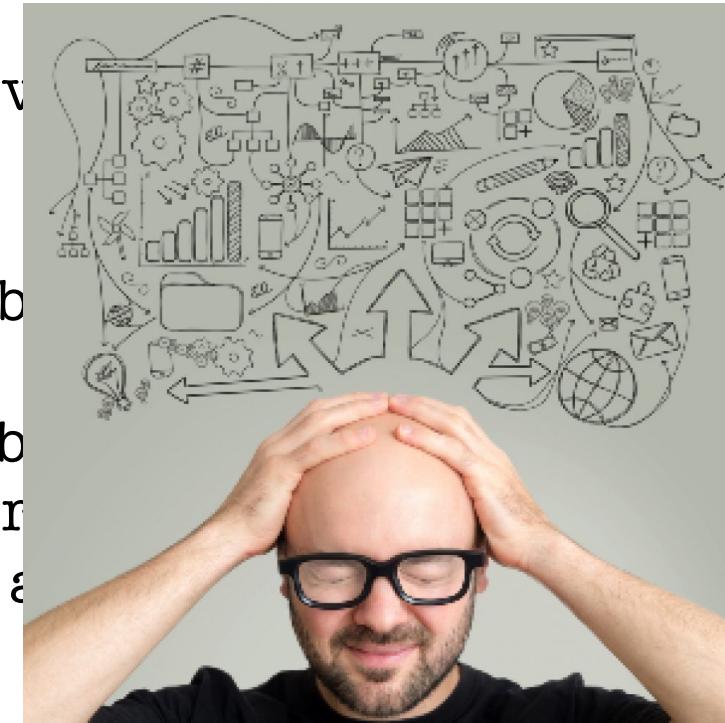
Decision Tree From Dataset

| Love Math | Love Art | Age | Love AI |
|-----------|----------|-----|---------|
| Yes | Yes | 7 | No |
| Yes | No | 12 | No |
| No | Yes | 18 | Yes |
| No | Yes | 35 | Yes |
| Yes | Yes | 38 | Yes |
| Yes | No | 50 | No |
| No | No | 83 | No |

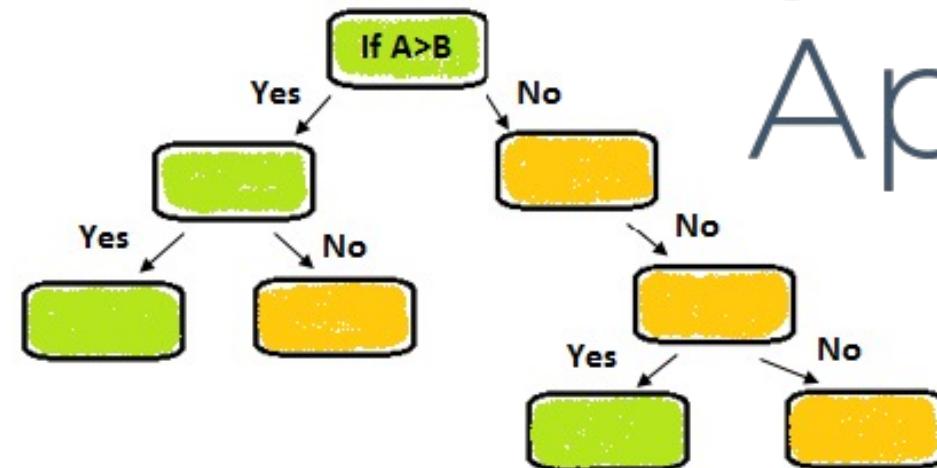
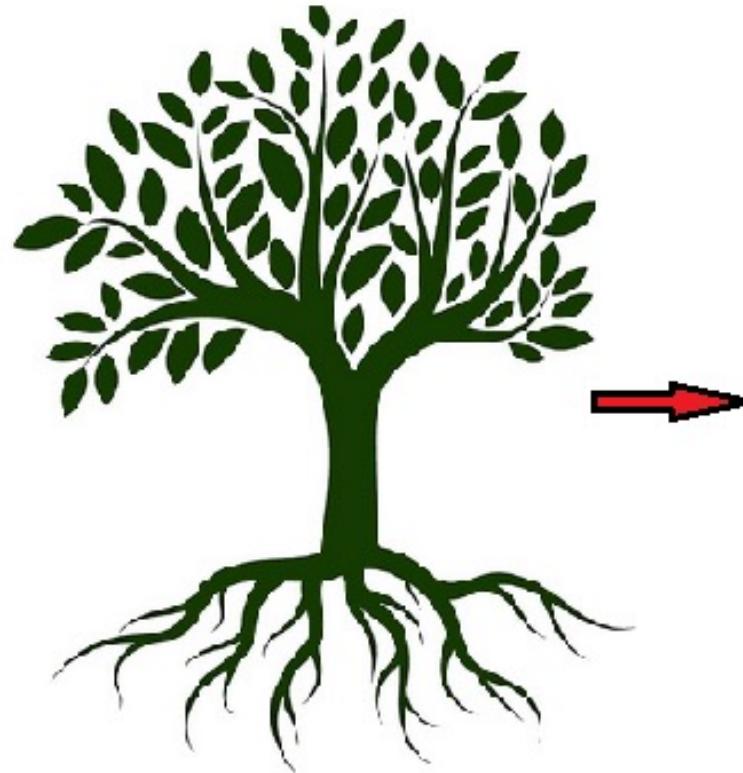


How to Build a Decision Tree

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**.
- **Step-3:** Divide the S into subsets that contains possible values of attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets created in step -3. Continue this process until a stage is reached where we cannot further classify the nodes and called the final node as a leaf node.



Decision Tree in Simple Concepts



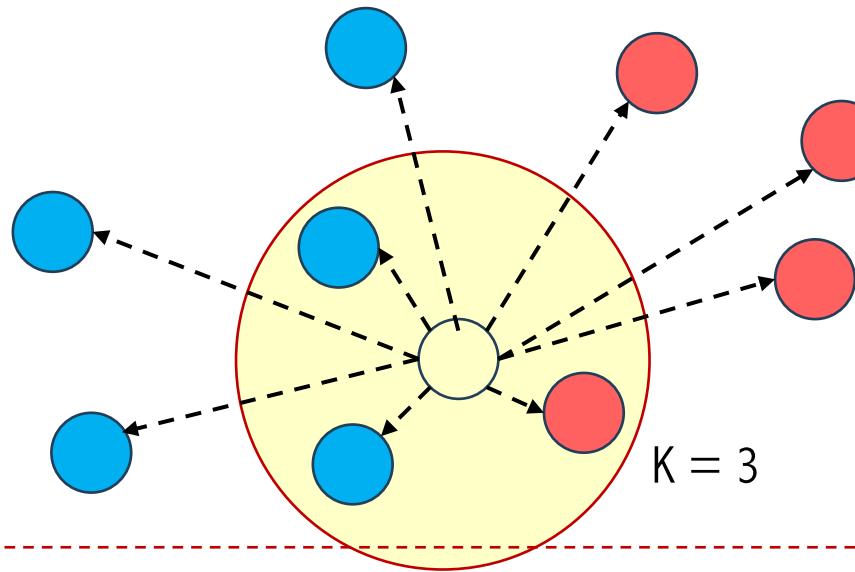
Simple
Approach



Decision Tree Motivation

KNN Limitations

Full dataset



Data cần phân loại



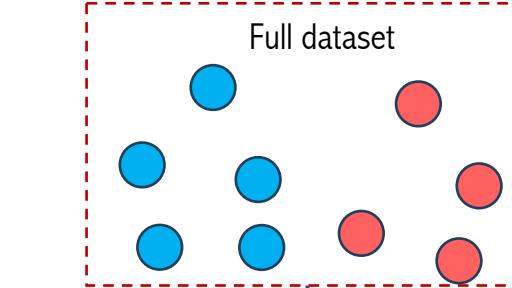
Class A



Class B

KNN has some drawbacks and challenges, such as computational expense, slow speed, for large datasets

Data cần phân loại



Decision Tree Idea

Sub-dataset

Điều kiện

Sub-dataset

Dự đoán class A

Sub-dataset

Điều kiện

Sub-dataset

Dự đoán class A

Sub-dataset

Dự đoán class B

How to Build Decision Tree

| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |

How to select the first node in the tree



Which one is a Root Node

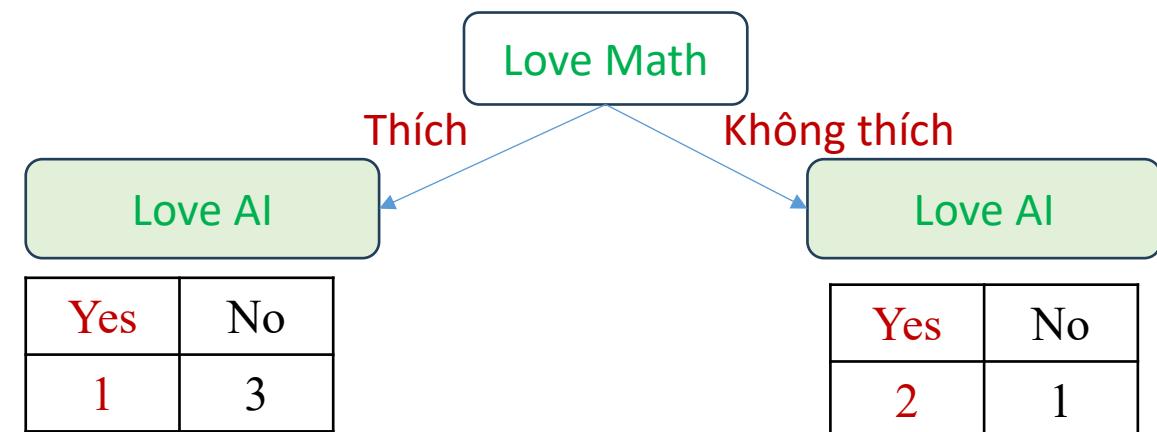
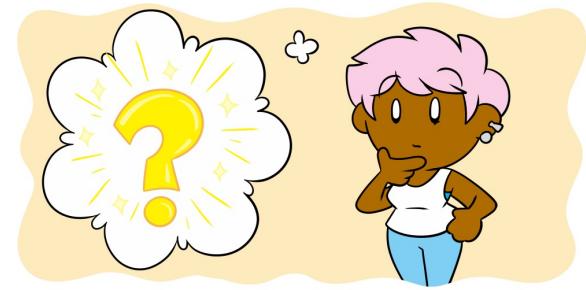


How to Build Decision Tree

| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |

How to select the first node in the tree

Vì hiện tại chúng ta chưa biết chọn thông tin nào làm root node. Nên cách đơn giản là giả sử lấy thông tin “love math” làm root node.



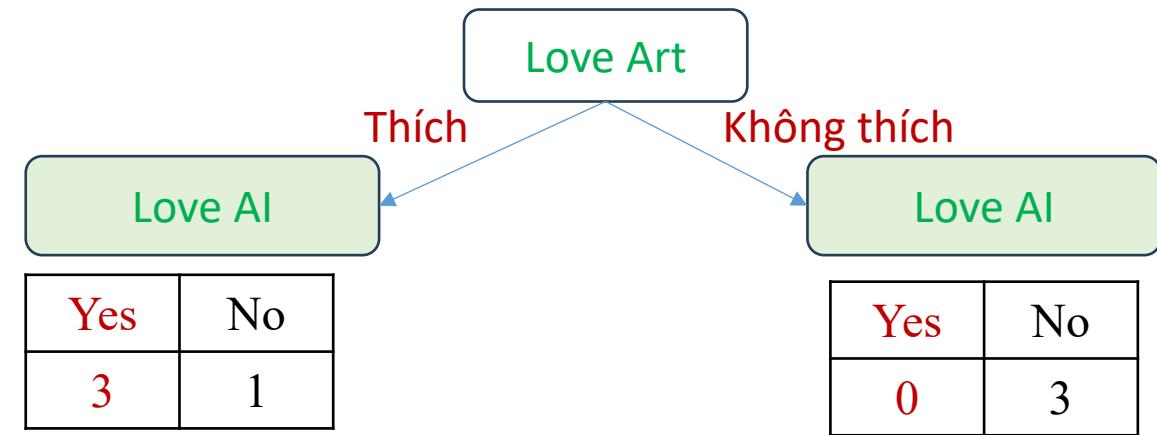
Để cho công bằng, chúng ta cũng nên thử chọn thêm Love Art hoặc Age làm root node nhé.

How to Build Decision Tree

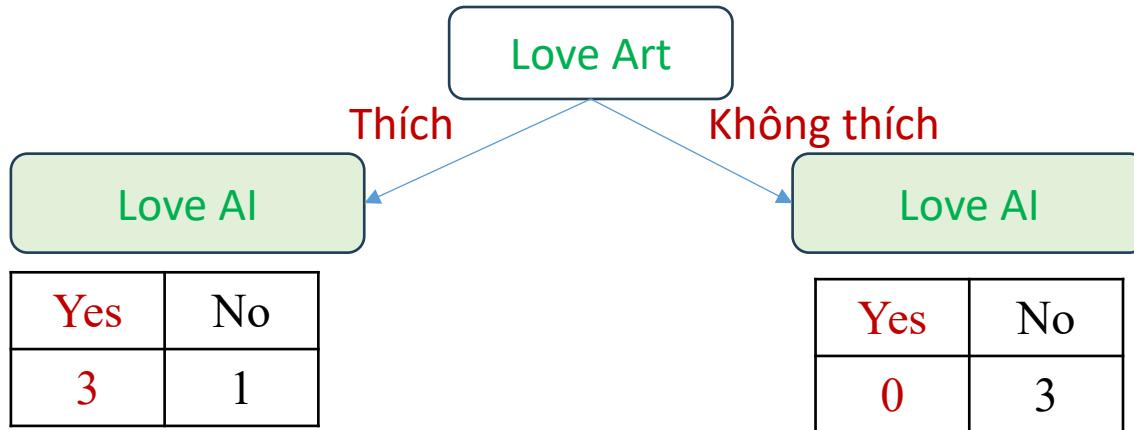
| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |

How to select the first node in the tree

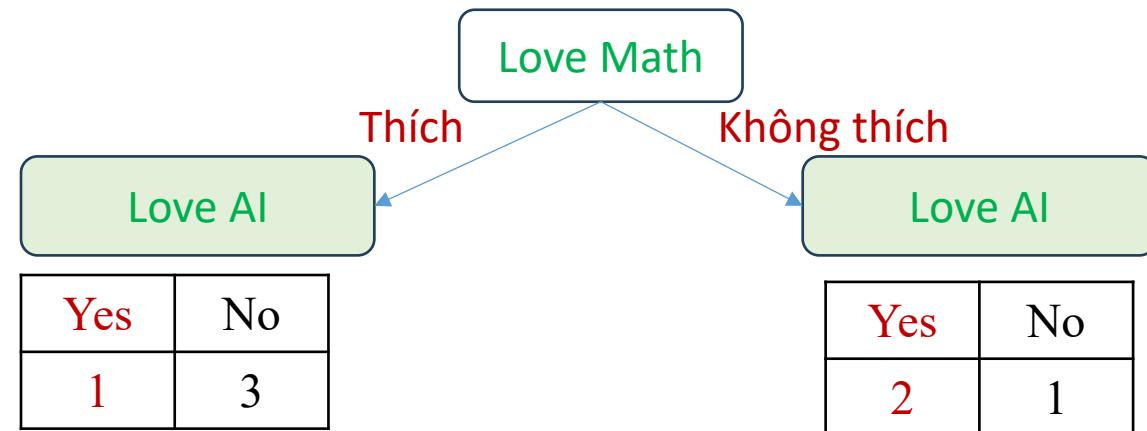
Chọn Love Art như là root node.



Review Two Approaches

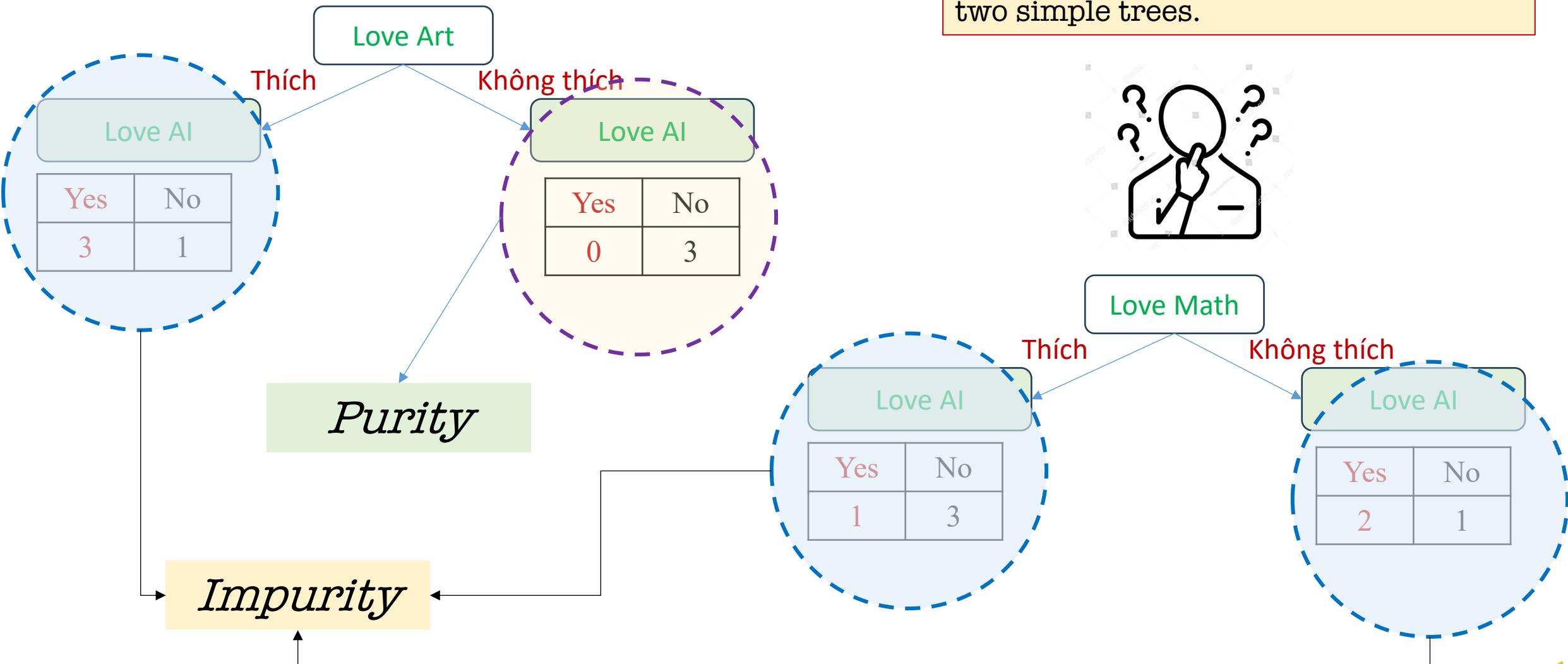


Do you have any **comments** on these two simple trees.

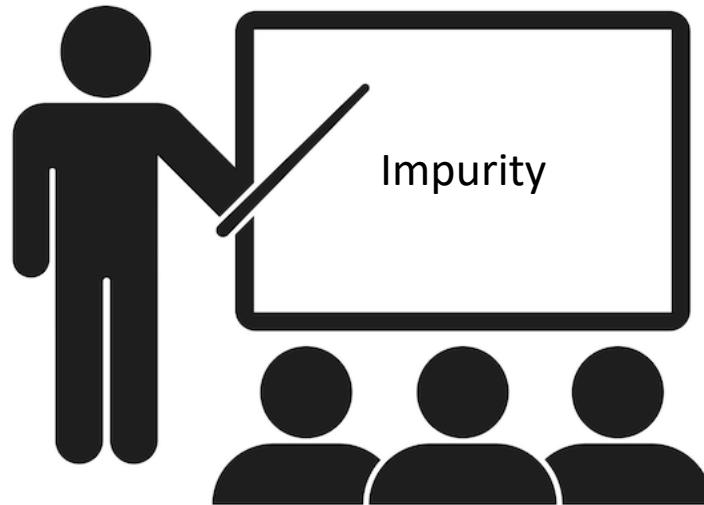


Review Two Approaches

Do you have any **comments** on these two simple trees.

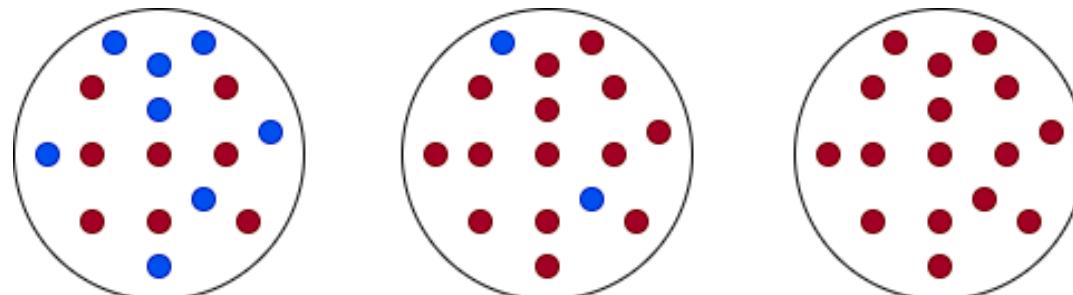


What is an Impurity

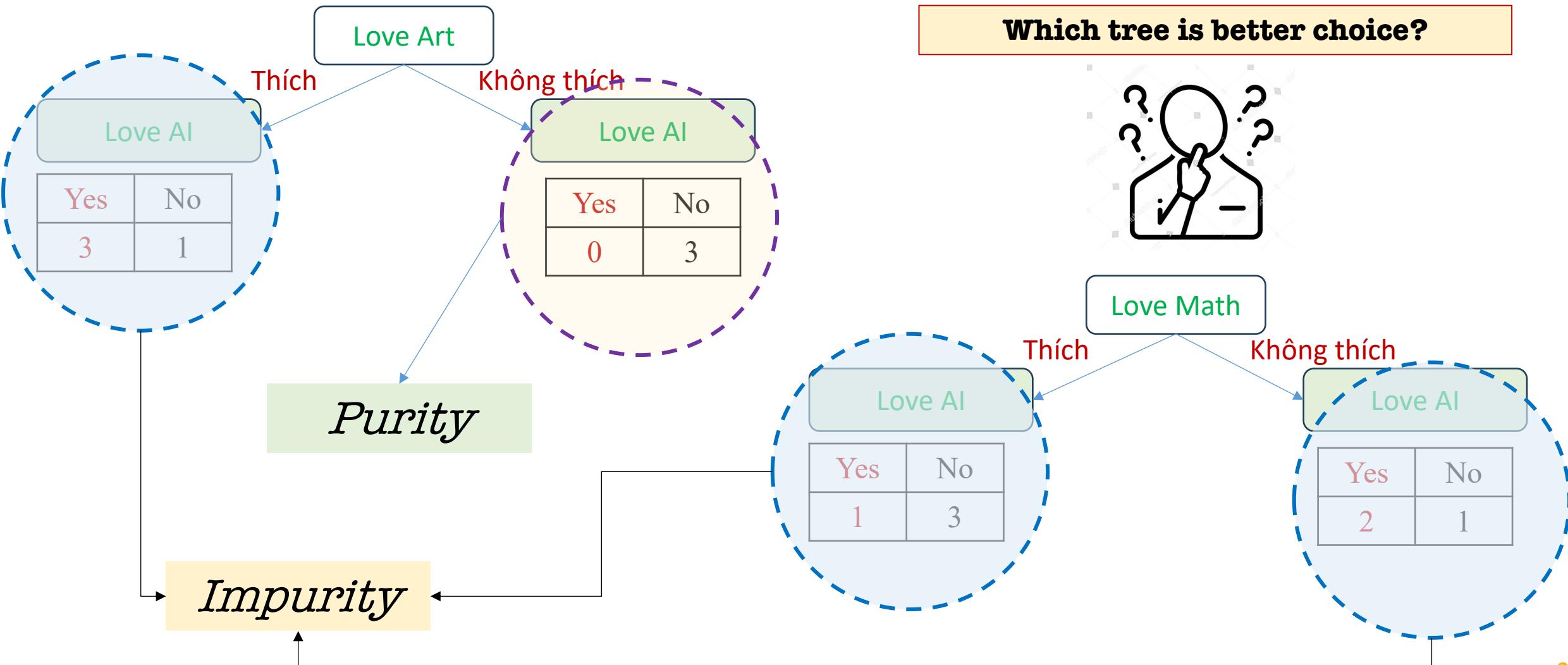


| | | | | | |
|-------|---|---|---|---|---|
| Set 1 | 1 | 1 | 1 | 1 | 1 |
| Set 2 | 1 | 1 | 2 | 1 | 1 |
| Set 3 | 1 | 2 | 4 | 6 | 7 |

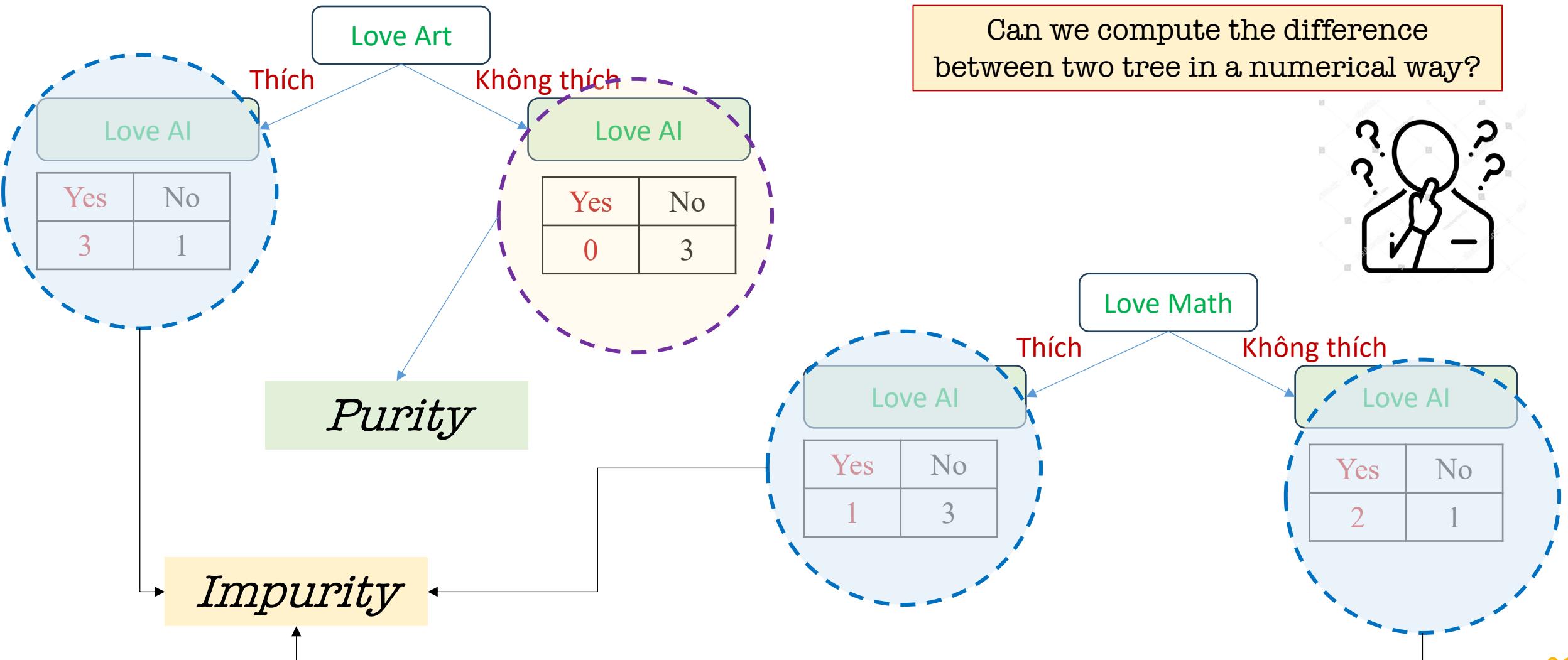
| | Impurity Score |
|-------|----------------|
| Set 1 | Thấp |
| Set 2 | Trung Bình |
| Set 3 | Cao |



Review Two Approaches



Review Two Approaches



Evaluation Metrics

Entropy – Information Gain

GNI IMPURITY



Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

Evaluation Metrics

Entropy – Information Gain

GNI IMPURITY



GINI Impurity

It was developed by statistician and sociologist Corrado Gini.

Consider a dataset D that contains samples from k classes. The probability of samples belonging to class i at a given node can be denoted as p_i . Then the Gini Impurity of D is defined as:

$$Gini(D) = 1 - \sum_{i=1}^k p_i^2$$

| | Probability | | | Gini Impurity | |
|--------|-------------|-------|-------|---------------|----------------------------|
| | n_1 | n_2 | p_1 | p_2 | $1 - p_1^2 - p_2^2$ |
| Node A | 0 | 10 | 0 | 1 | $1 - 0^2 - 1^2 = 0$ |
| Node B | 3 | 7 | 0.3 | 0.7 | $1 - 0.3^2 - 0.7^2 = 0.42$ |
| Node C | 5 | 5 | 0.5 | 0.5 | $1 - 0.5^2 - 0.5^2 = 0.5$ |

GINI Impurity

It was developed by statistician and sociologist Corrado Gini.

Consider a dataset D that contains samples from k classes. The probability of samples belonging to class i at a given node can be denoted as p_i . Then the Gini Impurity of D is defined as:

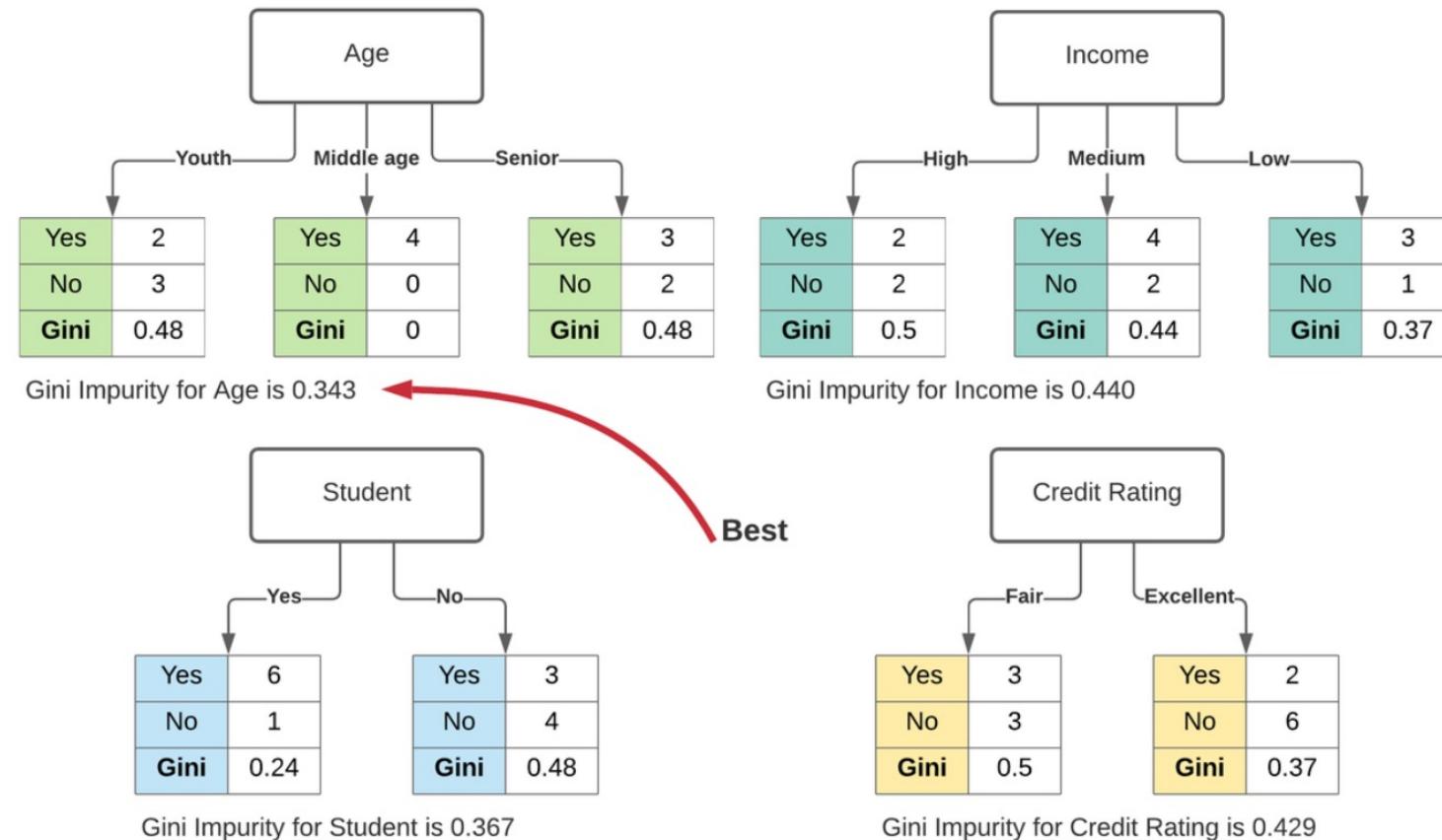
$$Gini(D) = 1 - \sum_{i=1}^k p_i^2$$

| | Count | Probability | Gini Impurity | | |
|--------|-------|-------------|---------------|-------|----------------------------|
| | n_1 | n_2 | p_1 | p_2 | $1 - p_1^2 - p_2^2$ |
| Node A | 0 | 10 | 0 | 1 | $1 - 0^2 - 1^2 = 0$ |
| Node B | 3 | 7 | 0.3 | 0.7 | $1 - 0.3^2 - 0.7^2 = 0.42$ |
| Node C | 5 | 5 | 0.5 | 0.5 | $1 - 0.5^2 - 0.5^2 = 0.5$ |

GINI Impurity

It was developed by statistician and sociologist Corrado Gini.

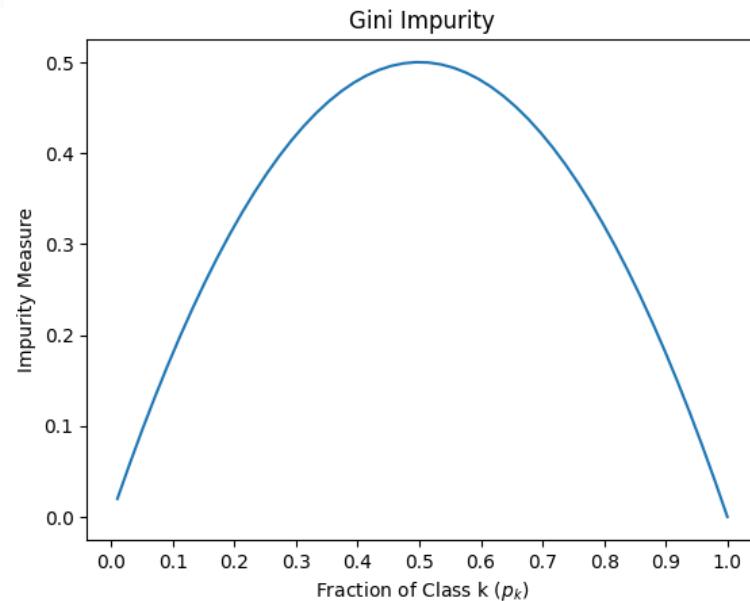
$$Gini(D) = 1 - \sum_{i=1}^k p_i^2$$



GINI Impurity

$$Gini(D) = 1 - \sum_{i=1}^k p_i^2$$

It was developed by [statistician](#) and [sociologist Corrado Gini](#).



```
#A figure is created to show Gini impurity measures
plt.figure()
x = np.linspace(0.01,1)
y = 1 - (x*x) - (1-x)*(1-x)
plt.plot(x,y)
plt.title('Gini Impurity')
plt.xlabel("Fraction of Class k ($p_k$)")
plt.ylabel("Impurity Measure")
plt.xticks(np.arange(0,1.1,0.1))

plt.show()
```

This figure shows that Gini impurity is maximum for the 50-50 sample ($p_1=0.5$) and minimum for the homogeneous sample ($p_1=0$ or $p_1=1$)

GINI Impurity

| | | | | | |
|-------|---|---|---|---|---|
| Set 1 | 1 | 1 | 1 | 1 | 1 |
| Set 2 | 1 | 1 | 2 | 1 | 1 |
| Set 3 | 1 | 2 | 4 | 6 | 7 |

| | No of Unique Element | Count of unique elements | Probability |
|-------|----------------------|--------------------------|---------------------|
| Set 1 | 1 | 5 | 5/5 |
| Set 2 | 1,2 | 4, 1 | 4/5, 1/5 |
| Set 3 | 1,2,4,6,7 | 1,1,1,1,1 | 1/5,1/5,1/5,1/5,1/5 |

GINI Impurity

| | | | | | |
|-------|---|---|---|---|---|
| Set 1 | 1 | 1 | 1 | 1 | 1 |
| Set 2 | 1 | 1 | 2 | 1 | 1 |
| Set 3 | 1 | 2 | 4 | 6 | 7 |

| | No of Unique Element | Count of unique elements | Probability |
|-------|----------------------|--------------------------|-------------|
| Set 1 | 1 | 5 | 5/5 |

$$n = 1, p_1 = 1$$

$$\text{GINI} = 1 - \sum_{i=1}^n (p_i)^2 \quad \text{Gini} = 1 - [1^2] = 0$$

GINI Impurity

| | | | | | |
|-------|---|---|---|---|---|
| Set 1 | 1 | 1 | 1 | 1 | 1 |
| Set 2 | 1 | 1 | 2 | 1 | 1 |
| Set 3 | 1 | 2 | 4 | 6 | 7 |

| | No of Unique Element | Count of unique elements | Probability |
|-------|----------------------|--------------------------|-------------|
| Set 3 | 2 | 4,1 | 4/5, 1/5 |

$$GINI = 1 - \sum_{i=1}^n (p_i)^2$$

$n = 2, p_1 = 4/5, p_2 = 1/5$

$$Gini = 1 - [4/5^2 + 1/5^2] = 0.32$$

GINI Impurity

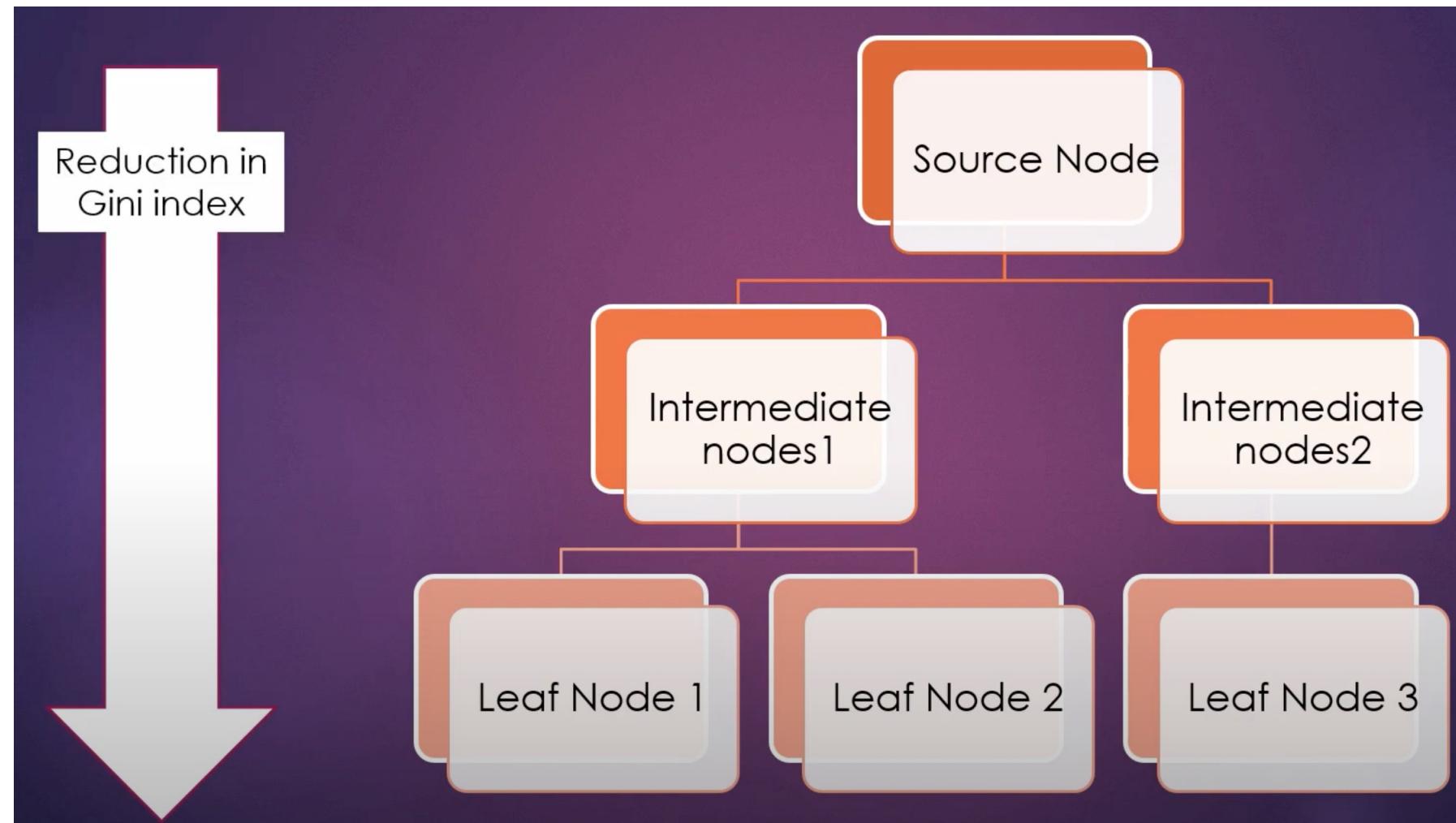
| | | | | | |
|-------|---|---|---|---|---|
| Set 1 | 1 | 1 | 1 | 1 | 1 |
| Set 2 | 1 | 1 | 2 | 1 | 1 |
| Set 3 | 1 | 2 | 4 | 6 | 7 |

| | No of Unique Element | Count of unique elements | Probability |
|-------|----------------------|--------------------------|-------------|
| Set 3 | 1 | 5 | 5/5 |

$$GINI = 1 - \sum_{i=1}^n (p_i)^2 \quad n = 5, \quad p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$$

$$Gini = 1 - [1/5^2 + 1/5^2 + 1/5^2 + 1/5^2 + 1/5^2] = 0.8$$

Idea of GNI Impurity

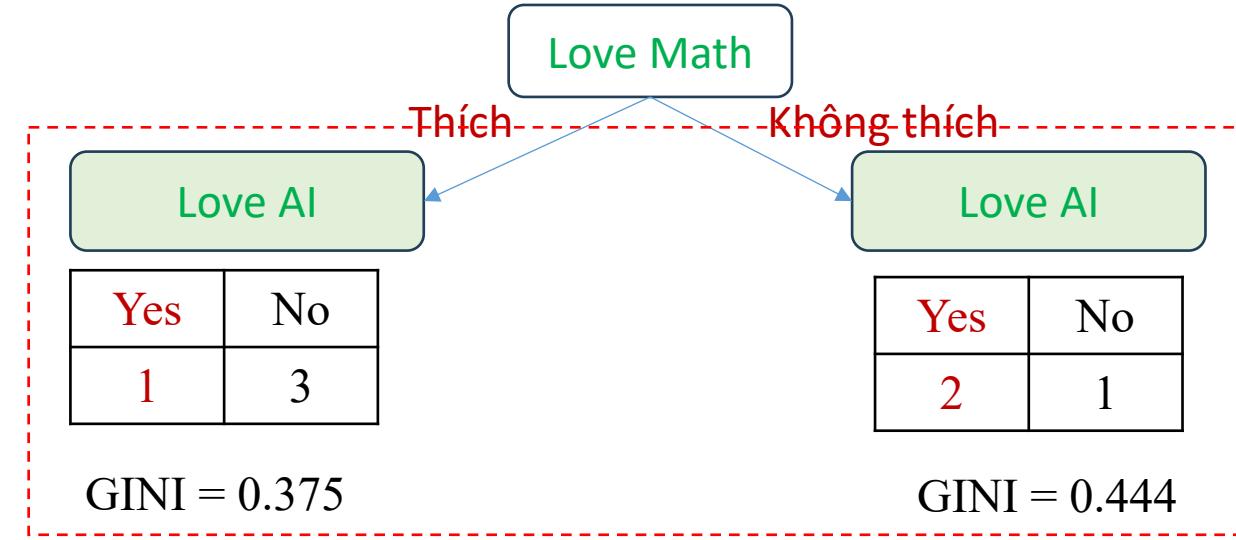


Gini Impurity for Leaf Nodes

$$\text{Gini} = 1 - \sum_{i=1}^n p_i^2$$

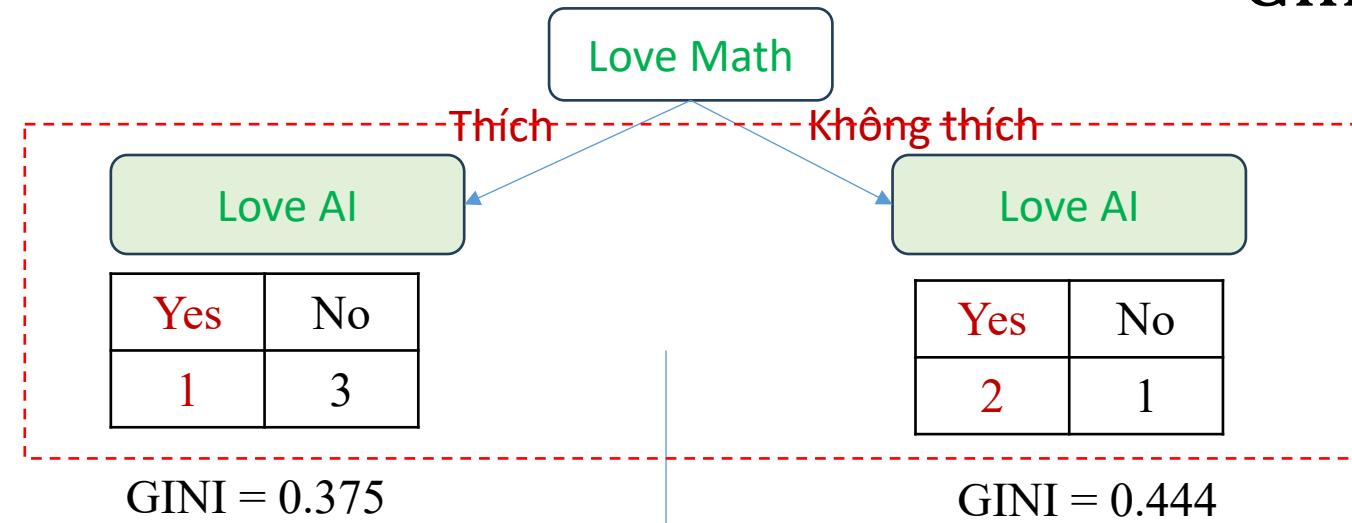
Cách tính Gini Impurity cho Leaf Nodes

1. Calculate the probabilities of all classes.
2. Square the calculated probabilities
3. Sum all the squared probabilities into a single integer
4. Subtract the single integer from 1



Gini Impurity for Leaf Nodes

$$\text{Gini} = 1 - \sum_{i=1}^n P_i^2$$

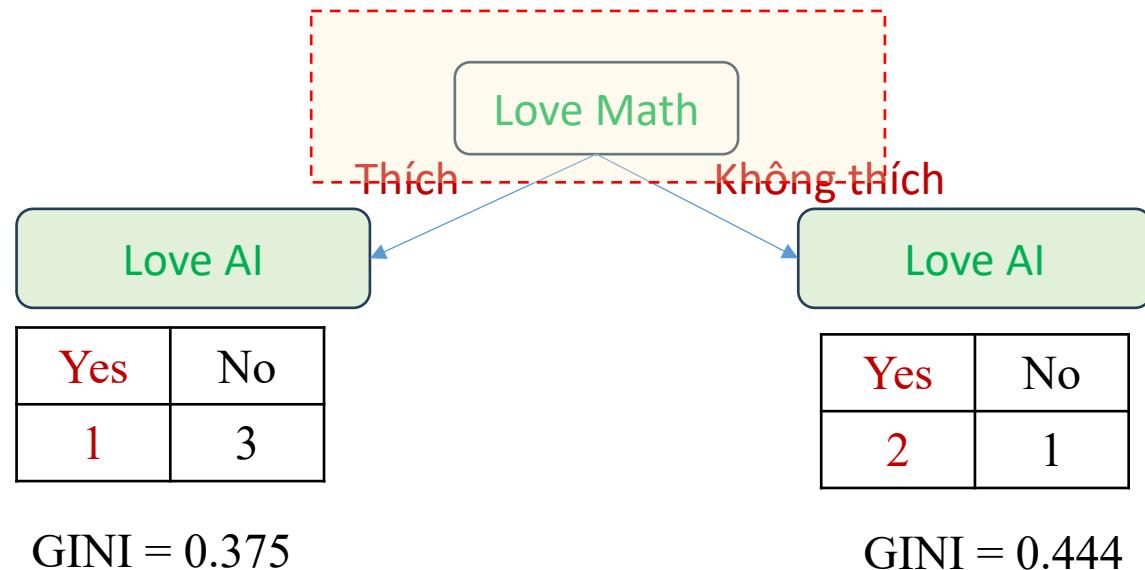


$$\text{Gini} = 1 - (\text{the probability of Yes})^2 - (\text{the probability of No})^2$$

$$GINI = 1 - [(1/4)^2 + (3/4)^2] = 0.375$$

$$GINI = 1 - [(2/3)^2 + (1/3)^2] = 0.444$$

Gini Impurity for Root/Internal Nodes

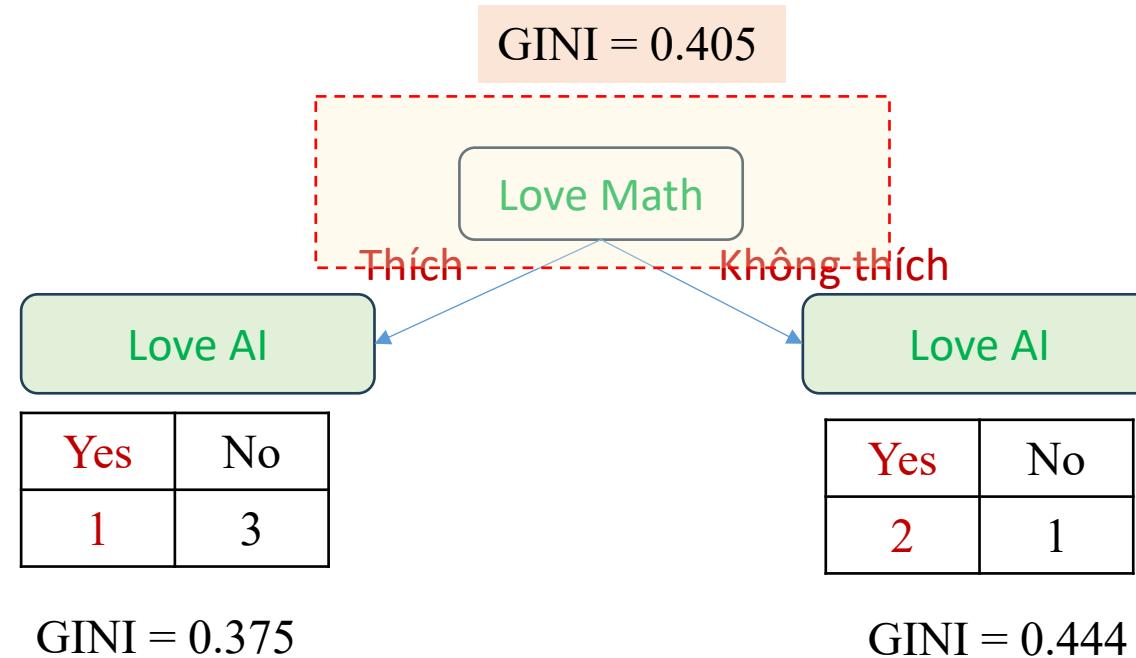


Cách tính Gini Impurity cho Root/Internal Nodes

1. Loop over the leaf node, from k = 1 all the way to the K leaf nodes
2. Find the gini impurity of the current k'th leaf
3. Count the number of observations in the k'th leaf
4. Divide by the total number of observation in the all leaf nodes
5. The result of each leaf is summed for a final gini impurity

$$Gini(Internal_j) = \sum_{k'th\ leaf=1}^{K\ leaf\ nodes} \left(\frac{count(L_k)}{count(L_1, \dots, L_k)} \right) Gini(L_k)$$

Gini Impurity for Root/Internal Nodes



Total Impurity = Weight average of Gini Impurities for The leaves

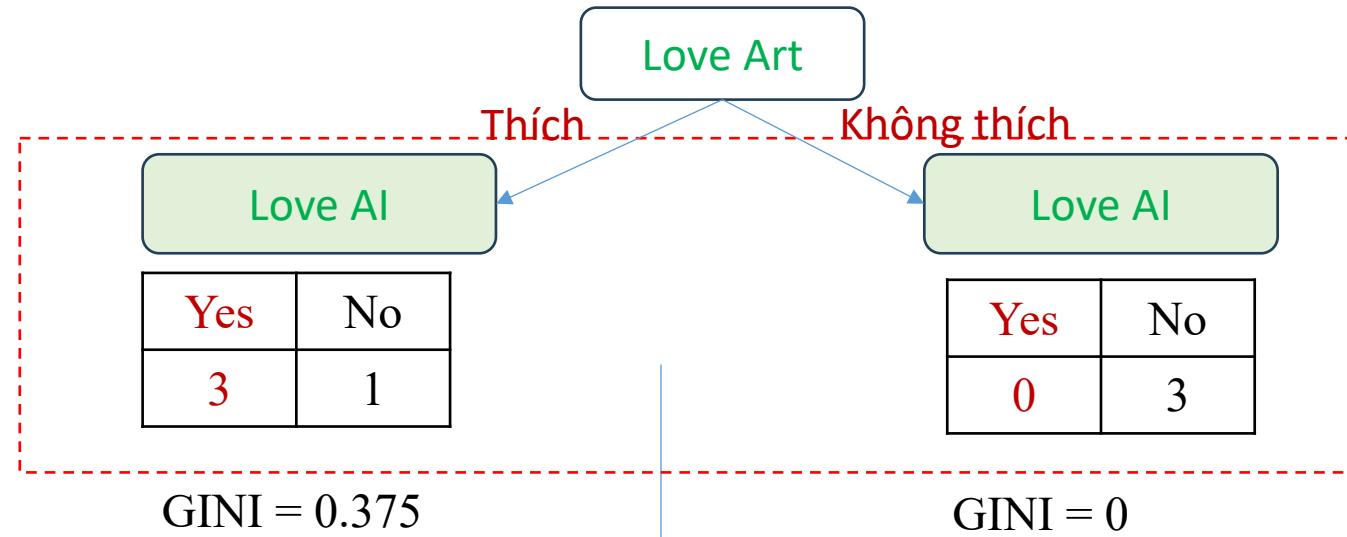
$$\text{Total Impurity} = \frac{4}{7} \times 0.375 + \frac{3}{7} \times 0.444 = 0.405$$

If a data set D is split on an attribute A into two subsets D_1 and D_2 with sizes n_1 and n_2 , respectively, the Gini Impurity can be defined as:

$$Gini_A(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2)$$

Gini Impurity for Leaf Nodes

$$GINI = 1 - \sum_{i=1}^n (p_i)^2$$

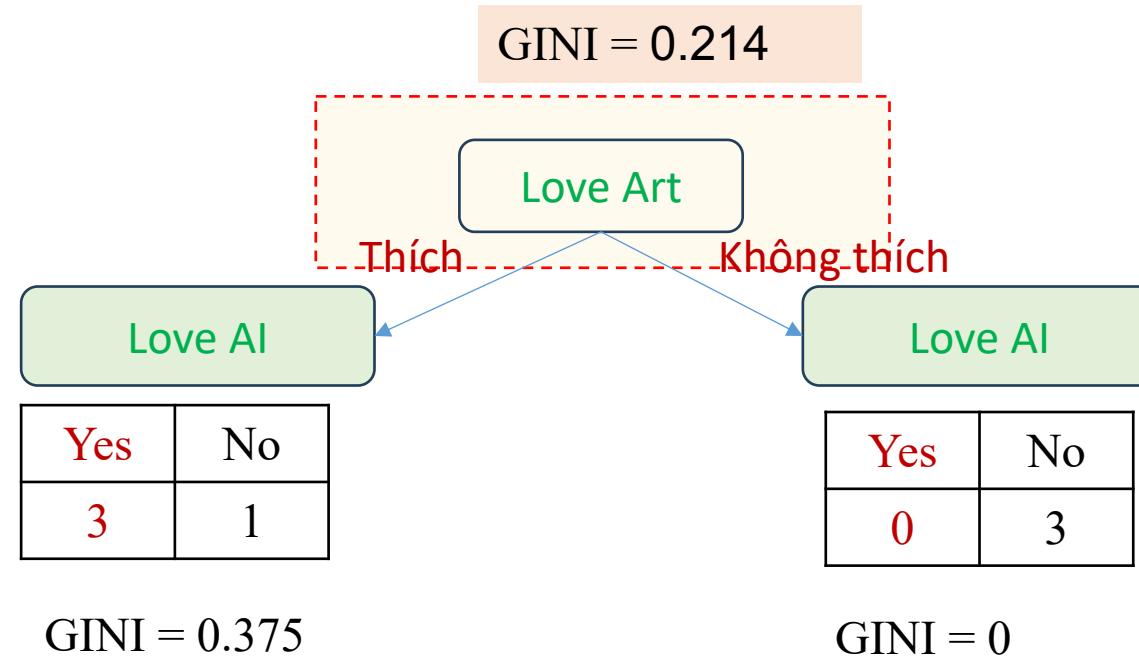


$$Gini = 1 - (\text{the probability of Yes})^2 - (\text{the probability of No})^2$$

$$GINI = 1 - [(3/4)^2 + (1/4)^2] = 0.375$$

$$GINI = 1 - [(3/3)^2] = 0$$

Gini Impurity for Root/Internal Nodes



Total Impurity = Weight average of Gini Impurities for The leaves

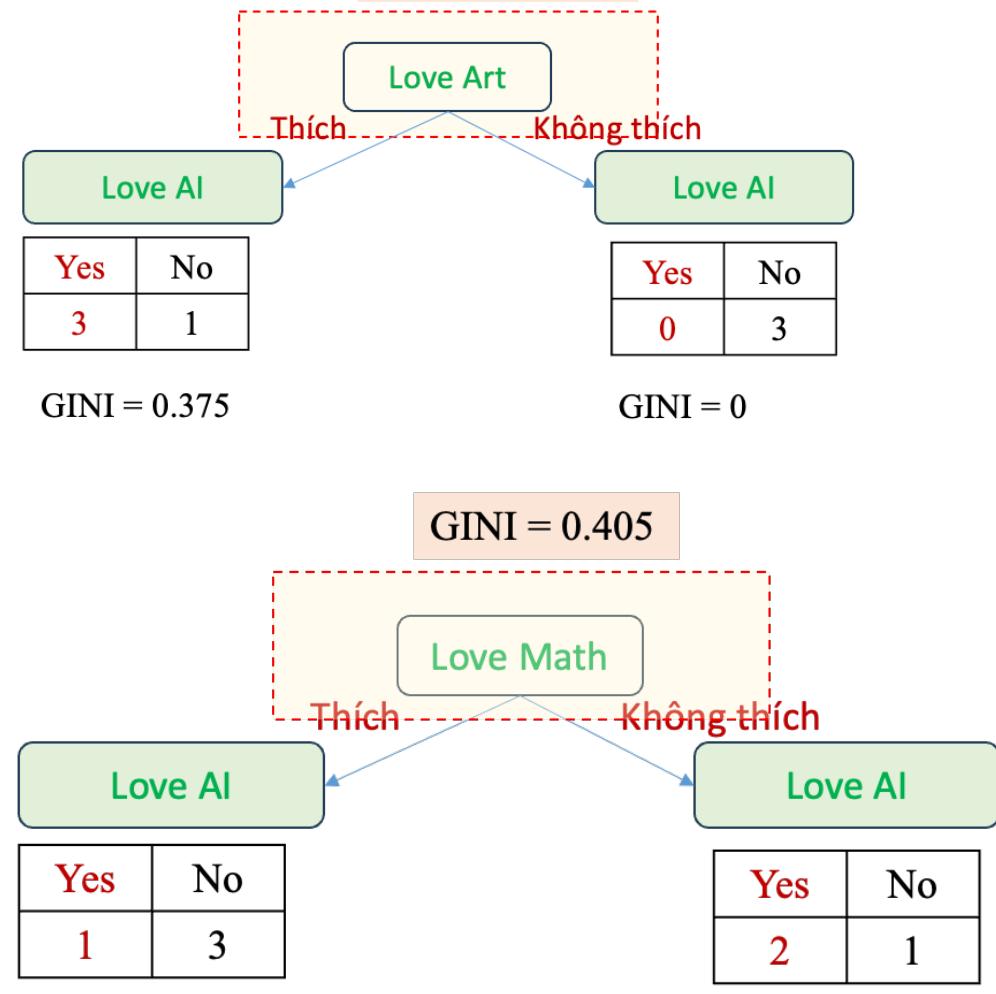
$$\text{Total Impurity} = \frac{4}{7} \times 0.375 + \frac{3}{7} \times 0 = 0.214$$

What is a GINI of “AGE”?

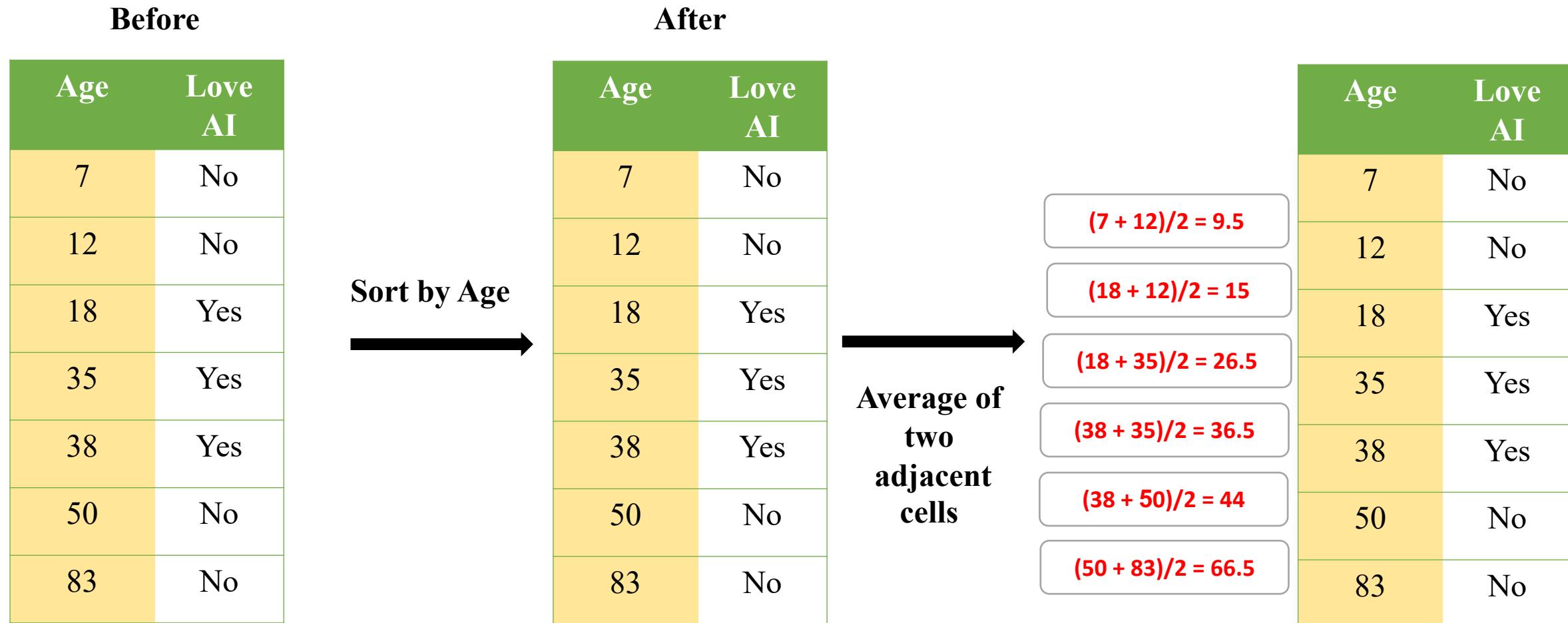
| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| | Yes | | 38 | Yes |
| | No | | 50 | No |
| | No | | 83 | No |



GINI = 0.214



Sort By Age



GINI IMPURITY FOR EACH OF AGE

| Age | Love AI |
|-----|---------|
| 7 | No |
| 12 | No |
| 18 | Yes |
| 35 | Yes |
| 38 | Yes |
| 50 | No |
| 83 | No |

$$(7 + 12)/2 = 9.5$$

$$(18 + 12)/2 = 15$$

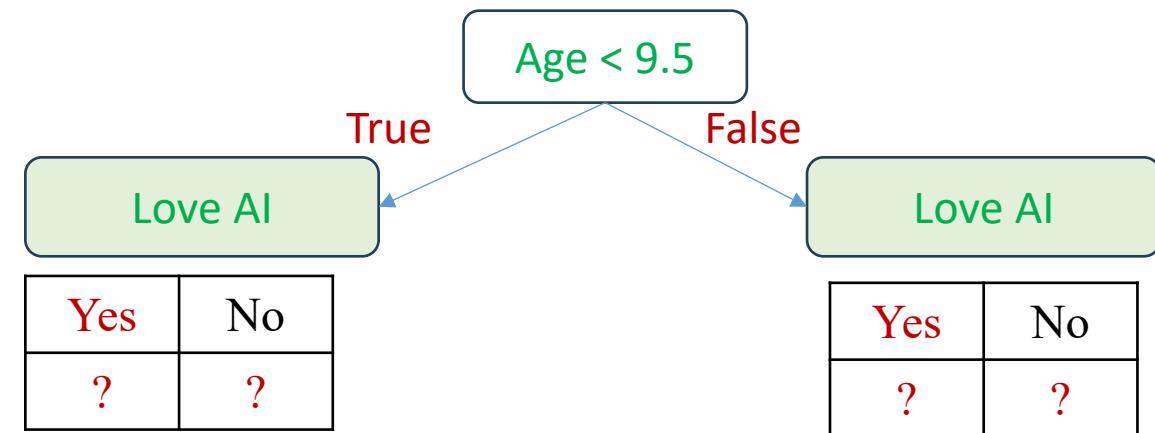
$$(18 + 35)/2 = 26.5$$

$$(38 + 35)/2 = 36.5$$

$$(38 + 50)/2 = 44$$

$$(50 + 83)/2 = 66.5$$

GINI = ???



GINI IMPURITY FOR EACH OF AGE

| Age | Love AI |
|-----|---------|
| 7 | No |
| 12 | No |
| 18 | Yes |
| 35 | Yes |
| 38 | Yes |
| 50 | No |
| 83 | No |

(7 + 12)/2 = 9.5

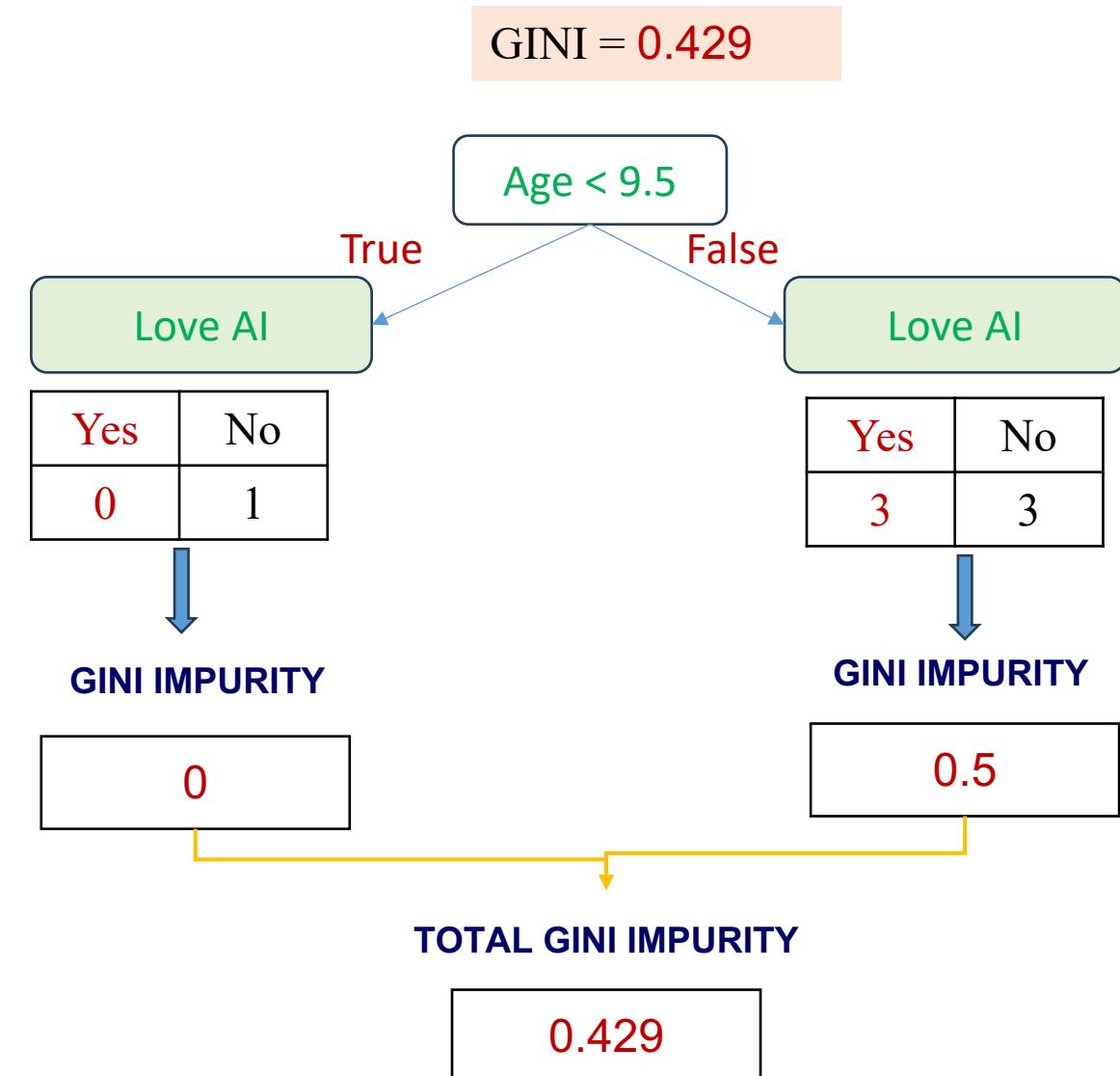
(18 + 12)/2 = 15

(18 + 35)/2 = 26.5

(38 + 35)/2 = 36.5

(38 + 50)/2 = 44

(50 + 83)/2 = 66.5



GINI IMPURITY FOR EACH OF AGE

| Age | Love AI |
|-----|---------|
| 7 | No |
| 12 | No |
| 18 | Yes |
| 35 | Yes |
| 38 | Yes |
| 50 | No |
| 83 | No |

(7 + 12)/2 = 9.5

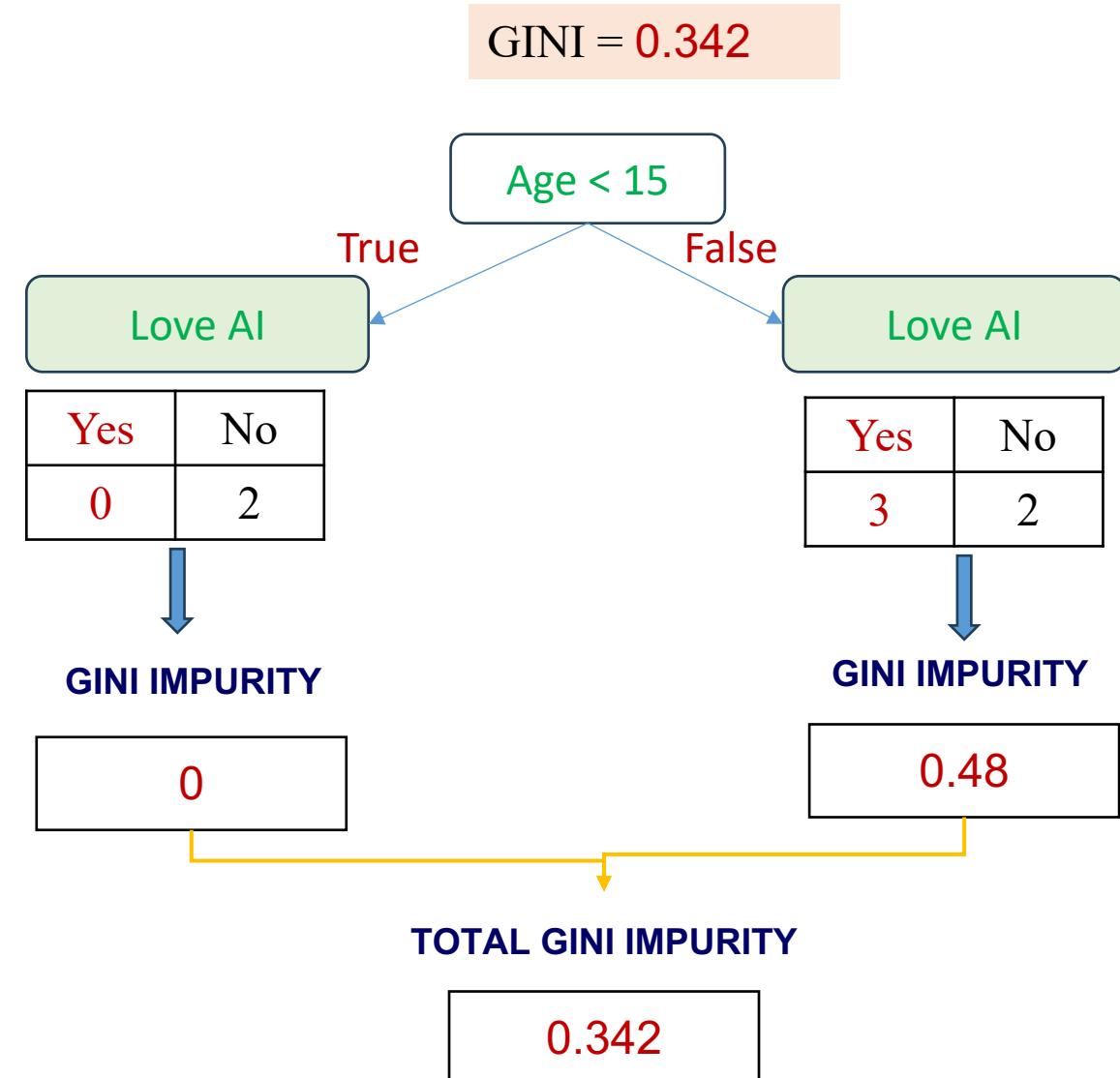
(18 + 12)/2 = 15

(18 + 35)/2 = 26.5

(38 + 35)/2 = 36.5

(38 + 50)/2 = 44

(50 + 83)/2 = 66.5



GINI IMPURITY FOR EACH OF AGE

| Age | Love AI |
|-----|---------|
| 7 | No |
| 12 | No |
| 18 | Yes |
| 35 | Yes |
| 38 | Yes |
| 50 | No |
| 83 | No |

$$(7 + 12)/2 = 9.5$$

$$(18 + 12)/2 = 15$$

$$(18 + 35)/2 = 26.5$$

$$(38 + 35)/2 = 36.5$$

$$(38 + 50)/2 = 44$$

$$(50 + 83)/2 = 66.5$$

GINI for each threshold



GINI

0.429

0.343

0.476

0.476

0.343

0.429

Best GINI



GINI

0.429

0.343

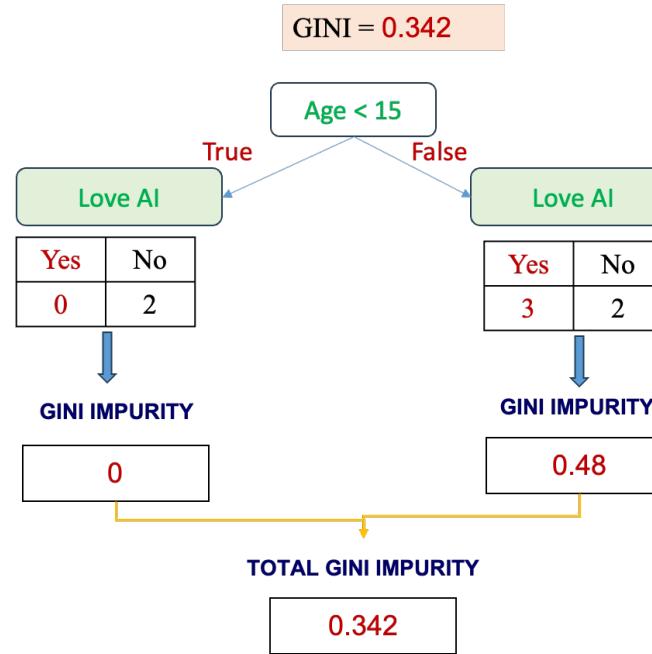
0.476

0.476

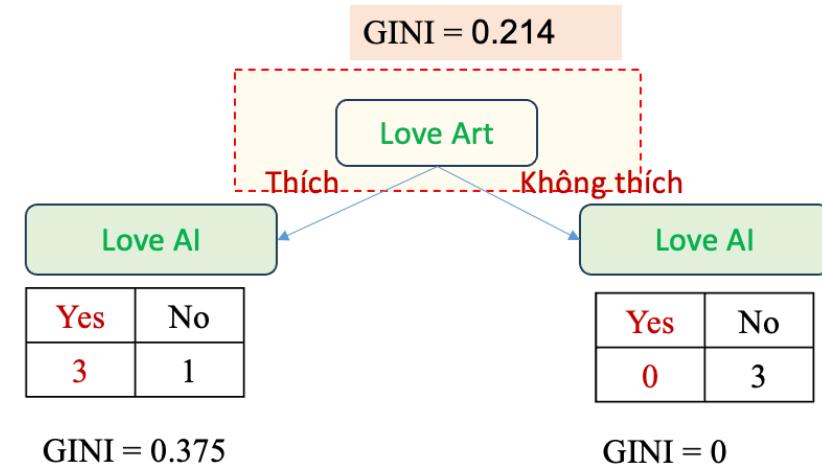
0.343

0.429

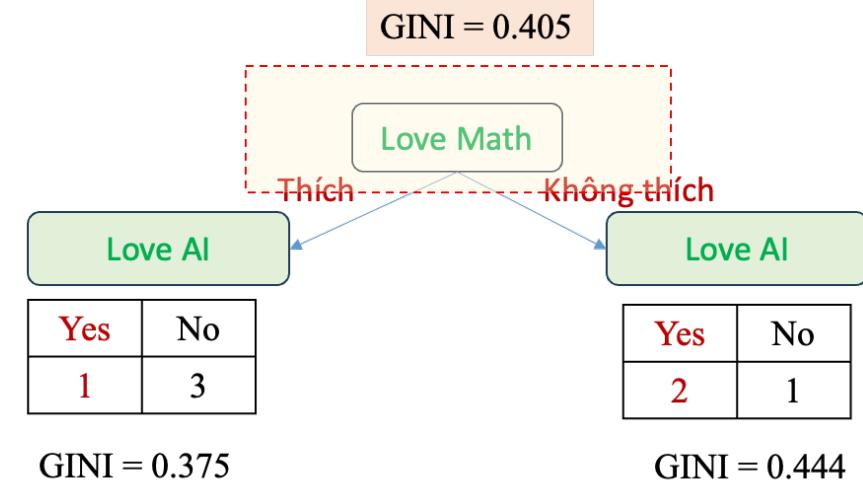
GINI for Three Attributes



Which attribute is in the first node?

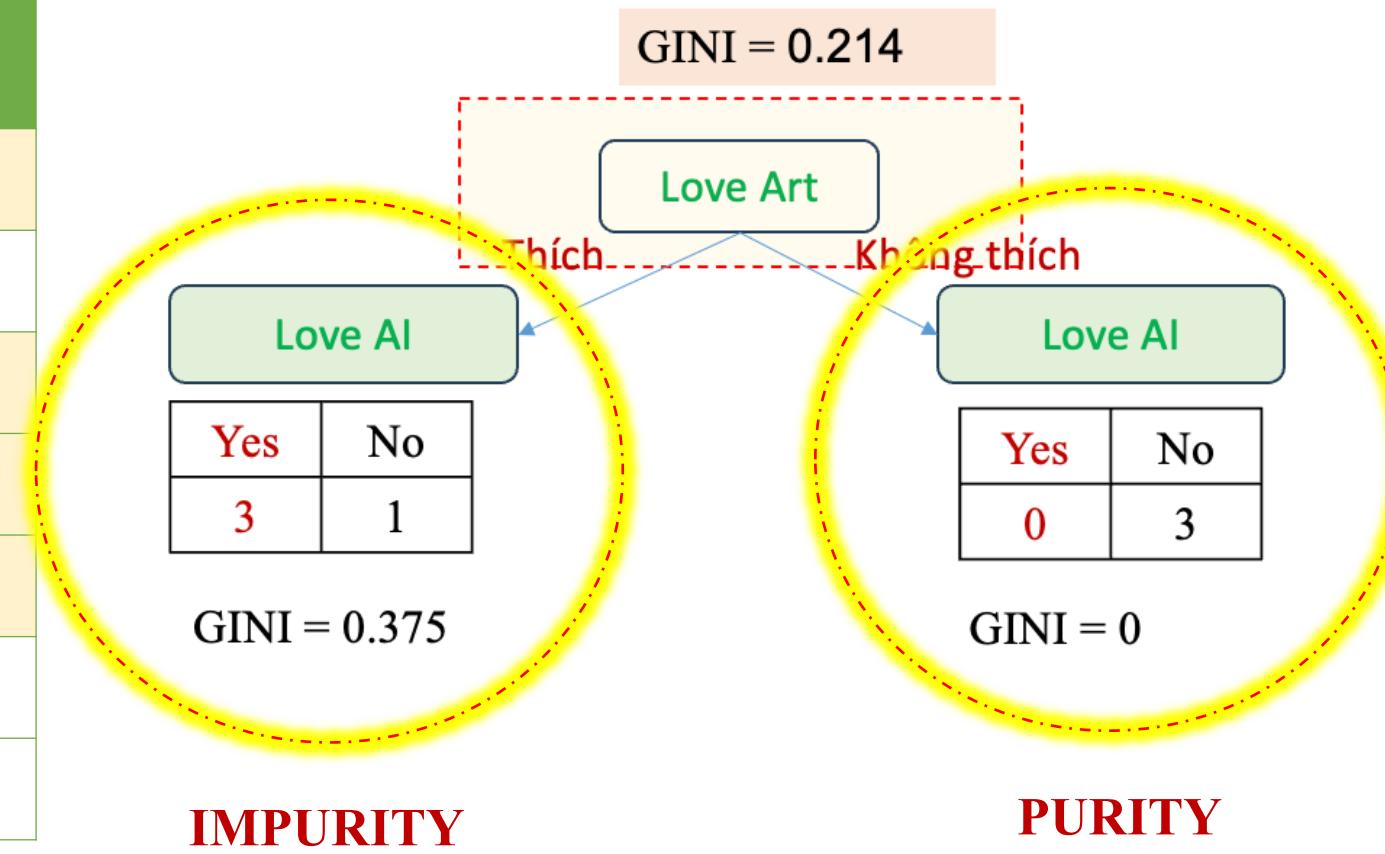


Love Art is the best one



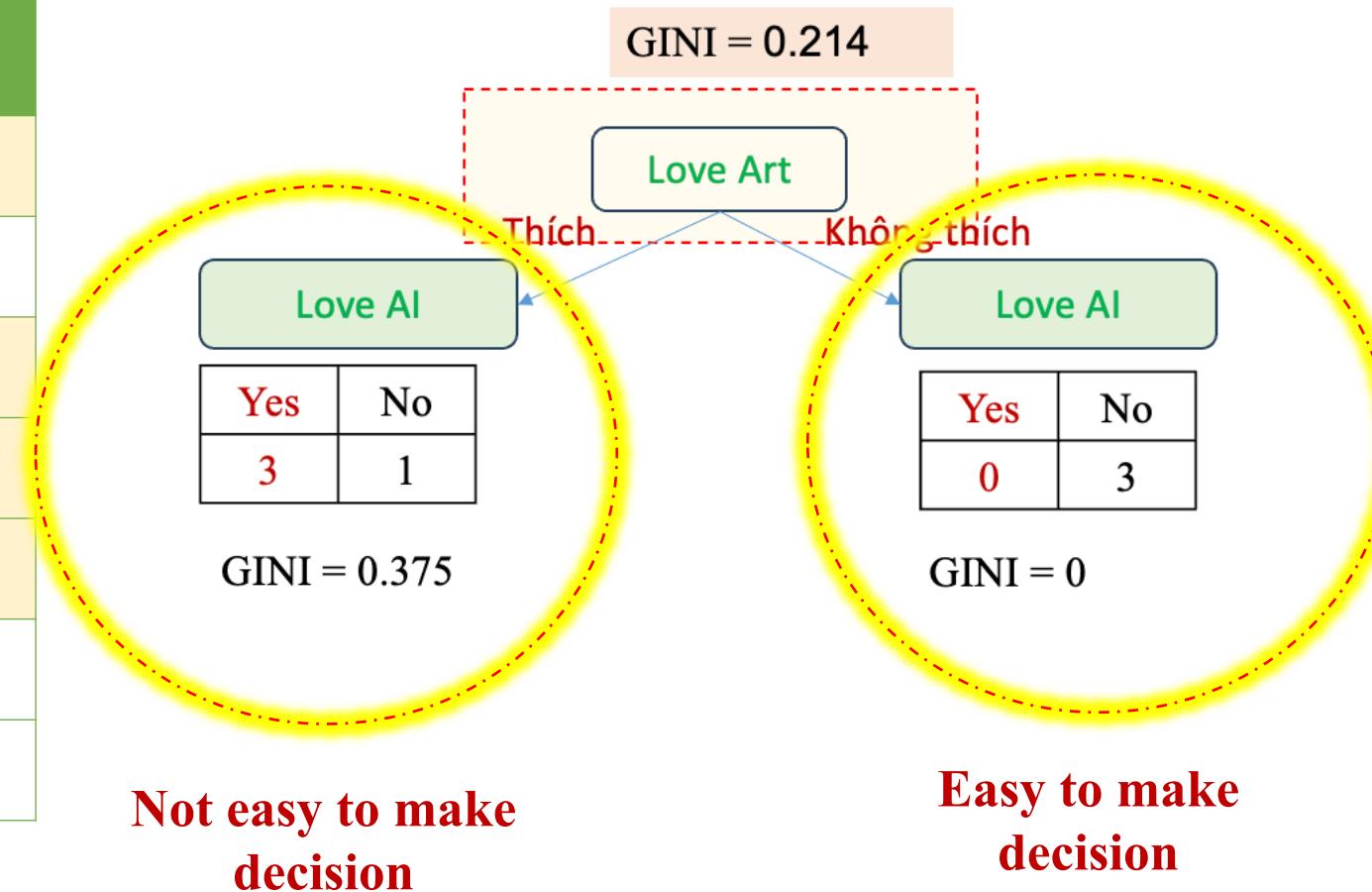
First Node in The Tree: Love Art

| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |



First Node in The Tree: Love Art

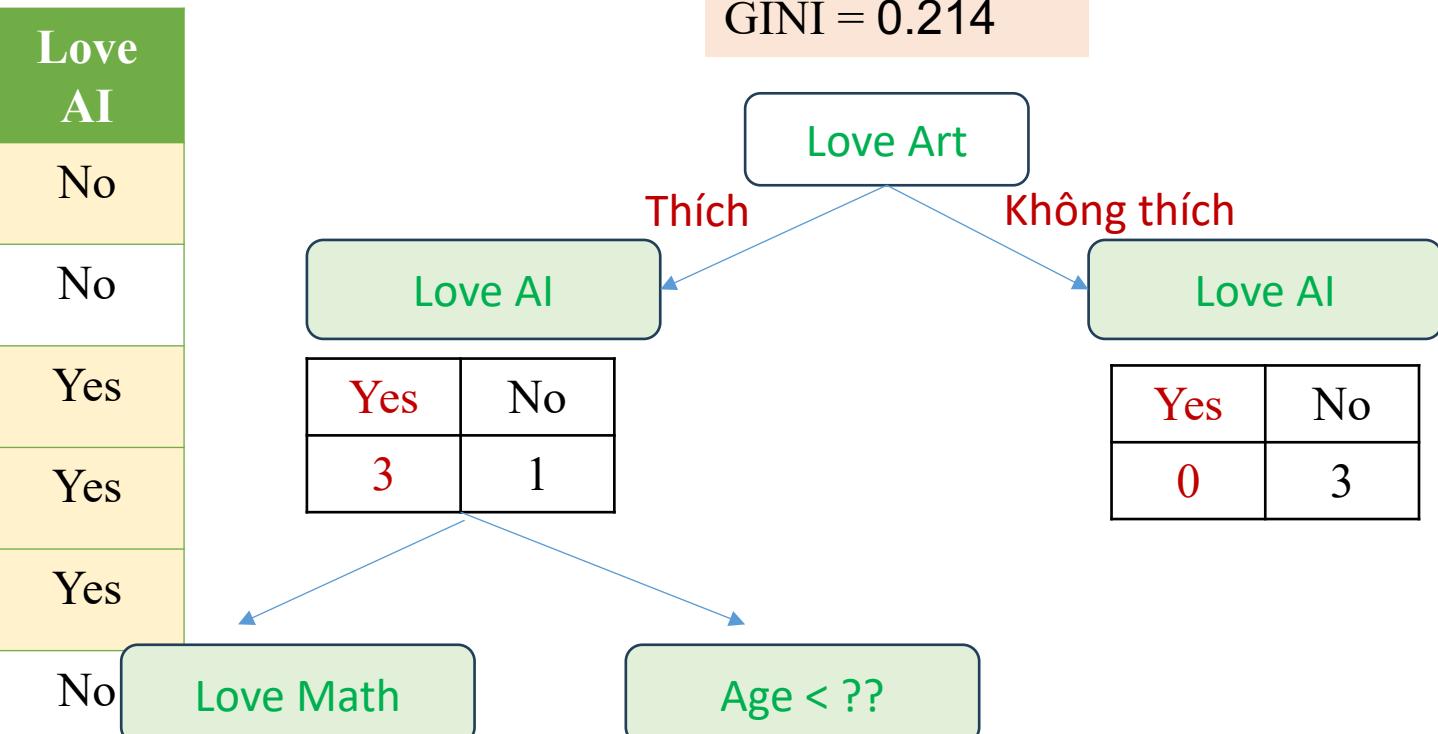
| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |



How to build a Tree

| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |

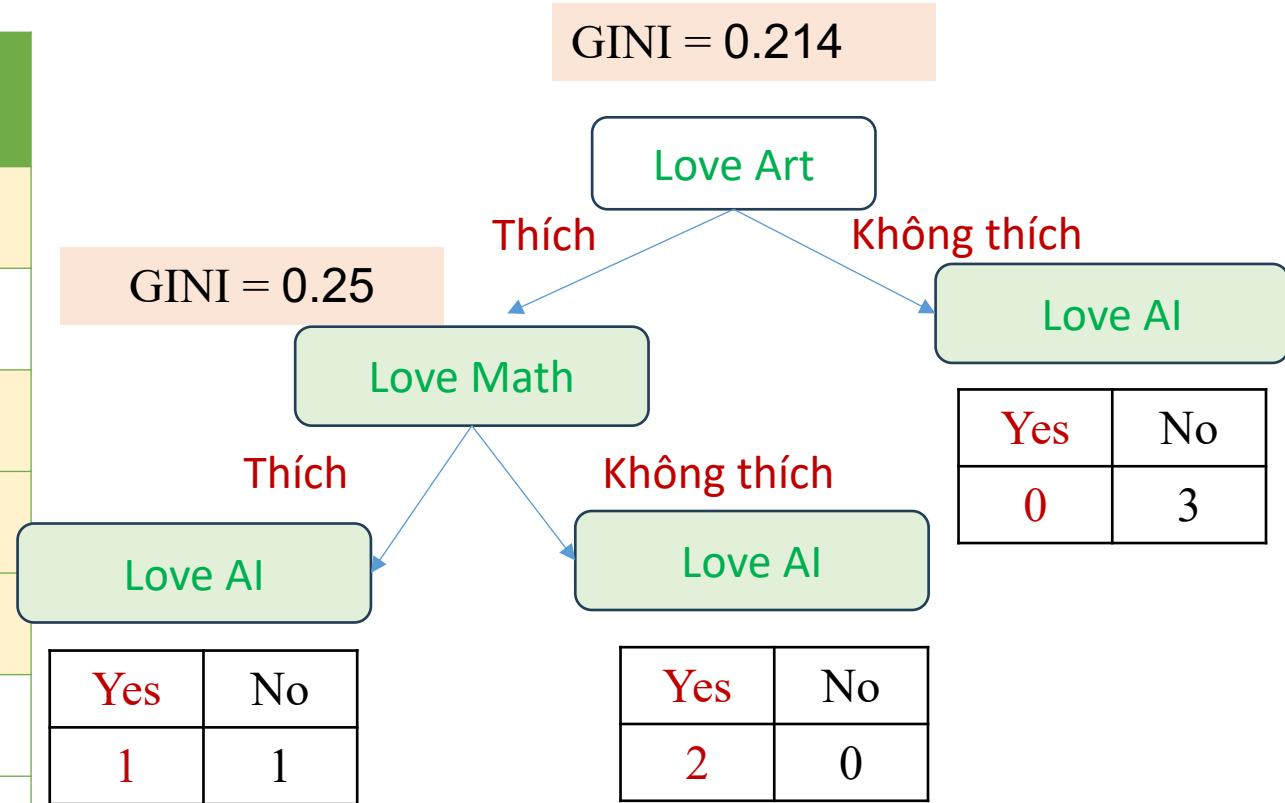
GINI = 0.214



WHICH ONE IS THE NEXT INTERNODE? Love Math or Age

How to build a Tree

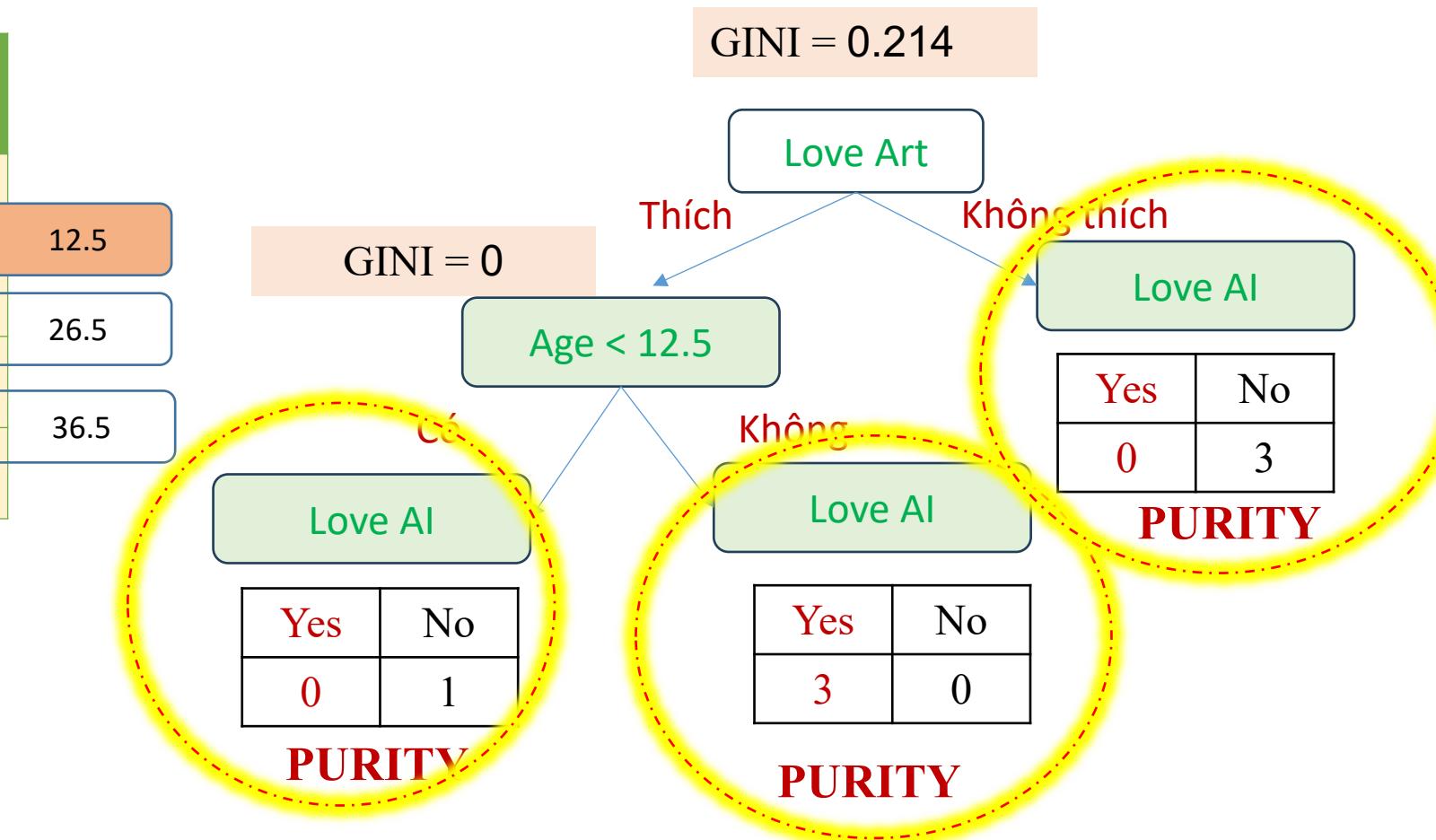
| No. | Love Math | Love Art | Age | Love AI |
|-----|-----------|----------|-----|---------|
| 1 | Yes | Yes | 7 | No |
| 2 | Yes | No | 12 | No |
| 3 | No | Yes | 18 | Yes |
| 4 | No | Yes | 35 | Yes |
| 5 | Yes | Yes | 38 | Yes |
| 6 | Yes | No | 50 | No |
| 7 | No | No | 83 | No |



Giả sử chúng ta chọn “Love Math” là node kế tiếp, cần tính GINI trong trường hợp này

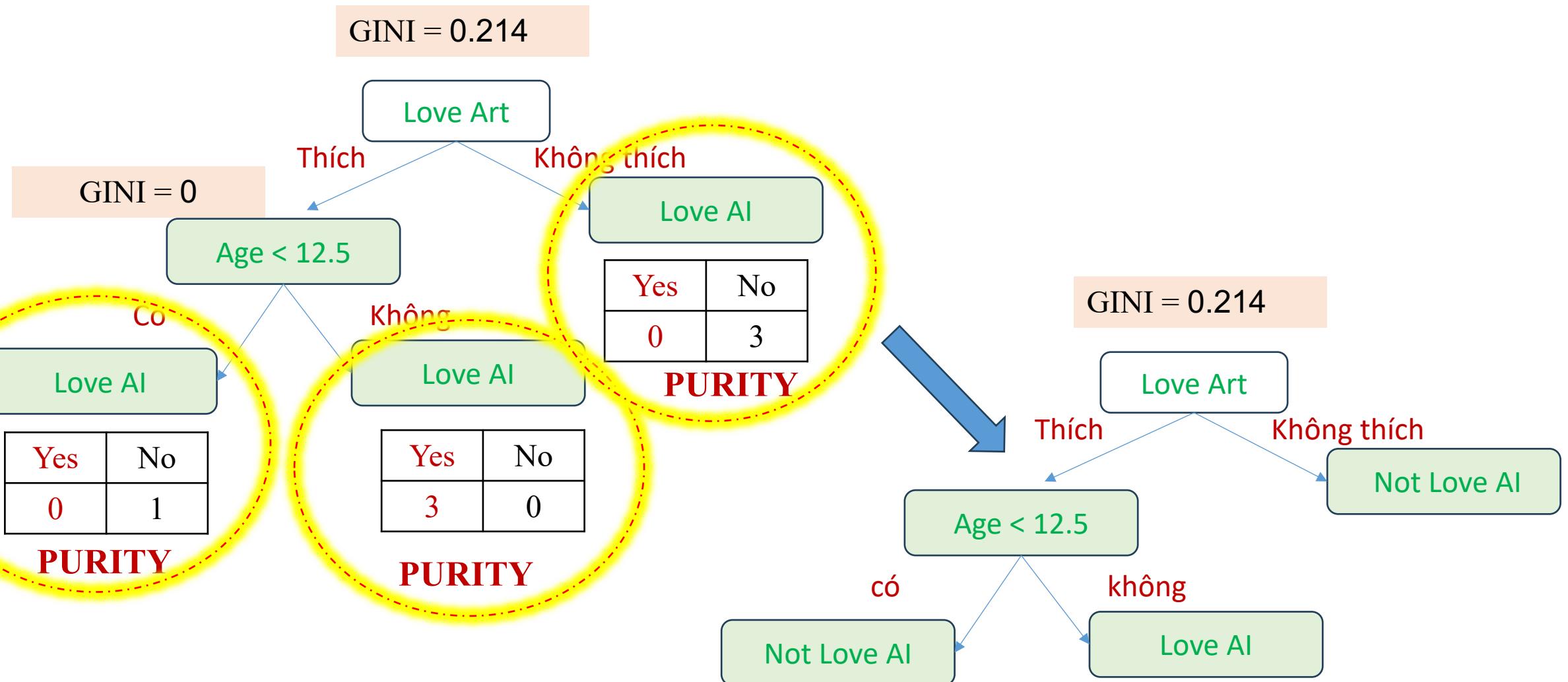
How to build a Tree

| Love Math | Love Art | Age | Love AI |
|-----------|----------|-----|---------|
| Yes | Yes | 7 | No |
| No | Yes | 18 | Yes |
| No | Yes | 35 | Yes |
| Yes | Yes | 38 | Yes |



Dễ dàng đưa ra quyết định trong trường hợp này. Không cần phải tiếp tục xây dựng Tree.

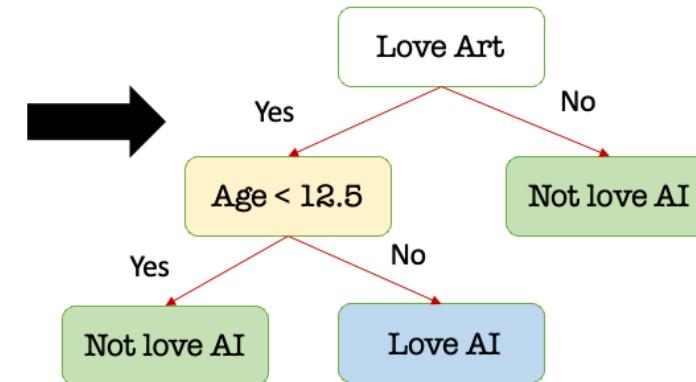
Final Tree



Slides at The First Look

Decision Tree From Dataset

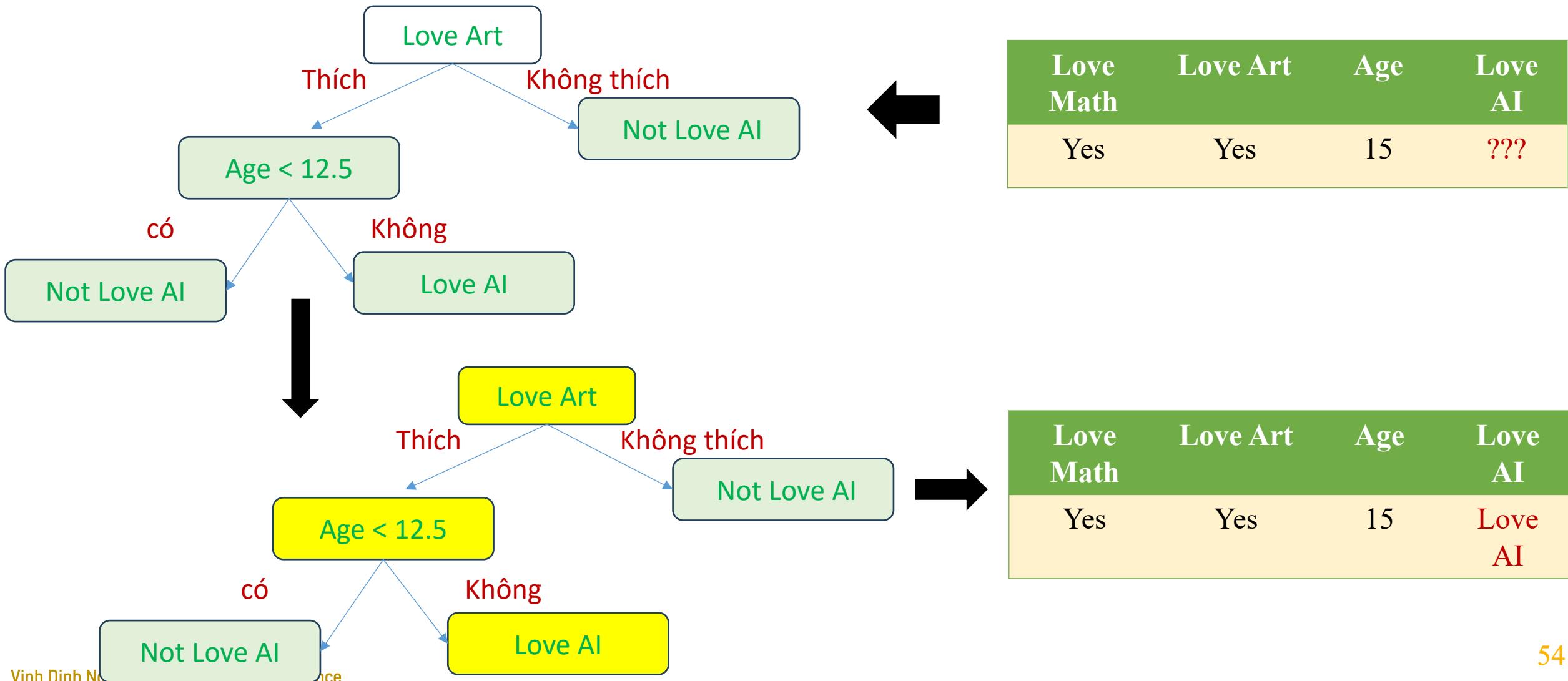
| Love Math | Love Art | Age | Love AI |
|-----------|----------|-----|---------|
| Yes | Yes | 7 | No |
| Yes | No | 12 | No |
| No | Yes | 18 | Yes |
| No | Yes | 35 | Yes |
| Yes | Yes | 38 | Yes |
| Yes | No | 50 | No |
| No | No | 83 | No |



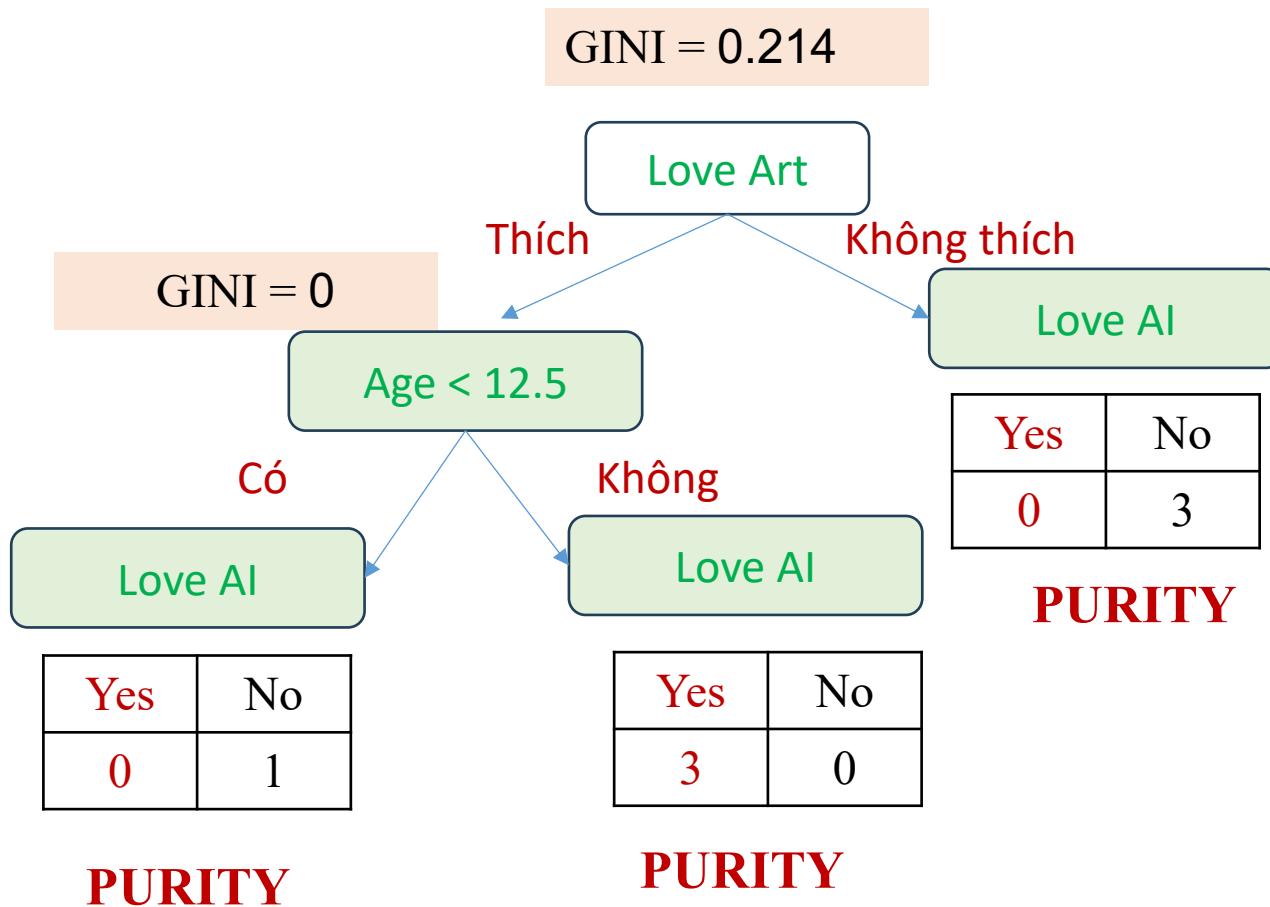
The StatQuest Illustrated Guide To Machine Learning, [Josh Starmer PhD](#)
Vinh Dinh Nguyen- PhD in Computer Science

12

Decision Tree: Prediction



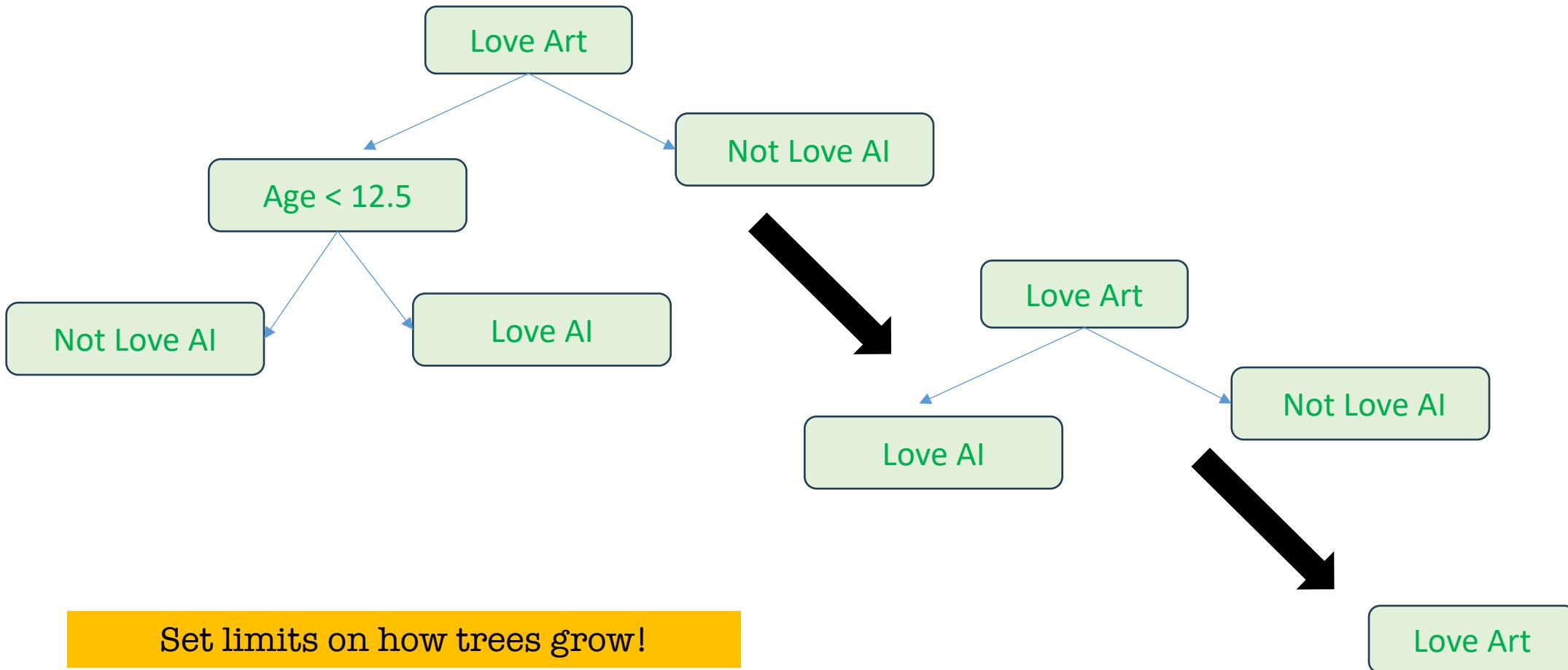
Overfitting in Tree



How to Reduce
overfitting in
Decision Tree

Set limits on how trees grow!

Pruning Algorithm



Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

Evaluation Metrics

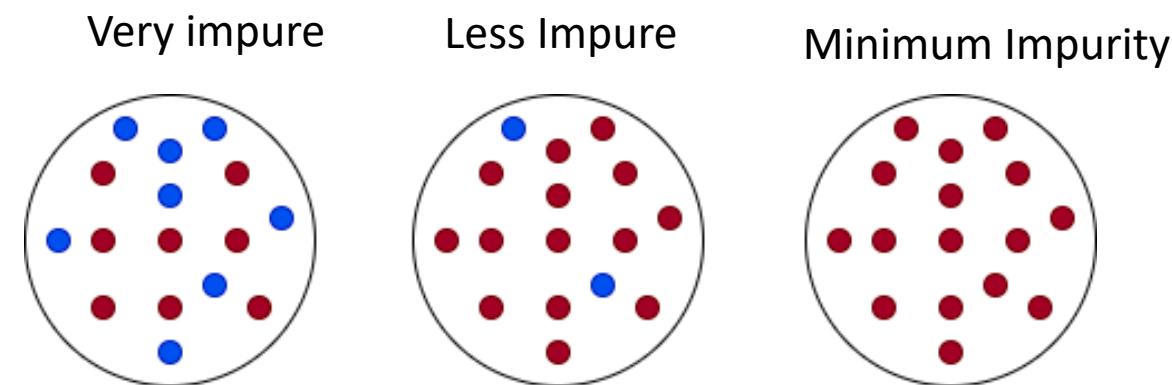
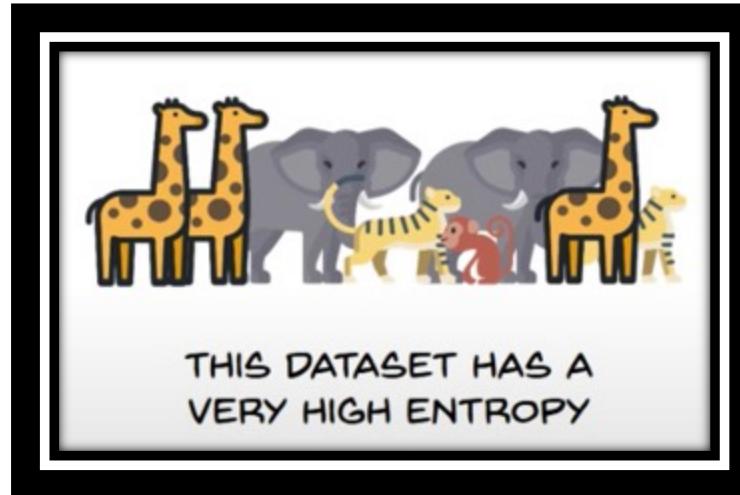
Entropy – Information Gain

GNI IMPURITY



What is Entropy

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations (dataset)



$$E = - \sum_{i=1}^N p_i \log_2 p_i$$



Này là gì vậy?
Nhìn rối quá

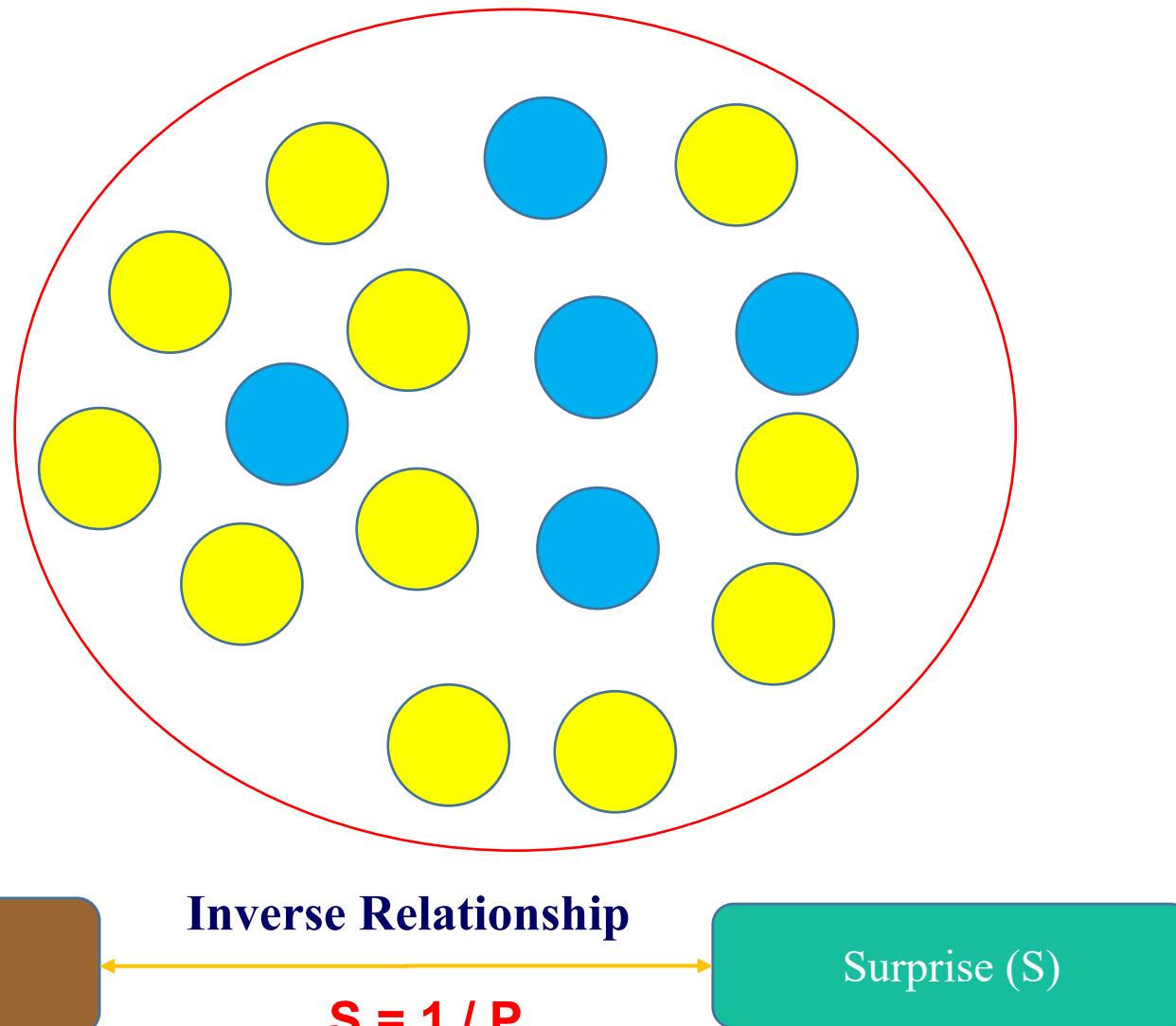
What is Entropy



$$E = - \sum_{i=1}^N p_i \log_2 p_i$$

What is Entropy

- Sample dataset



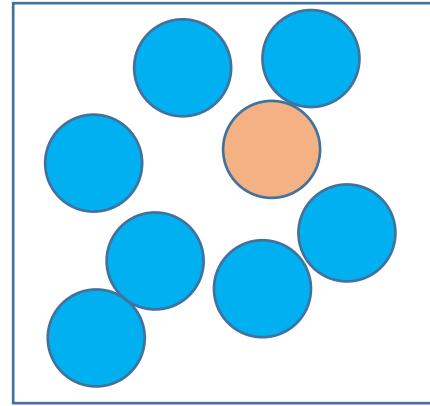
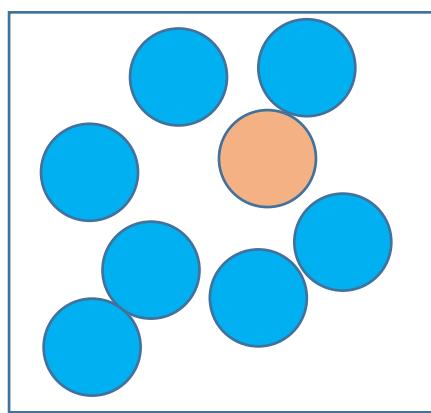
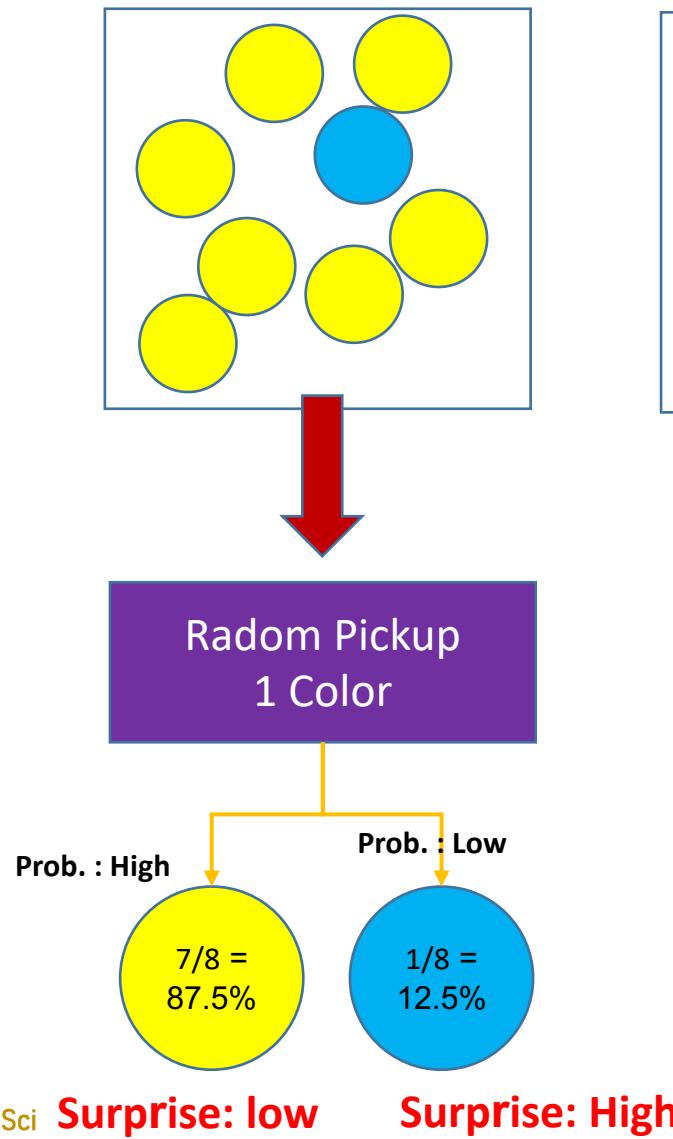
Probability (P)

Surprise (S)

Inverse Relationship

$$S = 1 / P$$

What is Entropy

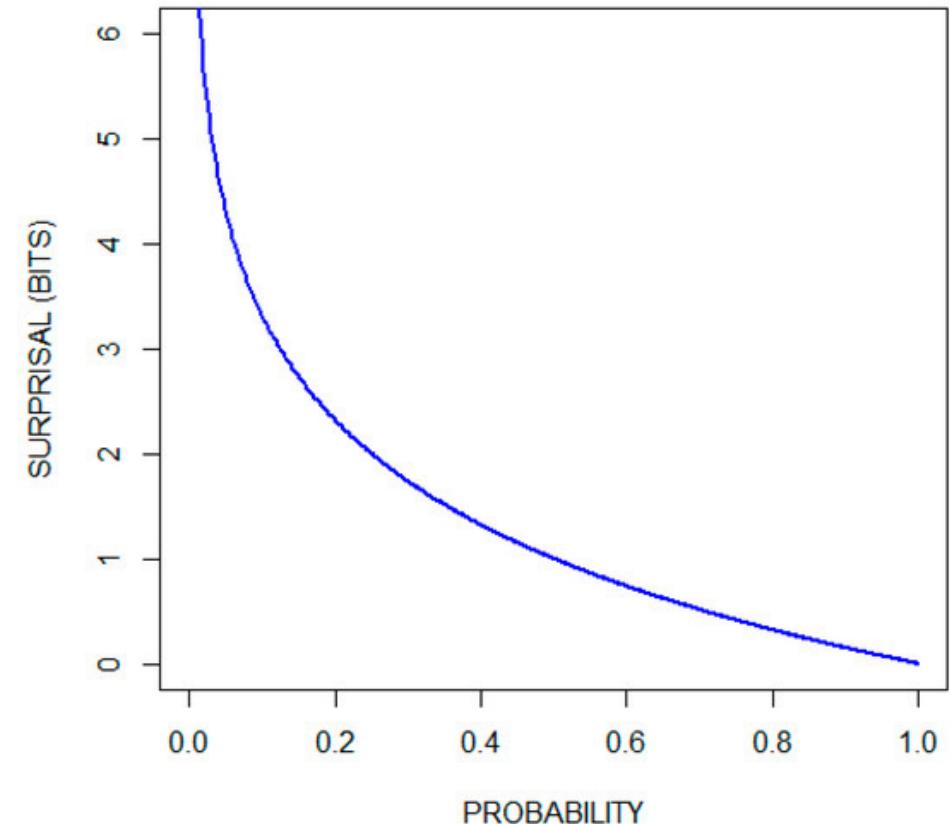


Inverse Relationship: $S = 1/P$

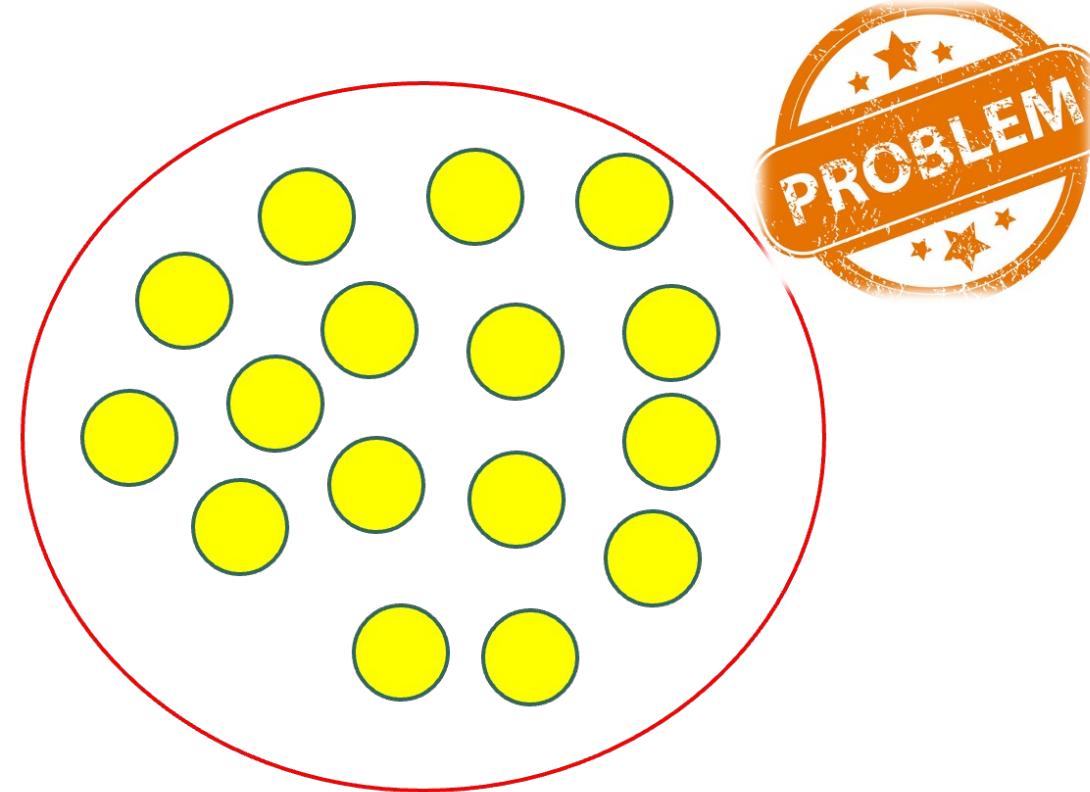


For Yellow: $P = 7/8 \Rightarrow S = 1/(7/8) = 8/7$
For Blue: $P = 1/8 \Rightarrow S = 1/(1/8) = 8$

Probability Vs Surprise



Probability Vs Surprise



$P = \text{to get } \bigcirc = 1$

$S = \text{to get } \bigcirc = 1/P = 1$

S should be 0

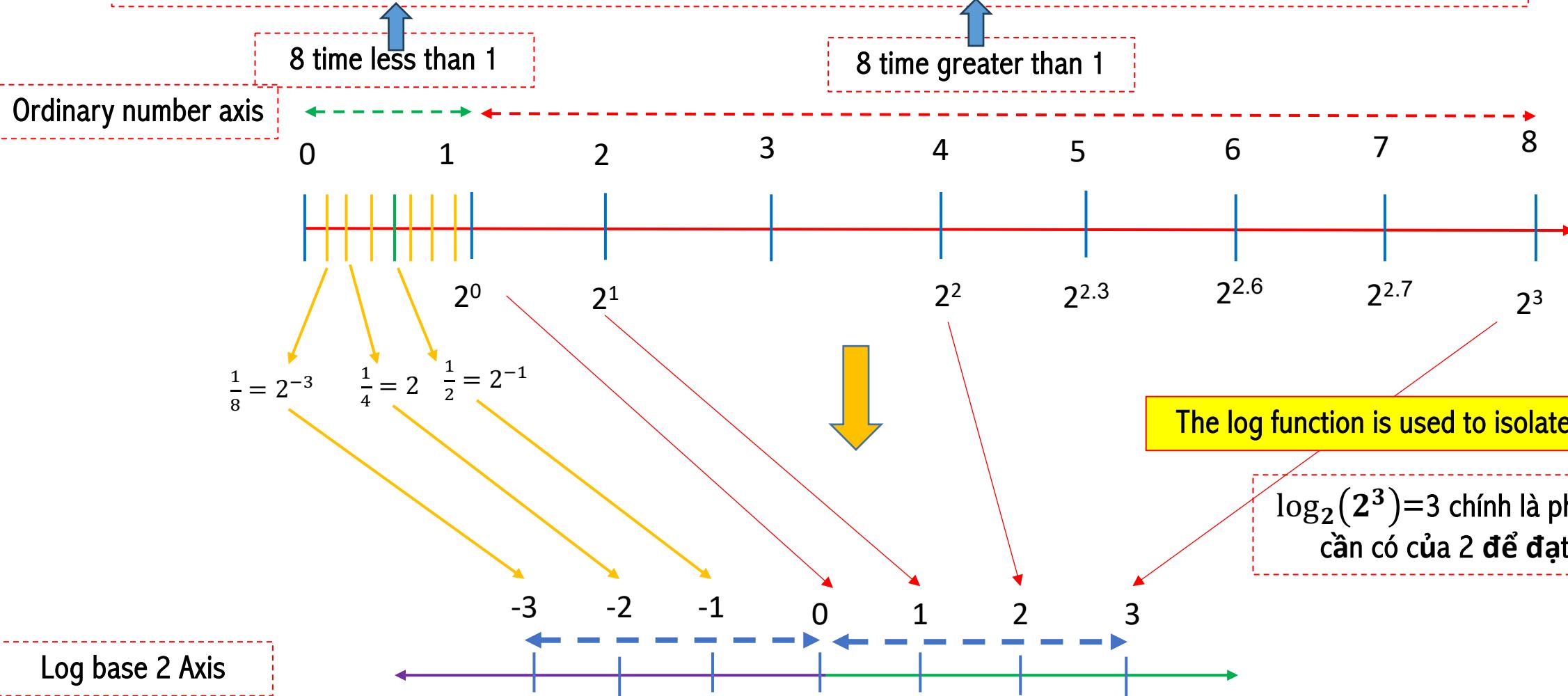
Logs (Logarithms)

SOLUTION



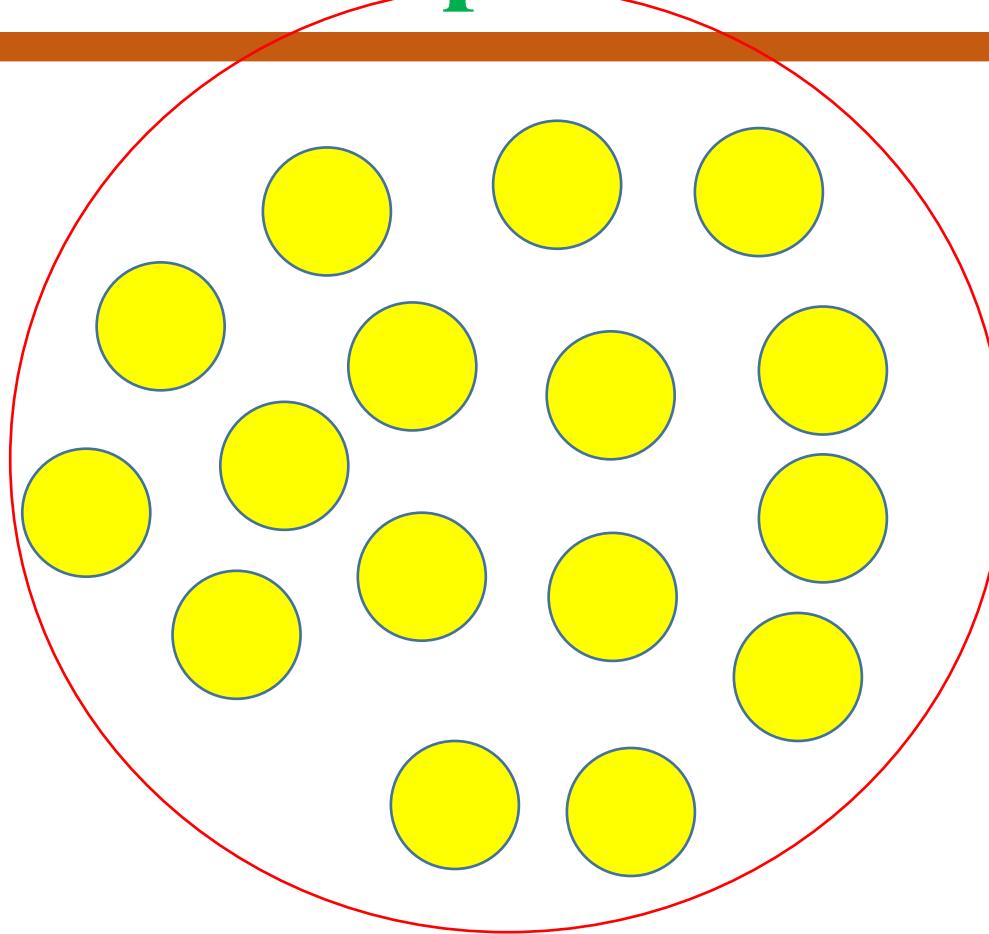
Logs (Logarithms)

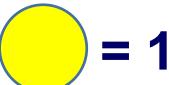
Nhận xét: cả 2 measurement đều thể hiện độ lớn so với 1. Nhưng distance không đối xứng tại 1

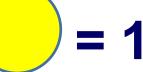


Nhận xét: cả 2 measurement đều thể hiện độ lớn so với 1. Nhưng distance đối xứng tại 1

Sample Dataset

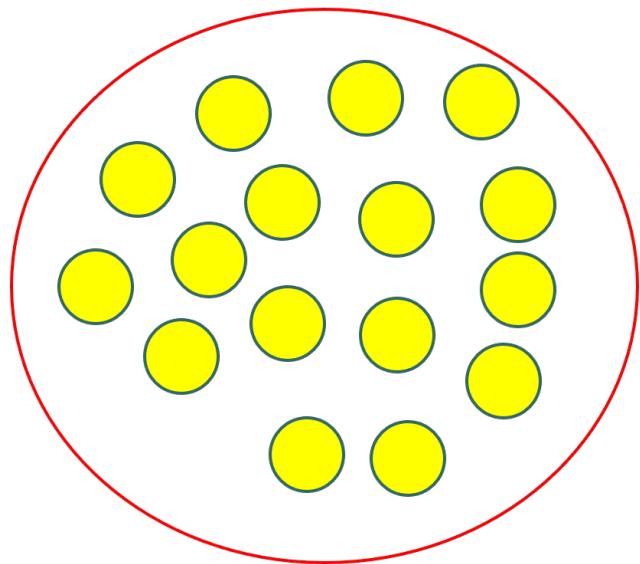


P = to get  = 1

S = to get  = $1/P = 1$

We use the log of $1/P$ to calculate S

Sample Dataset



$$P = \text{to get } \text{yellow circle} = 1$$

$$S = \text{to get } \text{yellow circle} = \log_2(1/1) = 0$$

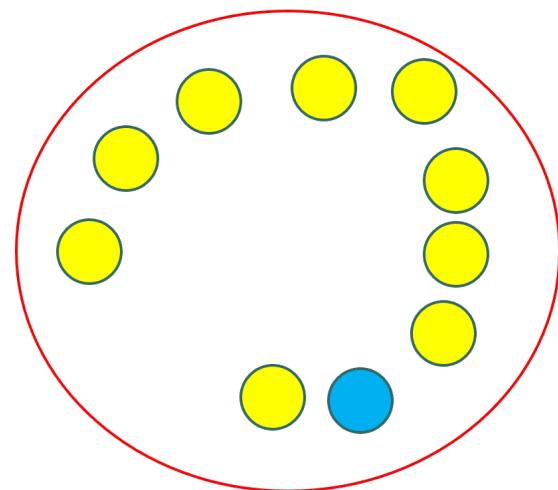
No surprise

$$P = \text{to get } \text{blue circle} = 0$$

$$S = \text{to get } \text{blue circle} = \log_2(1/0) = \log_2(1) - \log_2(0) = \text{Undefined}$$

Big surprise

Sample Dataset



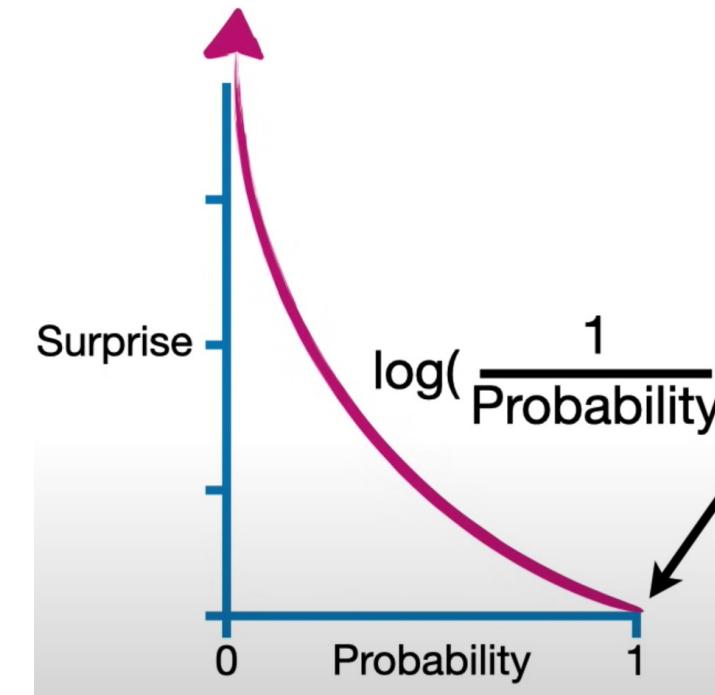
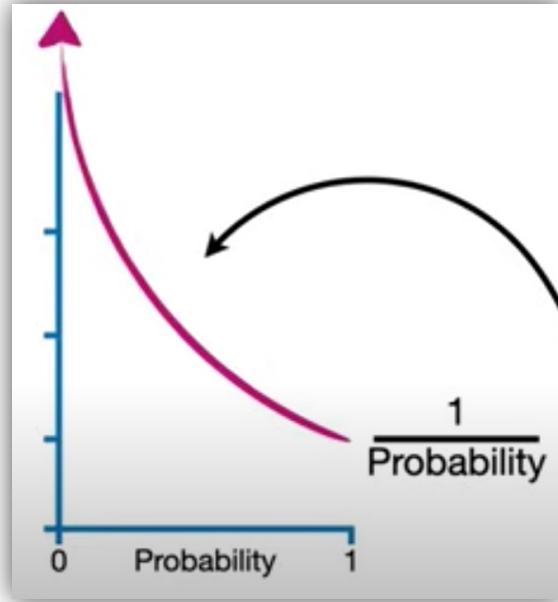
$$P = \text{to get } \text{yellow circle} = 0.9$$

$$S = \text{to get } \text{yellow circle} = \log_2(1/0.9) = 0.15$$

$$P = \text{to get } \text{blue circle} = 0.1$$

$$S = \text{to get } \text{blue circle} = \log_2(1/0.1) = 3.32$$

Probability vs Surprise

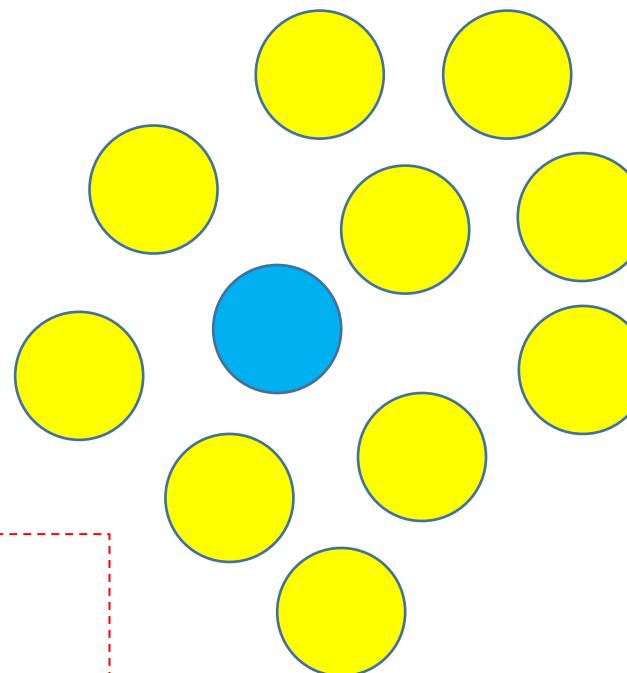


$$S = \log_2(1/P)$$

Sample Dataset

P to get  = 9/10

P to get  = 1/10



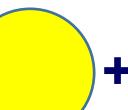
S to get  = $\log_2(1 / 0.9) = 0.15$

S to get  = $\log_2(1 / 0.1) = 3.32$

$$\begin{aligned} \log_2 \frac{1}{0.9 \times 0.9 \times 1} &= \log_2 1 - \log_2 (0.9 \times 0.9 \times 1) \\ &= 0 - \log_2 (0.9 \times 0.9 \times 1) \\ &= 0 - \log_2 (0.9) - \log_2 (0.9) - \log_2 (1.0) \\ &= 0.15 + 0.15 + 3.32 \end{aligned}$$

Flip the coin 3 times

P to get    = **0.9 * 0.9 * 0.1**

S to get    = **S**  + **S**  + **S** 

Gợi nhớ!

If two events A and B are independent , then the probability of happening of both A and B is:

$$P(A \cap B) = P(A) \cdot P(B)$$

Sample Dataset

| | | |
|-------------|---|---|
| |  |  |
| Probability | 9/10 | 1/10 |
| Surprise | 0.15 | 3.32 |

**HOW MUCH SURPRISE
FROM GETTING YELLOW COLORS IN 100 TIMES**

$$(0.9*100) * 0.15$$

Expected number of yellows

Total surprise from getting
Yellow

Sample Dataset

| | | |
|-------------|---|---|
| |  |  |
| Probability | 9/10 | 1/10 |
| Surprise | 0.15 | 3.32 |

**HOW MUCH SURPRISE
FROM GETTING BLUE COLORS IN 100 TIMES**

$$(0.1 * 100) * 3.32$$

Expected number of blues

Total surprise from getting Blue

AVERAGE SURPRISE FOR 100 TIMES

$S_B =$ HOW MUCH SURPRISE
FROM GETTING BLUE COLORS IN 100 TIMES

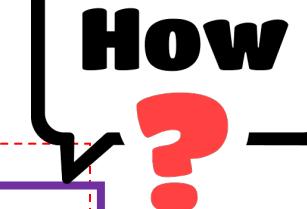
$$(0.1 * 100) * 3.32$$



$S_Y =$ HOW MUCH SURPRISE
FROM GETTING YELLOW COLORS IN 100 TIMES

$$(0.9 * 100) * 0.15$$

$$E = - \sum_{i=1}^N p_i \log_2 p_i$$



Total Surprise = $S_B + S_Y = (0.1 \times 100) \times 3.32 + (0.9 \times 100) \times 0.15 = 46.7$

Average = Total Surprise / 100 = $(S_B + S_Y = (0.1 \times 100) \times 3.32 + (0.9 \times 100) \times 0.15) / 100 = 0.467$

Entropy = Total Surprise / 100 = $(0.1 \times 3.32) + (0.9 \times 0.15) = 0.467$

$p(x)$

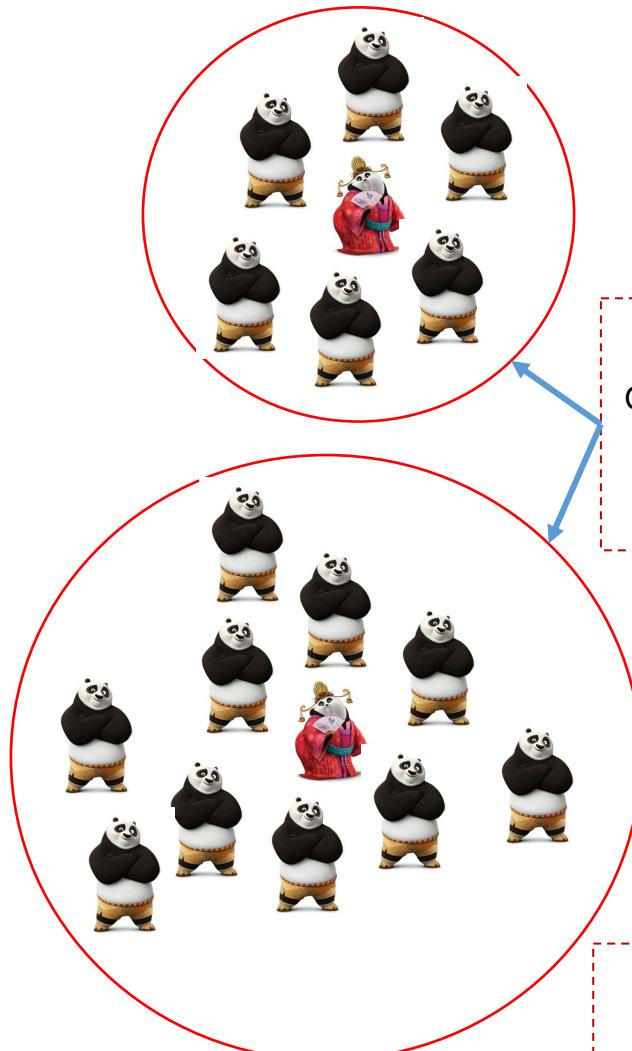
$\log\left(\frac{1}{p(x)}\right)$



Entropy = $\sum \log\left(\frac{1}{p(x)}\right) p(x)$

↑
Surprise ↑
The probability of the Surprise.

Sample Dataset

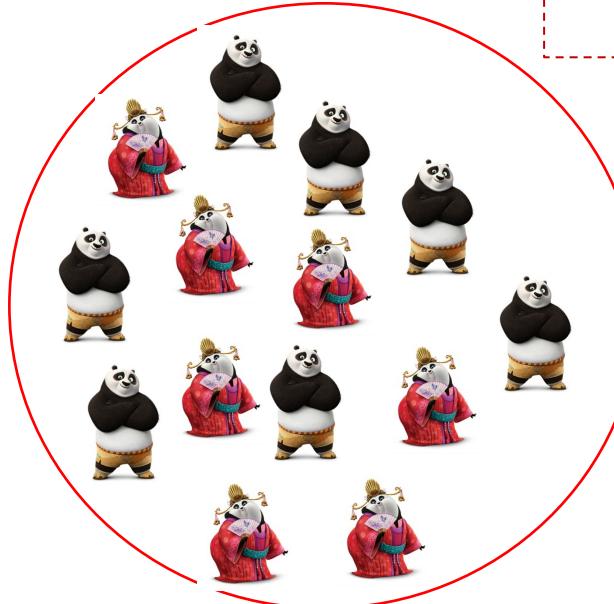


$$\text{Entropy} = \sum p(x) \log \left(\frac{1}{p(x)} \right) = \frac{6}{7} \log_2 \frac{7}{6} + \frac{1}{7} \log_2 7$$

Entropy = 0.59

When we increase the difference in the number of male and female panda.
Entropy is lower

Entropy = 0.44

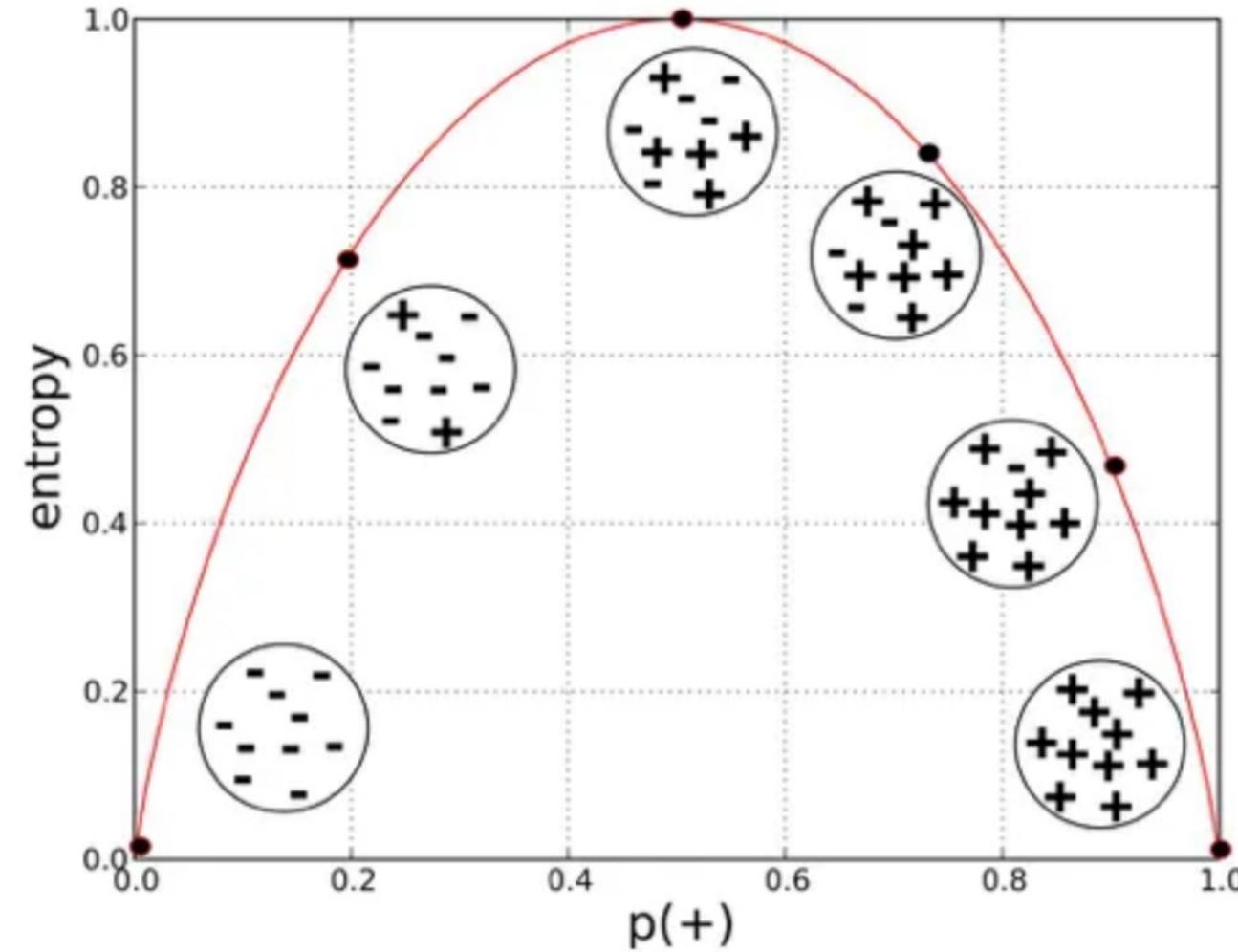


Entropy is highest when we have same number of female and male panda

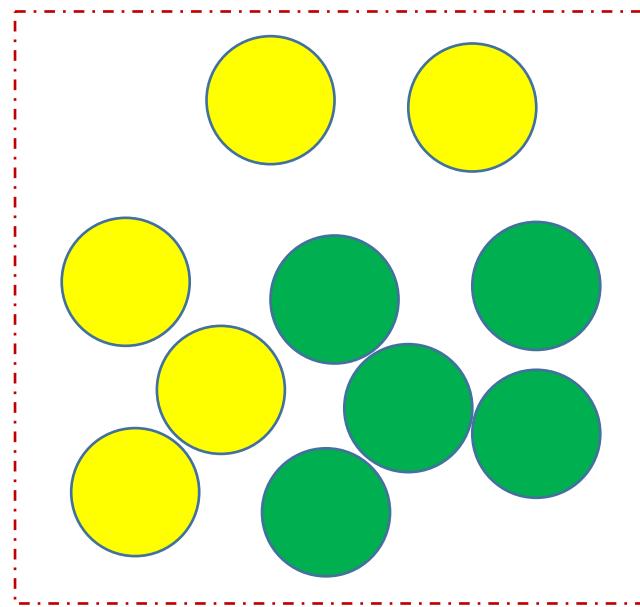
Entropy = 1.0

Entropy is used to quantify the similarity or difference in the dataset

Entropy and Probability

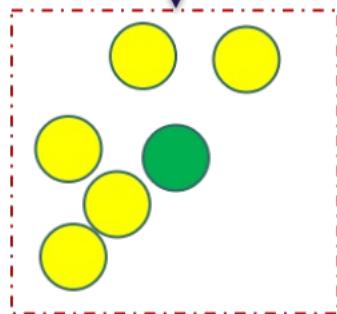
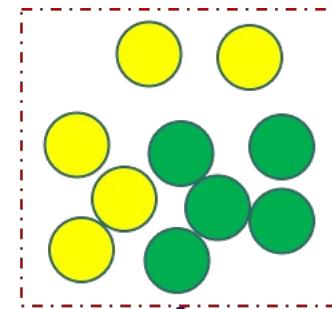


Sample Dataset

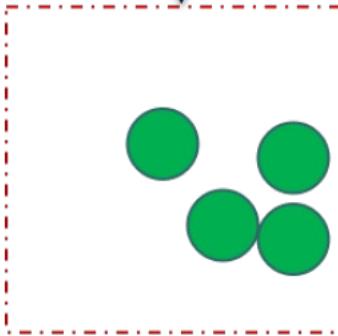


E= 0.65

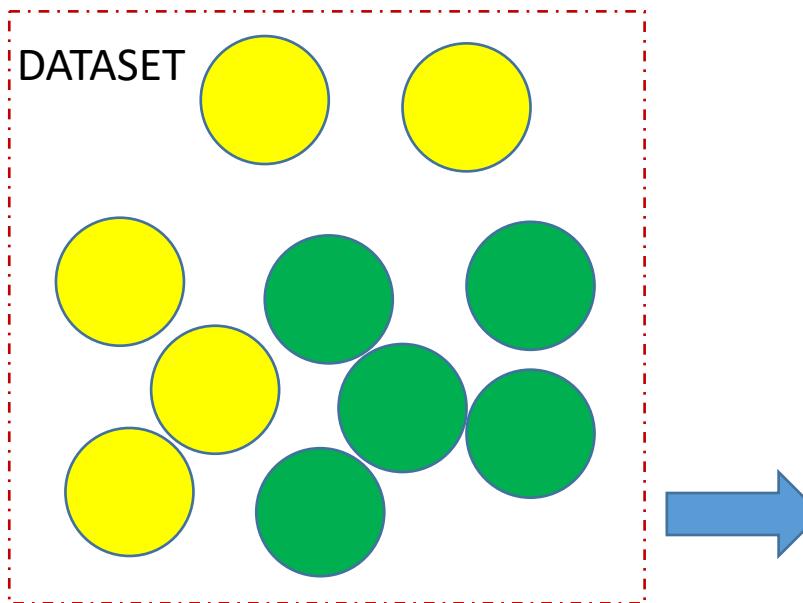
$$E_{initial} = -((0.5\log_2 0.5) + (0.5\log_2 0.5)) \\ = 1$$



E= 0



Information Gain (IG)



$$E_{\text{after_slit}} = 0.65 * 0.6 + 0.4 * 0 = 0.39$$

$$\text{Gain} = 0.61 = E_{\text{initial}} - E_{\text{after_slit}}$$

$$E_{\text{initial}} = -((0.5 \log_2 0.5) + (0.5 \log_2 0.5))$$

$$= 1$$

Maximum the information gain

$$IG(Y, X) = E(Y) - E(Y|X)$$

$$\text{Gain} = E_{\text{parent}} - E_{\text{children}}$$

$$E = 0.65$$

$$E = 0$$



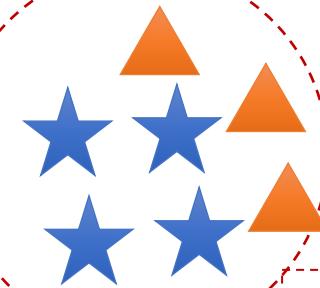
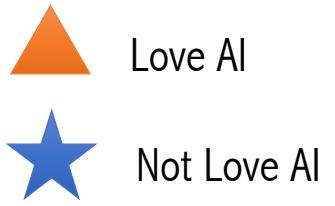
Example

| No. | Love Math | Love Art | Love AI |
|-----|-----------|----------|---------|
| 1 | Yes | Yes | No |
| 2 | Yes | No | No |
| 3 | No | Yes | Yes |
| 4 | No | Yes | Yes |
| 5 | Yes | Yes | Yes |
| 6 | Yes | No | No |
| 7 | No | No | No |

Cần xác định Root node
nên là Love Math hay Love Art?

Giả sử ta chọn root node
là Love Math?

Entire Population



$$\text{Entropy}_{\text{before}} = -\frac{3}{7} \log \left(\frac{3}{7} \right) - \frac{4}{7} \log \left(\frac{4}{7} \right) = 0.985$$

Love Math is Yes



$$\text{Entropy}_{\text{after}} = \frac{4}{7} \times 0.811 + \frac{3}{7} \times 0.918 = 0.856$$

Love Math is No



$$\text{Information Gain} = 0.985 - 0.856 = 0.129$$

$$\text{Entropy} = -\frac{3}{4} \log \left(\frac{3}{4} \right) - \frac{1}{4} \log \left(\frac{1}{4} \right) = 0.811$$

$$\text{Entropy} = -\frac{3}{4} \log \left(\frac{3}{4} \right) - \frac{1}{4} \log \left(\frac{1}{4} \right) = 0.918$$

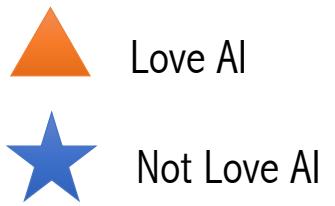
Example

| No. | Love Math | Love Art | Love AI |
|-----|-----------|----------|---------|
| 1 | Yes | Yes | No |
| 2 | Yes | No | No |
| 3 | No | Yes | Yes |
| 4 | No | Yes | Yes |
| 5 | Yes | Yes | Yes |
| 6 | Yes | No | No |
| 7 | No | No | No |

Cần xác định Root node
nên là Love Math hay Love Art?

Giả sử ta chọn root node
là Love Art?

Entire Population



$$\text{Entropy}_{\text{before}} = -\frac{3}{7} \log \left(\frac{3}{7} \right) - \frac{4}{7} \log \left(\frac{4}{7} \right) = 0.985$$

Love Art is Yes



Love Art is No



$$\text{Entropy}_{\text{after}} = \frac{4}{7} \times 0.811 + \frac{3}{7} \times 0 = 0.463$$

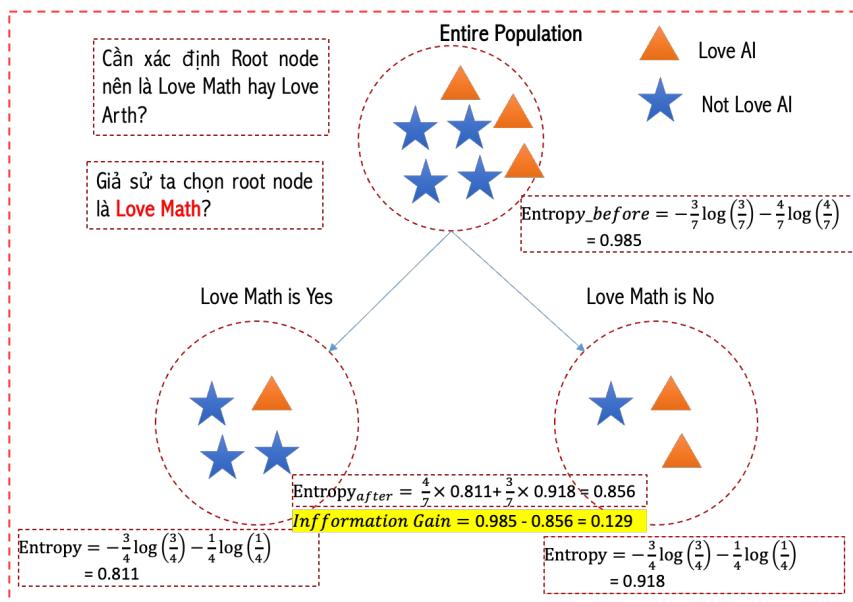
$$\text{Information Gain} = 0.985 - 0.463 = 0.522$$

$$\text{Entropy} = -\frac{3}{4} \log \left(\frac{3}{4} \right) - \frac{1}{4} \log \left(\frac{1}{4} \right) = 0.811$$

$$\text{Entropy} = -\frac{3}{3} \log \left(\frac{3}{3} \right) = 0$$

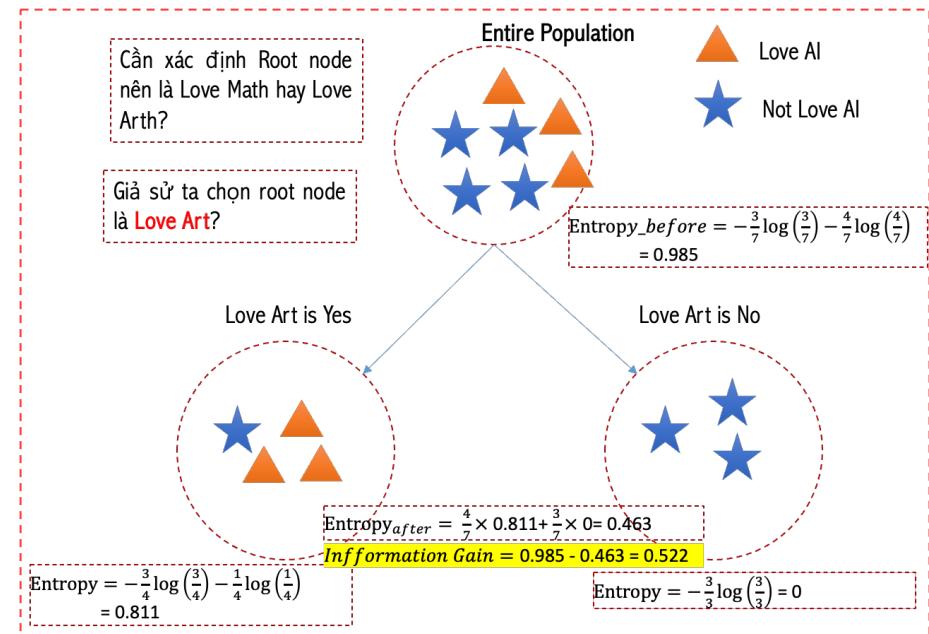
Example

Love Math is a Root Node



$$Information\ Gain = 0.129$$

Love Art is a Root Node



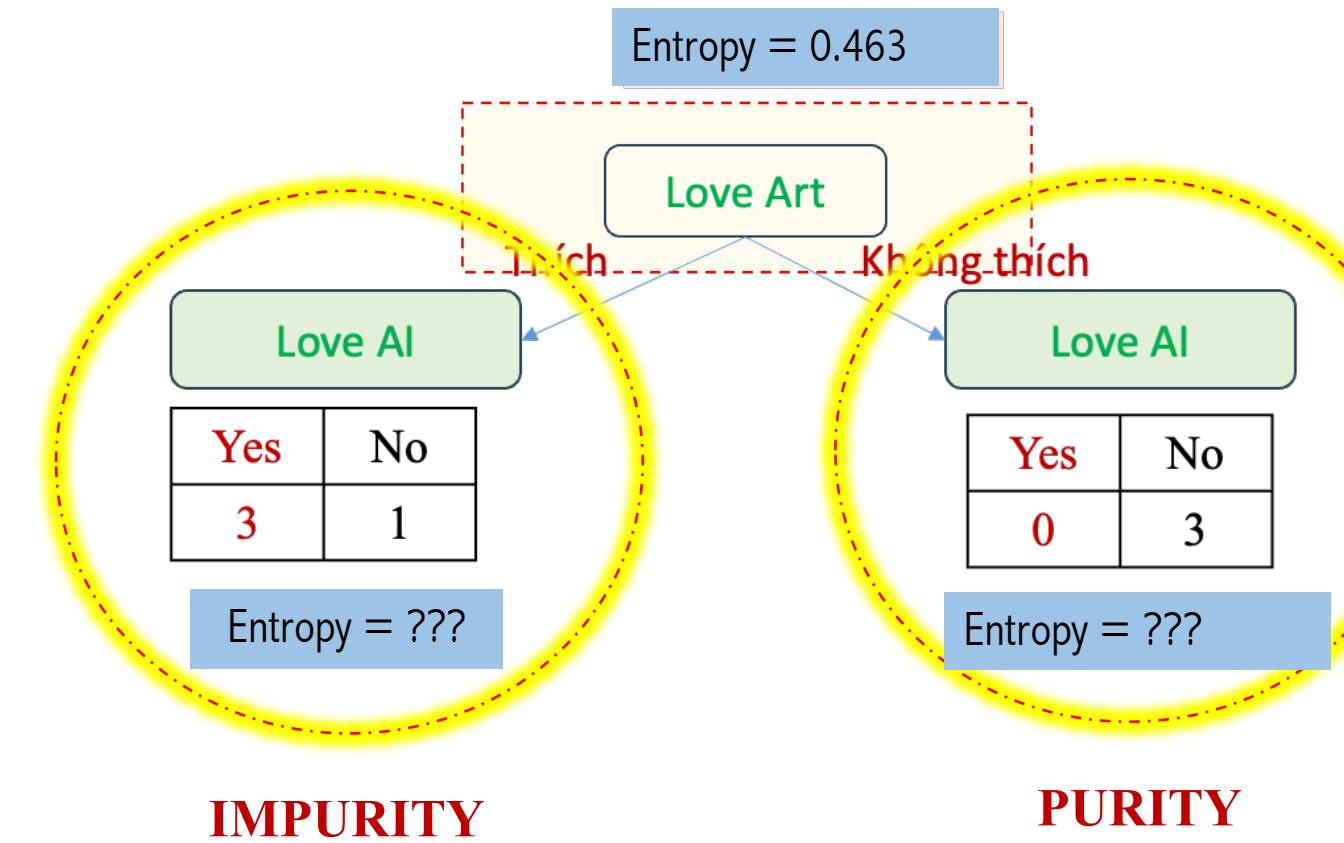
$$Information\ Gain = 0.522$$

Which attribute is in the first node?

Maximum the information gain

Example

| No. | Love Math | Love Art | Love AI |
|-----|-----------|----------|---------|
| 1 | Yes | Yes | No |
| 2 | Yes | No | No |
| 3 | No | Yes | Yes |
| 4 | No | Yes | Yes |
| 5 | Yes | Yes | Yes |
| 6 | Yes | No | No |
| 7 | No | No | No |



Continue to expand this tree using Entropy metric

Outline

- **Introduction to Tree**
- **Decision Tree**
- **Decision Tree with Gini**
- **Decision Tree with Entropy**
- **Several Examples**

Example

| | age | income | student | credit_rate | Default |
|----|------------|--------|---------|-------------|---------|
| 0 | youth | high | no | fair | no |
| 1 | youth | high | no | excellent | no |
| 2 | middle_age | high | no | fair | yes |
| 3 | senior | medium | no | fair | yes |
| 4 | senior | low | yes | fair | yes |
| 5 | senior | low | yes | excellent | no |
| 6 | middle_age | low | yes | excellent | yes |
| 7 | youth | medium | no | fair | no |
| 8 | youth | low | yes | fair | yes |
| 9 | senior | medium | yes | | |
| 10 | youth | medium | yes | | |
| 11 | middle_age | medium | no | | |
| 12 | middle_age | high | yes | | |
| 13 | senior | medium | no | | |

This data set is used to predict whether a person will default on their credit card. There are two classes (default = 'yes', no_default = 'no'):

```
# Defining a simple dataset
attribute_names = ['age', 'income', 'student', 'credit_rate']
class_name = 'default'
data1 = {
    'age' : ['youth', 'youth', 'middle_age', 'senior', 'senior', 'middle_age', 'youth', 'youth', 'senior',
             'income' : ['high', 'high', 'high', 'medium', 'low', 'low', 'low', 'medium', 'low', 'medium', 'medium', 'medium',
                         'student' : ['no', 'no', 'no', 'yes', 'yes', 'yes', 'no', 'yes', 'yes', 'no', 'yes', 'no'],
                         'credit_rate' : ['fair', 'excellent', 'fair', 'fair', 'fair', 'excellent', 'fair', 'fair', 'fair', 'fair',
                                         'default' : ['no', 'no', 'yes', 'yes', 'no', 'yes', 'no', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'no']]}
df1 = pd.DataFrame (data1, columns=data1.keys())
```

Example

```
# STEP 1: Calculate gini(D)
def gini_impurity (value_counts):
    n = value_counts.sum()
    p_sum = 0
    for key in value_counts.keys():
        p_sum = p_sum + (value_counts[key] / n ) * (value_counts[key] / n )
    gini = 1 - p_sum
    return gini

class_value_counts = df1[class_name].value_counts()
print(f'Number of samples in each class is:\n{class_value_counts}')

gini_class = gini_impurity(class_value_counts)
print(f'\nGini Impurity of the class is {gini_class:.3f}')
```

Number of samples in each class is:
yes 9
no 5
Name: default, dtype: int64

Gini Impurity of the class is 0.459

Example

```
# STEP 2:  
# Calculating gini impurity for the attributes  
def gini_split_a(attribute_name):  
    attribute_values = df1[attribute_name].value_counts()  
    gini_A = 0  
    for key in attribute_values.keys():  
        df_k = df1[df1[attribute_name] == key].value_counts()  
        n_k = attribute_values[key]  
        n = df1.shape[0]  
        gini_A = gini_A + ((n_k / n) * gini_impurity(df_k))  
    return gini_A  
  
gini_attiribute = {}  
for key in attribute_names:  
    gini_attiribute[key] = gini_split_a(key)  
print(f'Gini for {key} is {gini_attiribute[key]:.3f}')
```

Gini for age is 0.343
Gini for income is 0.440
Gini for student is 0.367
Gini for credit_rate is 0.429

Example

```
# STEP 3:  
# Compute Gini gain values to find the best split  
# An attribute has maximum Gini gain is selected for splitting.  
  
min_value = min(gini_attiribute.values())  
print('The minimum value of Gini Impurity : {0:.3} '.format(min_value))  
print('The maximum value of Gini Gain     : {0:.3} '.format(1-min_value))  
  
selected_attribute = min(gini_attiribute.keys())  
print('The selected attiribute is: ', selected_attribute)
```

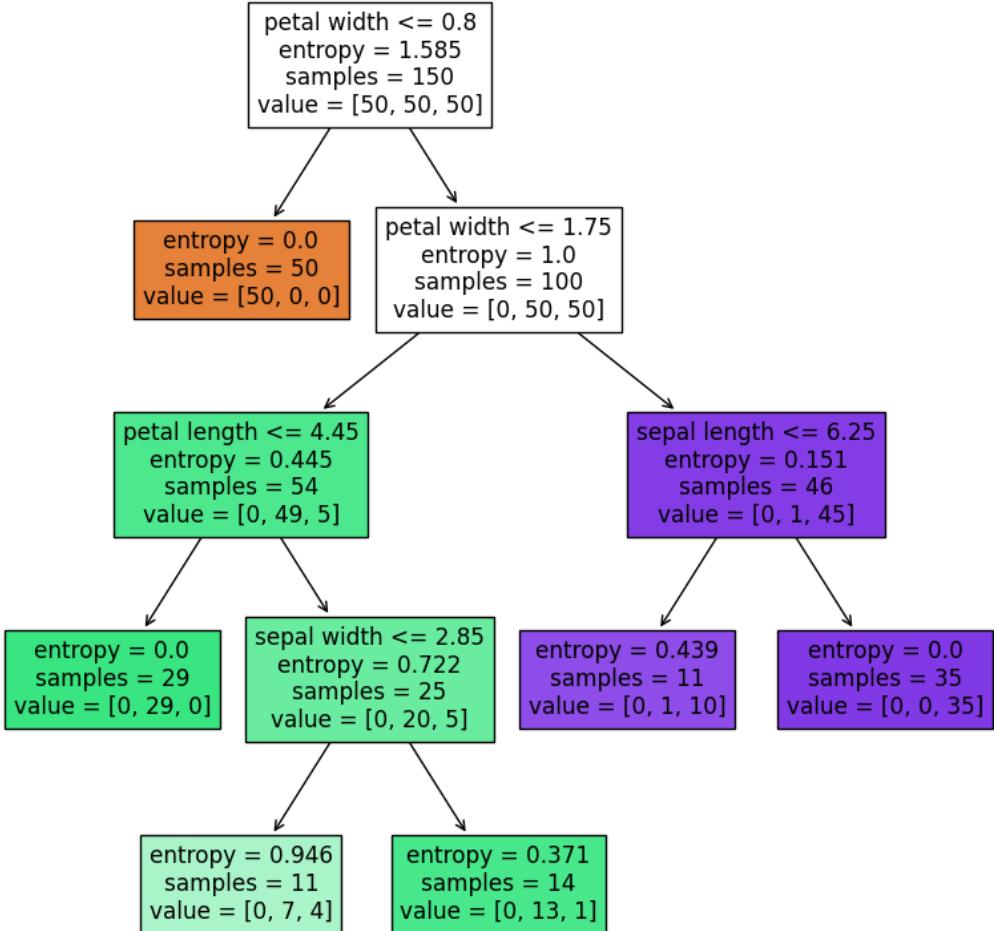
The minimum value of Gini Impurity : 0.343
The maximum value of Gini Gain : 0.657
The selected attiribute is : age

Iris Flower Classification



| | Sepal length | Sepal width | Petal length | Petal width | Class |
|-----|--------------|-------------|--------------|-------------|-----------|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| : | : | : | : | : | : |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 | virginica |

Iris Flower Classification (Entropy)



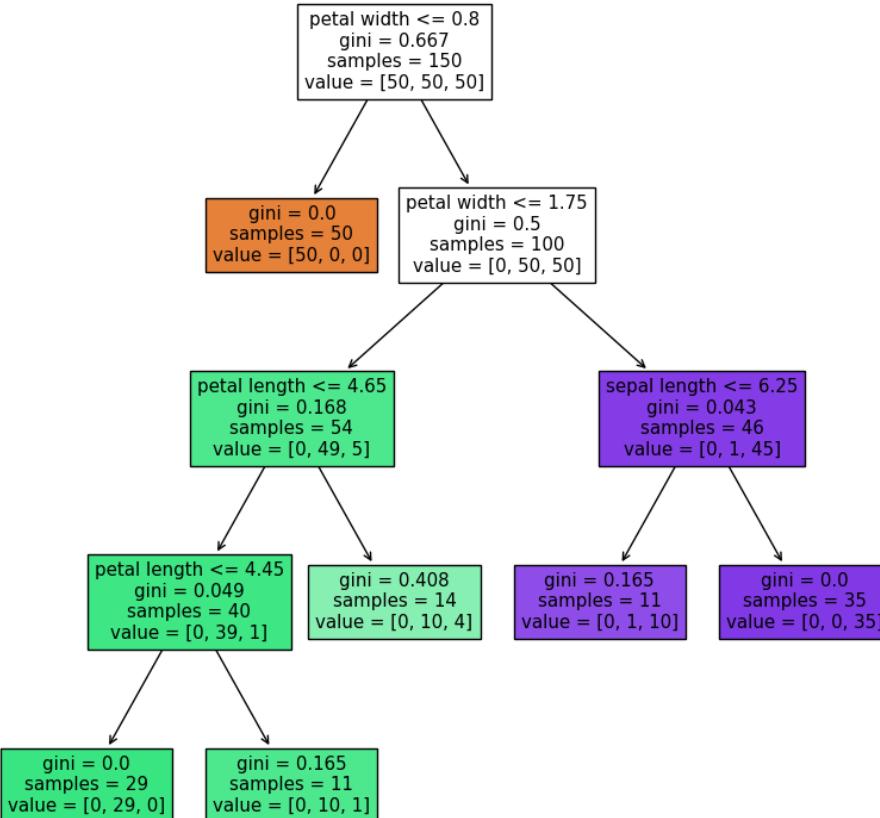
```
dataset = load_iris()
X = dataset.data
y = dataset.target

classifier = tree.DecisionTreeClassifier(criterion="entropy",
                                         max_depth=4, min_samples_leaf=10)

classifier.fit(X,y)
fig, ax = plt.subplots(figsize=(10,10))
tree.plot_tree(classifier,ax=ax, feature_names=["sepal length", "sepal width",
                                                "petal length", "petal width"], filled=True)

plt.show()
```

Iris Flower Classification (Gini)



```

dataset = load_iris()
X = dataset.data
y = dataset.target

classifier = tree.DecisionTreeClassifier(criterion="gini",
                                         max_depth=4, min_samples_leaf=10)

classifier.fit(X,y)
fig, ax = plt.subplots(figsize=(10,10))
tree.plot_tree(classifier,ax=ax, feature_names=["sepal length", "sepal width",
                                                "petal length", "petal width"],
               filled=True)

plt.show()
  
```

