AI VIETNAM
All-in-One Course
(TA Session)

Decision Tree

Extra Class

AI VIET NAM
@aivietnam.edu.vn

Nguyen-Thuan Duong – TA

Year 2024

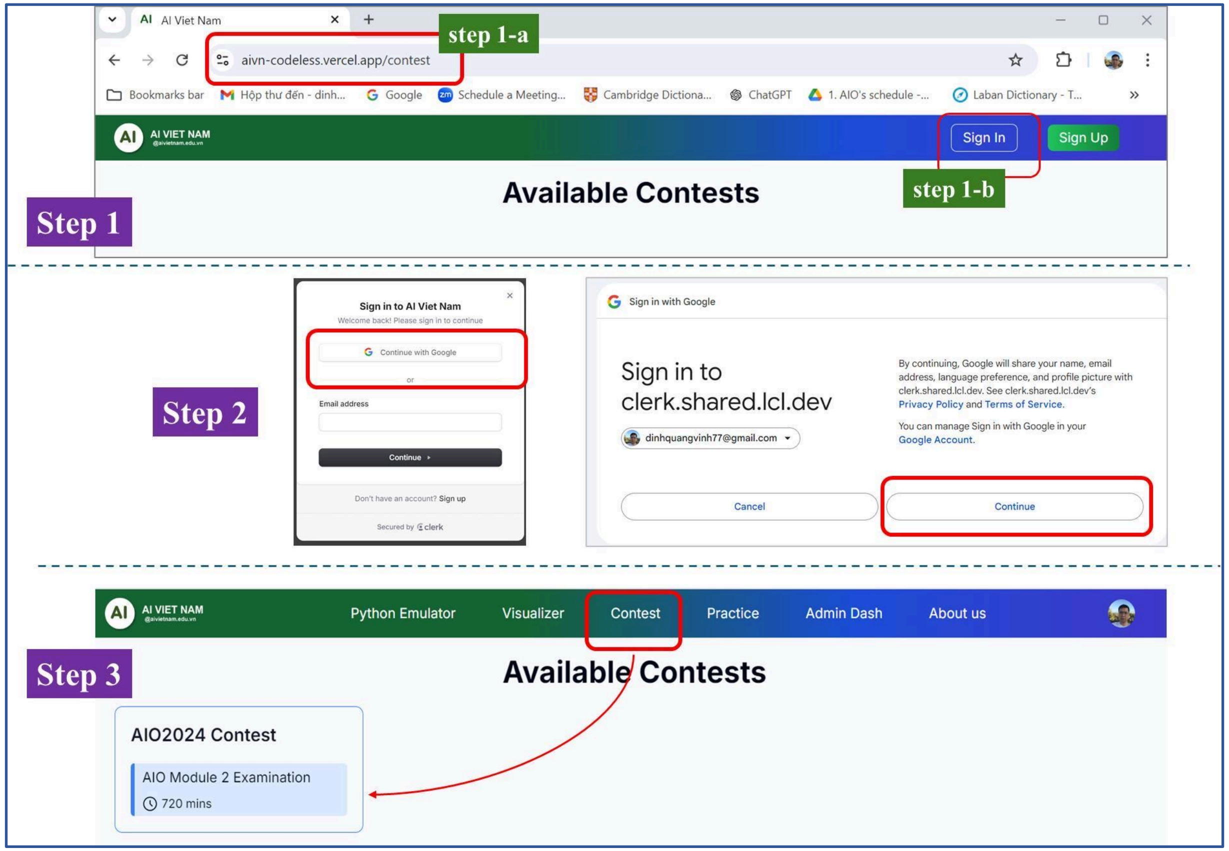
Outline

- Introduction
- Rule-based
- Decision Tree
- Classification
- Regression
- Question

Introduction

Introduction

❖ Getting Started



Multi-choices	Coding	In-time	Is-passed
30/44	3/8	True	True
40/44	0/8	True	True
20/44	8/8	False	False
35/44	3/8	False	False
10/44	2/8	True	False
15/44	1/8	True	True

Introduction

❖ Problem

Multi-choices	Coding	In-time
30/44	3/8	True
40/44	0/8	True
20/44	8/8	False
35/44	3/8	False
10/44	2/8	True
15/44	1/8	True

How to classify these data into categories

Is-passed
True
True
False
False
False
True

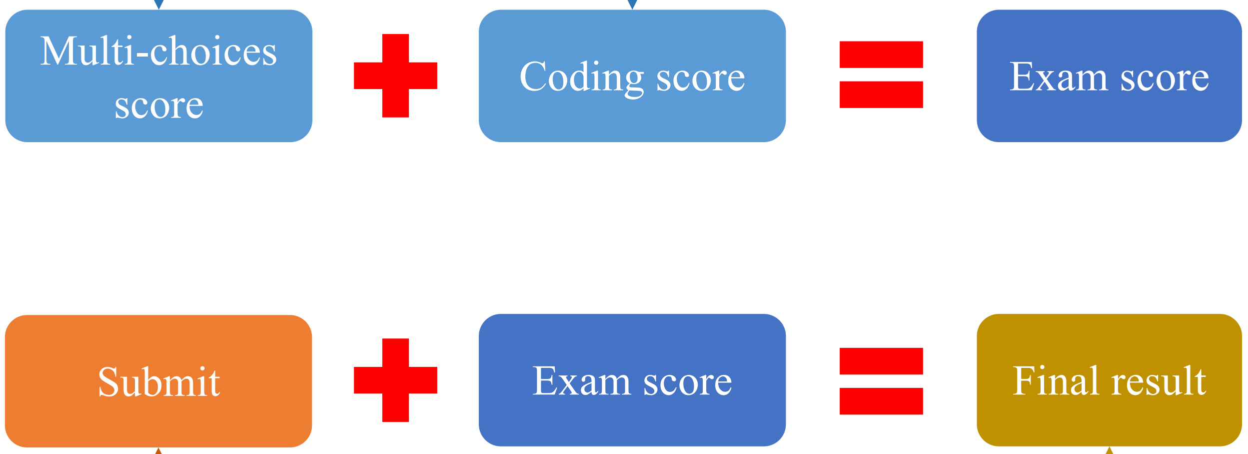
Data input → Program/Code → Target

Rule-based

Rule-based

❖ Analysis

Multi-choices	Coding	In-time	Is-passed
30/44	3/8	True	True
40/44	0/8	True	True
20/44	8/8	False	False
35/44	3/8	False	False
10/44	2/8	True	False
15/44	1/8	True	True



Rule-based

❖ Analysis

Multi-choices score + Coding score = Exam score

RULES:
- Exam score must be larger or equal to 0.5
- Each past must have a have a non-zero score

Submit + Exam score = Final result

RULES:
- Must be submitted in time

Rule-based

❖ Manual Rules Building

IF:
- Not submitted in time
- Multi-choices score == 0
- Coding score == 0

Then → Not Passed

ELSE ↓

IF:
- Exam score >= 0.5

Then → Passed

ELSE ↓

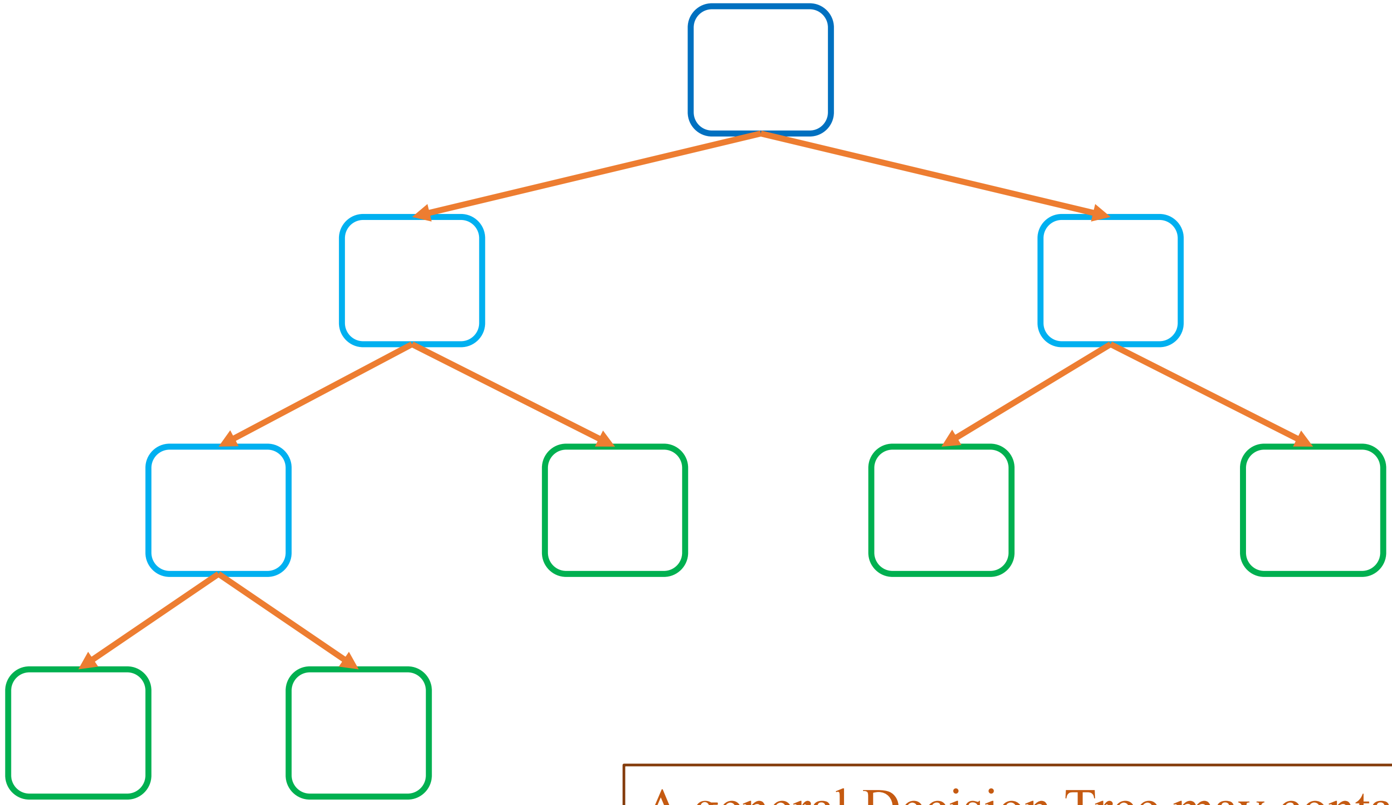
Not Passed

→ IF-Then-Else rules can be described as a tree (called decision tree)

Decision Tree

Decision Tree

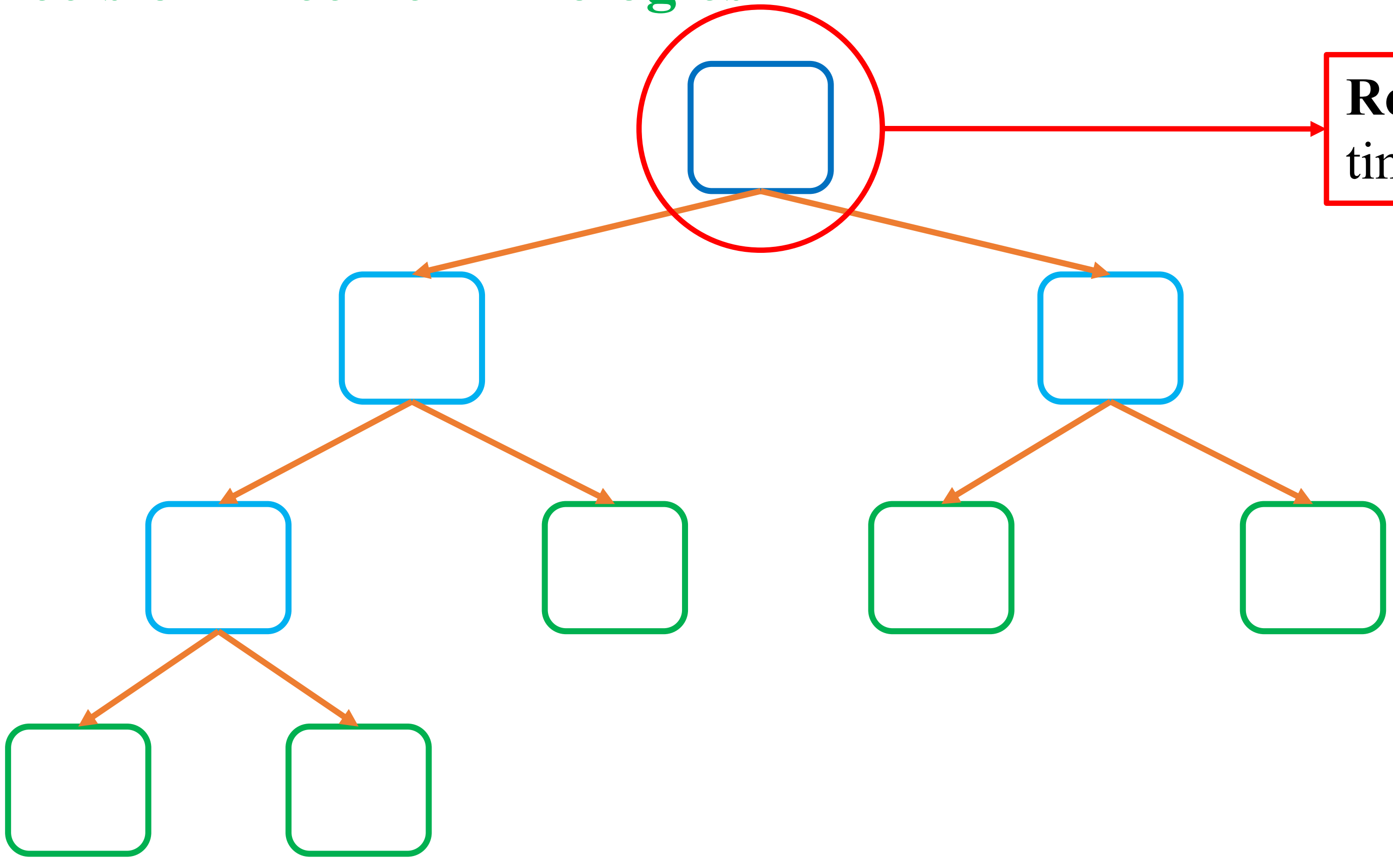
❖ Decision Tree Terminologies



A general Decision Tree may contain many <u>conditions</u> and <u>outcomes</u>.

Decision Tree

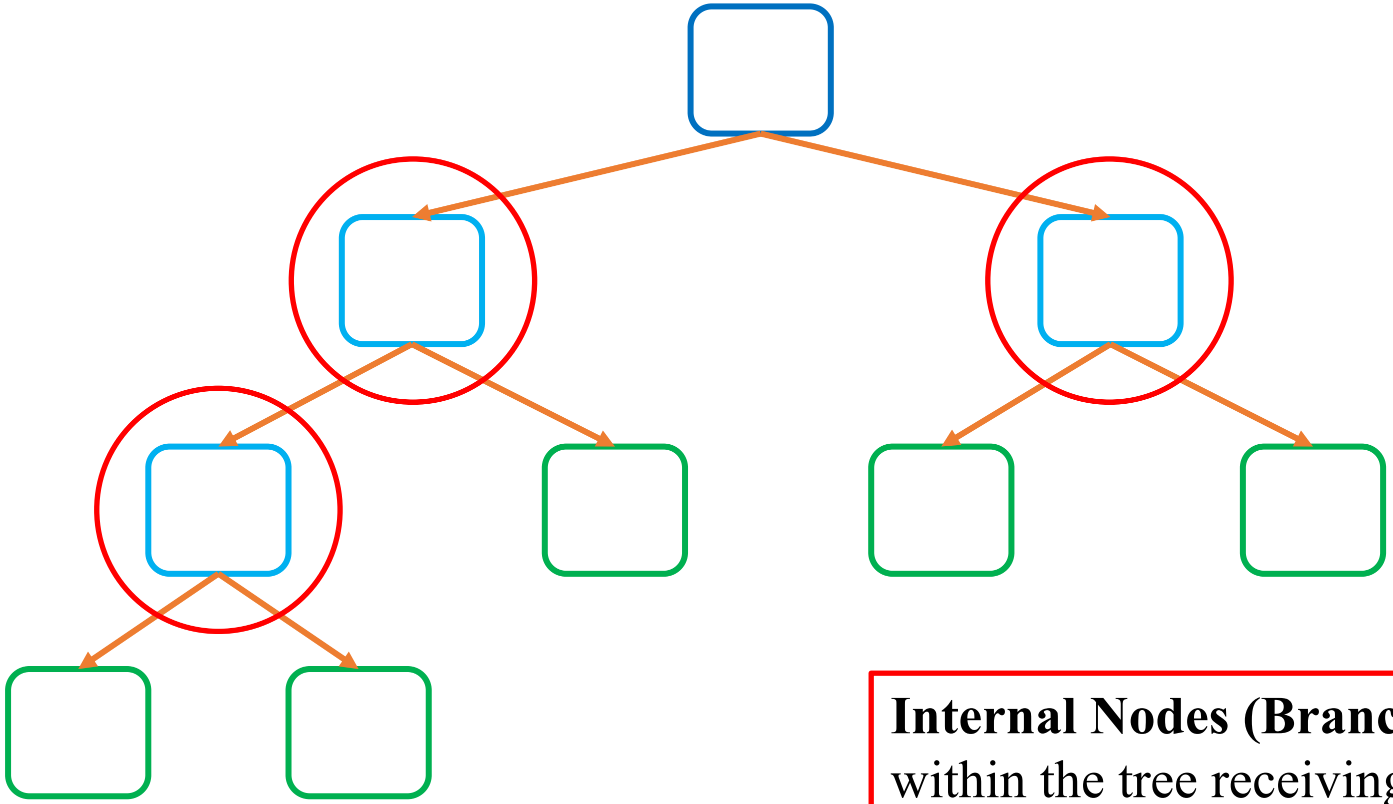
❖ Decision Tree Terminologies



Root Node: the first condition (first time split data) of the tree.

Decision Tree

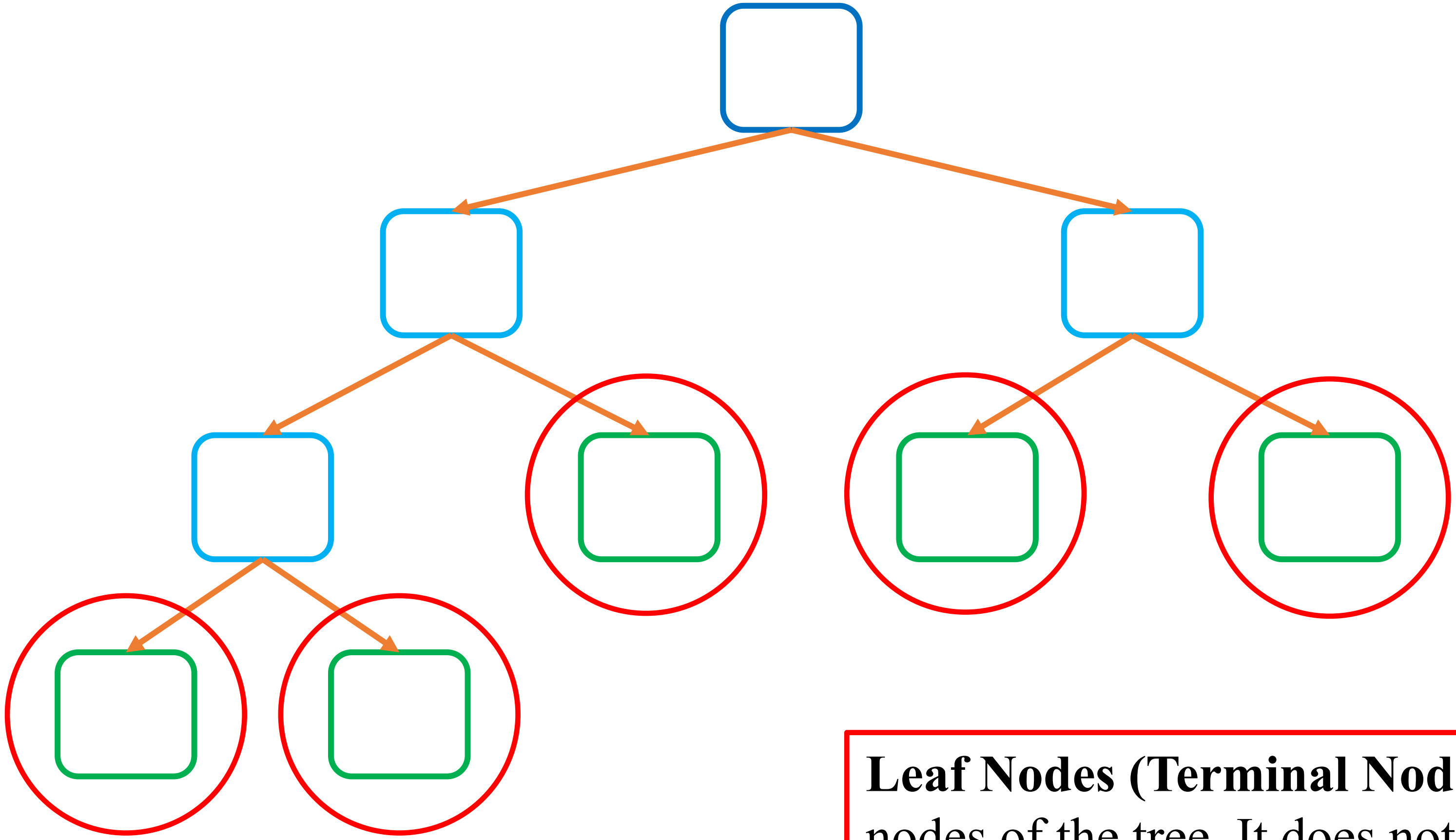
❖ Decision Tree Terminologies



Internal Nodes (Branch Nodes): the conditions within the tree receiving inputs from previous node and process (split) data to new nodes.

Decision Tree

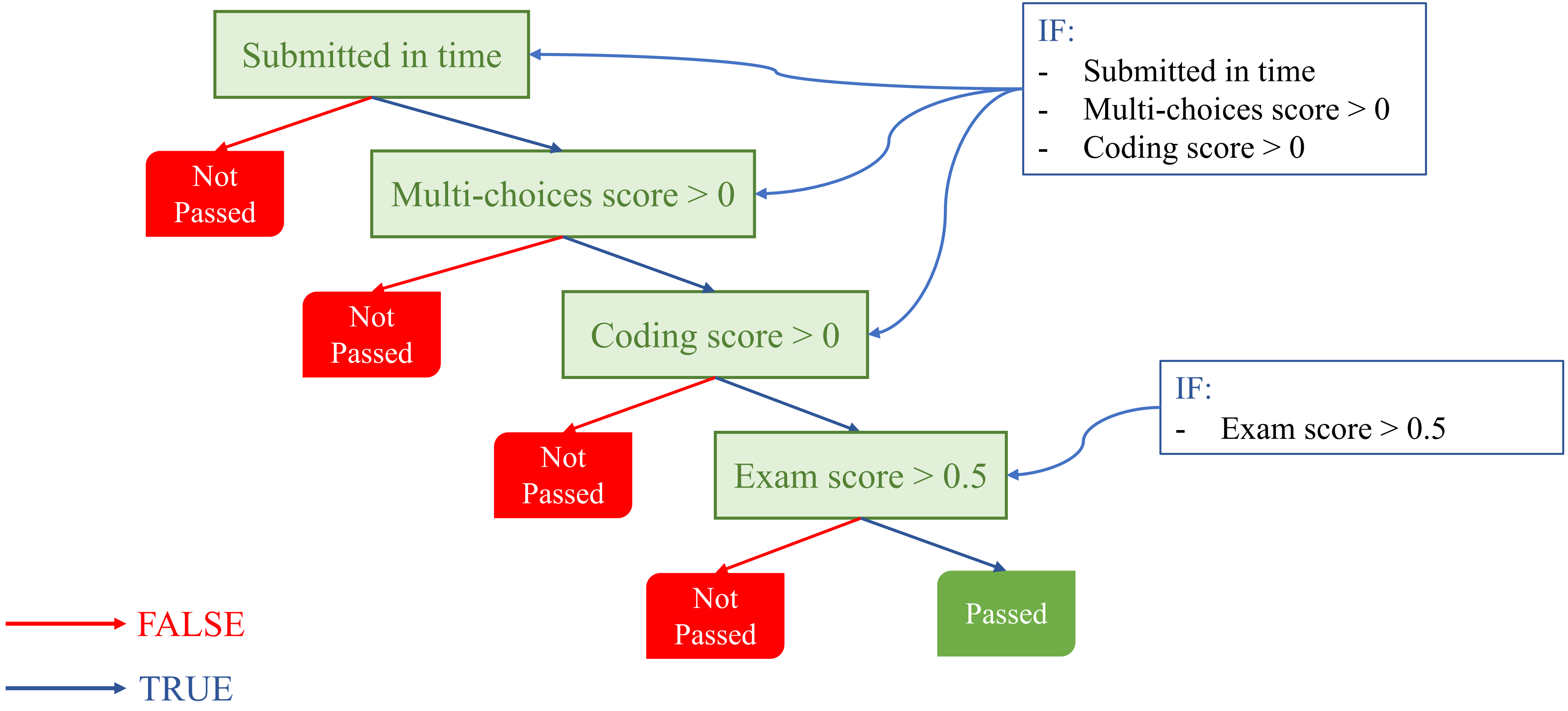
❖ Decision Tree Terminologies



Leaf Nodes (Terminal Nodes): the final decision nodes of the tree. It does not any further splits.

Decision Tree

❖ Manual Tree Building



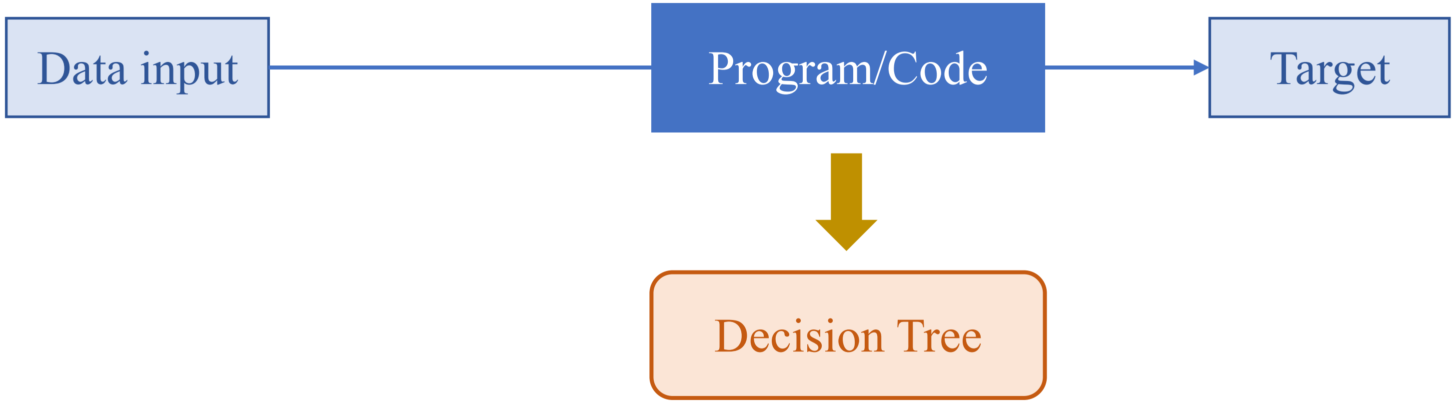
Decision Tree

❖ Problem is handled

Multi-choices	Coding	In-time
30/44	3/8	True
40/44	0/8	True
20/44	8/8	False
35/44	3/8	False
10/44	2/8	True
15/44	1/8	True

How to classify these data into categories

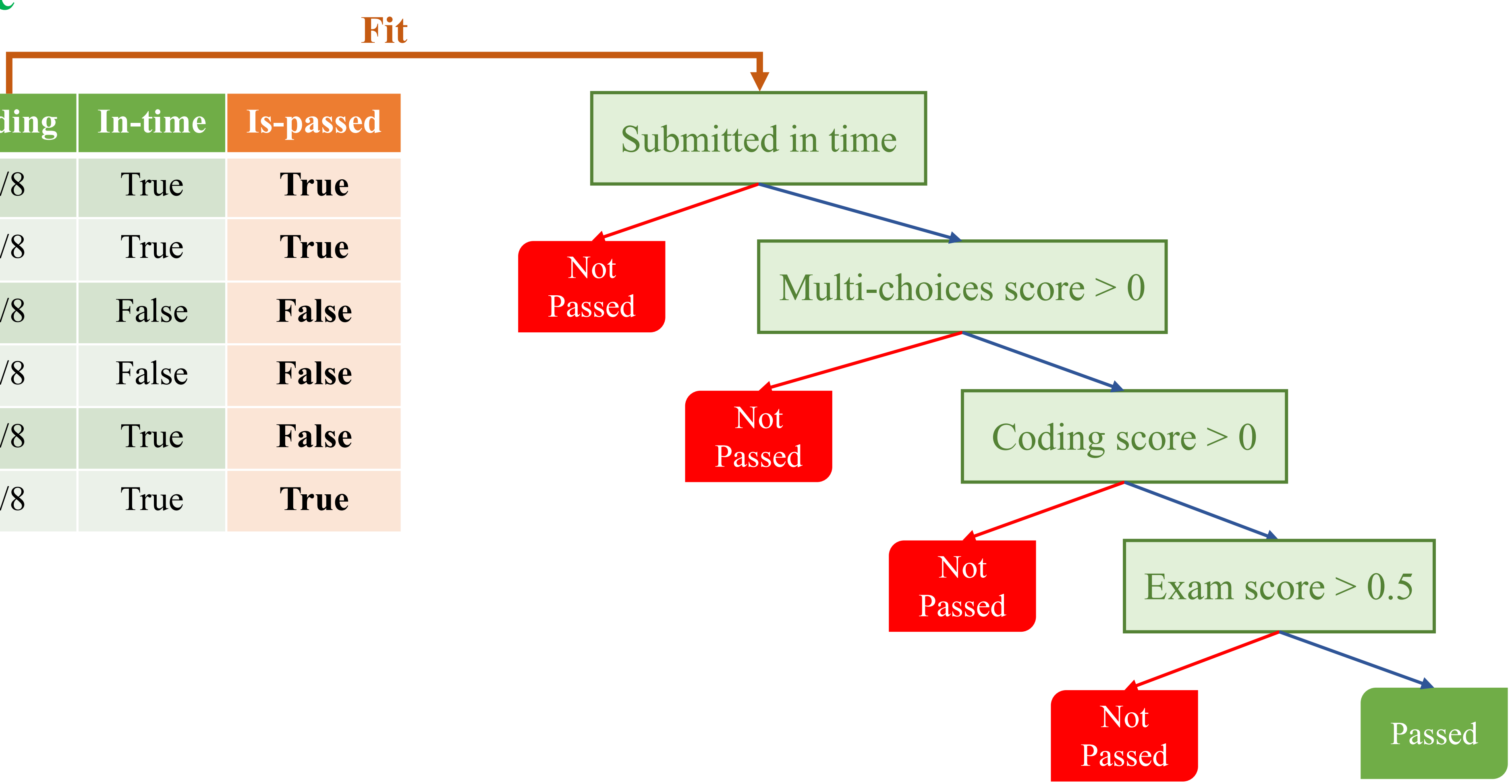
Is-passed
True
True
False
False
False
True



Decision Tree

❖ Fit data to tree

ID	Multi-choices	Coding	In-time	Is-passed
0	30/44	3/8	True	**True**
1	40/44	0/8	True	**True**
2	20/44	8/8	False	**False**
3	35/44	3/8	False	**False**
4	10/44	2/8	True	**False**
5	15/44	1/8	True	**True**



→ FALSE

→ TRUE

Decision Tree

❖ Fit data to tree

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
1	40/44	0/8	True
2	20/44	8/8	False
3	35/44	3/8	False
4	10/44	2/8	True
5	15/44	1/8	True

FULL DATA

Submitted in time

FALSE

ID	Multi-choices	Coding	In-time
2	20/44	8/8	False
3	35/44	3/8	False

Not Passed

TRUE

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
1	40/44	0/8	True
4	10/44	2/8	True
5	15/44	1/8	True

Multi-choices score > 0

Decision Tree

❖ Fit data to tree

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
1	40/44	0/8	True
4	10/44	2/8	True
5	15/44	1/8	True

Multi-choices score > 0

FALSE

ID	Multi-choices	Coding	In-time

Not Passed

TRUE

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
1	40/44	0/8	True
4	10/44	2/8	True
5	15/44	1/8	True

Coding score > 0

Decision Tree

❖ Fit data to tree

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
1	40/44	0/8	True
4	10/44	2/8	True
5	15/44	1/8	True

Coding score > 0

FALSE

ID	Multi-choices	Coding	In-time
1	40/44	0/8	True

Not Passed

TRUE

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
4	10/44	2/8	True
5	15/44	1/8	True

Exam score > 0.5

Decision Tree

❖ Fit data to tree

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True
4	10/44	2/8	True
5	15/44	1/8	True

ID	Multi-choices	Coding	In-time	Exam score
0	0.68	0.375	True	0.5275
4	0.23	0.25	True	0.24
5	0.34	0.125	True	0.2325

Exam score > 0.5

FALSE

ID	Multi-choices	Coding	In-time
4	10/44	2/8	True
5	15/44	1/8	True

Not Passed

TRUE

ID	Multi-choices	Coding	In-time
0	30/44	3/8	True

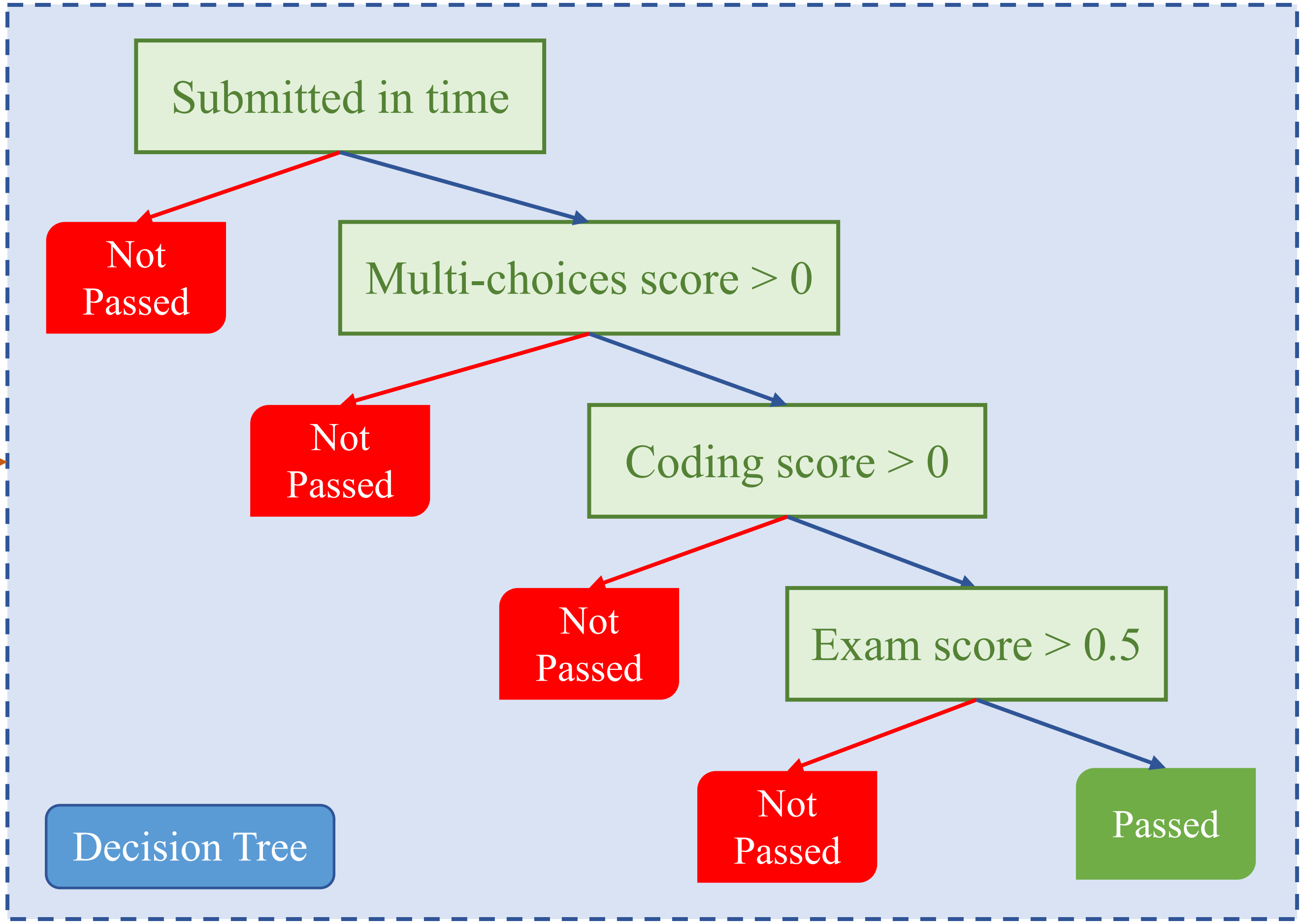
Passed

Note: weght of Multi-choices and Coding equal to 0.5

Decision Tree

❖ Manual Tree Building

Hand-definded rules



Each rule can build a corresponding tree.

Where is the best rule for dataset to build tree ?

Entropy

Entropy

❖ Mean– Variance – Std

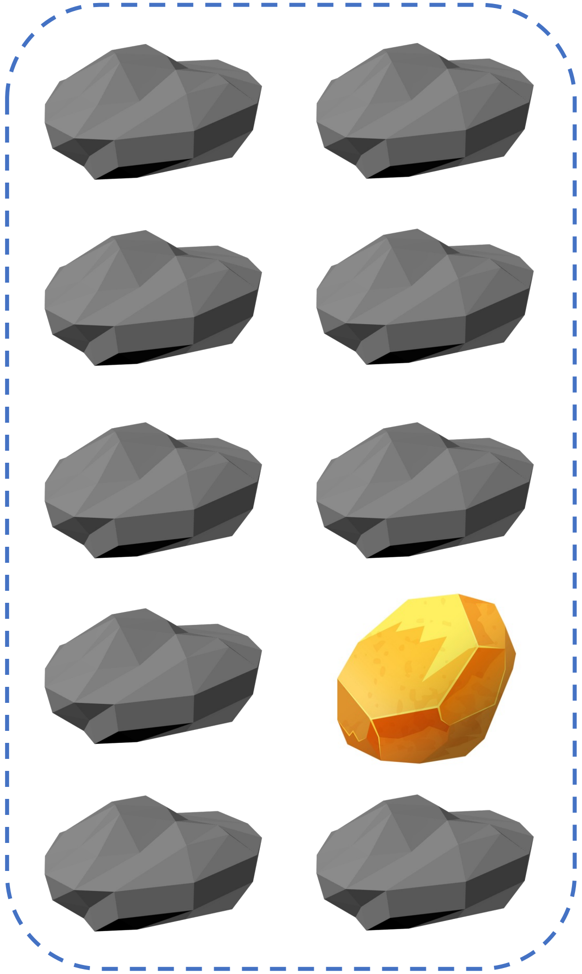
Formula	Example: $X = \{5, 3, 6, 7, 4\}$
mean $$E(X) = \sum_{i=1}^{N} X_i P_X(X_i)$$	$$E(X) = 5\times\frac{1}{5} + 3\times\frac{1}{5} + 6\times\frac{1}{5} + 7\times\frac{1}{5} + 4\times\frac{1}{5} = 5$$
variance $$var(X) = E\left(\left(X - E(X)\right)^2\right) = \sum_{i=1}^{N}\left(X_i - E(X)\right)^2 P_X(X_i)$$	$$var(X) = \frac{1}{5}[(5-5)^2 + (3-5)^2 + (6-5)^2 + (7-5)^2 + (4-5)^2] = \frac{1}{5}(0+4+1+4+1)=2$$
Standard deviation $$\sigma = \sqrt{var(X)}$$	$$\sigma = \sqrt{var(X)} = 1.41$$

Entropy

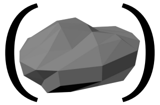
❖ Problem

Entropy

❖ Probability



$P($$) = \frac{1}{10} = 0.1$

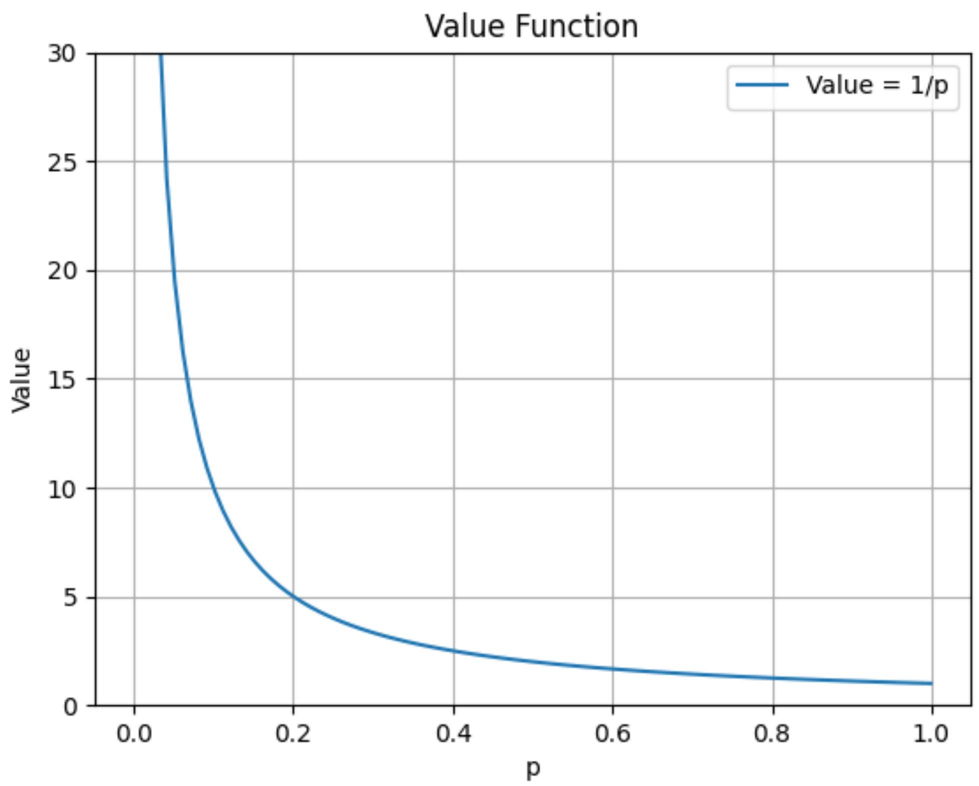
$P($$) = \frac{9}{10} = 0.9$

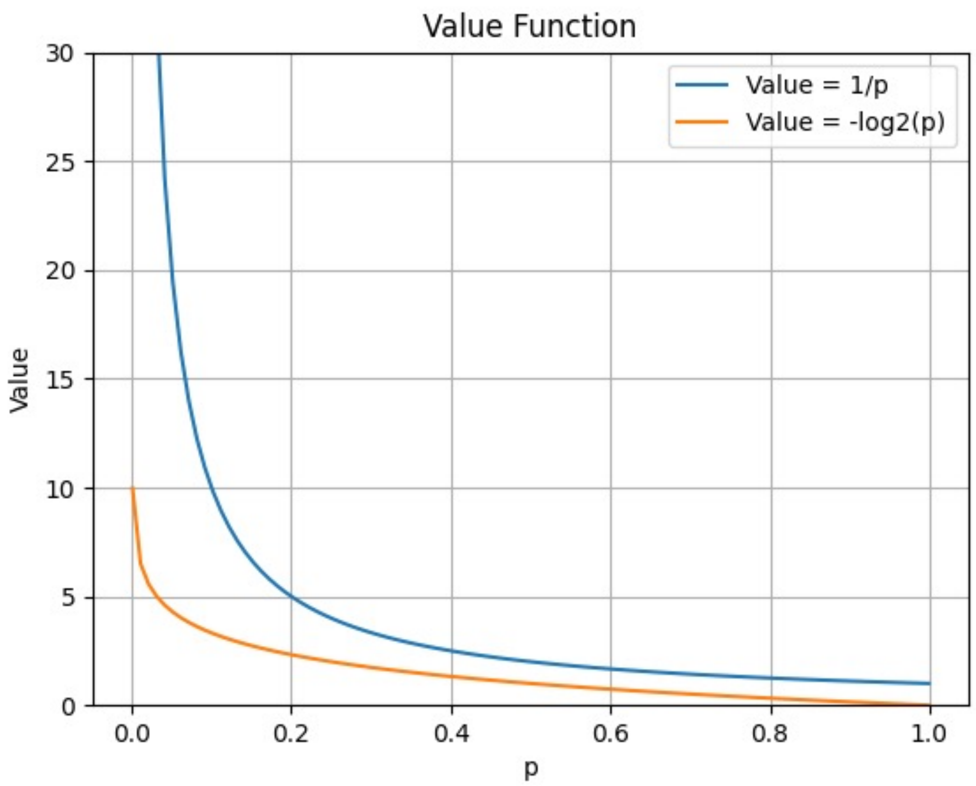
- Low probability
- High value (information)

- Value is scaled with probability

Entropy

❖ Scale with log2

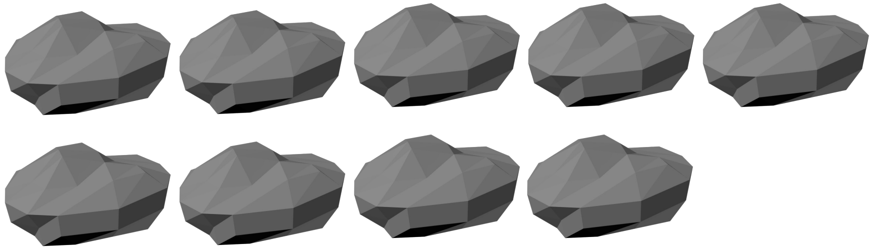




$$\text{Value}(x) = -log2(p(x))$$

Entropy

❖ Mean of Gain



$0.9 \times -log2(0.9) = 0.1368$

$0.1 \times -log2(0.1) = 0.332$

Total value:
$0.1368 + 0.332 = 0.4688$

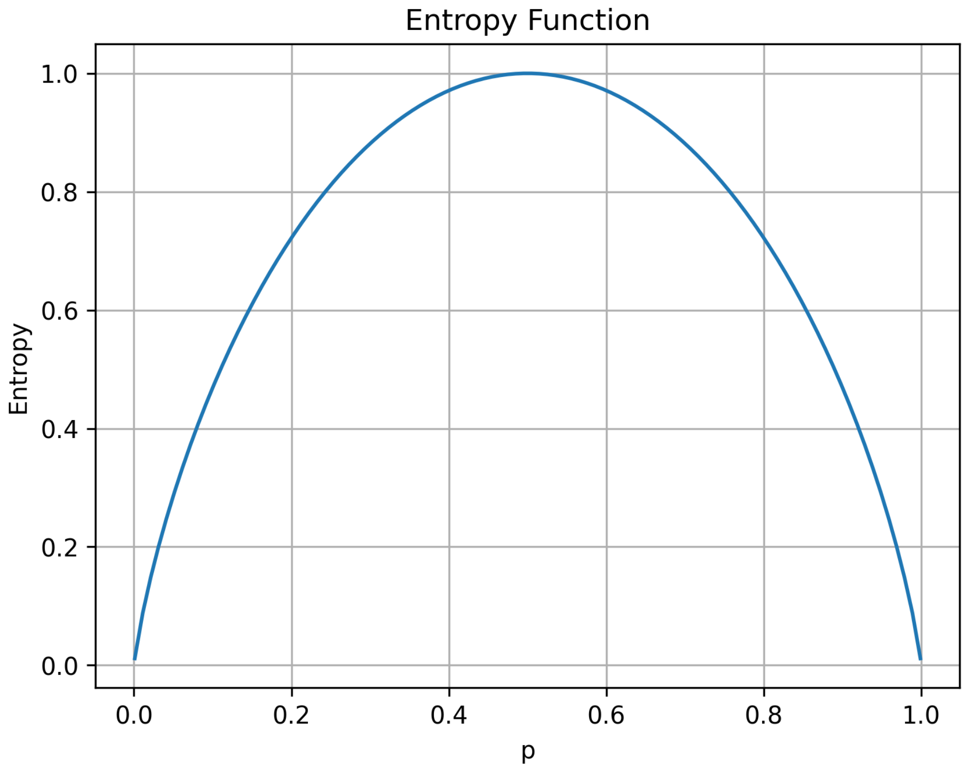
Entropy

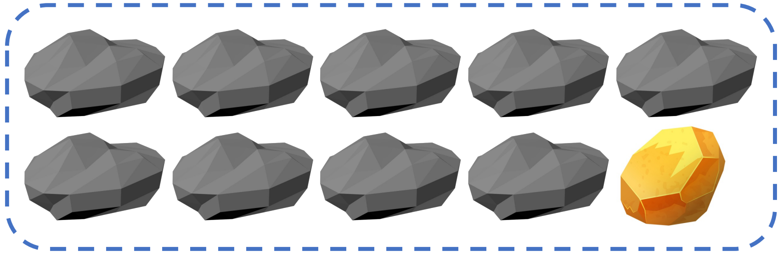
Entropy

❖ Calculate Entropy

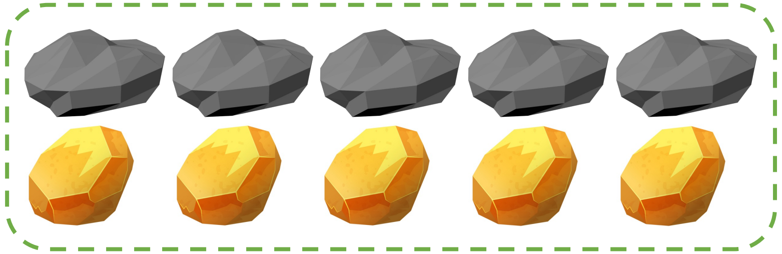
Entropy: Average of information

$$H(X) := -\sum_{x \in X} p(x)\, log(p(x))$$





Entropy = ?

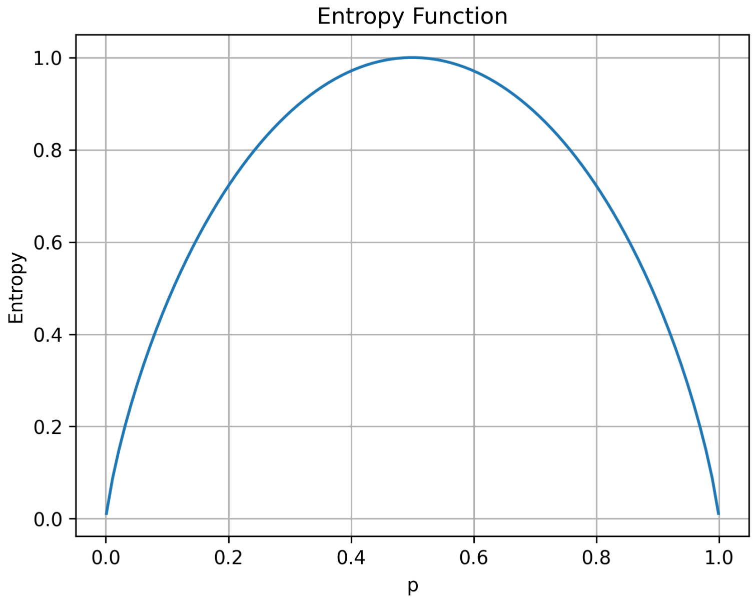


Entropy = ?

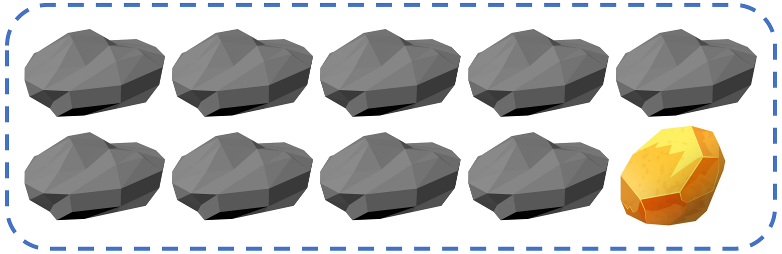
Which is larger ?

Entropy

❖ Calculate Entropy

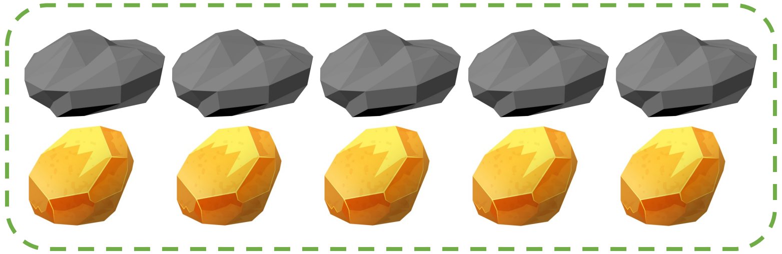


$$H(X) := -\sum_{x \in X} p(x)\, log(p(x))$$



$= -0.9log2(0.9) - 0.1log2(0.1)$
$= 0.4688$

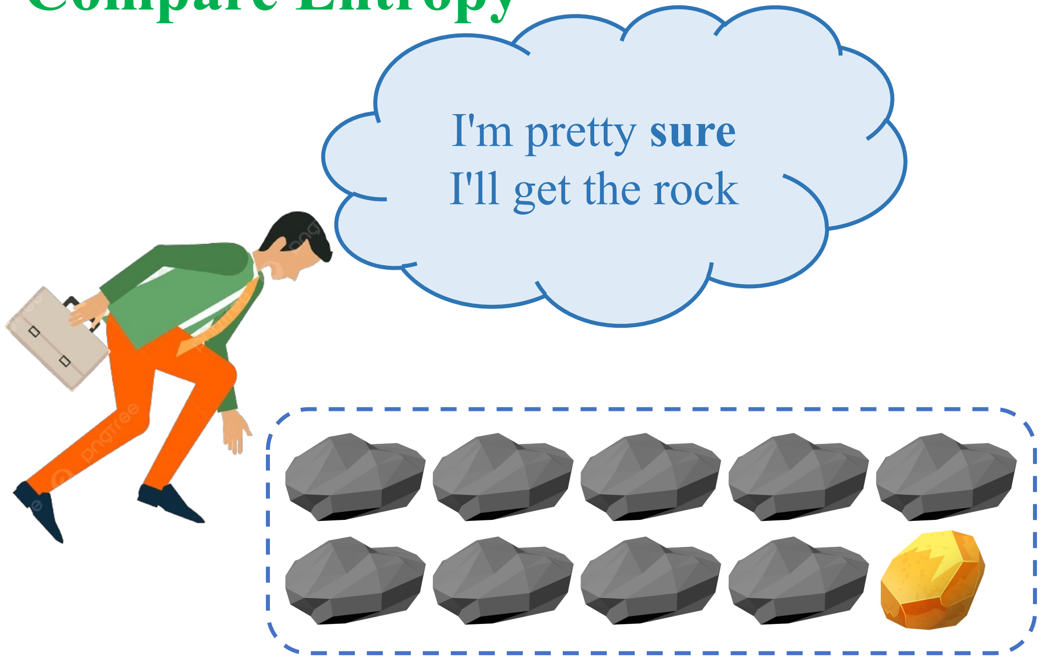
What exactly does Entropy describe?

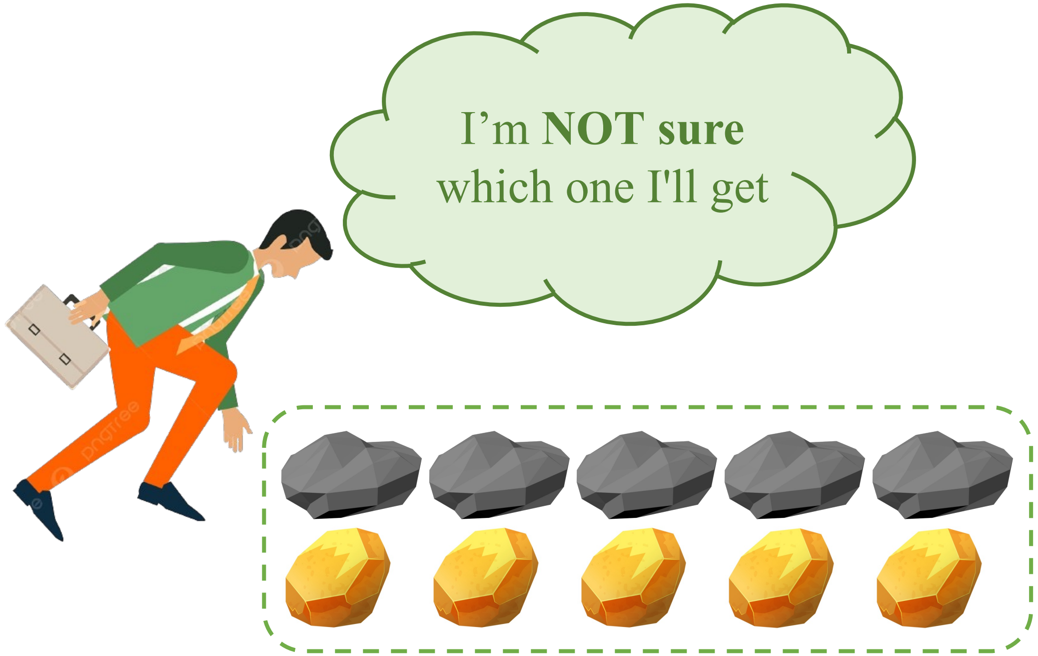


$= -0.5log2(0.5) - 0.5log2(0.5)$
$= 1.0$

Entropy

❖ Compare Entropy





➡ Entropy is used to measure disorder, randomness, or uncertainty.

Classification Tree

❖ Way to find the best rule

ID	Multi-choices	Coding	In-time	Is-passed
0	30/44	3/8	True	**True**
1	40/44	0/8	True	**True**
2	20/44	8/8	False	**False**
3	35/44	3/8	False	**False**
4	10/44	2/8	True	**False**
5	15/44	1/8	True	**True**

→ Best Rules → Best predictions

Optimal Decision Tree

Highest accuracy

Entropy

❖ Measure Information

Multi-choices
30/44
40/44
20/44
35/44
10/44
15/44

Is-passed
True
True
False
False
False
True

In-time
True
True
False
False
True
True

Coding
3/8
0/8
8/8
3/8
2/8
1/8

Entropy

❖ Information Gain (IG)

$$IG = Entropy(parent) - \left[\frac{Average}{Weighted}\right] Entropy(children)$$

$$IG(S,F) = E(S) - \sum_{f \in F} \frac{|S_f|}{S} E(S_f)$$

Classification Tree

Classification Tree

❖ Change dataset

Entropy:

$$E(S) = -\sum_{c \in C} p_c log_2 p_c$$

Information Gain

$$IG(S,F) = E(S) - \sum_{f \in F} \frac{|S_f|}{|S|} E(S_f)$$

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

$$S = \{9: Yes, 5: No\} \rightarrow E(S) = -\frac{9}{14} log_2\left(\frac{9}{14}\right) - \frac{5}{14} log_2\left(\frac{5}{14}\right) = 0.94$$

Parent Entropy

Classification Tree

❖ Children IG

Entropy:

$$E(S) = -\sum_{c \in C} p_c log_2 p_c$$

Information Gain

$$IG(S,F) = E(S) - \sum_{f \in F} \frac{|S_f|}{|S|} E(S_f)$$

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

$S_{weak} = \{6: Yes, 2: No\} \rightarrow E(S_{weak}) = -\frac{6}{8} log_2\left(\frac{6}{8}\right) - \frac{2}{8} log_2\left(\frac{6}{8}\right) = 0.811$

$S_{Strong} = \{3: Yes, 3: No\} \rightarrow E(S_{Strong}) = -\frac{3}{6} log_2\left(\frac{3}{6}\right) - \frac{3}{6} log_2\left(\frac{3}{6}\right) = 1$

$$\Rightarrow Gain(S, Wind) = E(S) - \frac{8}{14} E(S_{weak}) - \frac{6}{14} E(S_{Strong})$$

$$= 0.94 - \frac{8}{14} * 0.811 - \frac{6}{14} * 1 = 0.048$$

Category = 2

Classification Tree

❖ Children IG

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Entropy:

$$E(S) = -\sum_{c \in C} p_c \log_2 p_c$$

Information Gain

$$IG(S,F) = E(S) - \sum_{f \in F} \frac{|S_f|}{|S|} E(S_f)$$

Category = 3 > 2 → **Combine** →
Option_1: Sunny - (Overcast, Rain)
Option_2: Overcast - (Sunny, Rain)
Option_3: Rain – (Sunny, Overcast)

$$S_{Sunny} = \{2: Yes, 3: No\} \longrightarrow E(S_{Sunny}) = 0.97$$

$$S_{Overcast,Rain} = \{7: Yes, 2: No\} \longrightarrow E(S_{Overcast,Rain}) = 0.764$$

$$IG(S, Option\_1)$$
$$= E(S) - \frac{5}{14}E(S_{Sunny}) - \frac{9}{14}E(S_{Overcast,Rain})$$
$$= 0.94 - \frac{5}{14} * 0.97 - \frac{9}{14} * 0.764 = 0.102$$

$$\text{Gain}(S, \text{Outlook}) = \max \begin{cases} IG(S, Option\_1) = 0.102 \\ IG(S, Option\_2) = 0.226 \\ IG(S, Option\_3) = 0.003 \end{cases}$$

Classification Tree

❖ Children IG

Entropy:

$$E(S) = -\sum_{c \in C} p_c log_2 p_c$$

Information Gain

$$IG(S,F) = E(S) - \sum_{f \in F} \frac{|S_f|}{|S|} E(S_f)$$

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Gain(S, Outlook) = 0.226

Gain(S, Temp) = 0.015

Gain(S, Humidity) = 0.151

Gain(S, Wind) = 0.048

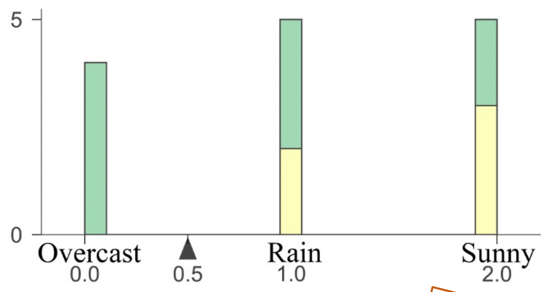
Choose Outlook with highest Gain score for root node

Option_2 is used to split

Classification Tree

❖ Build classfication tree

Play_Tennis
No
Yes



overcast

rain or sunny

Outlook	Temp	Humidity	Wind	Label
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Outlook	Temp	Humidity	Wind	Label
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Rain	Mild	High	Strong	No

Gain(S, Outlook) = 0.029

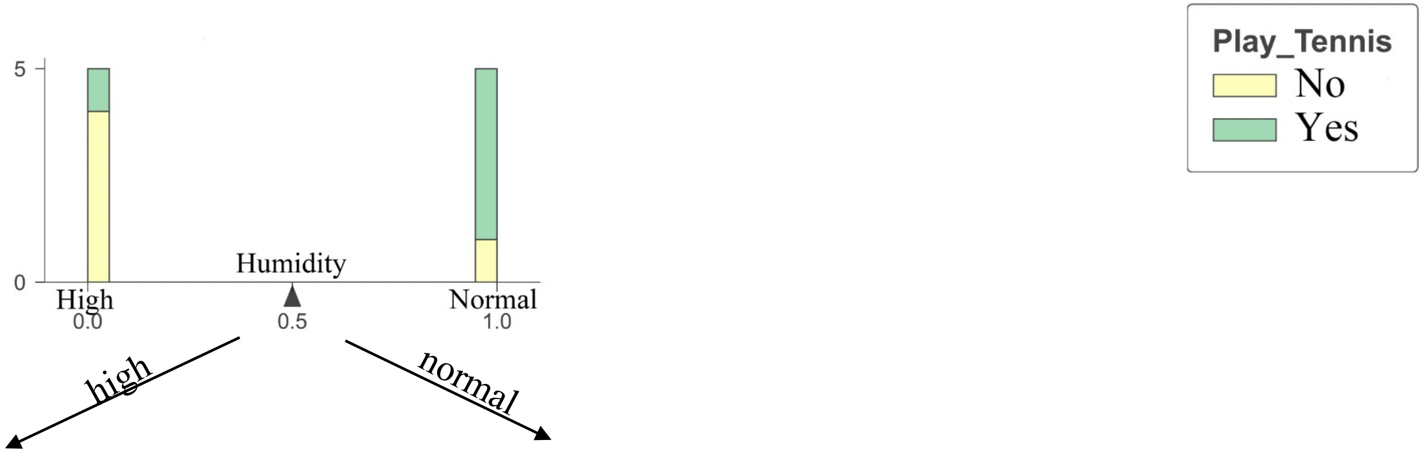
Gain(S, Temp) = 0.034

Gain(S, Humidity) = 0.278

Gain(S, Wind) = 0.124

Classification Tree

❖ Build classfication tree



Outlook	Temp	Humidity	Wind	Label
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Rain	Mild	High	Weak	Yes
Sunny	Mild	High	Weak	No
Rain	Mild	High	Strong	No

Gain(S, Outlook) = 0.321

Gain(S, Temp) = 0.17

Gain(S, Humidity) = 0.00

Gain(S, Wind) = 0.17

Outlook	Temp	Humidity	Wind	Label
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Gain(S, Outlook) = 0.17

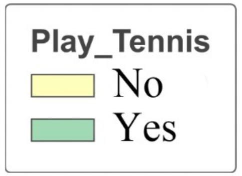
Gain(S, Temp) = 0.17

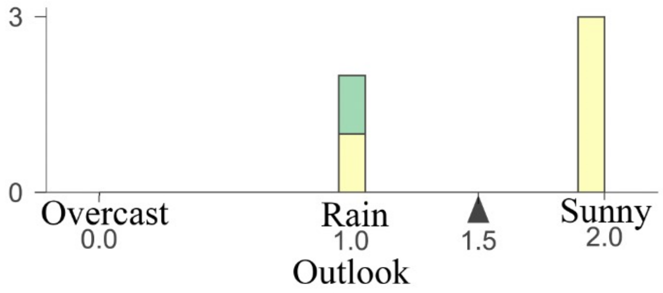
Gain(S, Humidity) = 0.00

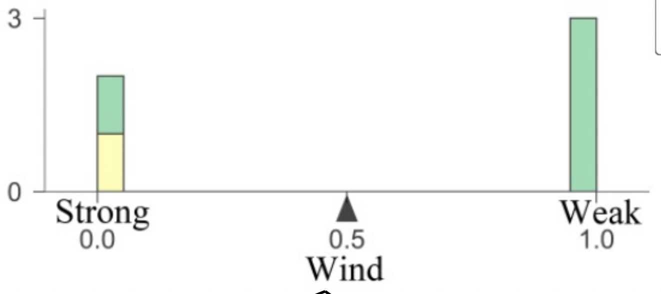
Gain(S, Wind) = 0.321

Classification Tree

❖ Build classfication tree







rain

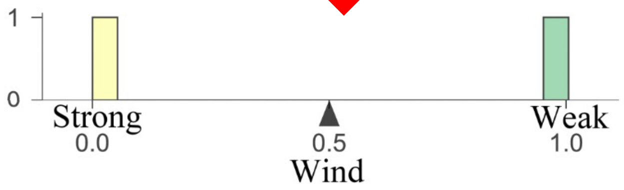
Outlook	Temp	Humidity	Wind	Label
Rain	Mild	High	Weak	Yes
Rain	Mild	High	Strong	No

Gain(S, Outlook) = 0.00

Gain(S, Temp) = 0.00

Gain(S, Humidity) = 0.00

Gain(S, Wind) = 1



sunny

Outlook	Temp	Humidity	Wind	Label
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No

strong

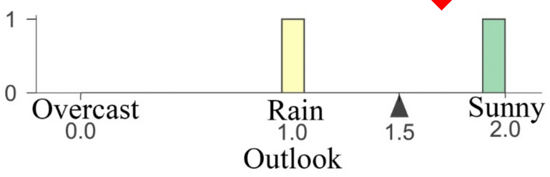
Outlook	Temp	Humidity	Wind	Label
Rain	Cool	Normal	Strong	No
Sunny	Mild	Normal	Strong	Yes

Gain(S, Outlook) = 1

Gain(S, Temp) = 1

Gain(S, Humidity) = 0.00

Gain(S, Wind) = 0.00

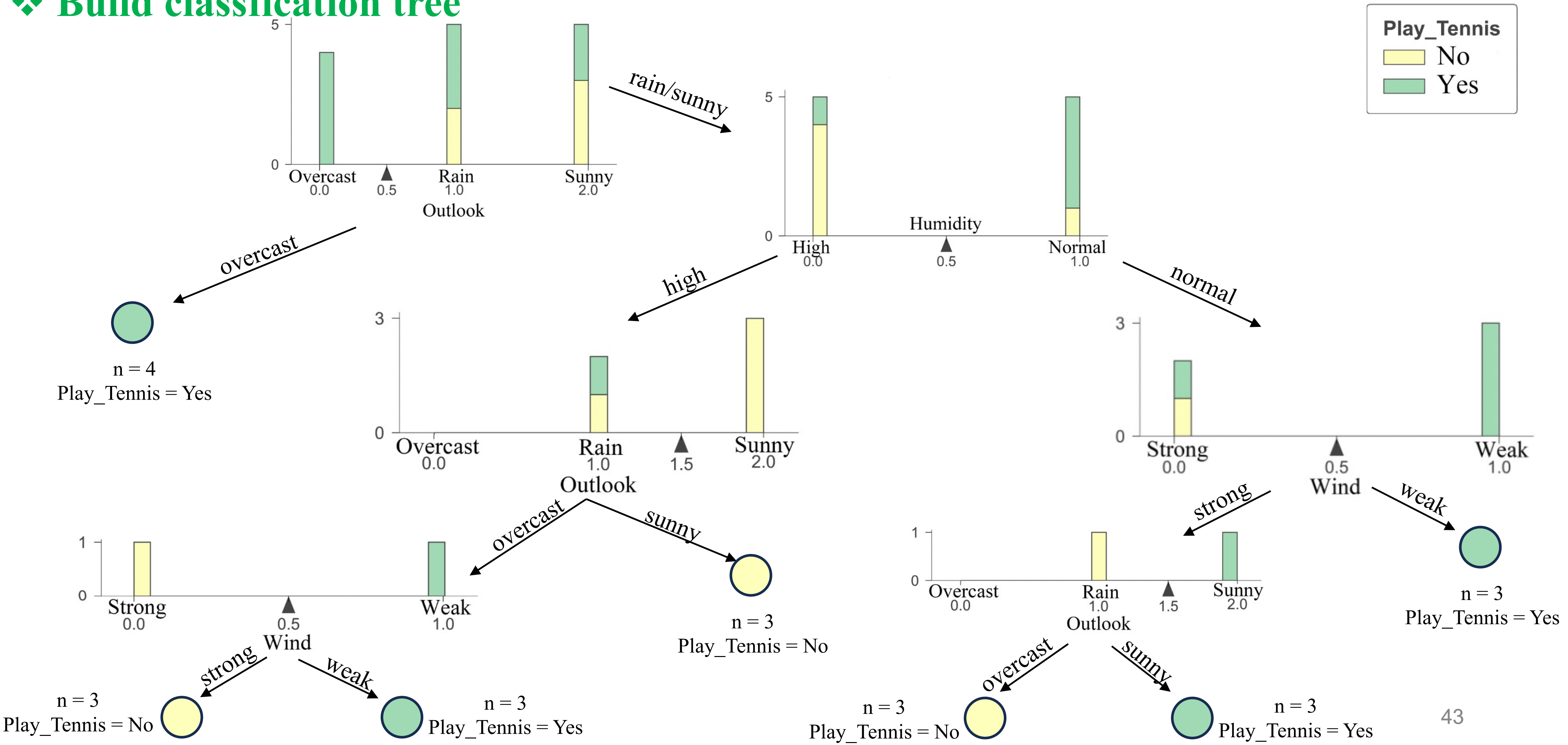


weak

Outlook	Temp	Humidity	Wind	Label
Rain	Cool	Normal	Weak	Yes
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes

Classification Tree

❖ Build classfication tree

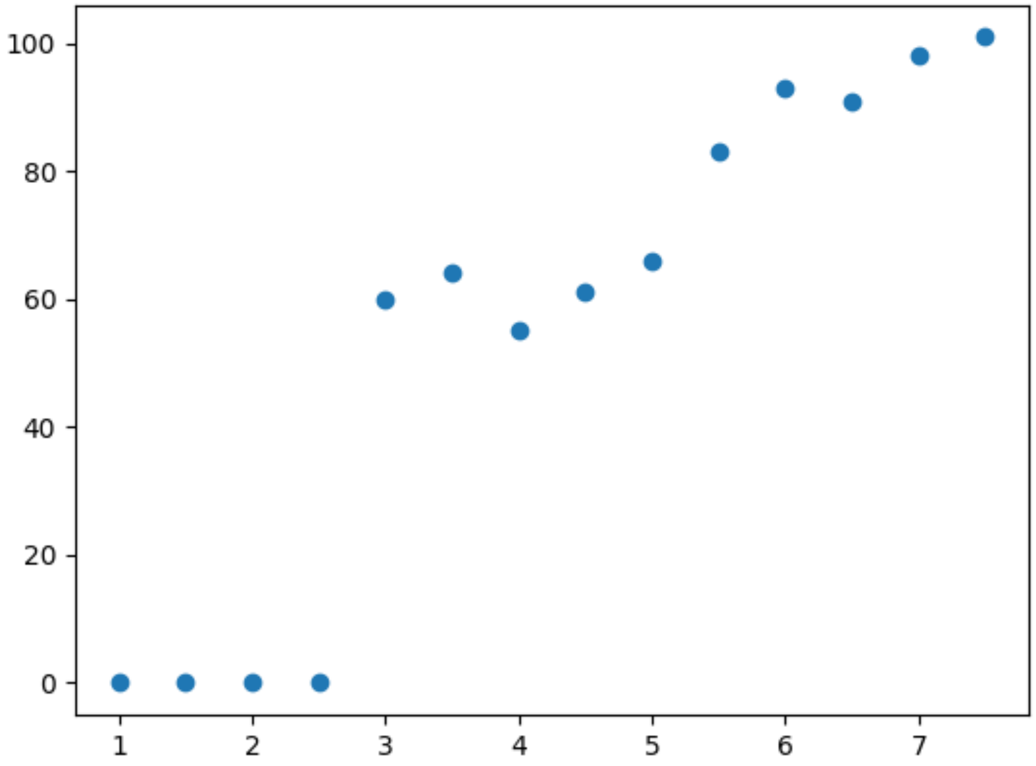


Regression Tree

Regression Tree

❖ Getting Started

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101



When Experience = 5.3,

Salary = ?

Regression Tree

❖ Compute mean and error

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

$$\mu = \frac{1}{|S|}\sum_i S_i = 55.14$$

$$mse = \frac{1}{|S|}\sum_i (S_i - \mu)^2$$
$$= 1417.97$$

Experience	Salary
1	0

$$\mu_L = \frac{1}{|L|}\sum_i L_i = 0$$

$$mse_L = \frac{1}{|L|}\sum_i (L_i - \mu)^2 = 0$$

Experience	Salary
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

$$a_{mse} = \frac{|L|}{|S|} mse_L + \frac{|R|}{|S|} mse_R$$
$$= \frac{1}{14} * 0 + \frac{13}{14} * 1275.15$$
$$= 1184.07$$

$$\mu_R = \frac{1}{|R|}\sum_i R_i = 59.38$$

$$mse_R = \frac{1}{|R|}\sum_i (R_i - \mu)^2 = 1275.15$$

Regression Tree

❖ Compute mean and error

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

$$\mu = \frac{1}{|S|}\sum_i S_i = 55.14$$

$$mse = \frac{1}{|S|}\sum_i (S_i - \mu)^2 = 1417.97$$

Experience	Salary
1	0
1.5	0
2	0
2.5	0

$$\mu_L = \frac{1}{|L|}\sum_i L_i = 0$$

$$mse_L = \frac{1}{|L|}\sum_i (L_i - \mu)^2 = 0$$

$$a_{mse} = \frac{|L|}{|S|} mse_L + \frac{|R|}{|S|} mse_R$$

$$= \frac{4}{14} * 0 + \frac{10}{14} * 282.35$$

$$= 201.68$$

Experience	Salary
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

$$\mu_R = \frac{1}{|R|}\sum_i R_i = 77.2$$

$$mse_R = \frac{1}{|R|}\sum_i (R_i - \mu)^2 = 282.35$$

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

$a_{mse} = 1184.07$

$a_{mse} = 911.19$

$a_{mse} = 588.68$

$\boxed{a_{mse} = 201.68}$

$a_{mse} = 383.92$

$a_{mse} = 526.52$

$a_{mse} = 543.51$

$a_{mse} = 575.09$

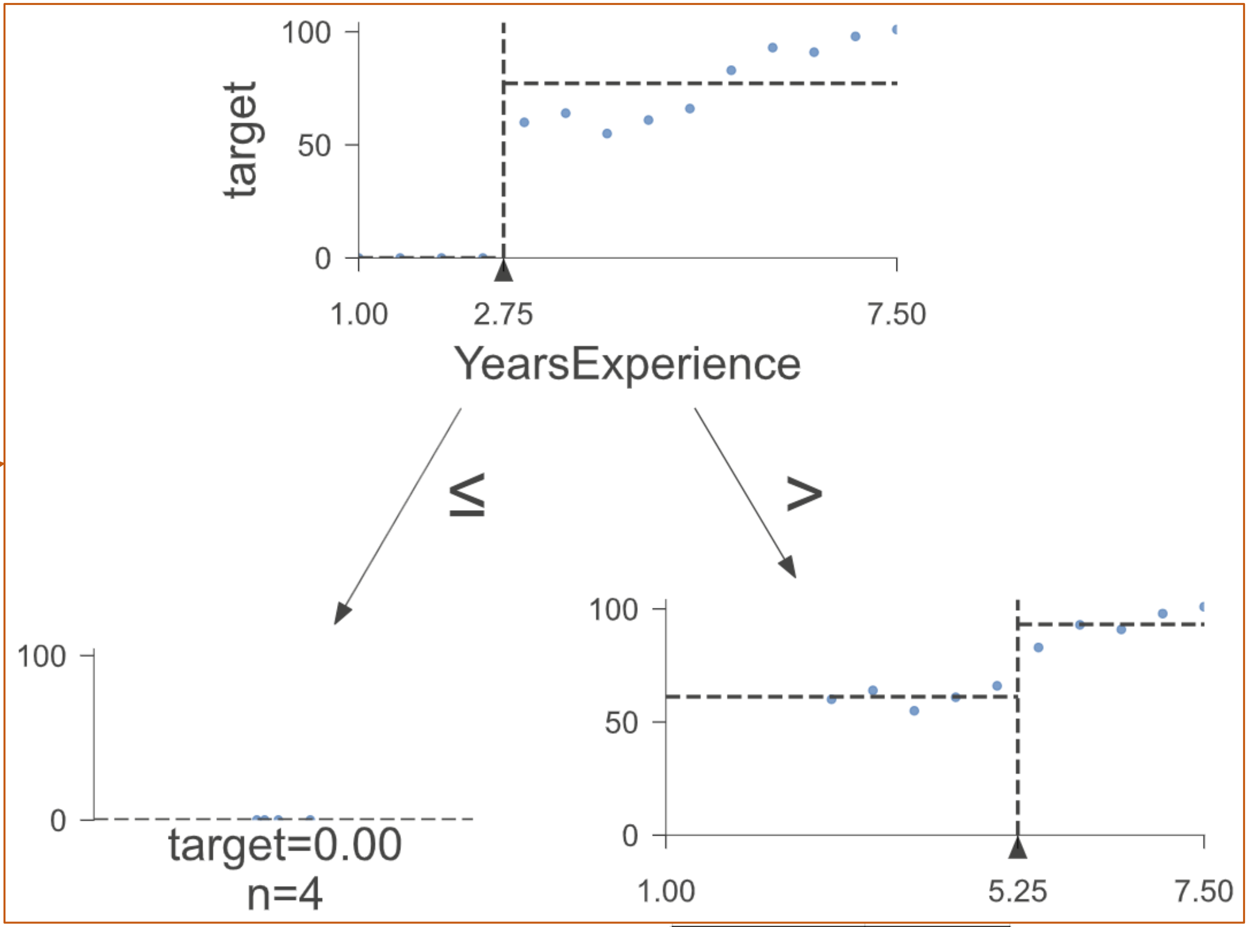
$a_{mse} = 613.34$

$a_{mse} = 758.4$

$a_{mse} = 947.73$

$a_{mse} = 1090.05$

$a_{mse} = 1256.21$



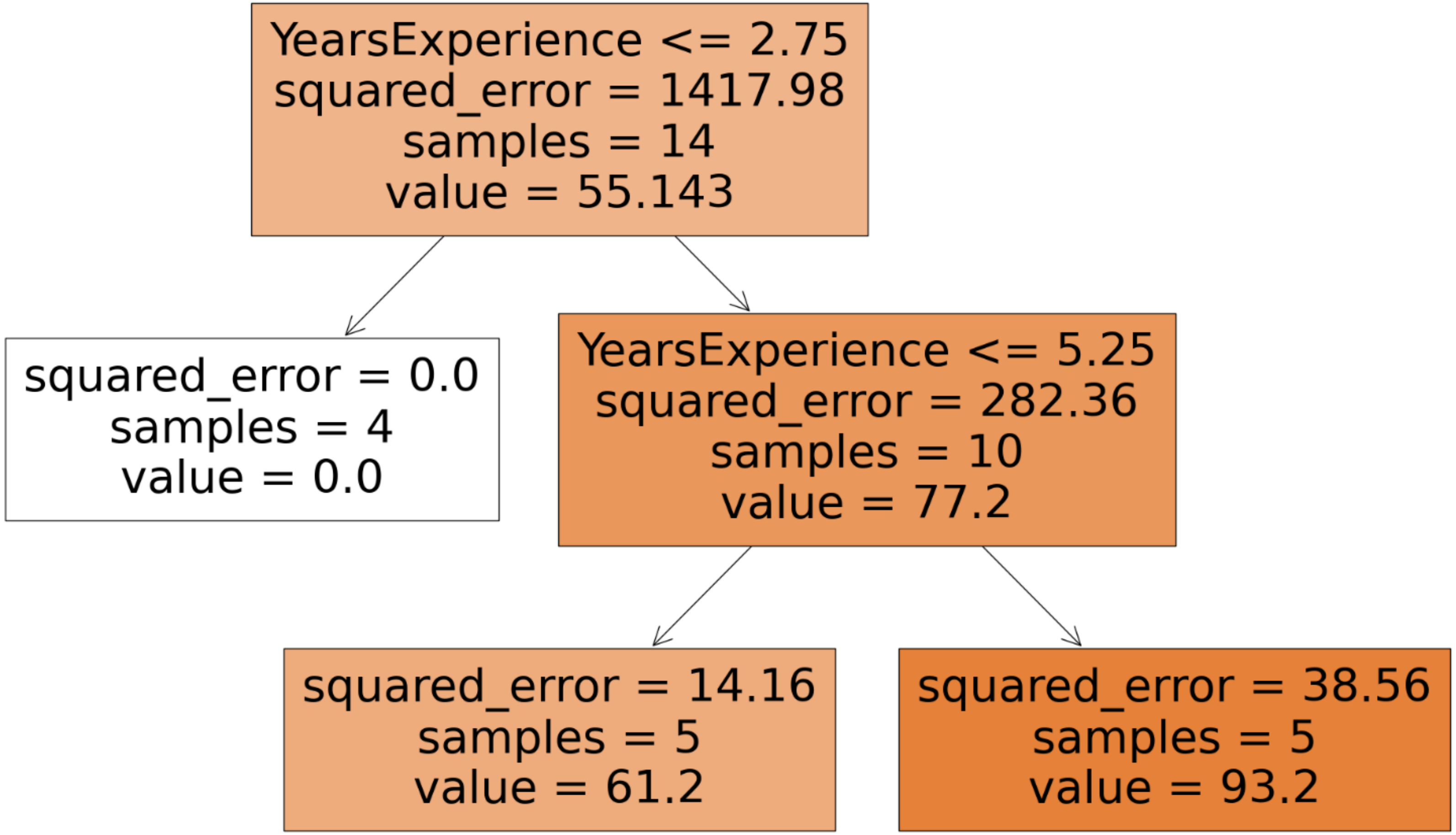
Experience	Salary
1	0
1.5	0
2	0
2.5	0

Experience	Salary
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

Regression Tree

❖ Plot tree

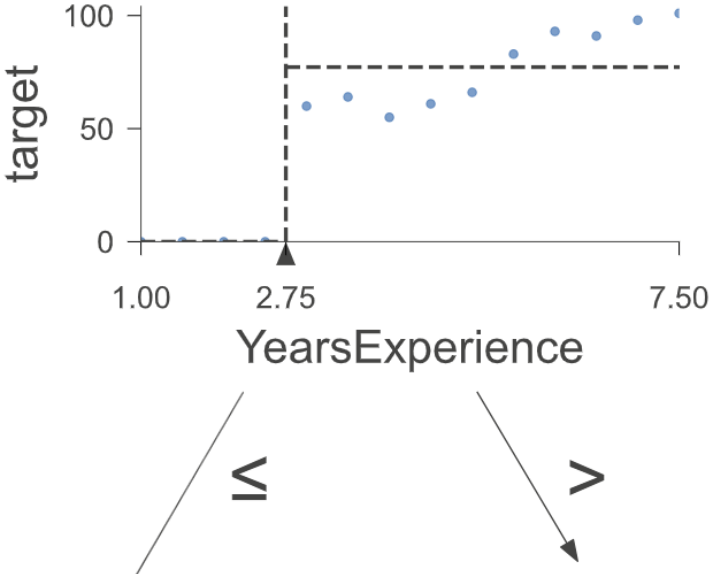
Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101



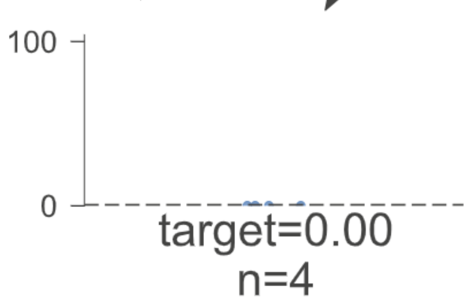
AI VIETNAM
All-in-One Course
(TA Session)

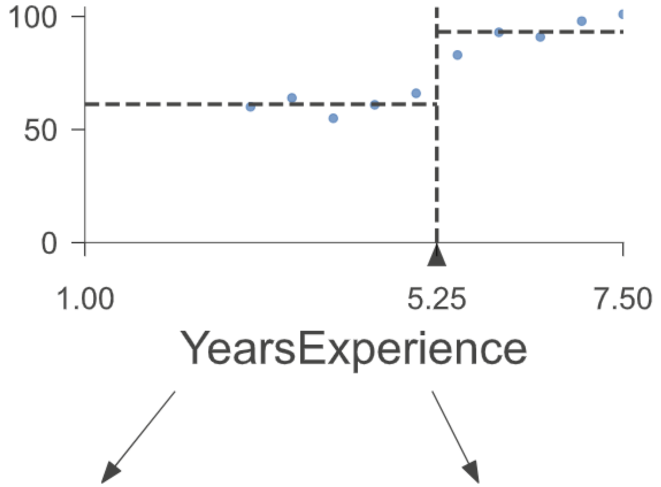
❖ Getting

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101



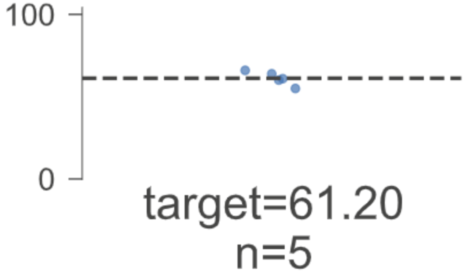
Experience	Salary
1	0
1.5	0
2	0
2.5	0

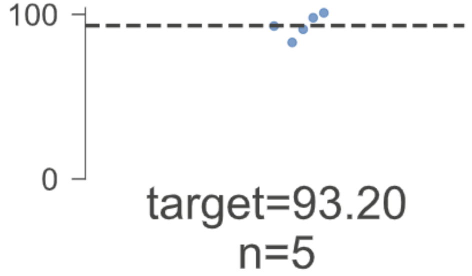




Experience	Salary
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

Experience	Salary
3	60
3.5	64
4	55
4.5	61
5	66



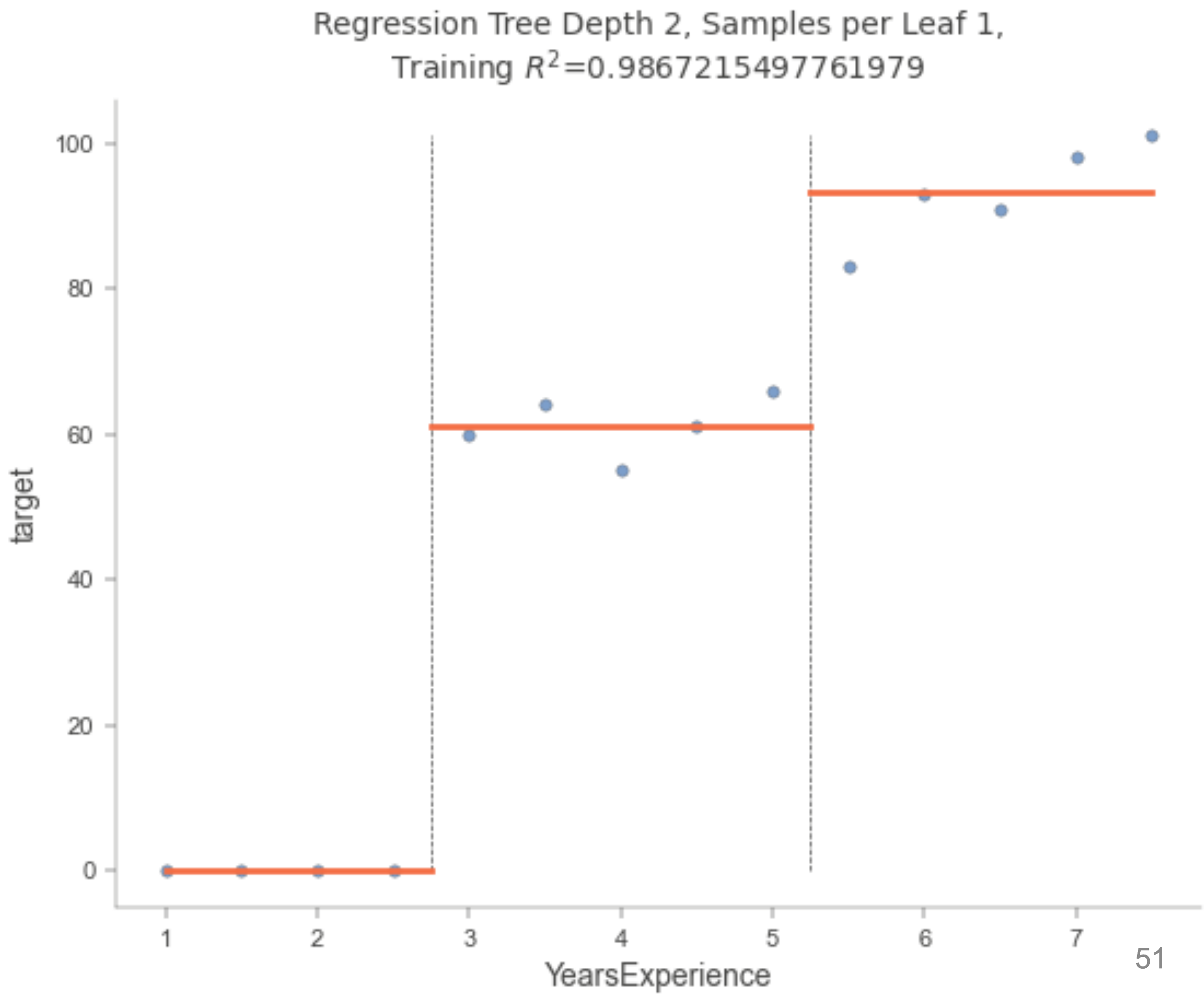


Experience	Salary
5.5	83
6	93
6.5	91
7	98
7.5	101

Regression Tree

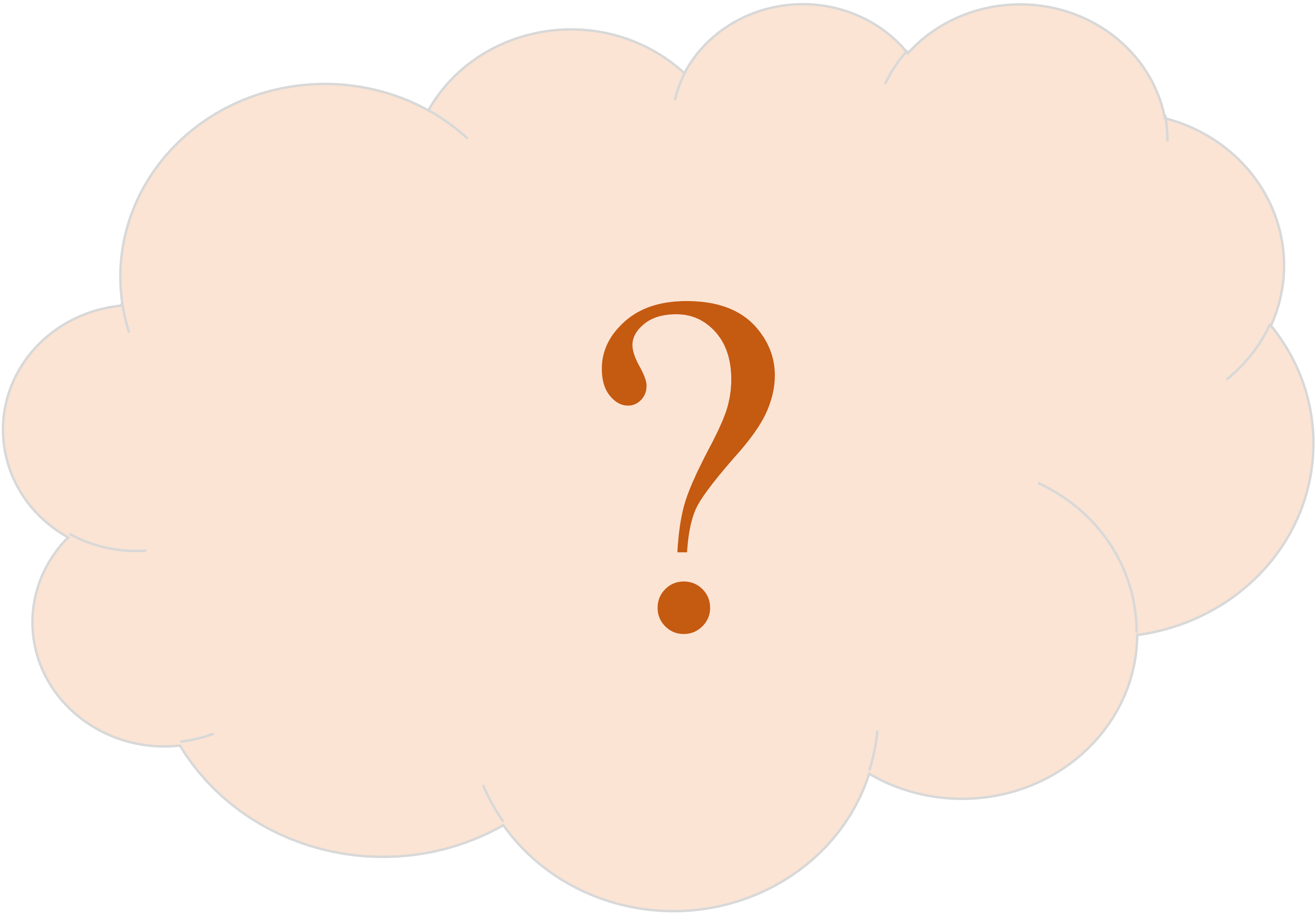
❖ Visualize regression line

Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101



Code Implementation

Question




Thank you!