AI VIETNAM
All-in-One Course

Decision Tree
(Warm-up Class)

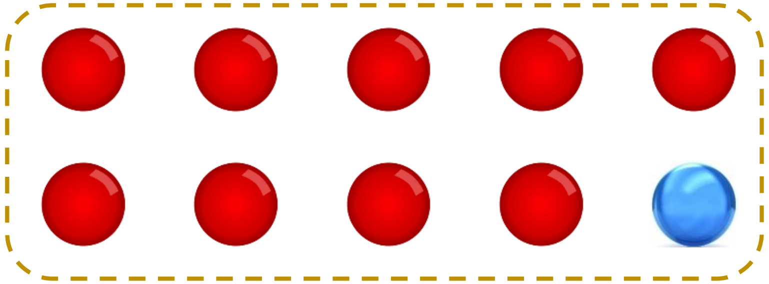
Quang-Vinh Dinh
Ph.D. in Computer Science

Year 2023

AI VIETNAM
All-in-One Course

Entropy

❖ Motivation



A: Get a red ball

B: Get a blue ball

$$p(A) = \frac{9}{10} = 0.9$$

$$p(B) = \frac{1}{10} = 0.1$$

E: Pick a ball from the basket

Experiment 1

Got a red ball 

Experiment 2

Got a blue ball 

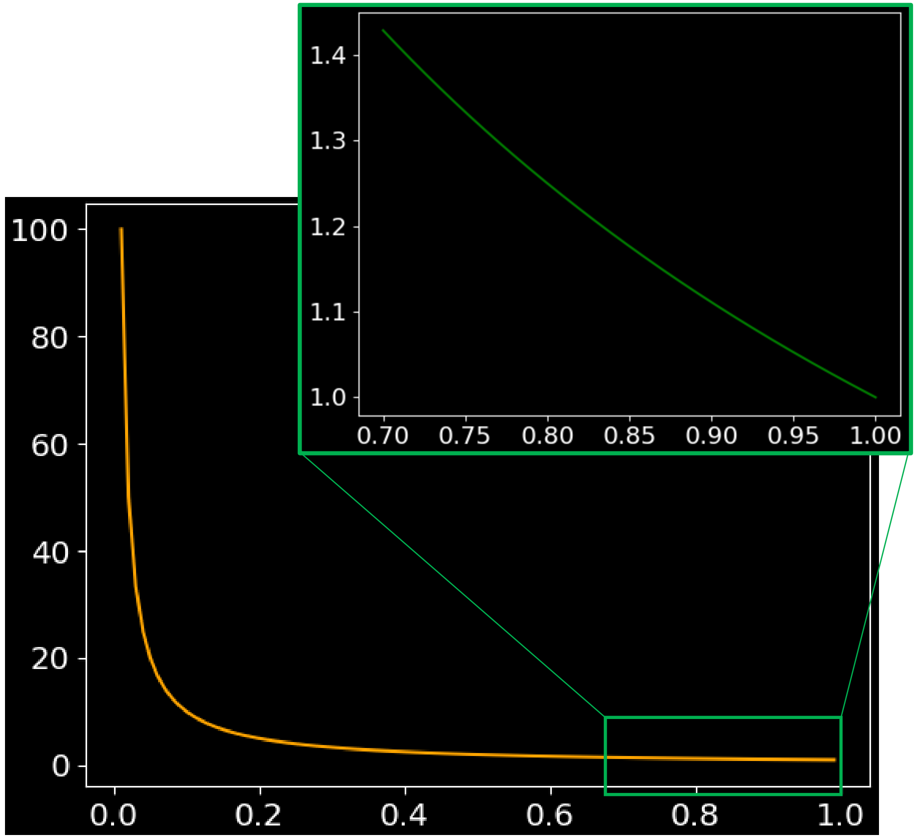
Which experiment makes you more surprised?

How to measure the surprises?

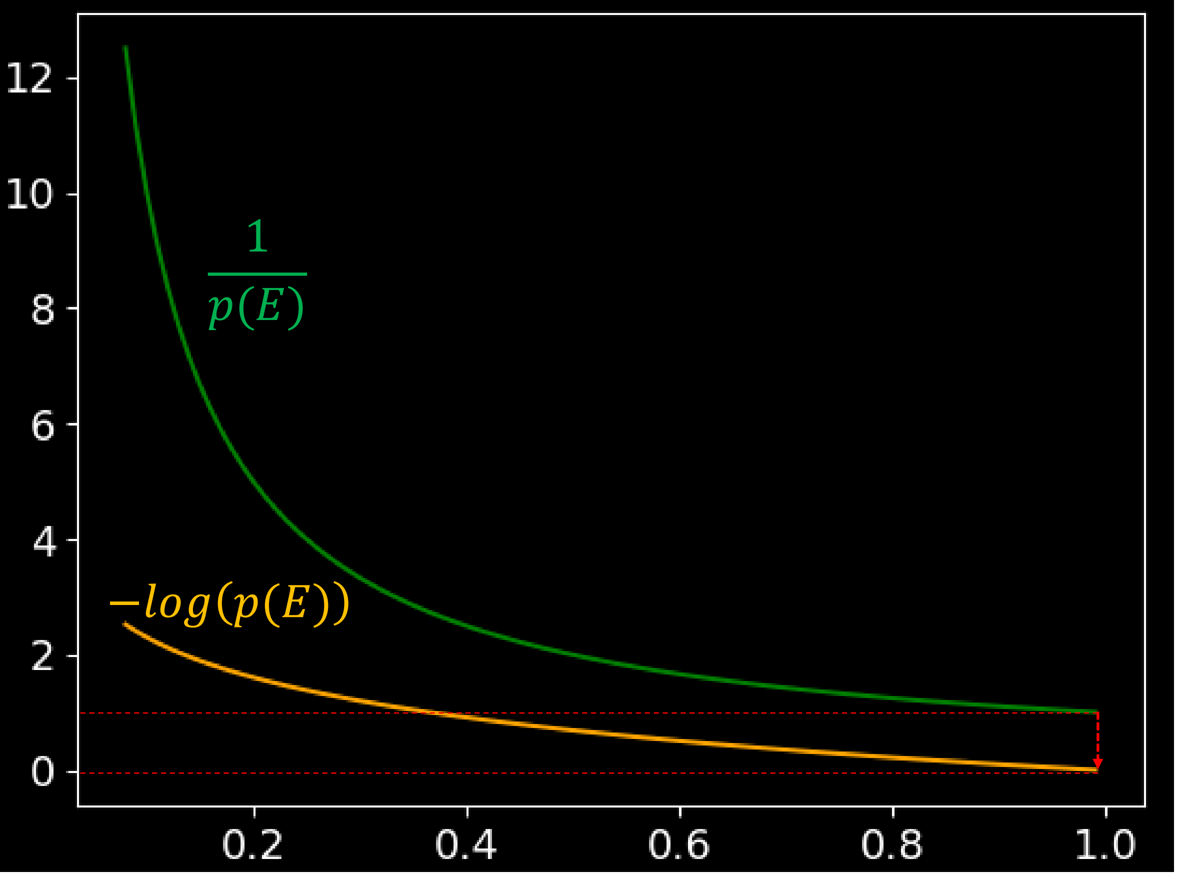
Observation

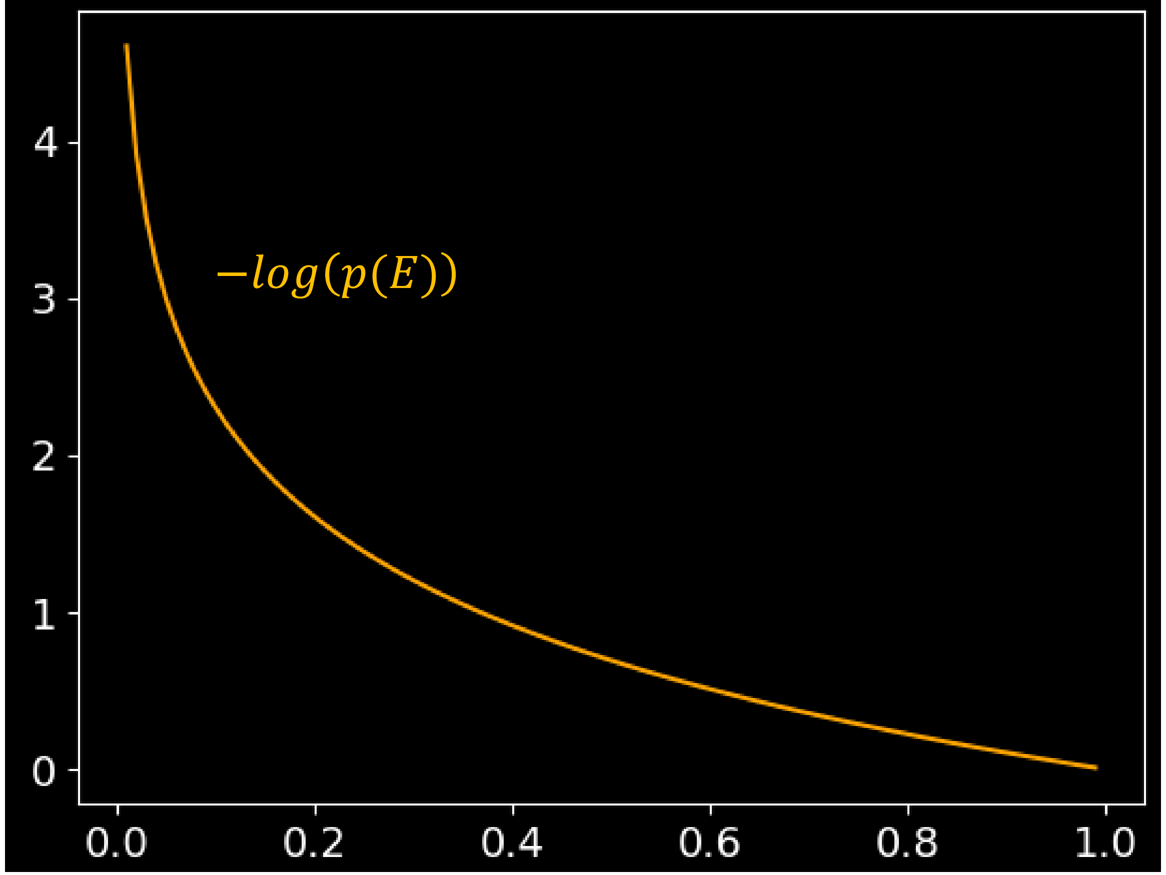
$$Surprise(E) \downarrow \quad \uparrow p(E)$$

$$\Rightarrow Surprise(E) = \frac{1}{p(E)}$$



Problem?





Monotonic decrease of the function surprise(E)

$$log(Surprise(E)) = log\left(\frac{1}{p(E)}\right)$$

$$= -log(p(E))$$

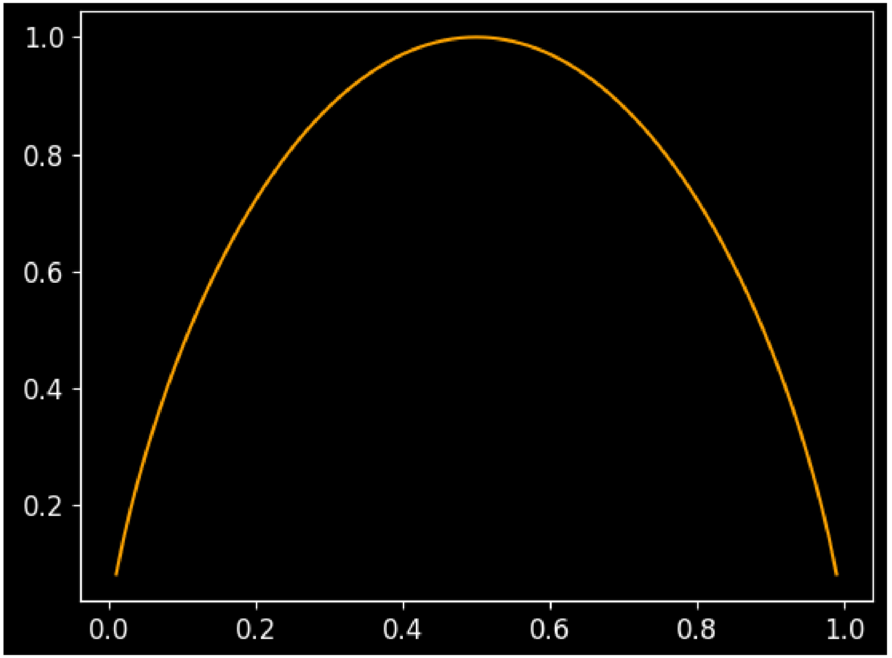
In information theory

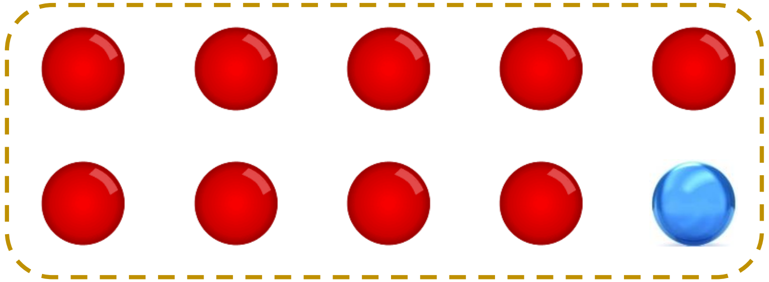
$$\text{Information}(x) = -log(p(x))$$

Entropy

Entropy: Average of information

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





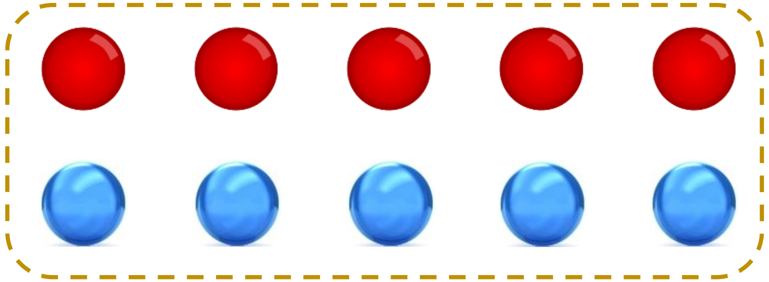
$$p(X = 0) = \frac{9}{10} = 0.9$$

$$p(X = 1) = \frac{1}{10} = 0.1$$

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

$$= -0.9log(0.9) - 0.1log(0.1)$$

$$= 0.468$$



$$p(X = 0) = \frac{5}{10} = 0.5$$

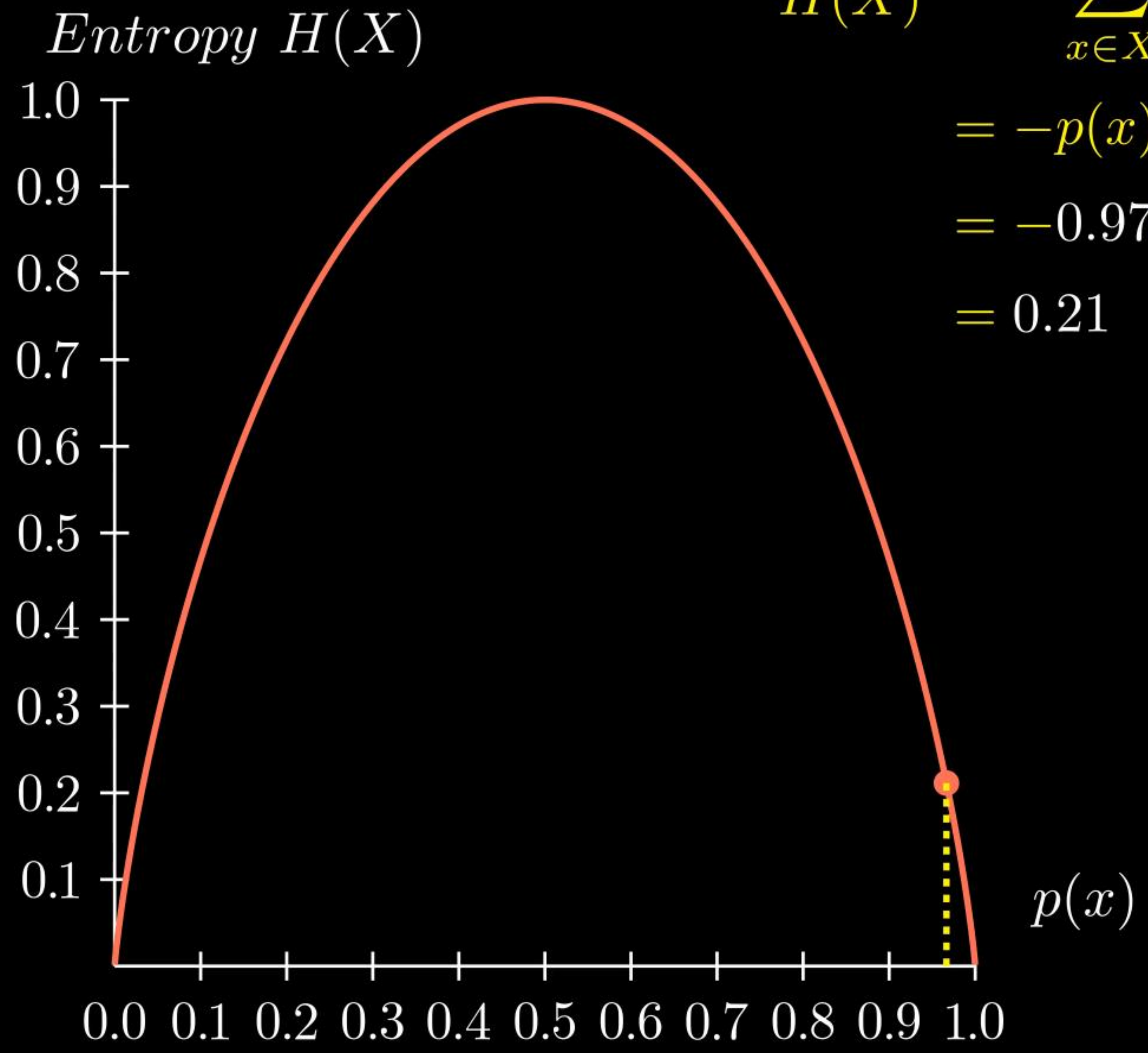
$$p(X = 1) = \frac{5}{10} = 0.5$$

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

$$= -0.5log(0.5) - 0.5log(0.5)$$

$$= 1.0$$

AI VIETNAM

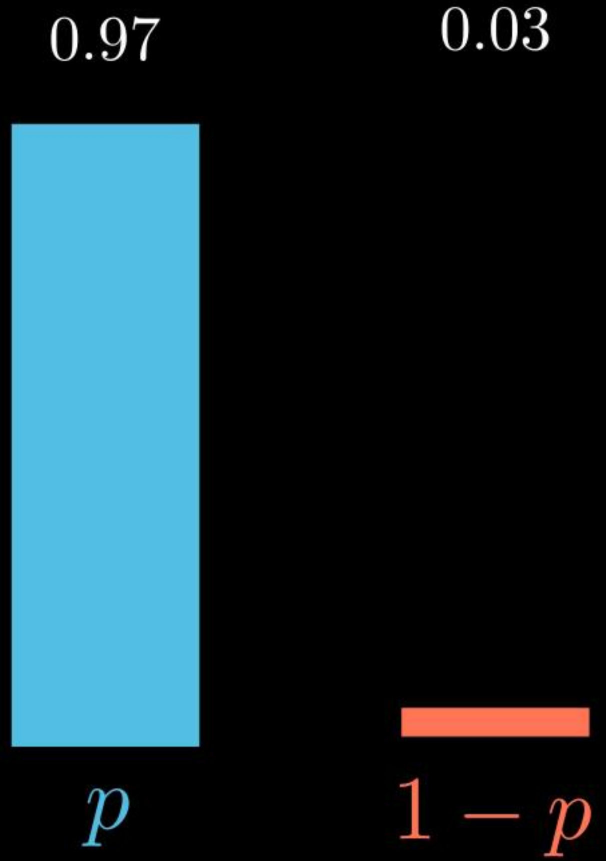


$$H(X) = -\sum_{x \in X} p(x) log_2(p(x))$$

$$= -p(x) log_2(p(x)) - (1 - p(x)) log_2(1 - p(x))$$

$$= -0.97 log_2(0.97) - (1 - 0.97) log_2(1 - 0.97)$$

$$= 0.21$$



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

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

Training phase

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

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

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

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

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

Category = 2

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

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

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

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

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

$$= 0.94 - \frac{5}{14} * 0.97 - \frac{9}{14} * 0.764 = 0.102$$

<u>Gain(S, Outlook) = 0.226</u>

Gain(S, Temp) = 0.015

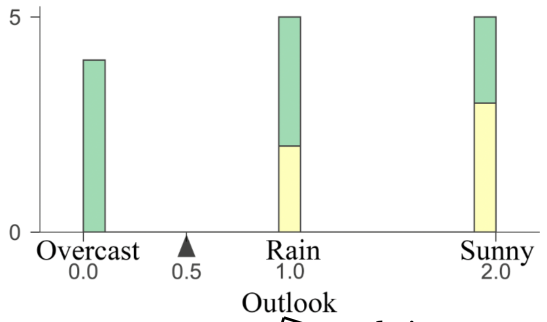
Gain(S, Humidity) = 0.151

Gain(S, Wind) = 0.048

Choose Outlook with highest Gain score for root node

Option_2 is used to split

Training phase



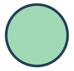
Play_Tennis
No
Yes

rain / sunny

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

overcast

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



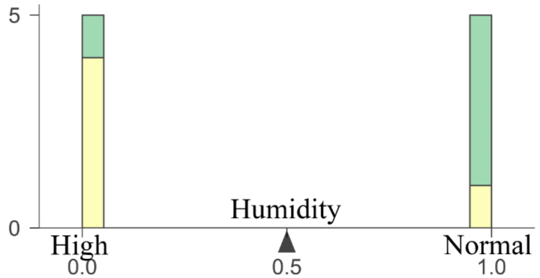
n = 4
Play_Tennis = Yes

Gain(S, Outlook) = 0.029

Gain(S, Temp) = 0.034

<u>Gain(S, Humidity) = 0.278</u>

Gain(S, Wind) = 0.124



high

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

normal

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

<u>Gain(S, Outlook) = 0.321</u>

Gain(S, Temp) = 0.17

Gain(S, Humidity) = 0.00

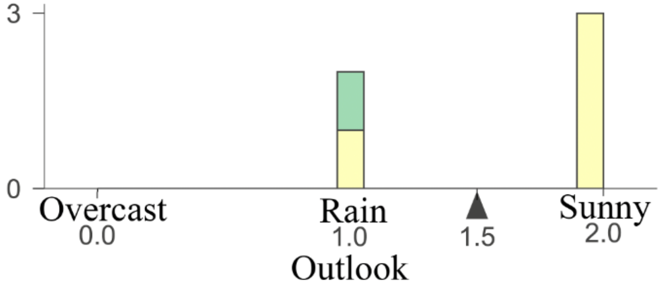
Gain(S, Wind) = 0.17

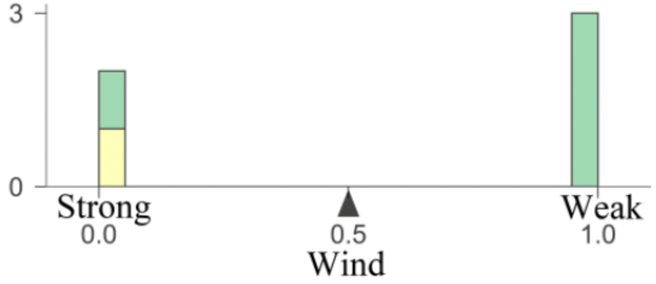
Gain(S, Outlook) = 0.17

Gain(S, Temp) = 0.17

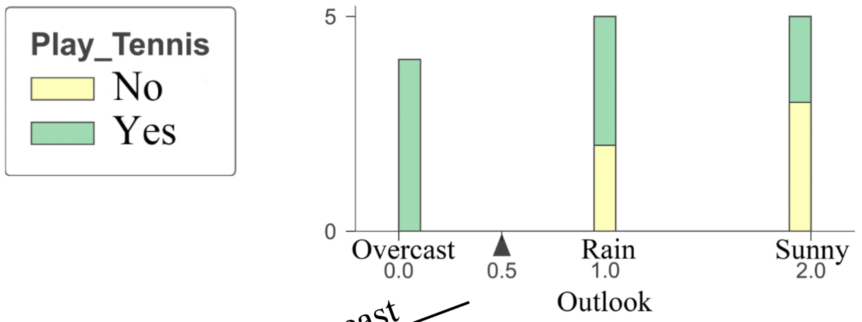
Gain(S, Humidity) = 0.00

<u>Gain(S, Wind) = 0.321</u>

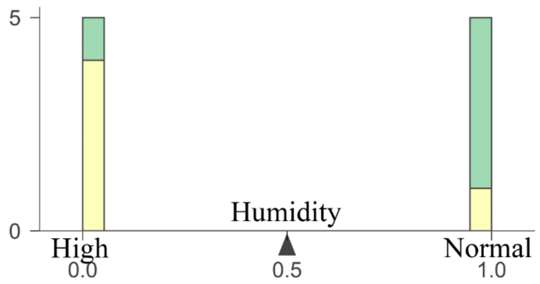




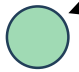
Training phase



rain/sunny



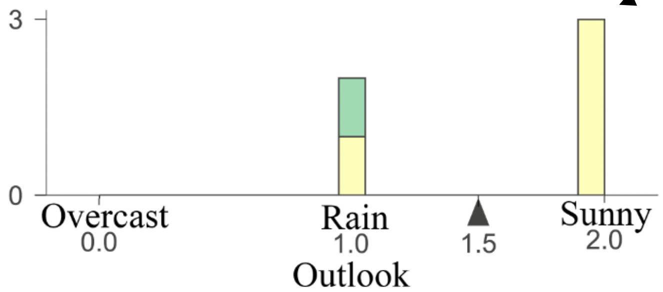
overcast

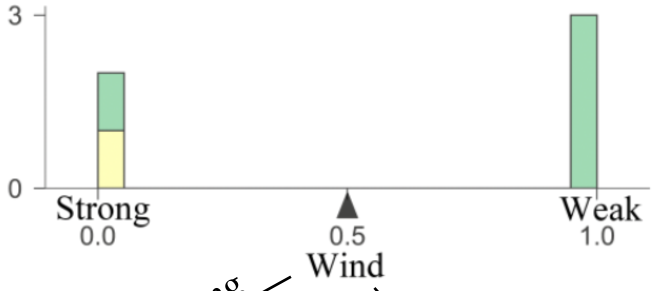


n = 4
Play_Tennis = Yes

high

normal





rain

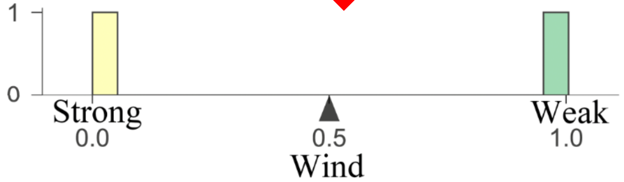
Outlook	Temp	Humidity	Wind	Label
Rain	Mild	High	Weak	Yes
Rain	Mild	High	Strong	No

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

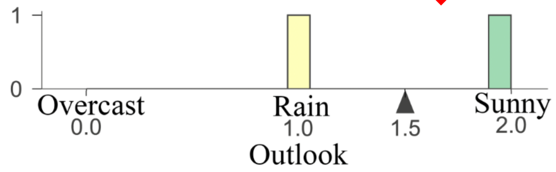
Outlook	Temp	Humidity	Wind	Label
Rain	Cool	Normal	Strong	No
Sunny	Mild	Normal	Strong	Yes

Gain(S, Outlook) = 1

Gain(S, Temp) = 1

Gain(S, Humidity) = 0.00

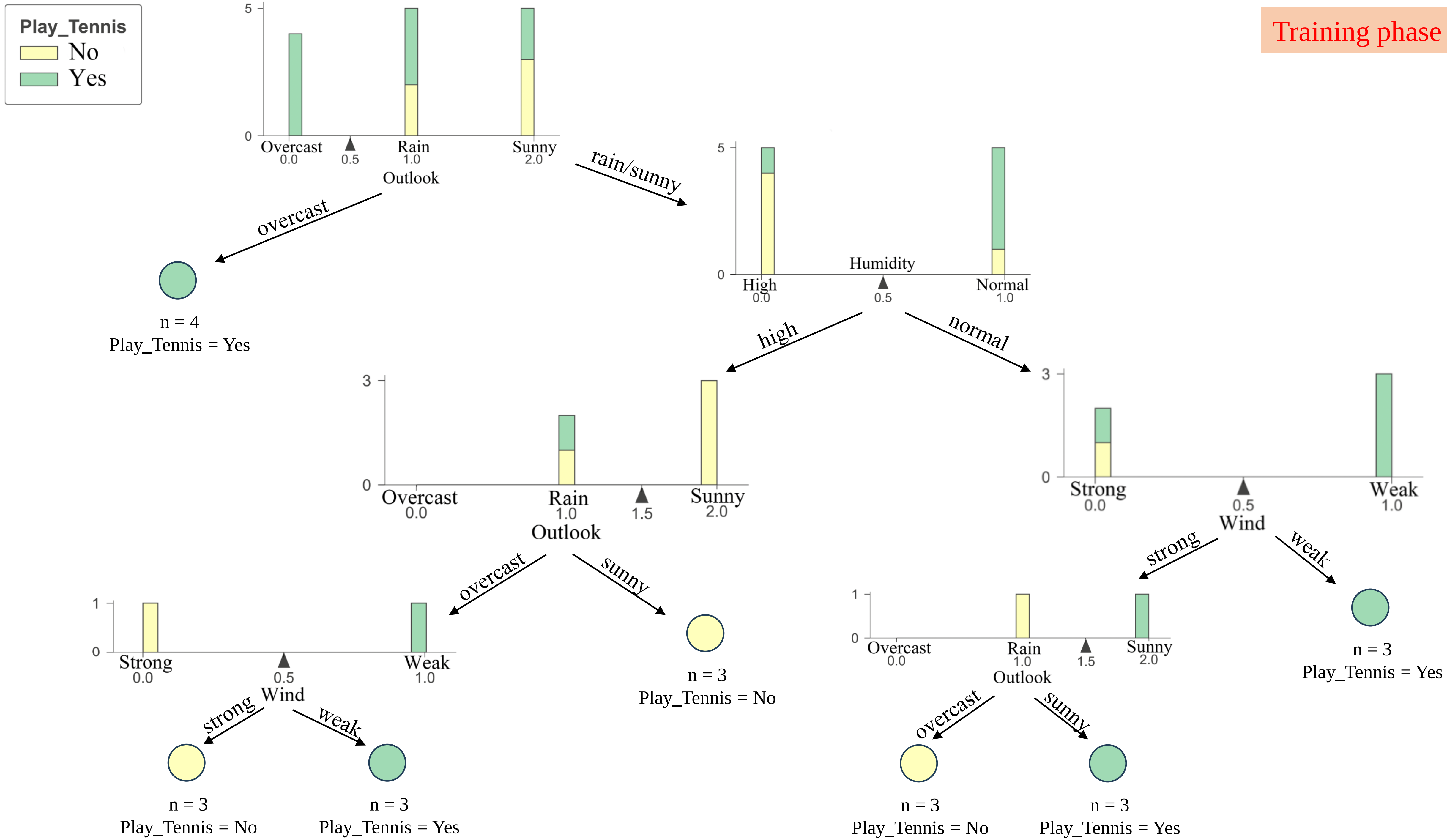
Gain(S, Wind) = 0.00



weak

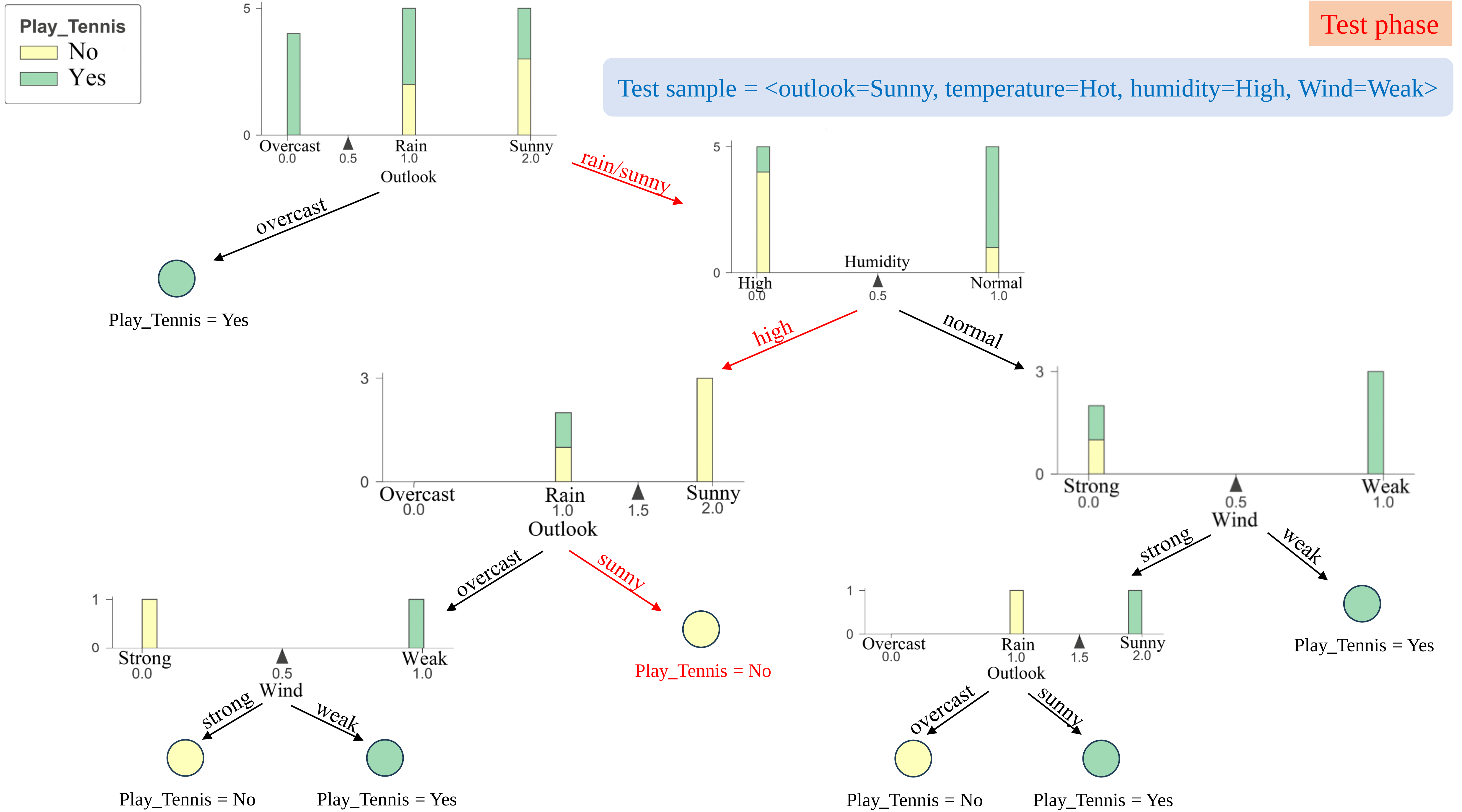
Outlook	Temp	Humidity	Wind	Label
Rain	Cool	Normal	Weak	Yes
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes

Training phase



Test phase

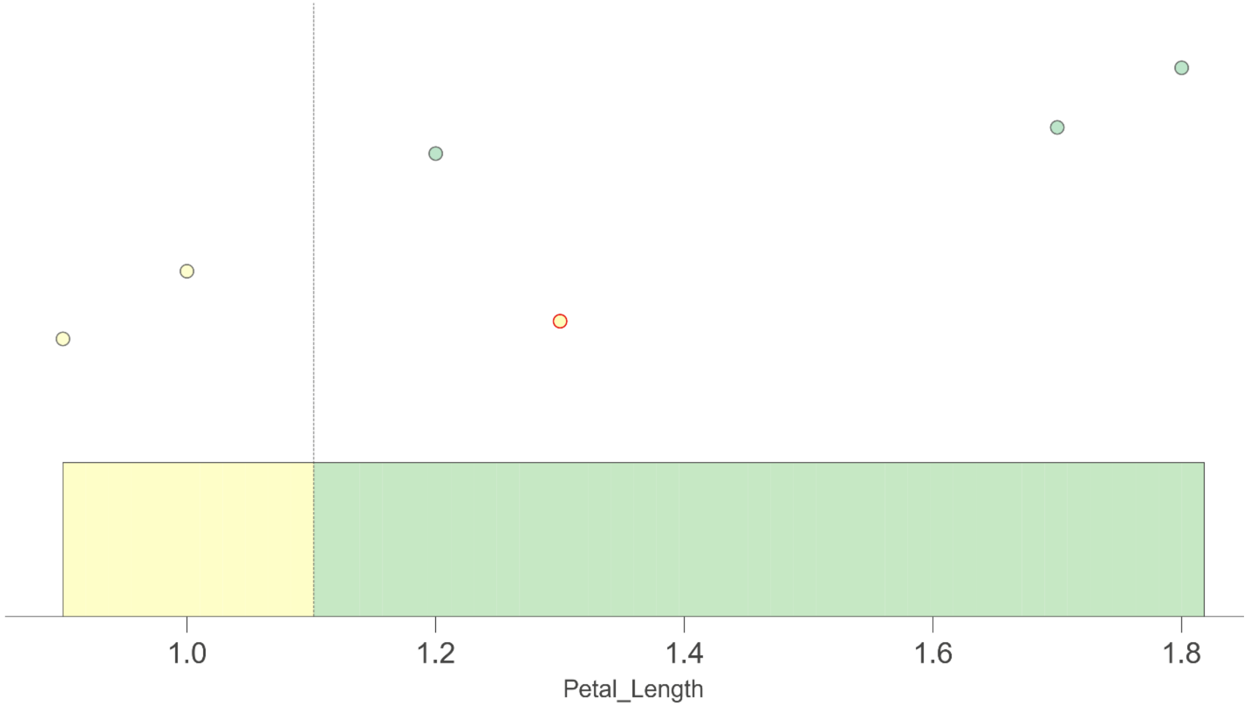
Test sample = <outlook=Sunny, temperature=Hot, humidity=High, Wind=Weak>

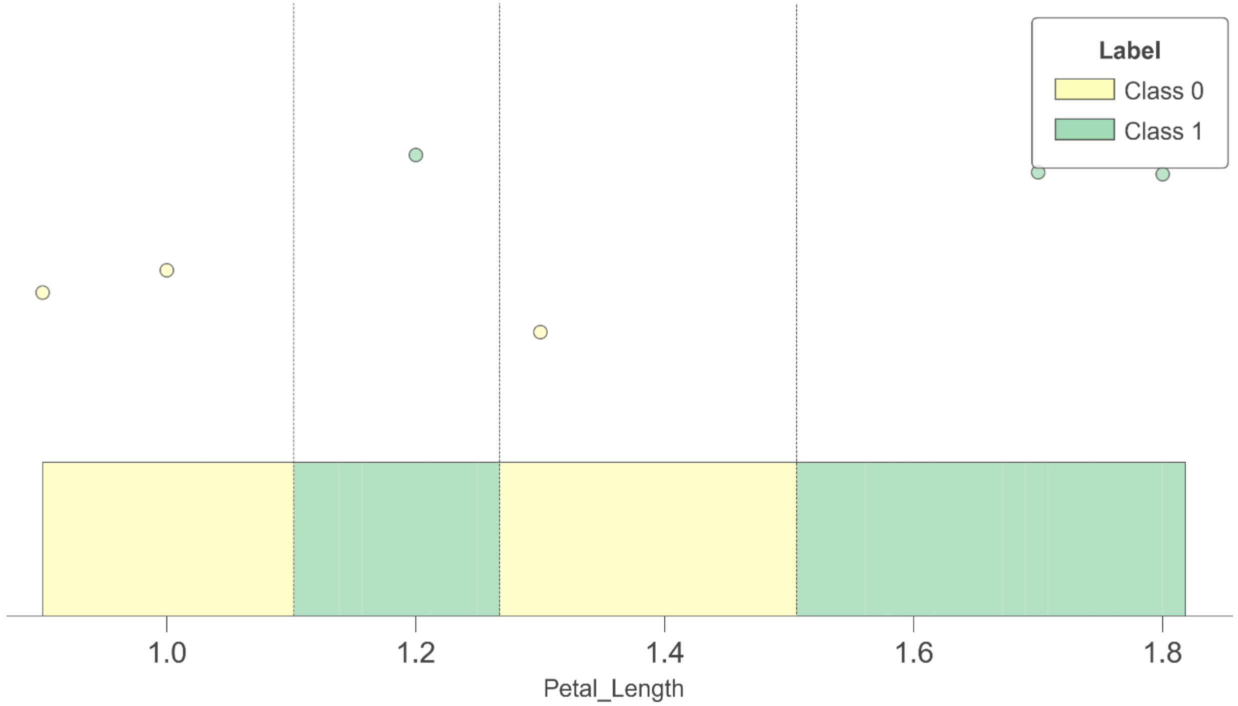


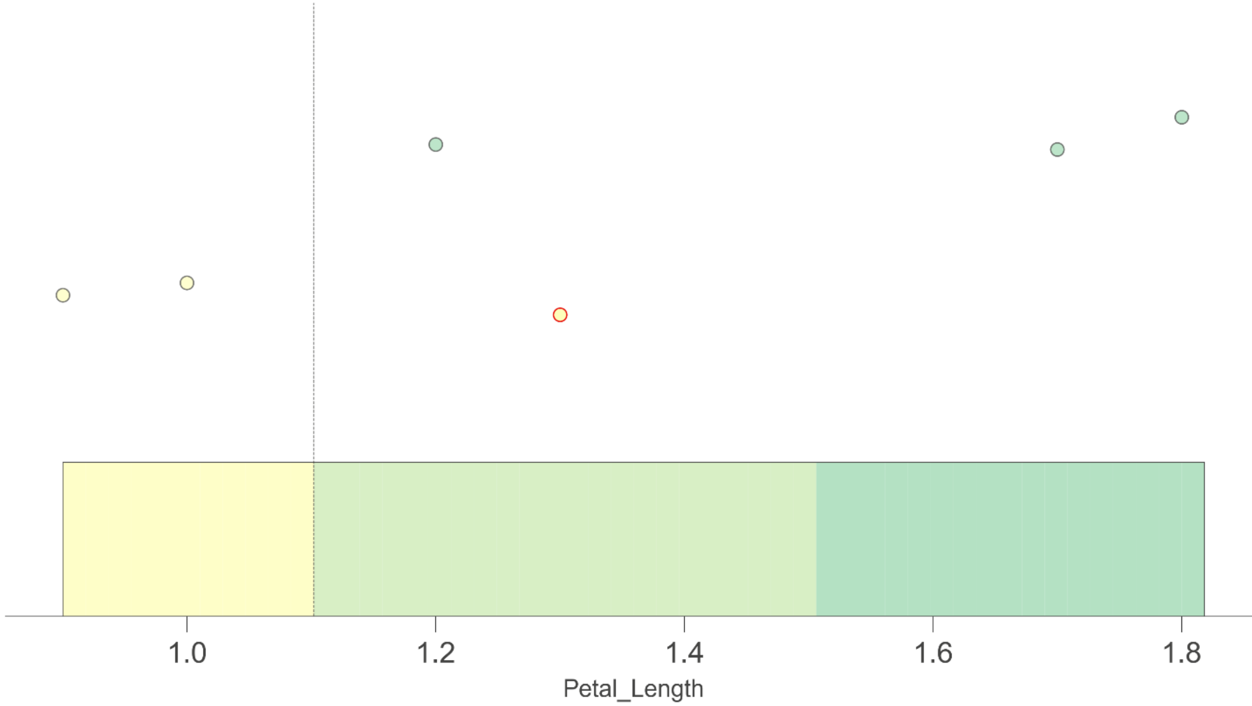
Classification

❖ 1D-IRIS

Petal_Length	Label
1	0
1.3	0
0.9	0
1.7	1
1.8	1
1.2	1



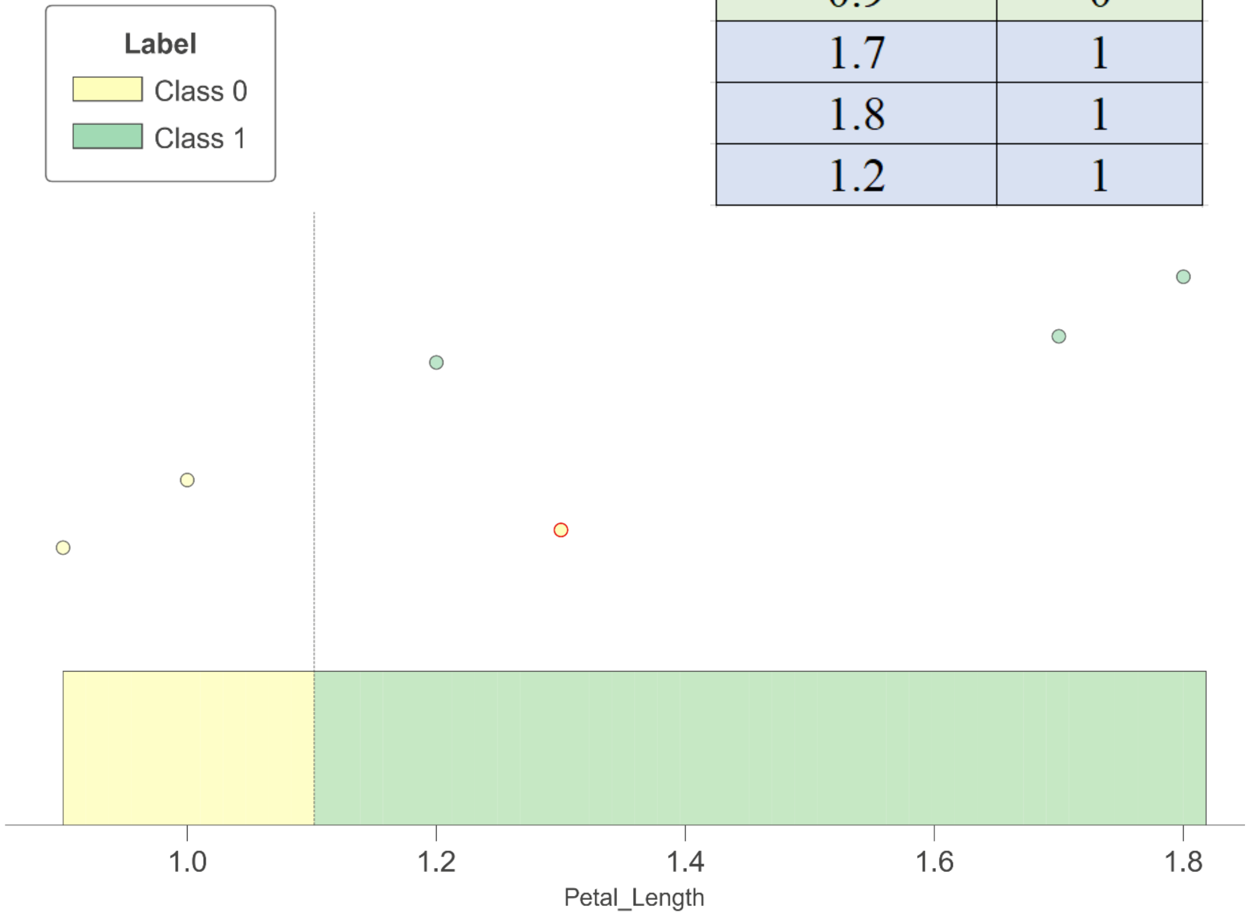


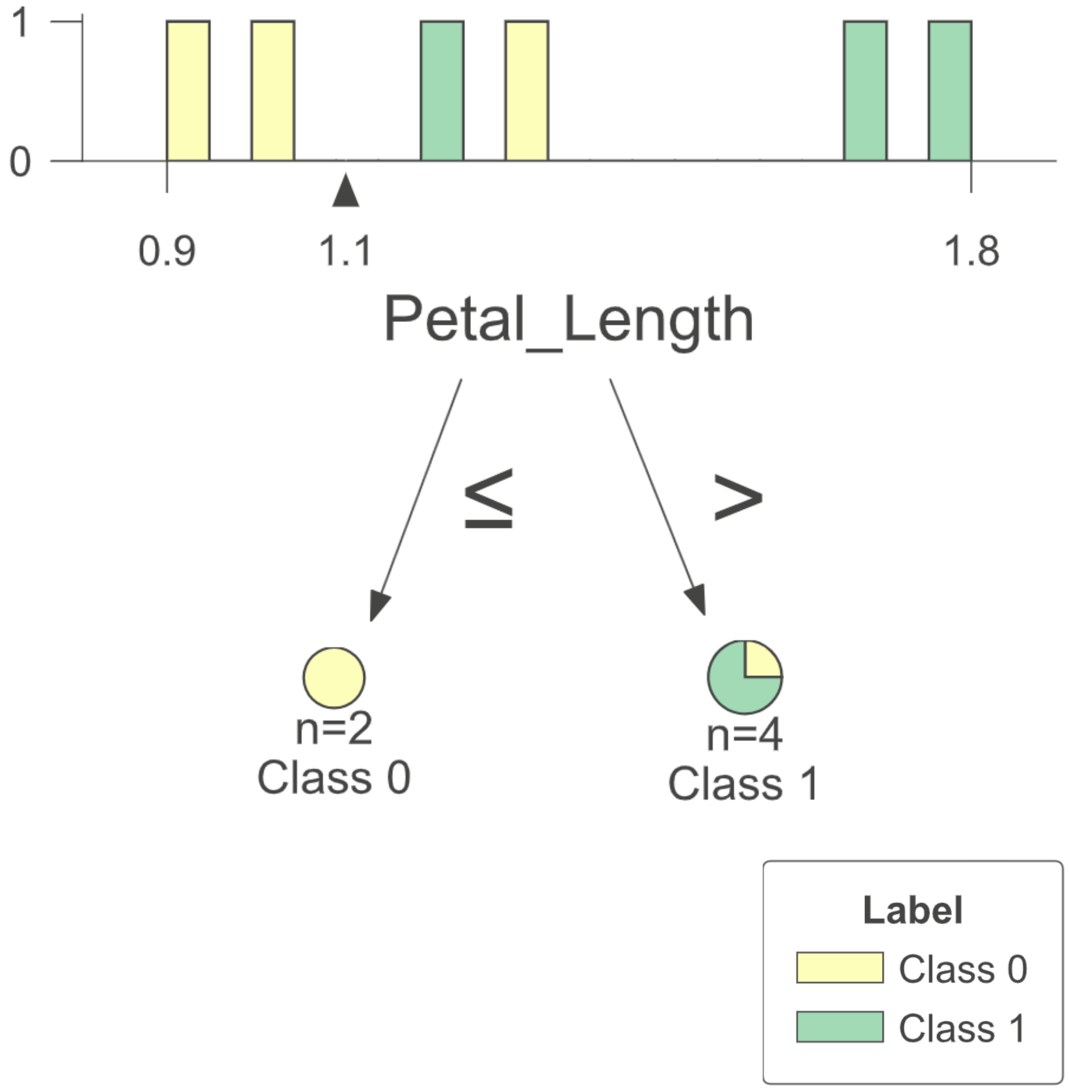


Classification

❖ 1D-IRIS

Petal_Length	Label
1	0
1.3	0
0.9	0
1.7	1
1.8	1
1.2	1

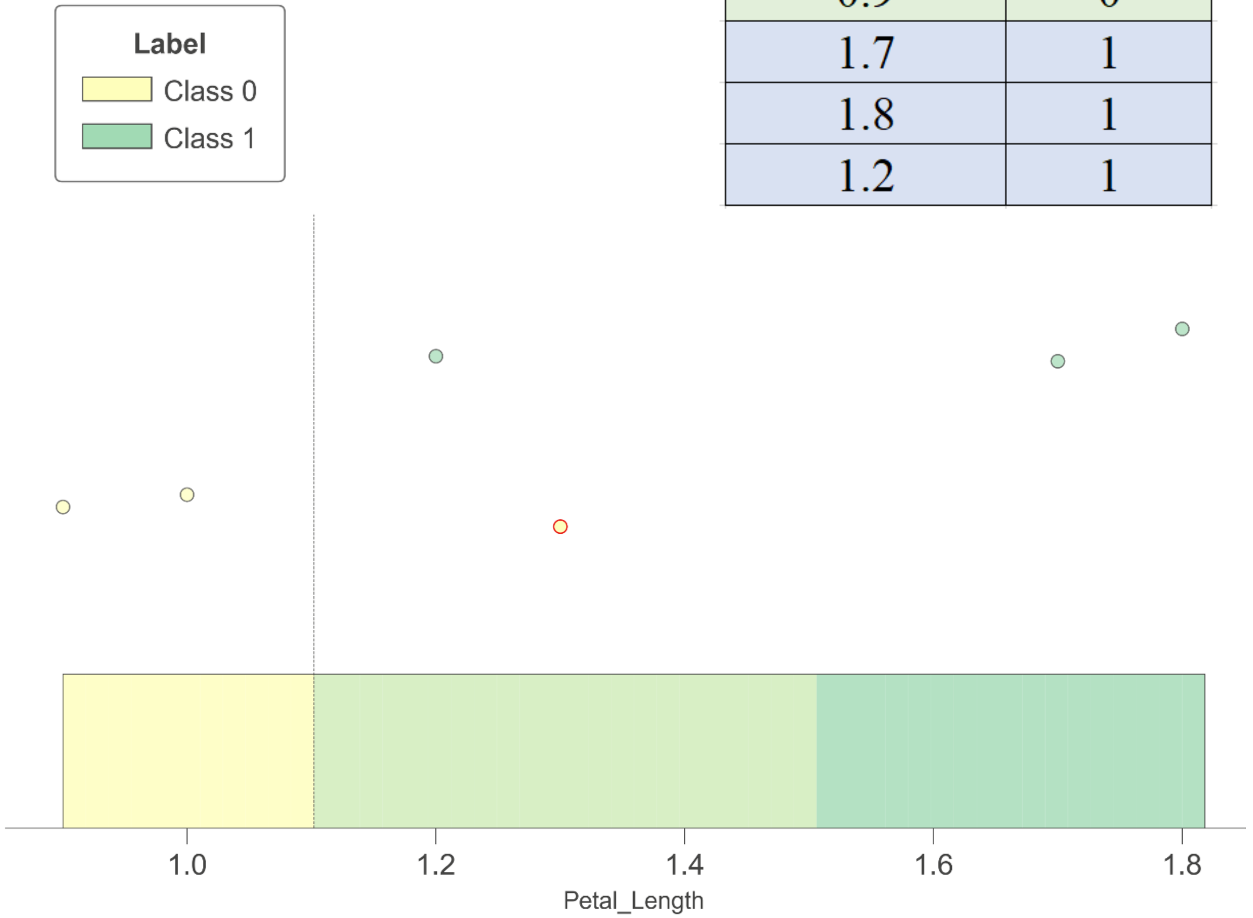


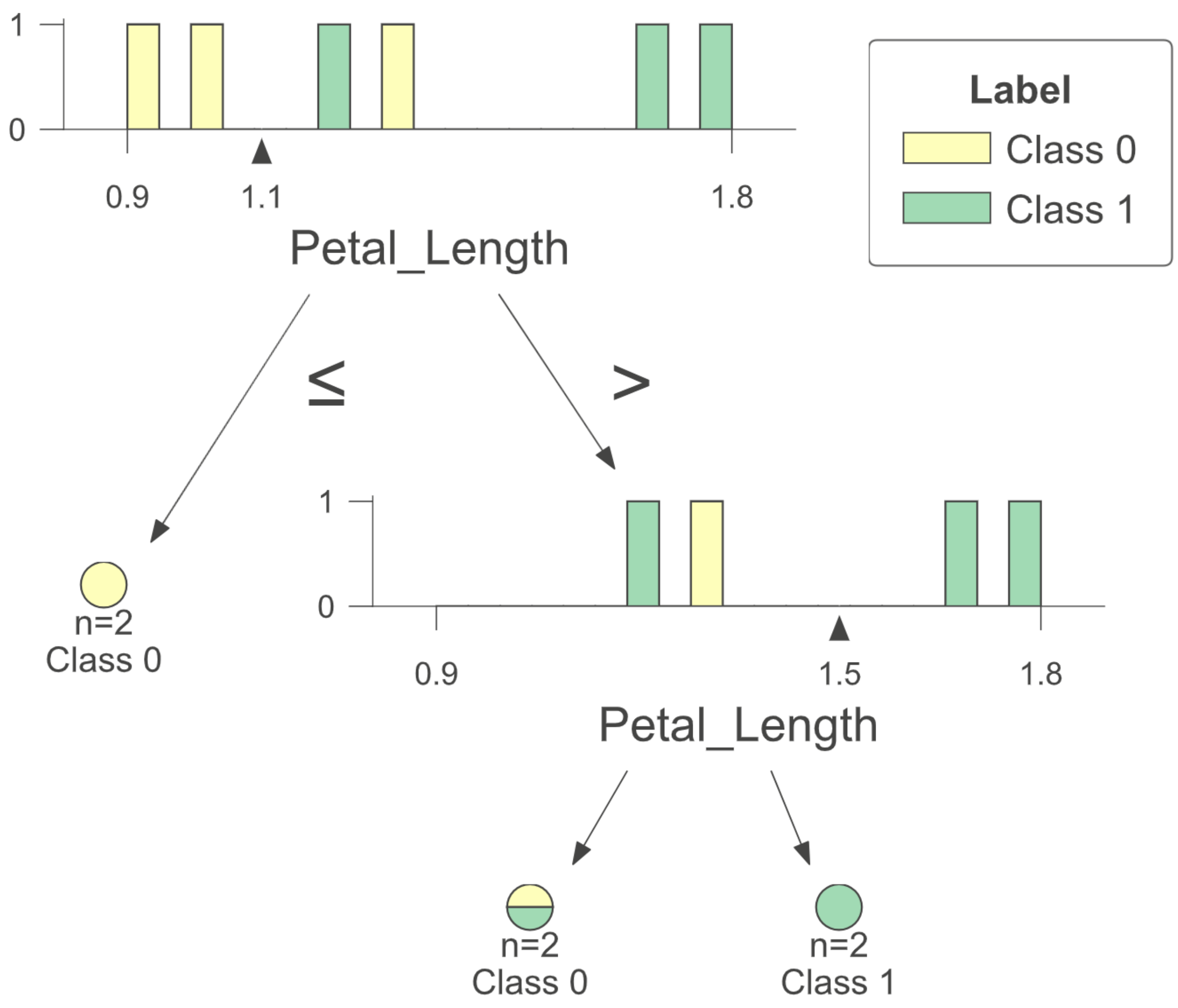


Classification

❖ 1D-IRIS

Petal_Length	Label
1	0
1.3	0
0.9	0
1.7	1
1.8	1
1.2	1



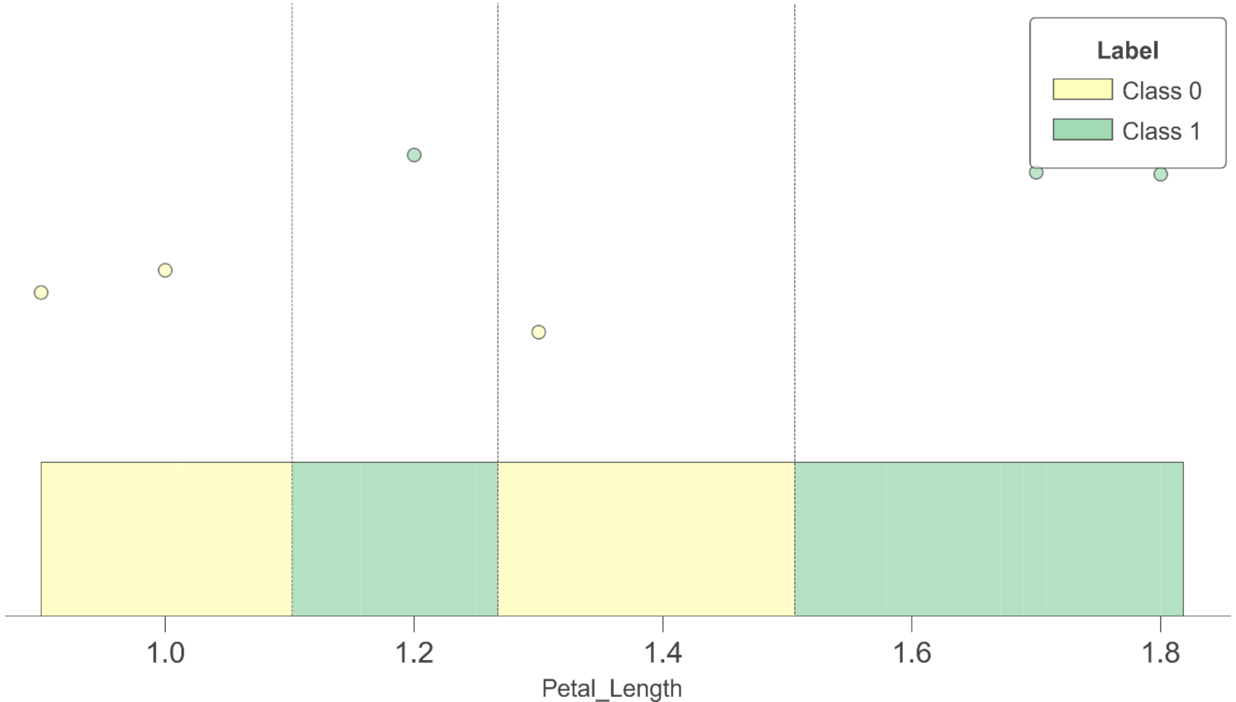


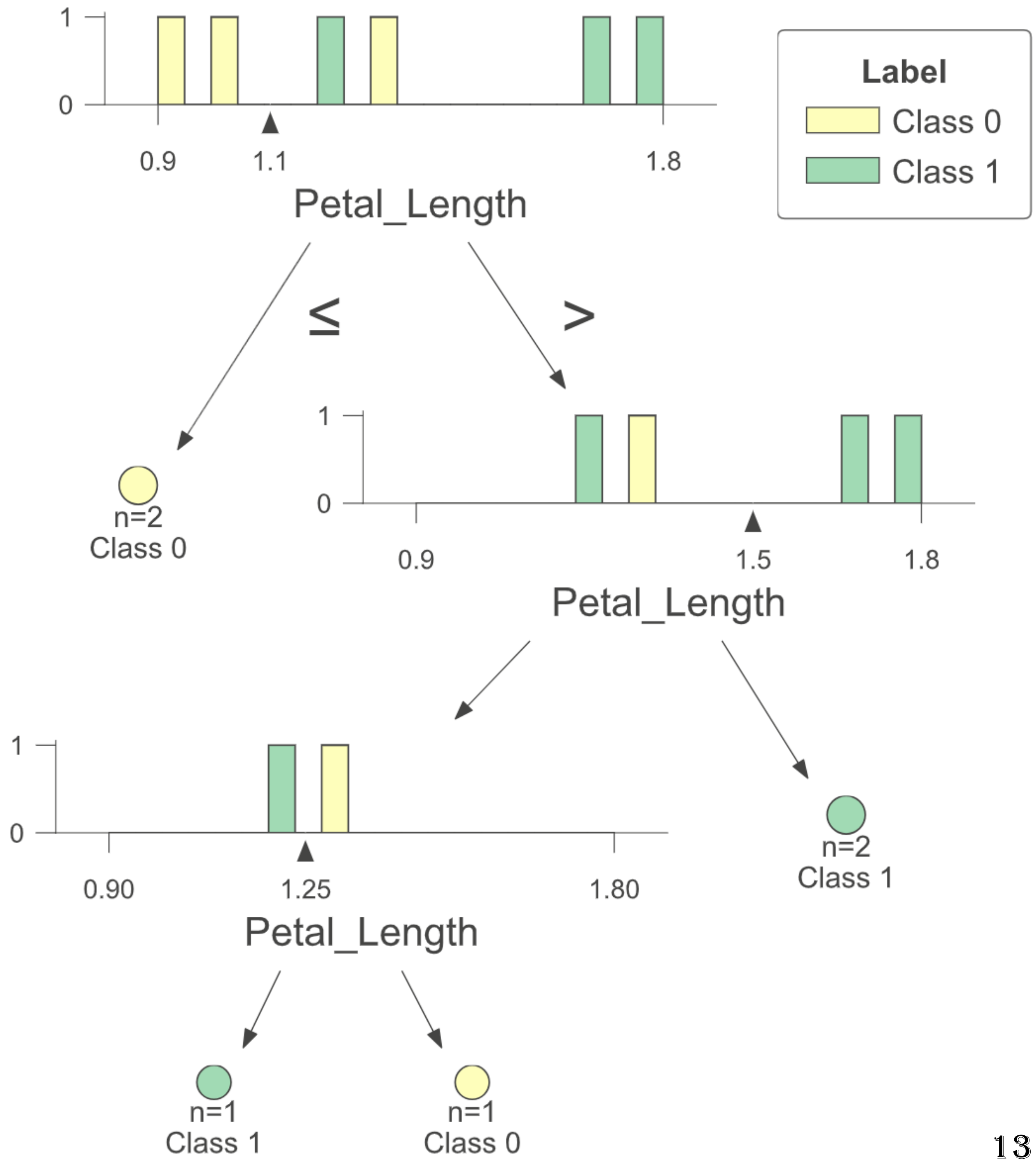
AI VIETNAM
All-in-One Course

Classification

❖ 1D-IRIS

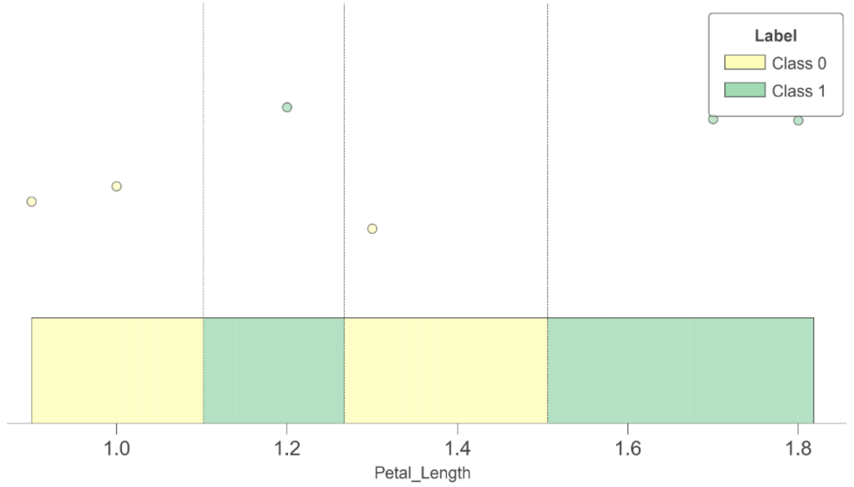
Petal_Length	Label
1	0
1.3	0
0.9	0
1.7	1
1.8	1
1.2	1



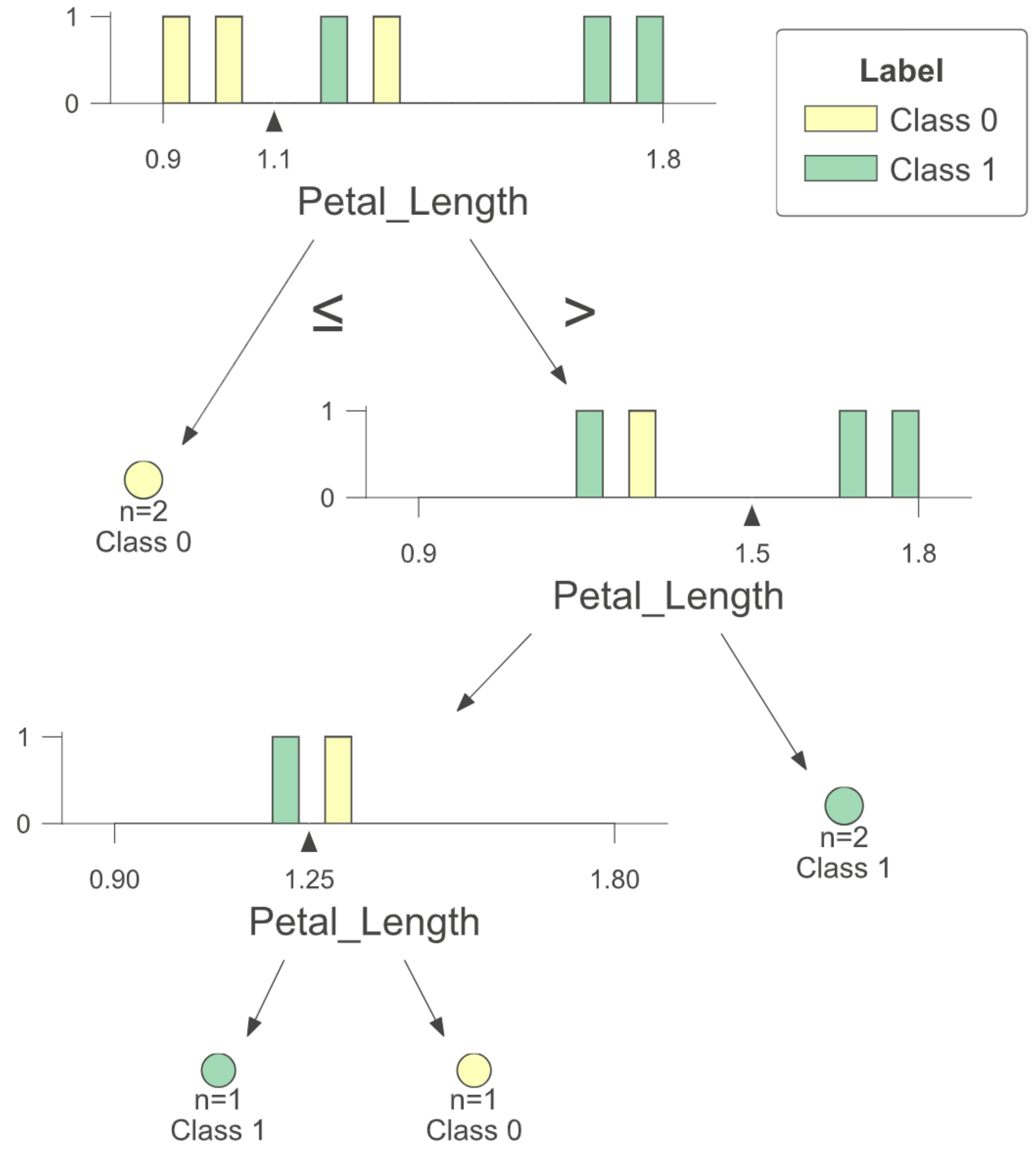


DT - Classification

❖ 1D-IRIS

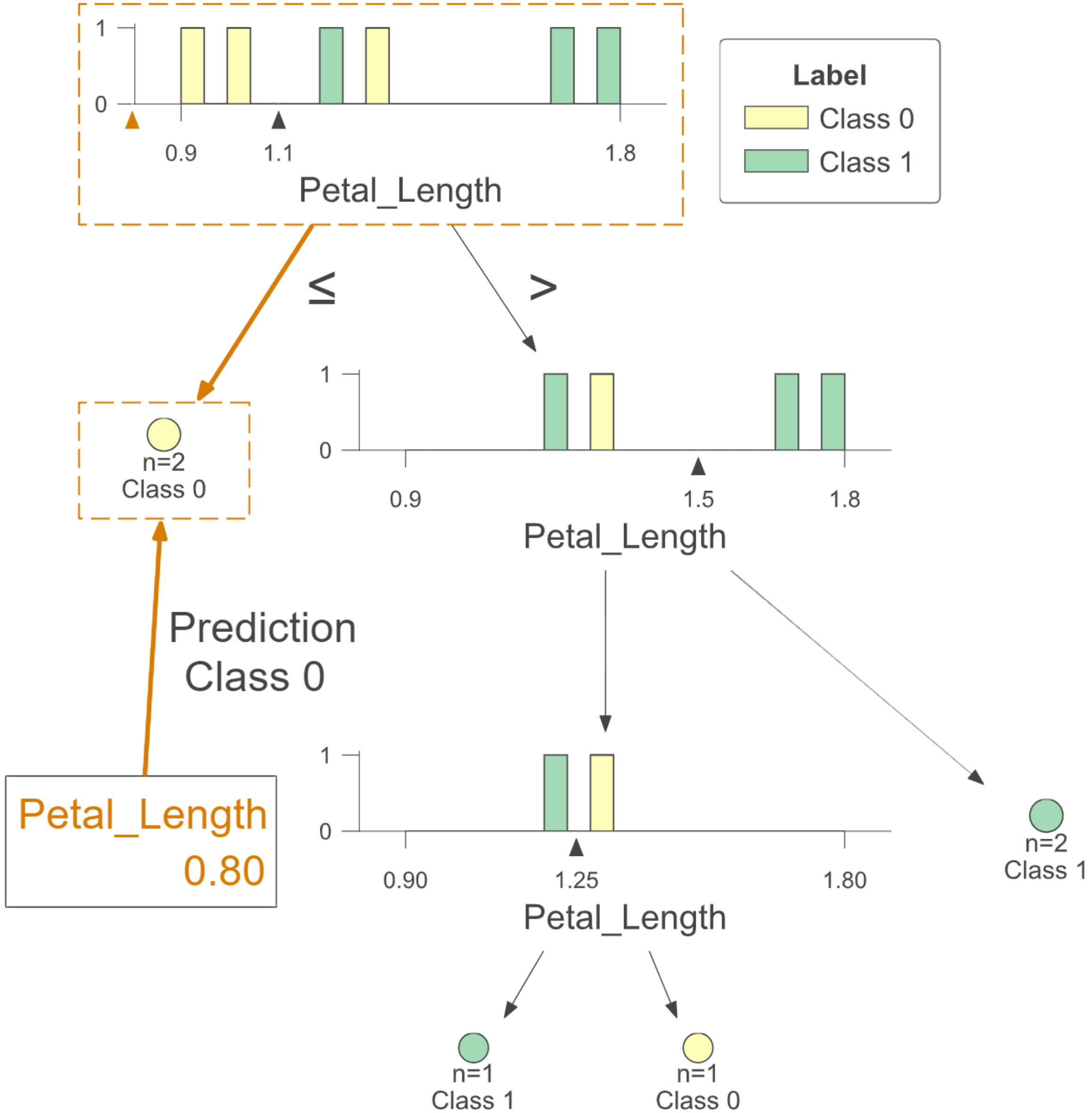


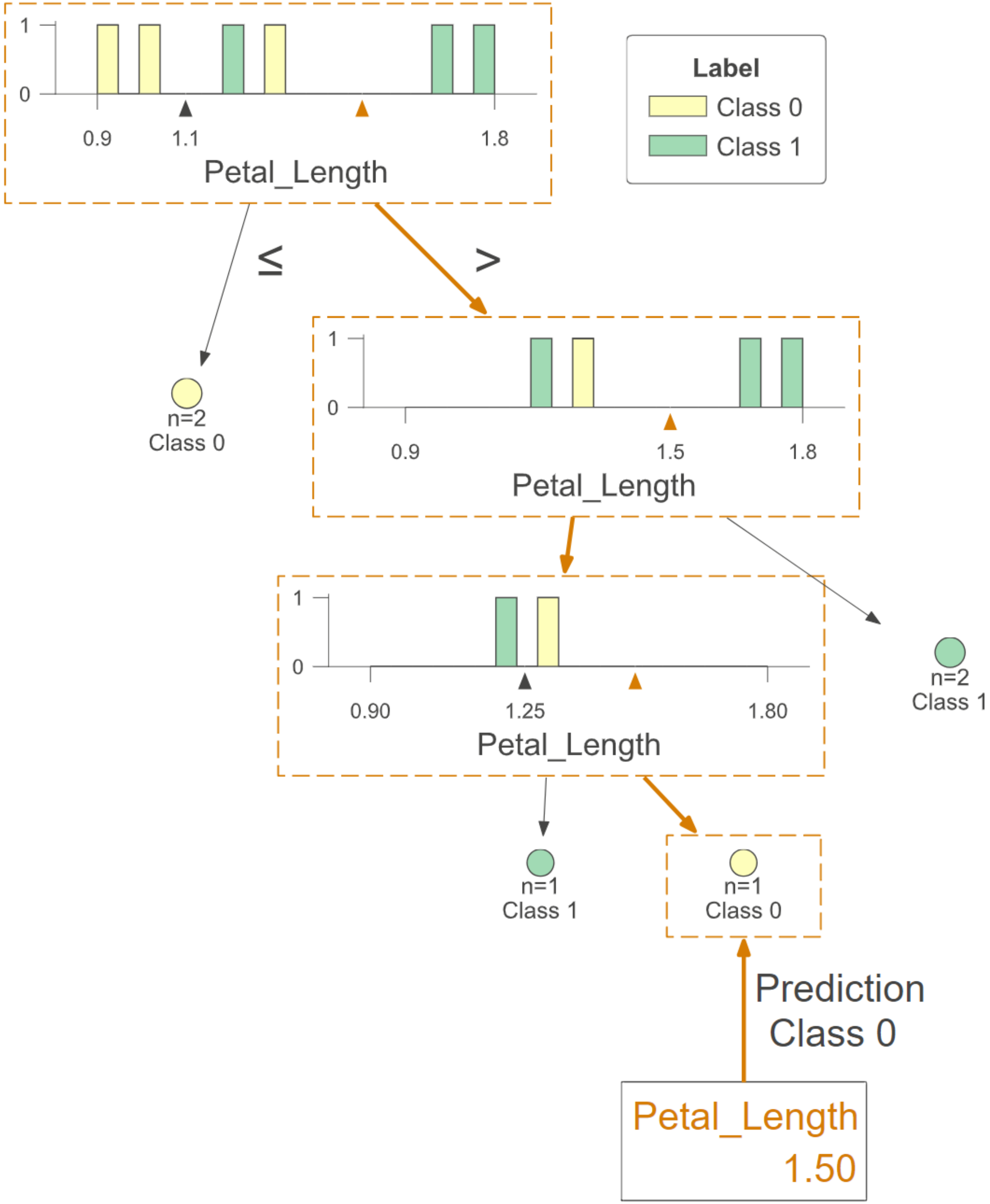
Petal_Length	Label
1	0
1.3	0
0.9	0
1.7	1
1.8	1
1.2	1




```
X = data['Petal_Length'].values.reshape(-1,1)
y = data['Label']

clf = DecisionTreeClassifier(max_depth=3,
                             criterion='entropy')
clf.fit(X, y)
```

Classification

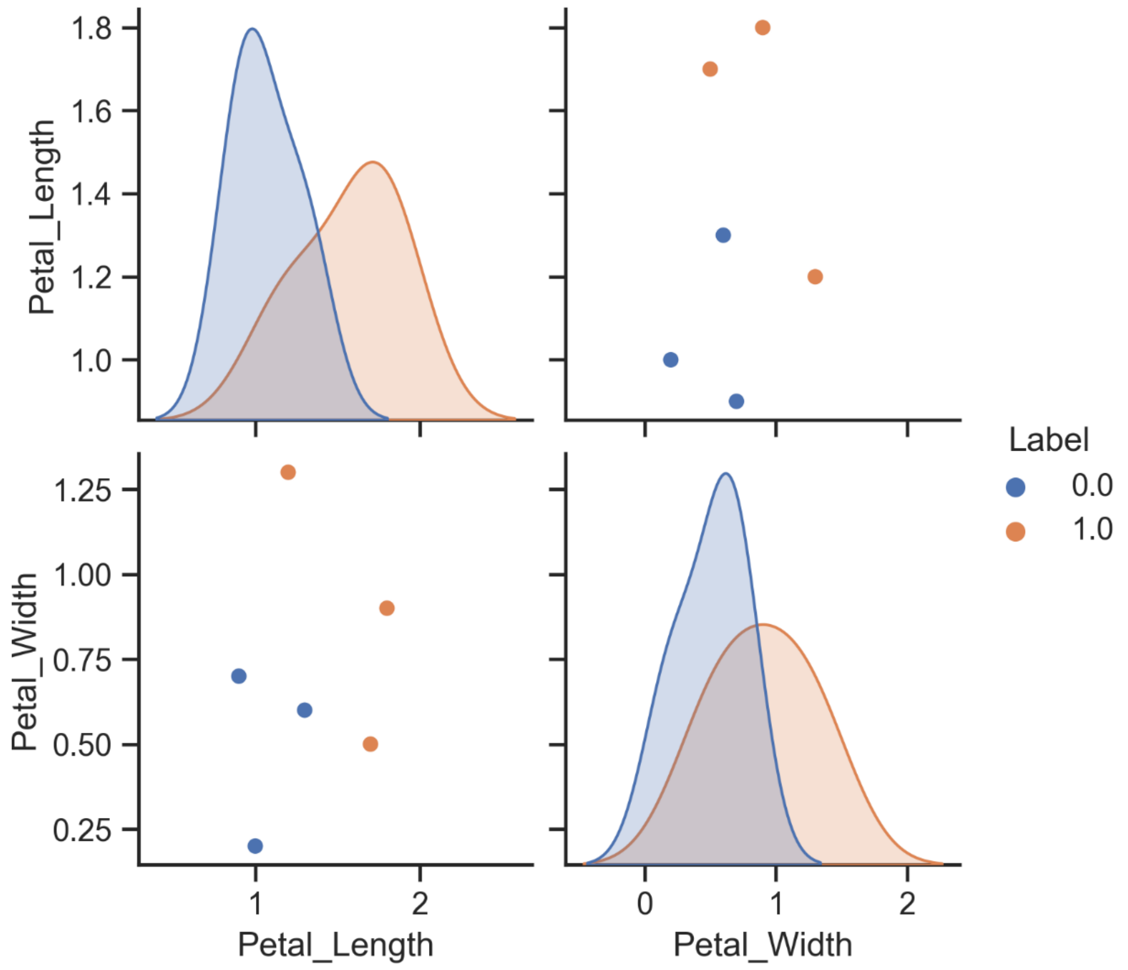


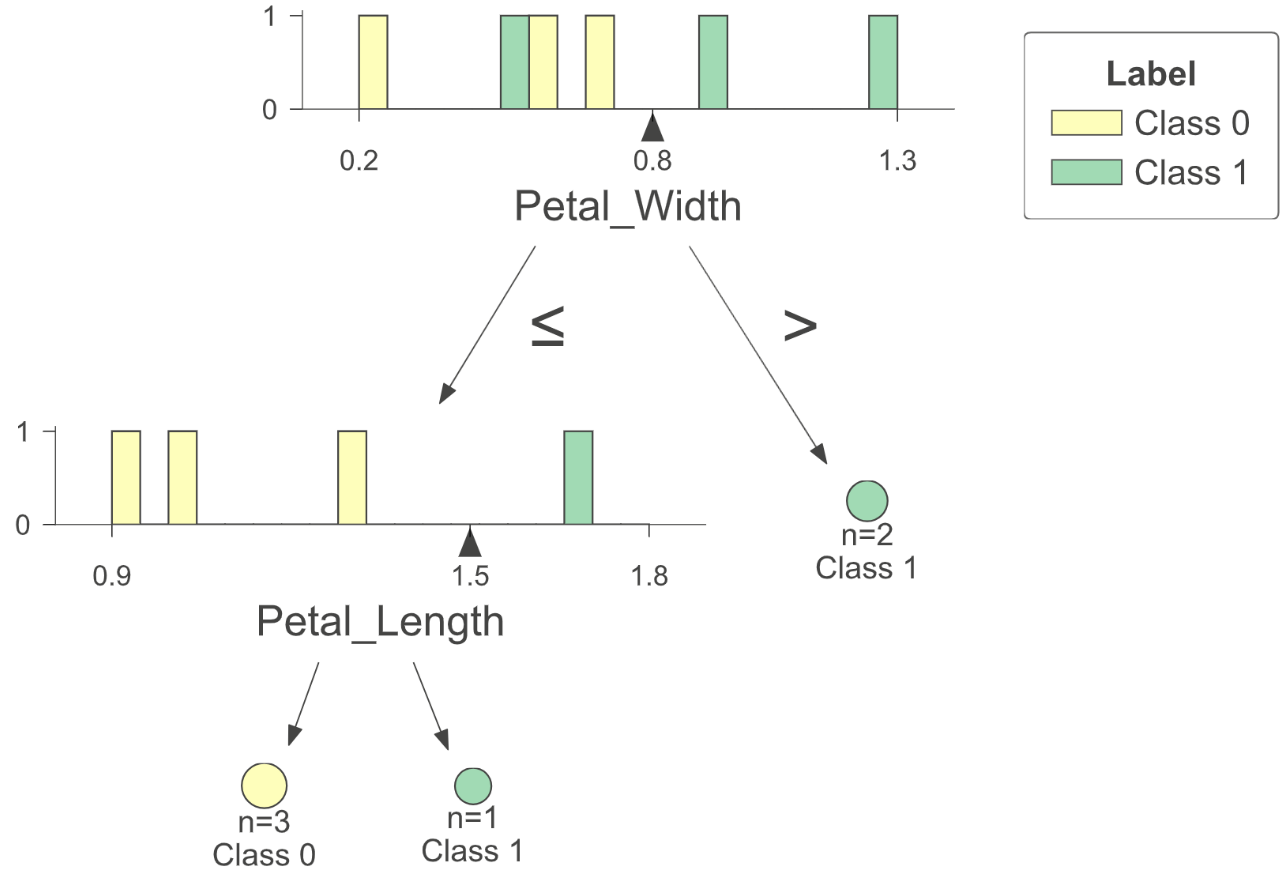


Classification

Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1
1.8	0.9	1
1.2	1.3	1

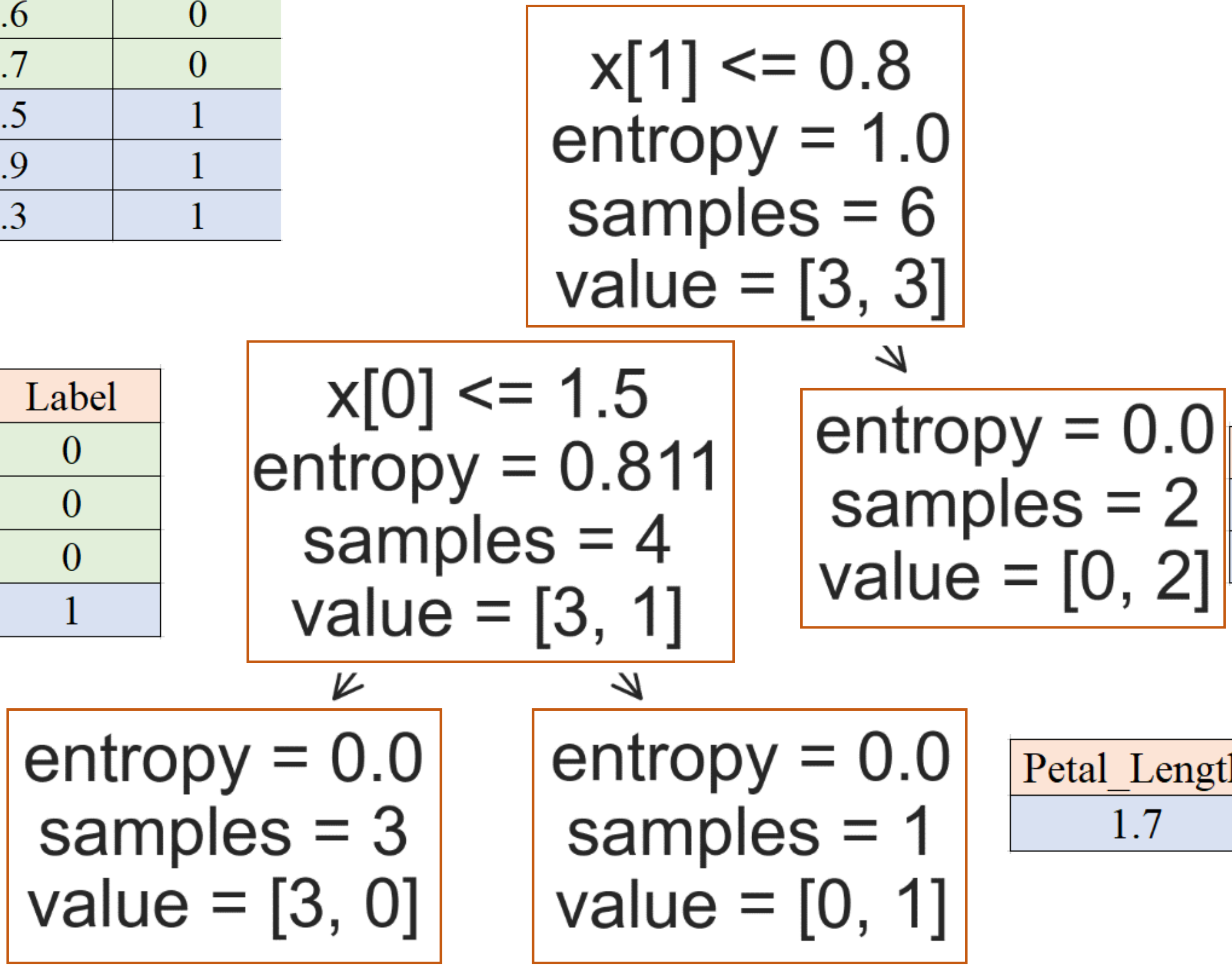
❖ Simple IRIS





Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1
1.8	0.9	1
1.2	1.3	1

```
|--- feature_1 <= 0.80
|   |--- feature_0 <= 1.50
|   |   |--- class: 0
|   |--- feature_0 >  1.50
|   |   |--- class: 1
|--- feature_1 >  0.80
|   |--- class: 1
```



Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1

Petal_Length	Petal_Width	Label
1.8	0.9	1
1.2	1.3	1

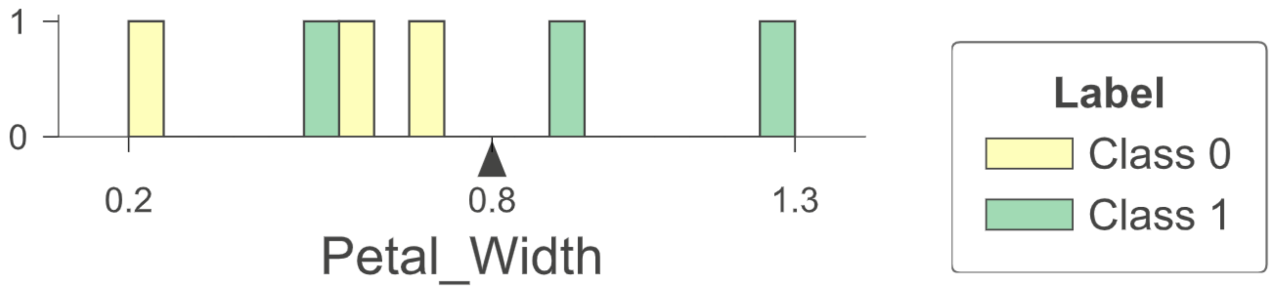
Petal_Length	Petal_Width	Label
1.7	0.5	1

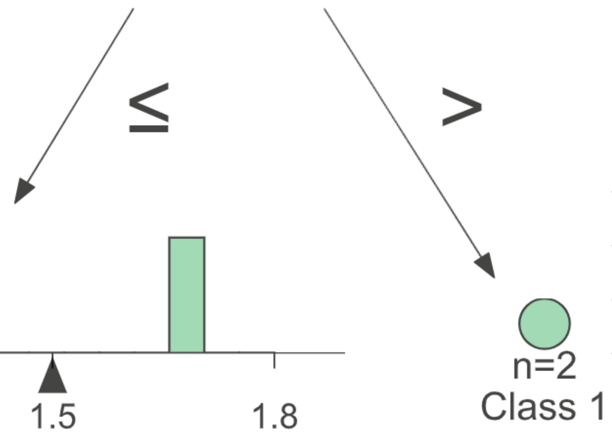
Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0

Classification Tree

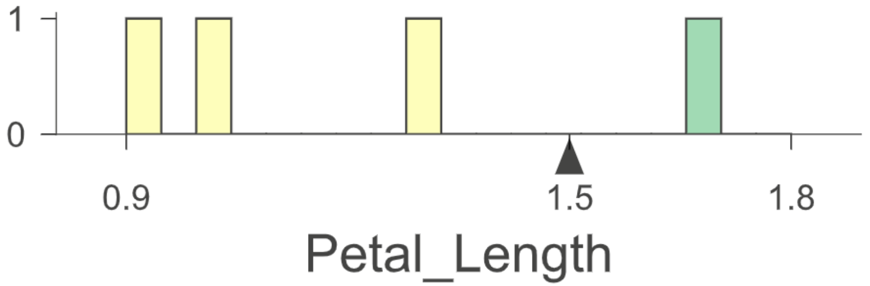
Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1
1.8	0.9	1
1.2	1.3	1

Simple IRIS

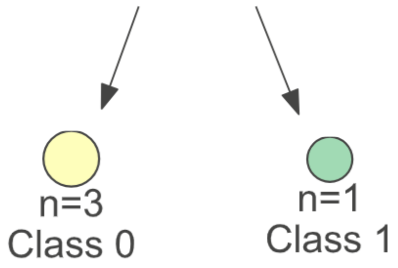




Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1



Petal_Length	Petal_Width	Label
1.8	0.9	1
1.2	1.3	1



Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0

Petal_Length	Petal_Width	Label
1.7	0.5	1

Classification

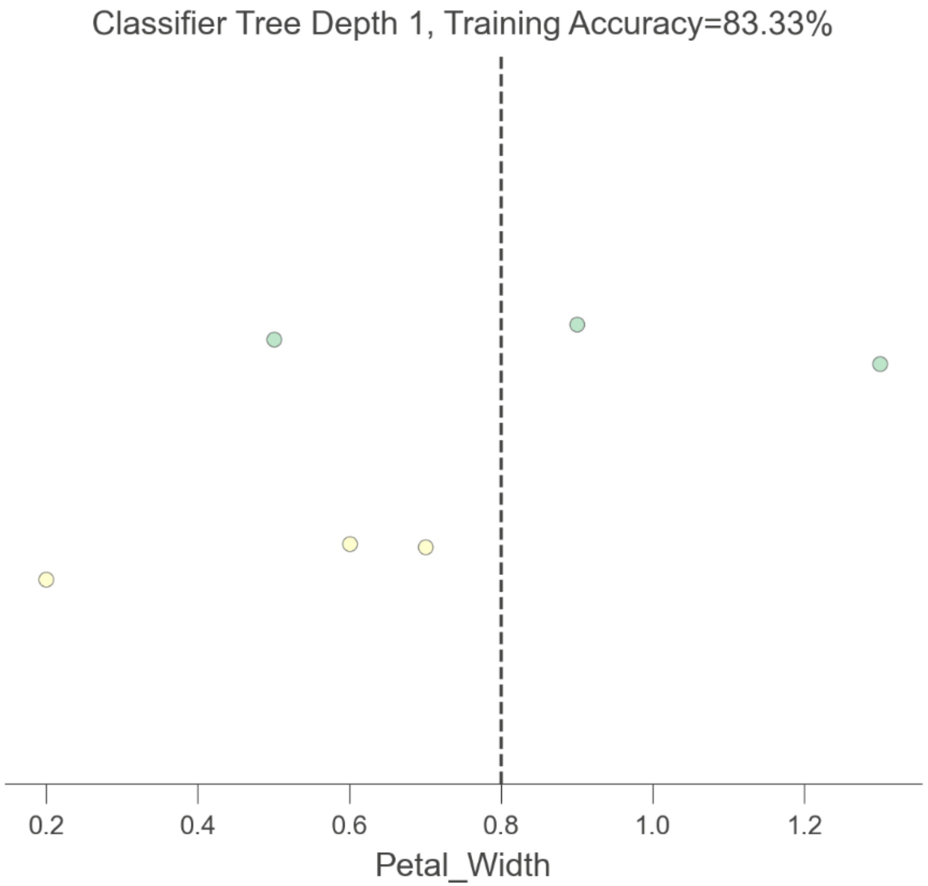
❖ Simple IRIS

Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1
1.8	0.9	1
1.2	1.3	1

0.8

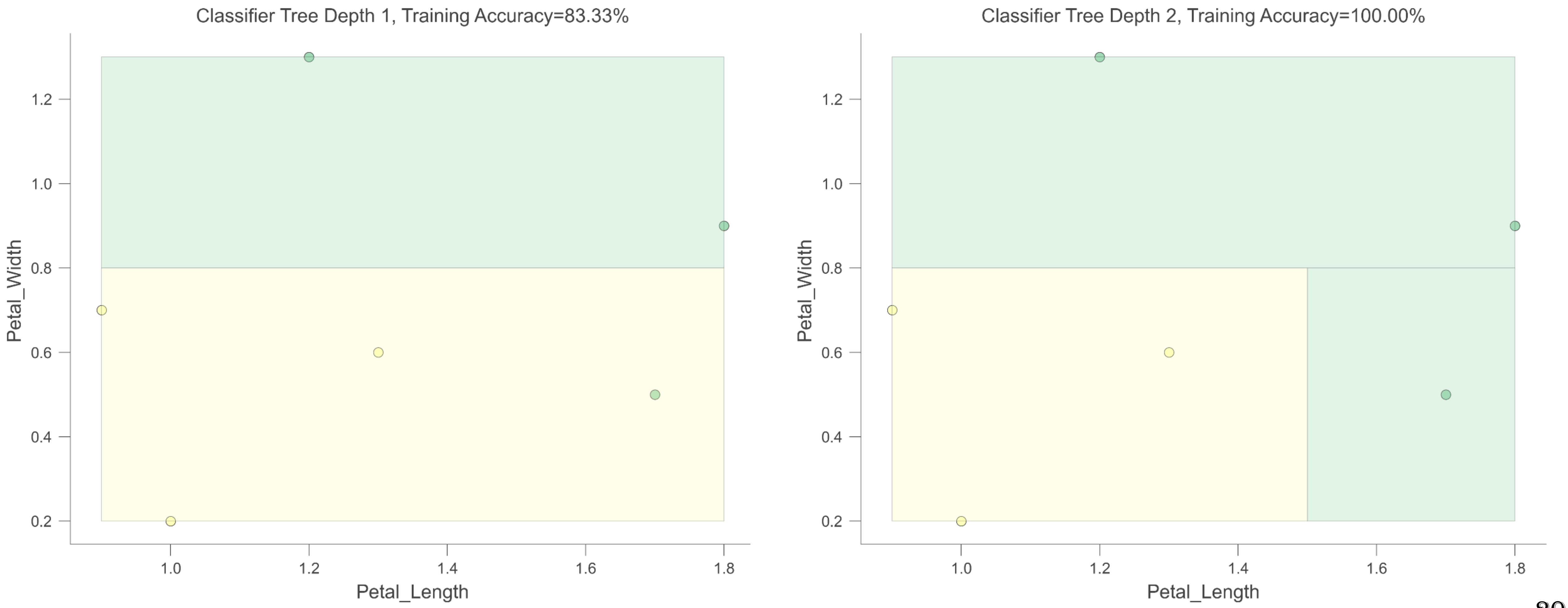
Petal_Length	Petal_Width	Label
1	0.2	0
1.3	0.6	0
0.9	0.7	0
1.7	0.5	1

Petal_Length	Petal_Width	Label
1.8	0.9	1
1.2	1.3	1



Classification

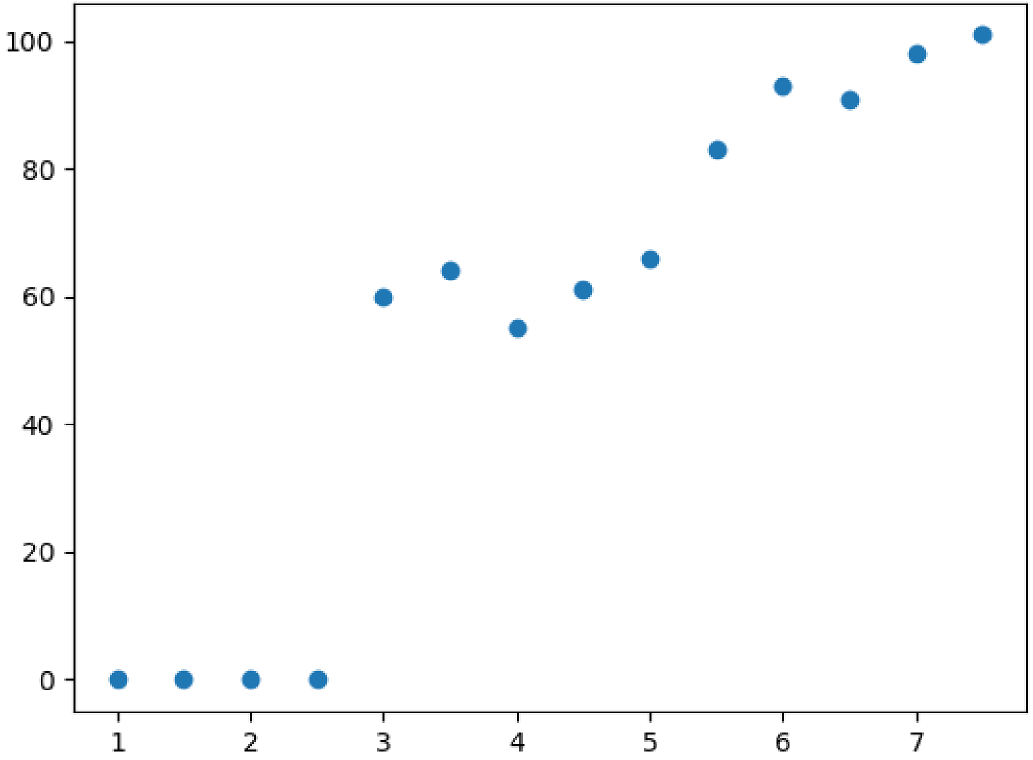
❖ Simple IRIS



Regression

❖ Salary prediction

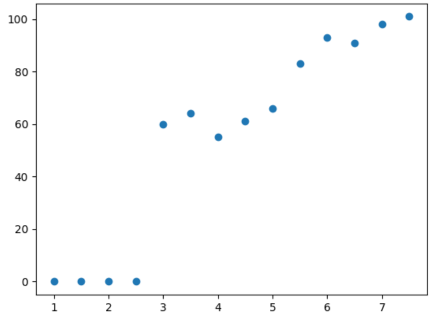
Experience	Salary
1	0
1.5	0
2	0
2.5	0
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101



When Experience = 5.3,

Salary = ?

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

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

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

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

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

$$= \frac{1}{14} * 0 + \frac{13}{14} * 1275.15$$

$$= 1184.07$$

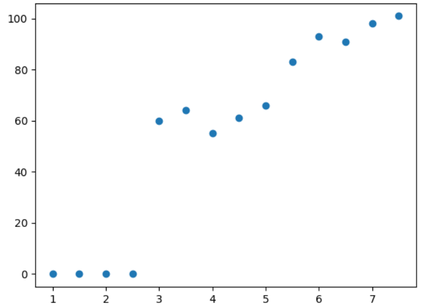
$$\mu_R = \frac{1}{|R|}\sum_i R_i = 59.38$$

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

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

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

Experience	Salary
1	0
1.5	0
2	0
2.5	0
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

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

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

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

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

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

$$= 201.68$$

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

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

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

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

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

$a_{mse} = 1184.07$

$a_{mse} = 911.19$

$a_{mse} = 588.68$

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

$a_{mse} = 383.92$

$a_{mse} = 526.52$

$a_{mse} = 543.51$

$a_{mse} = 575.09$

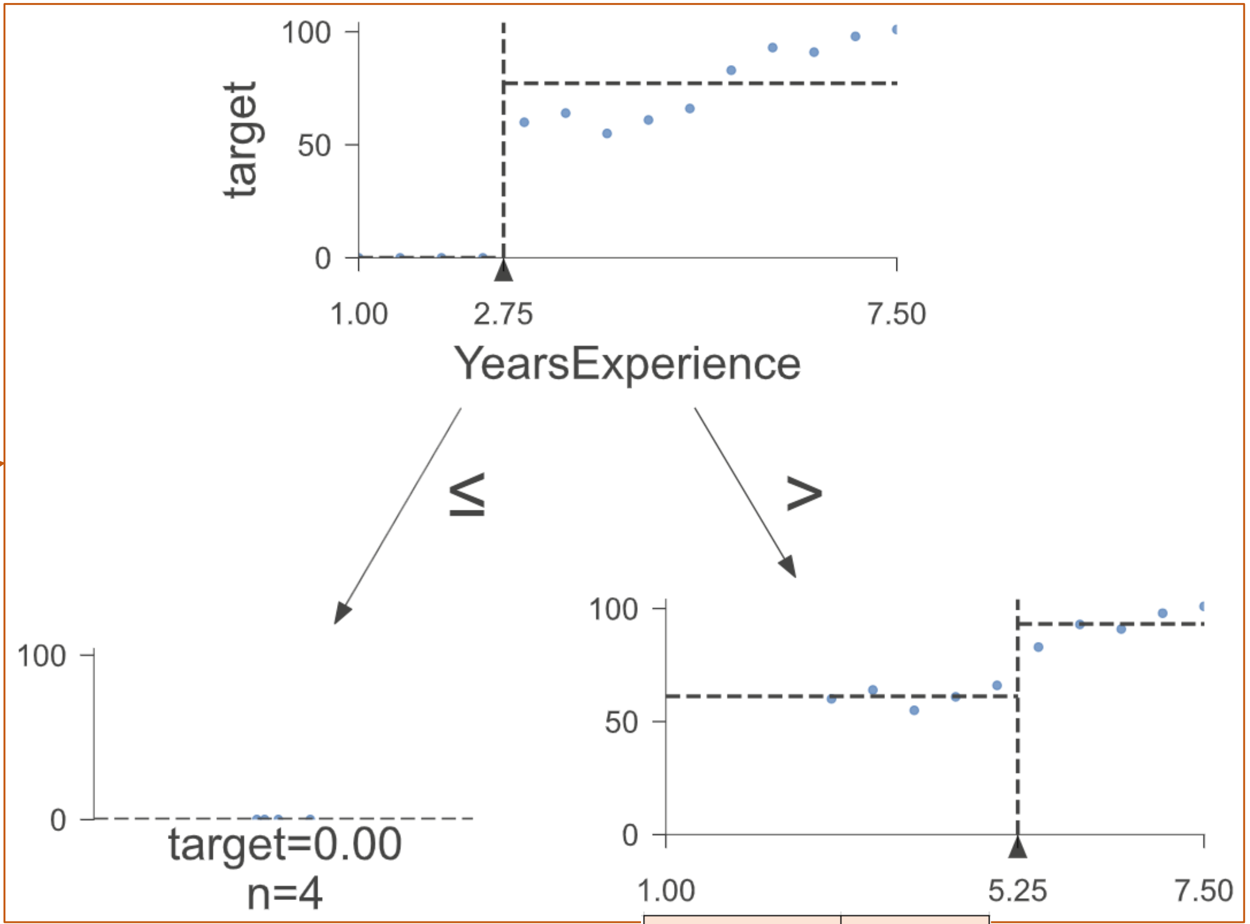
$a_{mse} = 613.34$

$a_{mse} = 758.4$

$a_{mse} = 947.73$

$a_{mse} = 1090.05$

$a_{mse} = 1256.21$



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

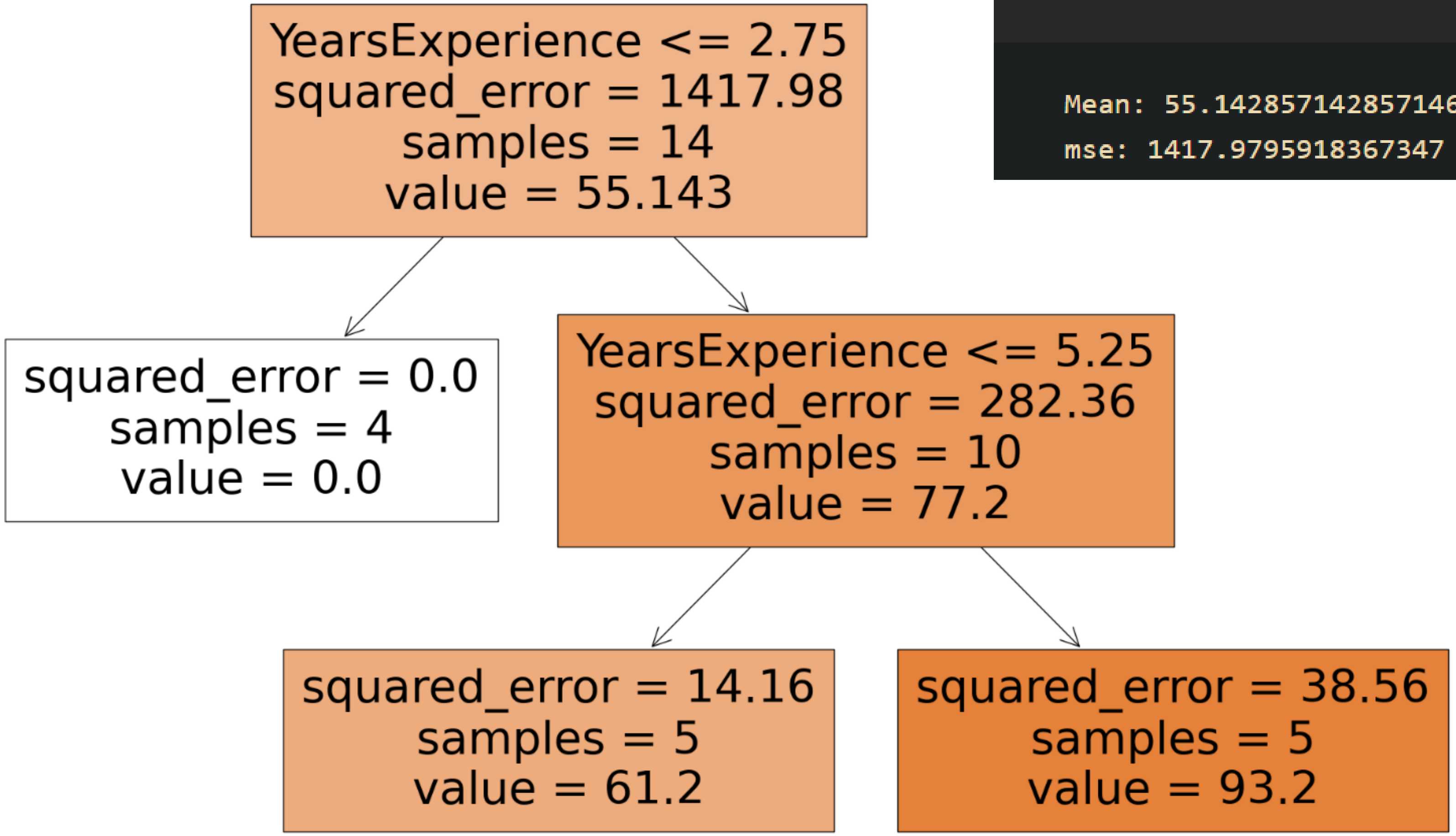
Regression

❖ Salary prediction

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

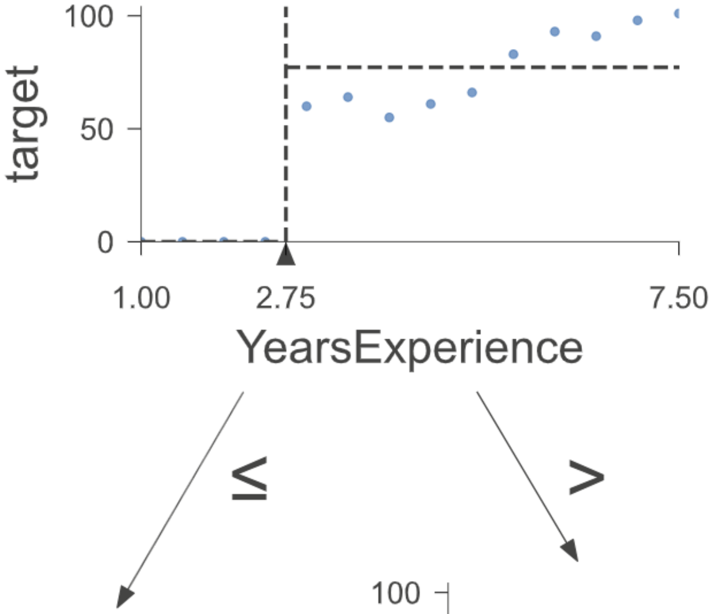
```
y_mean = y.mean()
print('Mean:', y_mean)

diff = (y - y_mean)**2
mse = diff.sum()/14
print('mse:', mse)
```

```
Mean: 55.142857142857146
mse: 1417.9795918367347
```

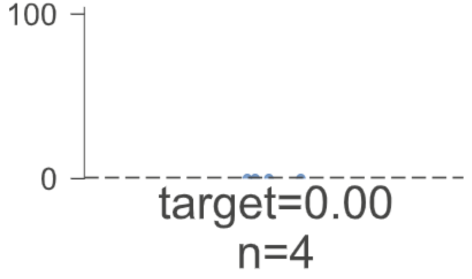


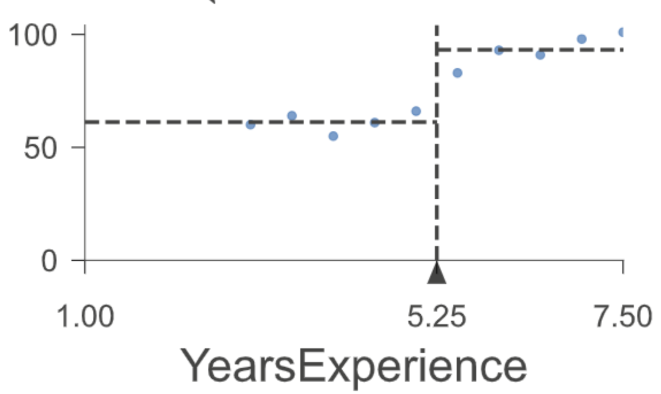
DT - Regression

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



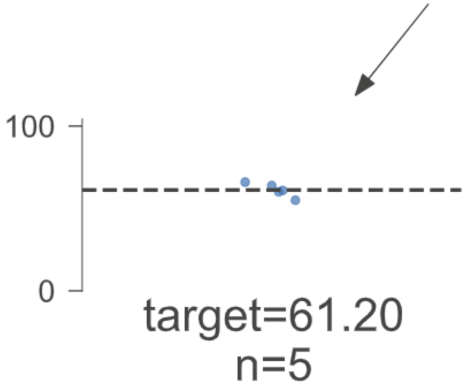
Experience	Salary
1	0
1.5	0
2	0
2.5	0

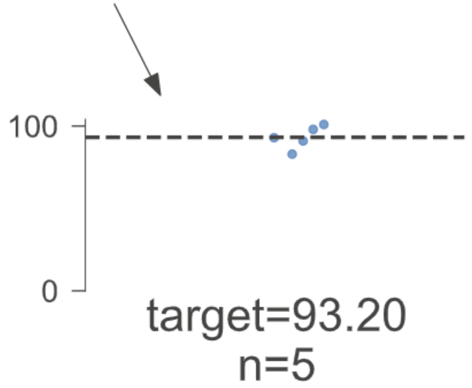




Experience	Salary
3	60
3.5	64
4	55
4.5	61
5	66
5.5	83
6	93
6.5	91
7	98
7.5	101

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

❖ Salary

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