TAUTOLOGY
INNOVATION
SCHOOL

Decision Tree

CT RT

TAUTOLOGY

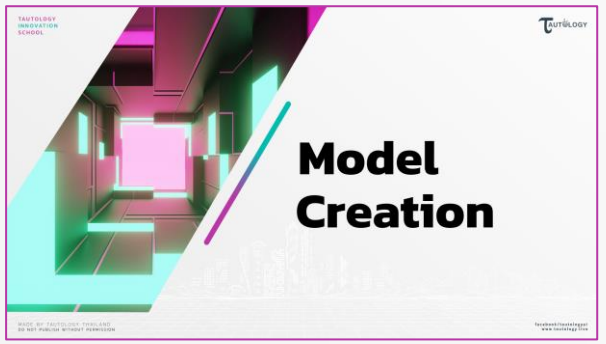
CLASSIFICATION TREE

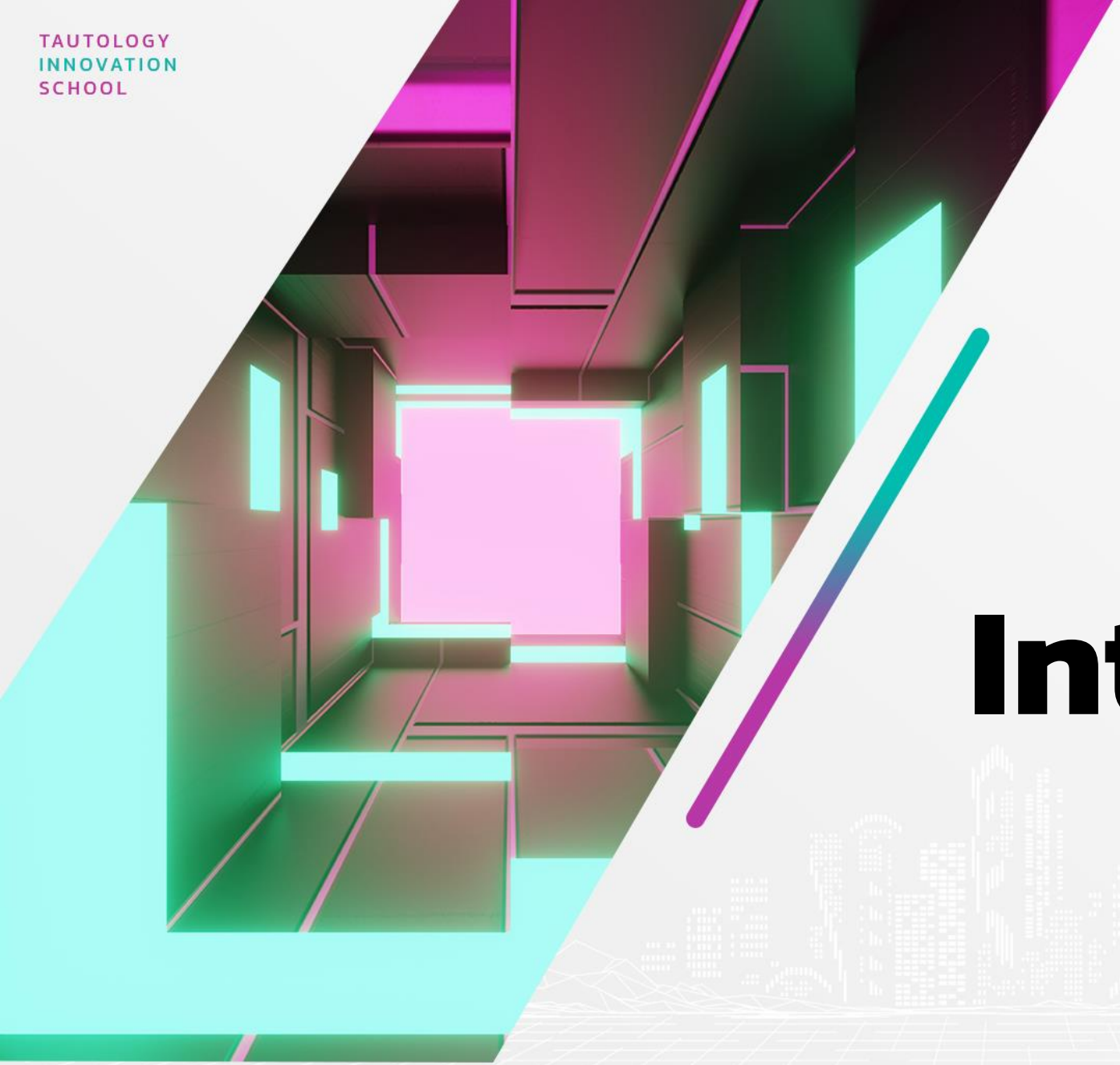
CLASSIFICATION TREE

BY TAUTOLOGY

facebook/tautologyai
www.tautology.live

Classification Tree





Introduction

Introduction

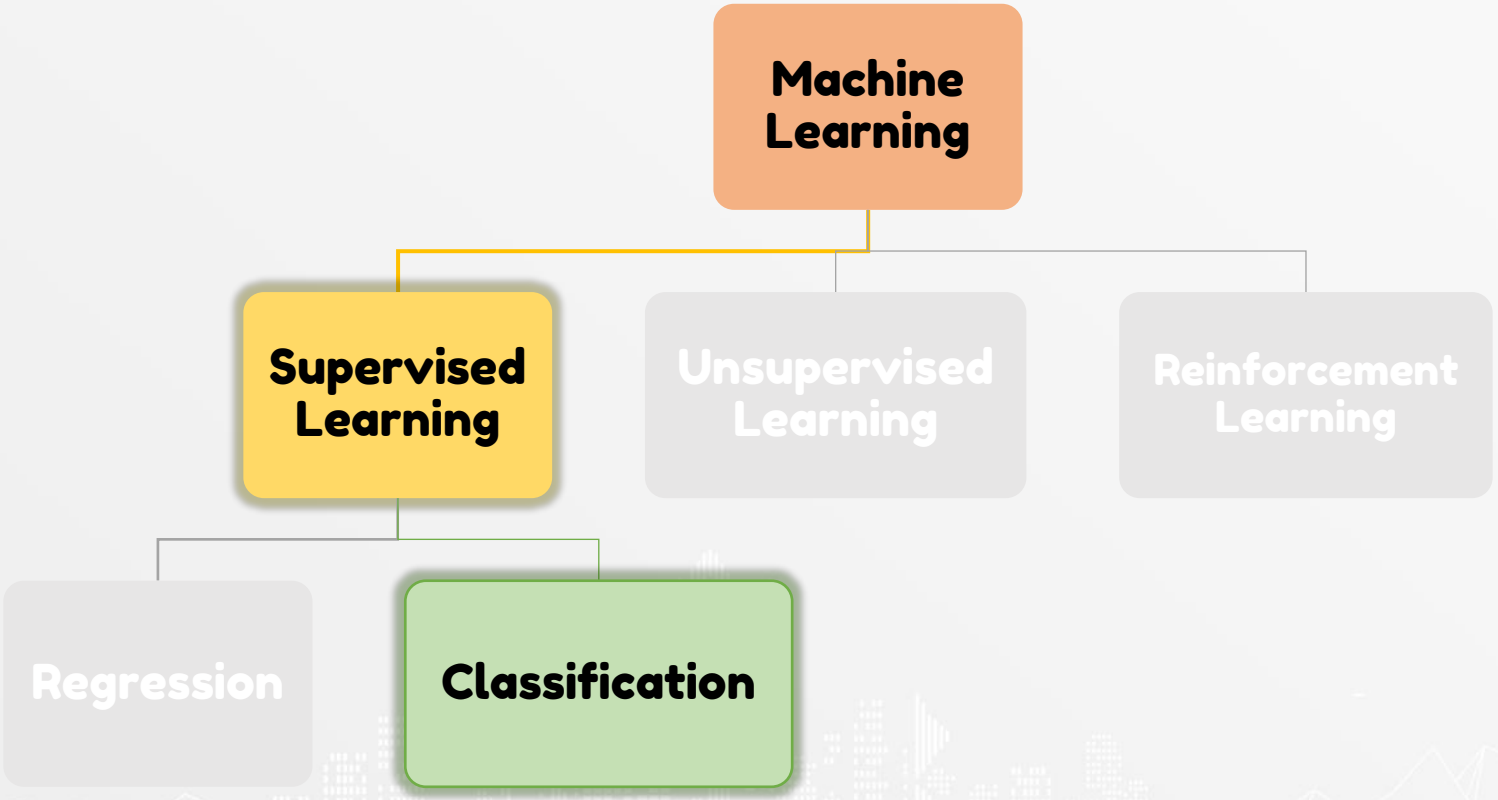
- What is Classification Tree?
- Data for Classification Tree
- Pros & Cons
- Real World Application

What is Classification Tree?

เป็น algorithm ตัวเดียว (decision tree) ที่ให้ผลลัพธ์ของ model
อยู่ในรูปของเหตุผล หรือ เราเรียกมันว่ากฎนั่นเอง

Classification Tree เป็นหนึ่งใน algorithm ประเภท supervised learning ที่ใช้สำหรับแก้ปัญหา classification โดยมีหลักการทำงาน คือ การสร้างชุดของกฎเพื่อจำแนกประเภทของข้อมูล

What is Classification Tree?



What is Classification Tree?

ผลการตรวจผู้ป่วยเบาหวาน

sex	BMI	target
0	26	diabetes
1	26	normal
1	28	diabetes
1	30	diabetes
0	28	normal
0	30	normal

sex <= 0.5

BMI <= 27

predict = ‘diabetes’

BMI > 27

predict = ‘normal’

sex > 0.5

BMI <= 27

predict = ‘normal’

BMI > 27

predict = ‘diabetes’

What is Classification Tree?

sex <= 0.5

BMI <= 27

predict = 'diabetes' => If sex <= 0.5 and BMI <= 27, then 'diabetes'

BMI > 27

predict = 'normal' => If sex <= 0.5 and BMI > 27, then 'normal'

sex > 0.5

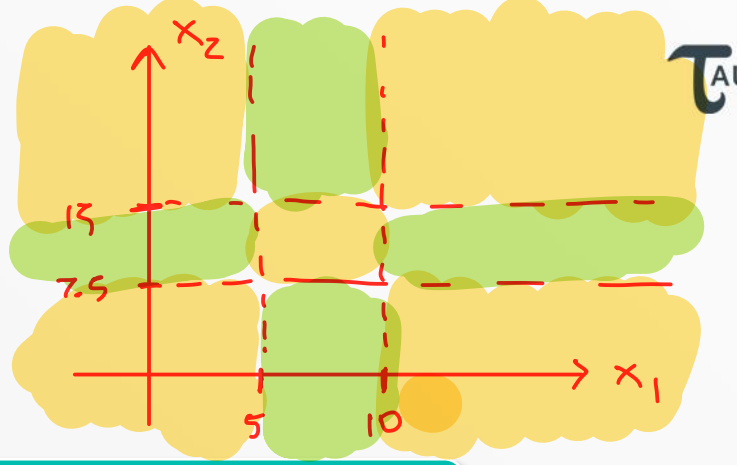
BMI <= 27

predict = 'normal' => If sex > 0.5 and BMI <= 27, then 'normal'

BMI > 27

predict = 'diabetes' => If sex > 0.5 and BMI > 27, then 'diabetes'

Introduction



What is Classification Tree?

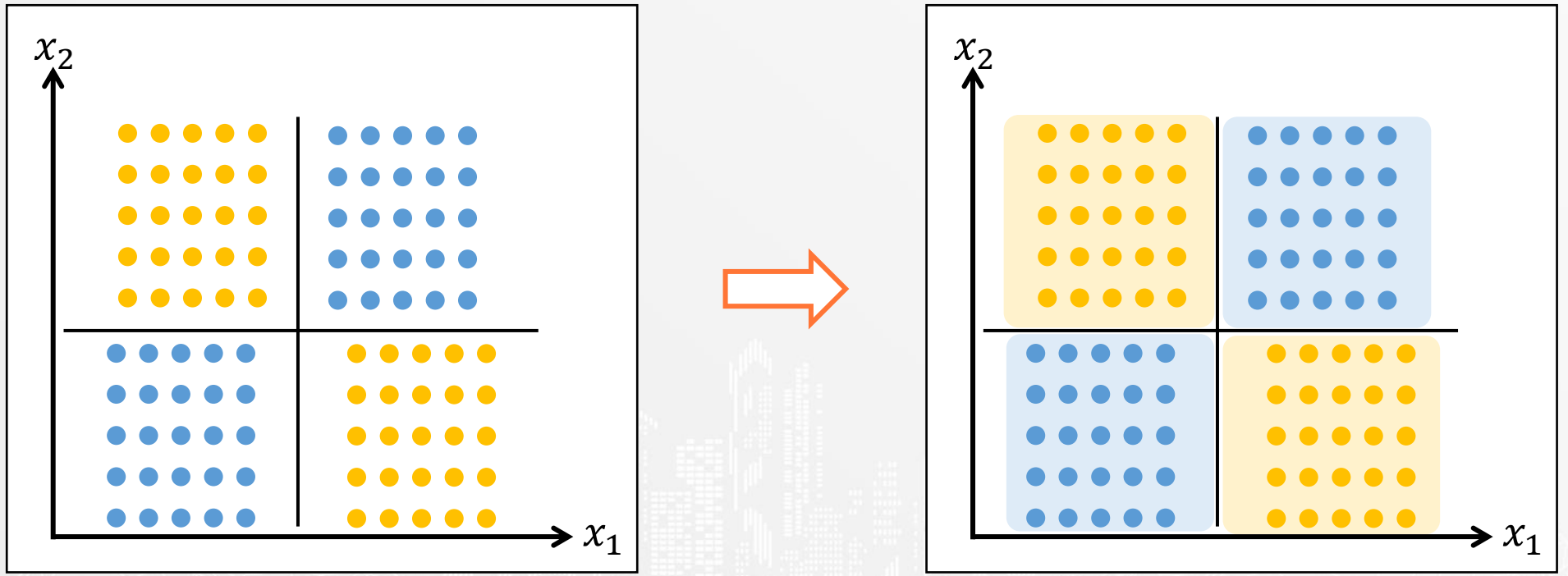
Data for Classification Tree

Pros & Cons

Real World Application

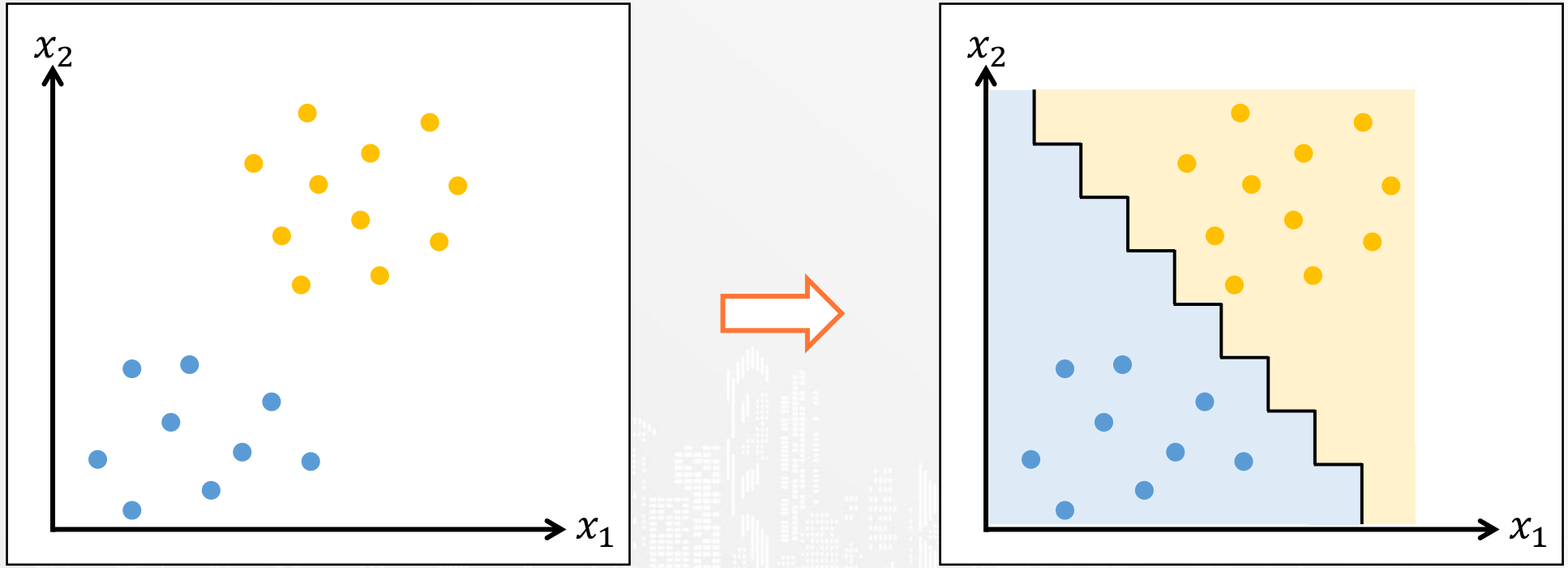
Data for Classification Tree

ตัวอย่างของข้อมูลที่เหมาะกับ Classification Tree



Data for Classification Tree

ตัวอย่างของข้อมูลที่ไม่เหมาะกับ Classification Tree



Introduction

What is Classification Tree?

Data for Classification Tree

Pros & Cons

Real World Application

Pros & Cons

ข้อดี

- หลักการของ algorithm เรียบง่าย
- สามารถตีความผลลัพธ์ได้ของ model ได้ง่าย (model อยู่ในรูปของกฎ)

ข้อเสีย

- ง่ายต่อการเกิด overfitting
- การเปลี่ยนแปลงข้อมูลเพียงเล็กน้อยใน training อาจส่งผลให้ model เปลี่ยนแปลงอย่างมาก

ข้อจำกัด

- decision boundary ที่ได้จะขนานกับแกนเสมอ

Introduction

- What is Classification Tree?
- Data for Classification Tree
- Pros & Cons
- Real World Application

Real World Application

การจำแนกผู้ป่วยโรคหัวใจ
โดยพิจารณาจาก อายุ เพศ ความดันโลหิต คอลเลสเตอรอล ประเภทการเจ็บหน้าอก เป็นต้น

อ้างอิง : [2023, Ozcan & Peker] A classification and regression tree algorithm for heart disease modeling and prediction

Real World Application

การระบุ spam e-mail
โดยพิจารณาจาก e-mail ที่ถูกส่งเข้าของ University Utara Malaysia's Computer Center ใน 1 สัปดาห์

อ้างอิง : [2022, Abdulrahman & Salim] Using Decision Tree Algorithms in Detecting Spam Emails Written in Malay: A Comparison Study

Real World Application

ทุก model สามารถตีความได้หมด
แต่ ค. ยาก/ง่าย ในการตีความจะแตกต่างกัน

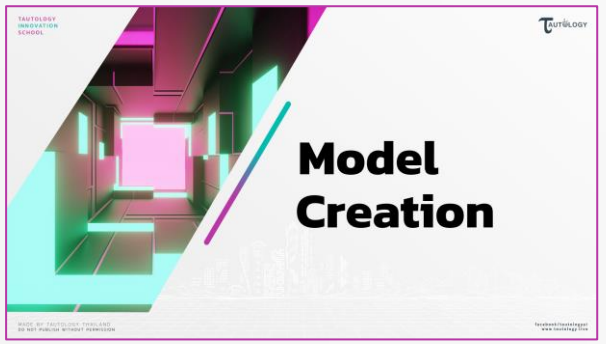
การจำแนกประเภทลูกค้าที่สามารถกู้เงินได้
โดยพิจารณาจาก เพศ อายุ ประเภทของบริษัทที่ทำงาน อาชีพ ระดับการศึกษา การแต่งงาน รายได้ จำนวนปีที่จะกู้

อ้างอิง : [2004, Xiu Li et al.] Applications of Classification Trees to Consumer Credit Scoring Methods in Commercial Banks

Introduction

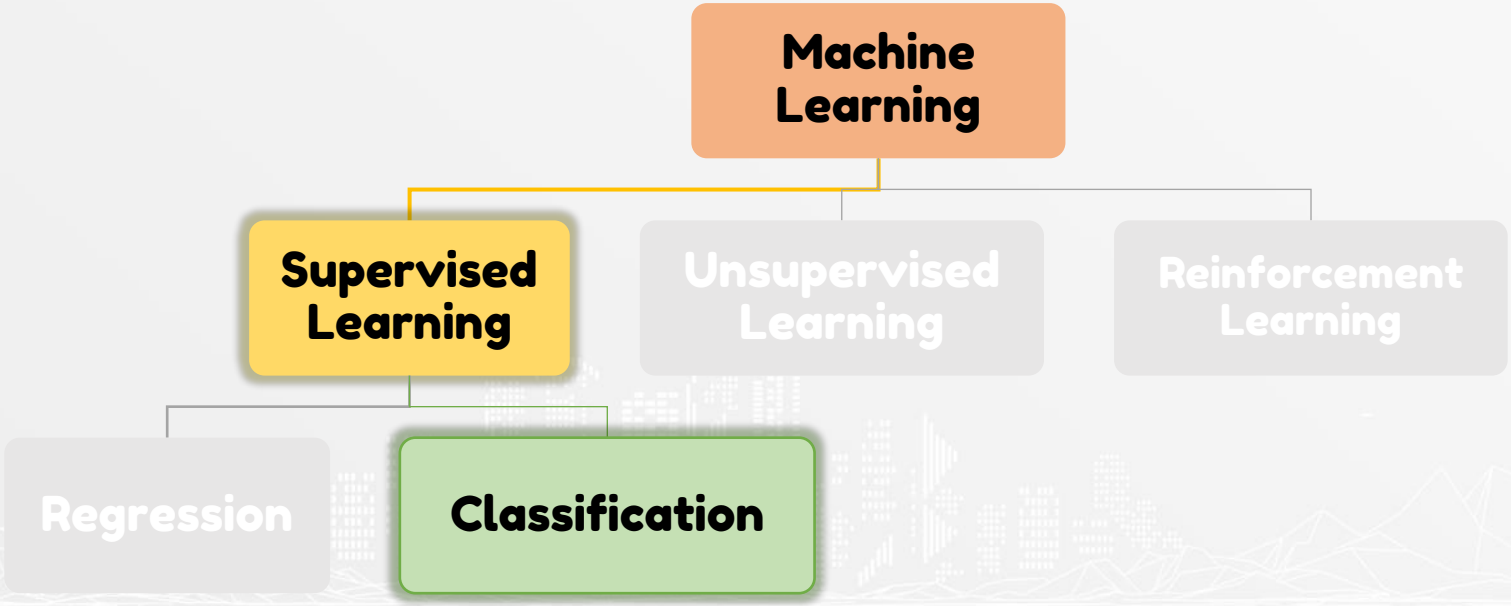
- What is Classification Tree?
- Data for Classification Tree
- Pros & Cons
- Real World Application

Classification Tree



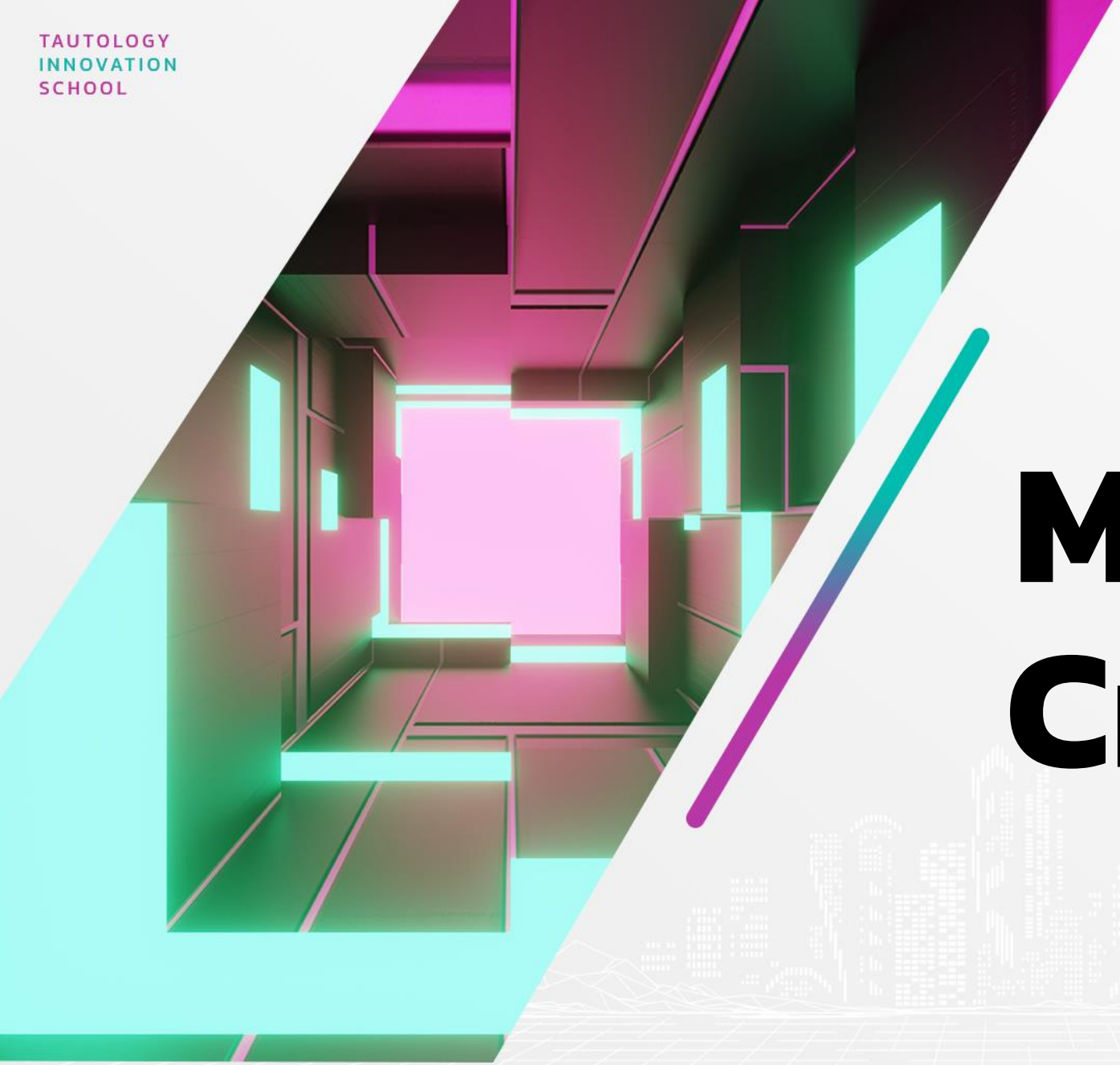
Classification Tree

Classification Tree เป็นหนึ่งใน algorithm ประเภท supervised learning



Concept of Supervised Learning

Data ⇨ Model ⇨ Prediction



Model Creation

Model Creation

- Assumption
- Real Face of the Model
- How to Create Model (Math)
- How to Create Model (Code)
- Further Reading

Assumption

- No Missing Features

Model Creation

Assumption

Real Face of the Model

How to Create Model (Math)

How to Create Model (Code)

Further Reading

Real Face of the Model

Classification Tree คือ ชุดของกฎเพื่อจำแนกประเภทของข้อมูล

sex <= 0.5

BMI <= 27

predict = 'diabetes' => If sex <= 0.5 and BMI <= 27, then 'diabetes'

BMI > 27

predict = 'normal' => If sex <= 0.5 and BMI > 27, then 'normal'

sex > 0.5

BMI <= 27

predict = 'normal' => If sex > 0.5 and BMI <= 27, then 'normal'

BMI > 27

predict = 'diabetes' => If sex > 0.5 and BMI > 27, then 'diabetes'

Model Creation

Assumption

Real Face of the Model

How to Create Model (Math)

How to Create Model (Code)

Further Reading

How to Create Model (Math)

ยิ่งลึก → ยิ่งซับซ้อน → overfit

- ☐ Step 1 : พิจารณา unique values ของ feature ทุกตัวใน dataset
- ☐ Step 2 : ตั้งคำถามจาก unique values
- ☐ Step 3 : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

เราใช้ metric ที่ใช้วัดคุณภาพของคำถาม

information gain = ค. ยุ่งเหยิงของระบบก่อนถามคำถาม − ค. ยุ่งเหยิงของระบบหลังถามคำถาม

ค. ยุ่งเหยิงของระบบสามารถวัดได้ด้วย ⇒ Gini Impurity

1. กำหนดความลึก (ในตัวอย่างเราลึก 2 ชั้น)
2. จำนวนตัวอย่างน้อยที่สุดที่ไม่ต้องการให้แบ่ง

How to Create Model (Math)

ตัวอย่างการคำนวณ Classification Tree

sex	BMI	target
0	26	diabetes
1	26	normal
1	28	diabetes
1	30	diabetes
0	28	normal
0	30	normal

ตารางแสดงข้อมูลผู้ป่วยที่เป็นโรคเบาหวาน

How to Create Model (Math)

☑ **Step 1** : พิจารณา unique values ของ feature ทุกตัวใน dataset

sex	BMI	target
0	26	diabetes
1	26	normal
1	28	diabetes
1	30	diabetes
0	28	normal
0	30	normal

✓ ➔ unique_values(sex) = {0, 1}

➔ unqiue_values(BMI) = {26, 28, 30}

✗

✗

How to Create Model (Math)

☑ **Step 2** : ตั้งคำถามจาก unique values

unique_values(sex) = {0, 1}

unqiue_values(BMI) = {26, 28, 30}

sex = 0 ?

sex = 1 ?

Question1 : sex <= 0.5 ?

Question2 : BMI <= 27 ?

Question3 : BMI <= 29 ?

3 ข้อ

ใช้กับ feature ที่เป็น numerical ได้

BMI = 26 ?

BMI = 28 ?

BMI = 30 ?

How to Create Model (Math)

information gain = ค่าความยุ่งเหยิงของข้อมูลก่อนแบ่ง − ค่าความยุ่งเหยิงของข้อมูลหลังแบ่ง

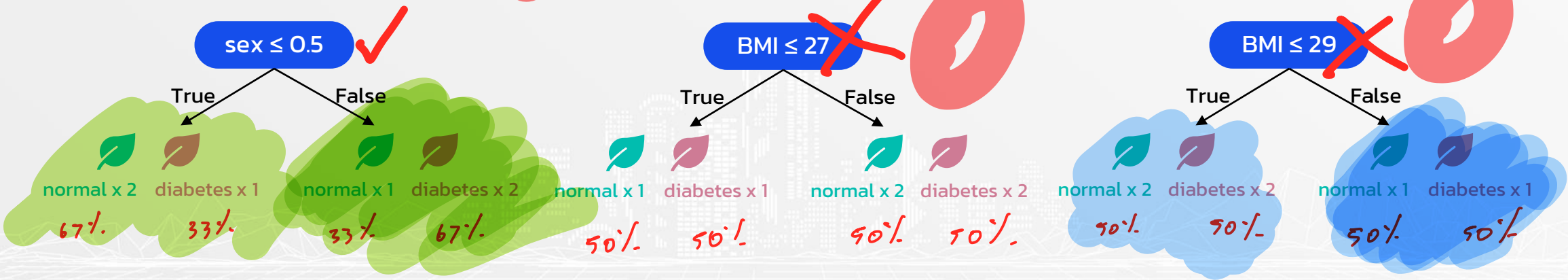
☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

Root Node

$$\frac{1}{2} - \frac{4}{9} = \frac{9-8}{18} = \frac{1}{8}$$

normal x 3 (50%) diabetes x 3 (50%)

$$Gini = 1 - \sum_{k \in K} p_k^2$$



Gini ก่อนอ่ากาม $= 1 - P_{normal}^2 - P_{diabetes}^2$

$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 1 - \frac{1}{4} - \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$

Information Gain = Gini ก่อน − Gini หลัง

Gini หลัง $= \frac{4}{9}$

$\frac{3}{6}\left(1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right) +$

$\frac{3}{6}\left(1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right)$

Gain $= \frac{1}{18}$

Gain $= 0$

Gini หลังอ่ากาม $= \frac{1}{2}$

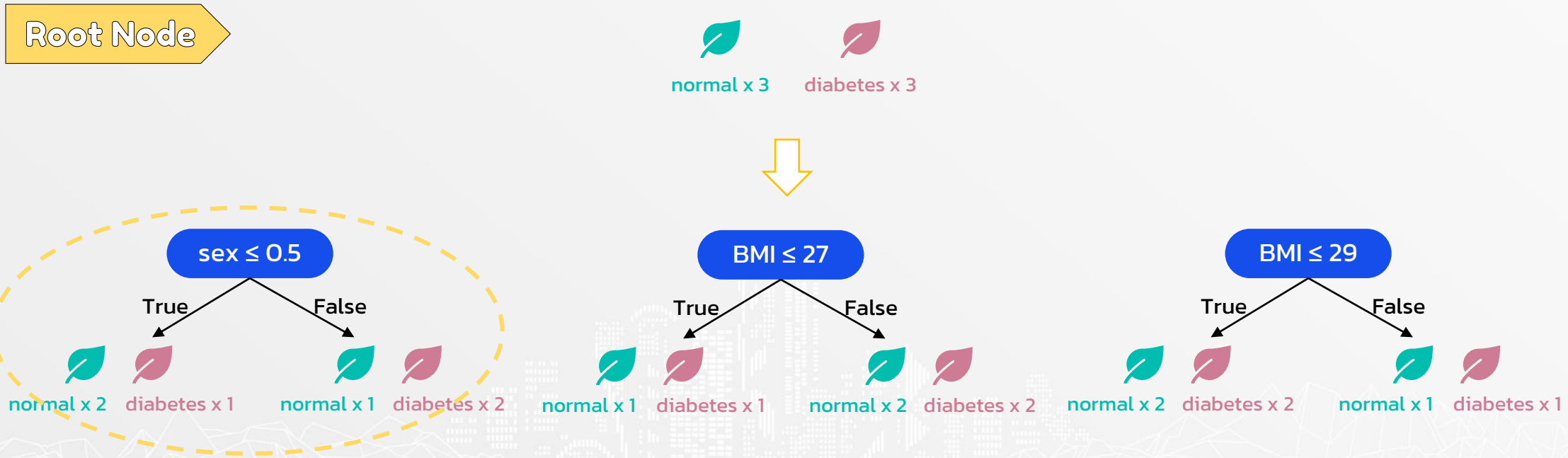
$\frac{4}{6}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) +$

$\frac{2}{6}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)$

Gain $= 0$

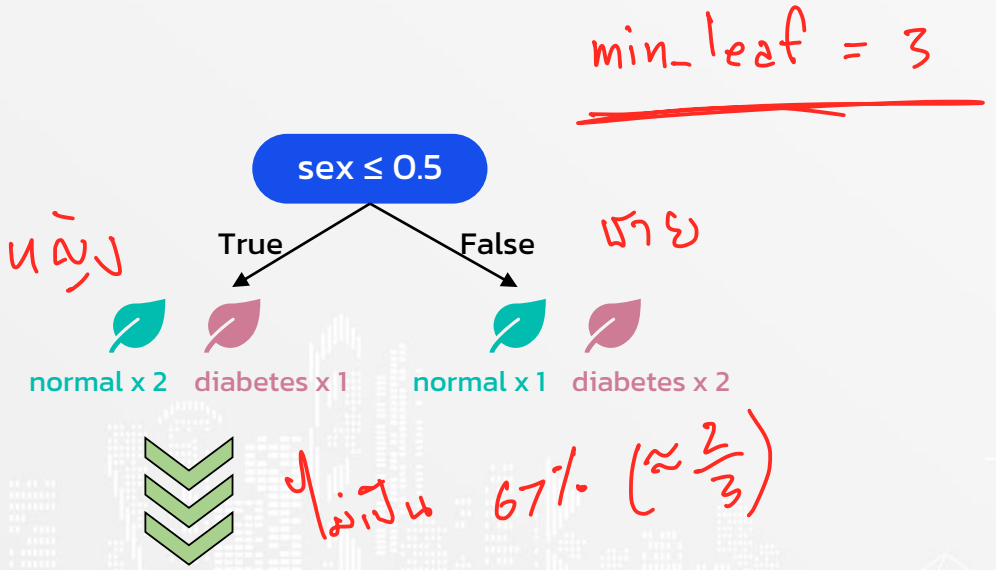
How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด



How to Create Model (Math)

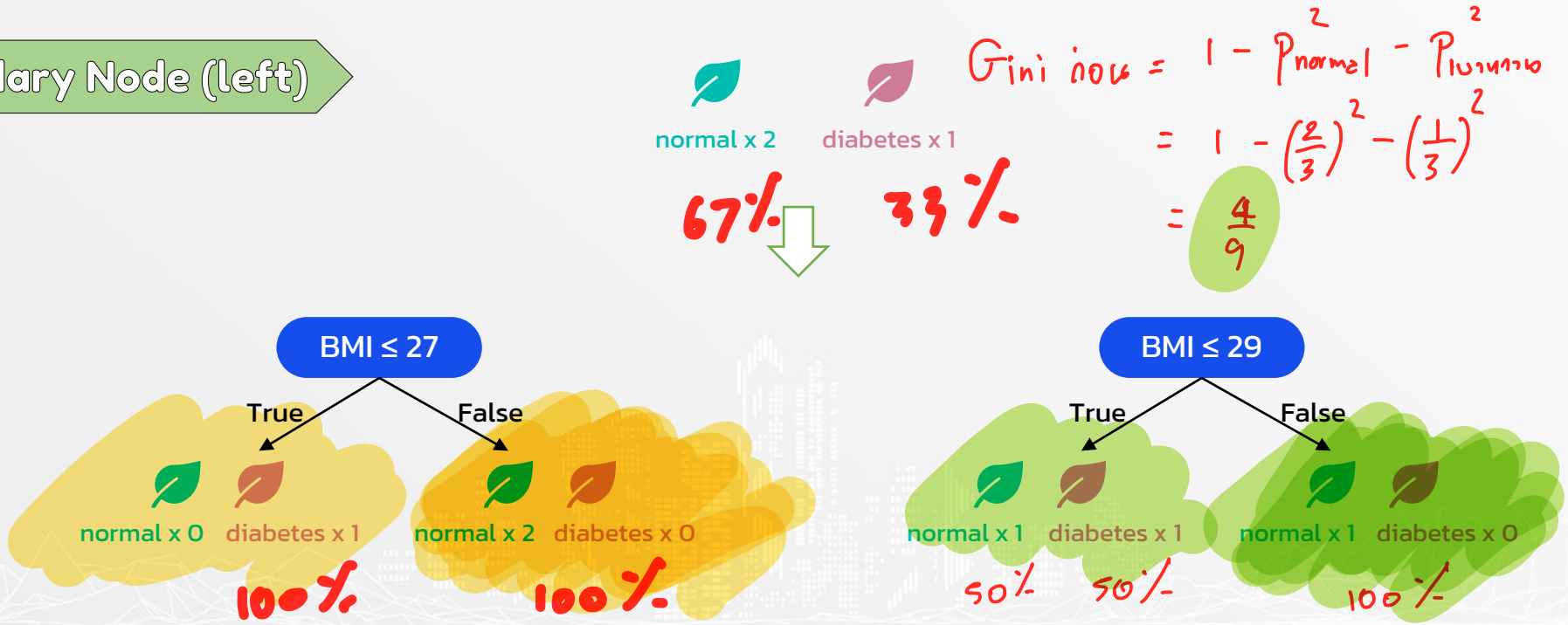
☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

Secondary Node (left)



Information Gain = Gini ก่อน − Gini หลัง

Gini หลัง $= \frac{1}{3}\left(1 - 0^2 - 1^2\right) + \frac{2}{3}\left(1 - 1^2 - 0^2\right)$

$= 0$

Gain $= \frac{4}{9} - 0$

$= \frac{4}{9}$

Gini หลัง $= \frac{2}{3}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) + \frac{1}{3}\left(1 - (1)^2 - 0^2\right)$

$= \frac{2}{3}\left(\frac{1}{2}\right) + \frac{1}{3}(0)$

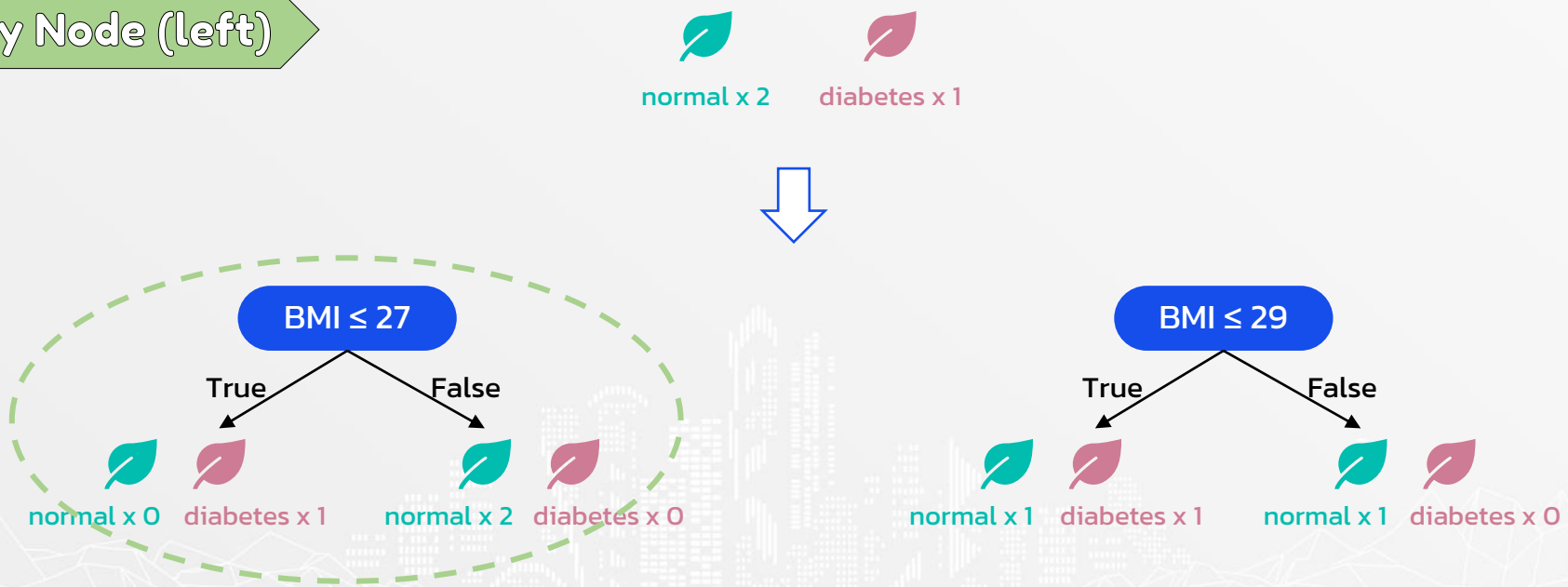
$= \frac{1}{3} + 0 = \frac{1}{3}$

Gain $= \frac{4}{9} - \frac{1}{3} = \frac{1}{9}$

How to Create Model (Math)

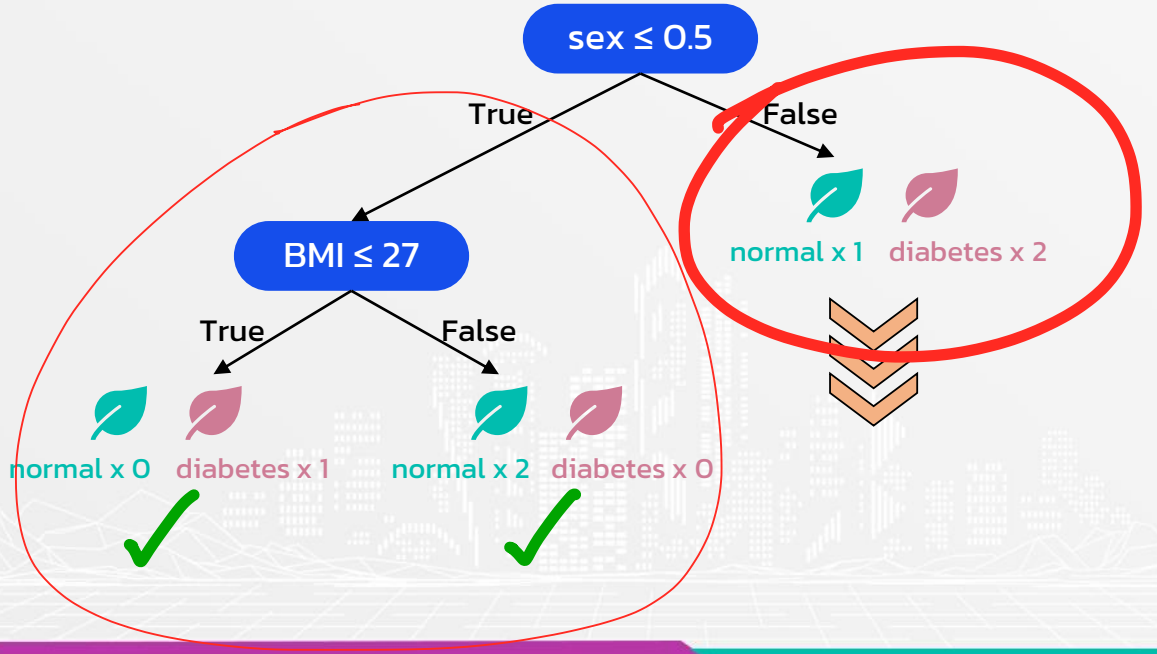
☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

Secondary Node (left)



How to Create Model (Math)

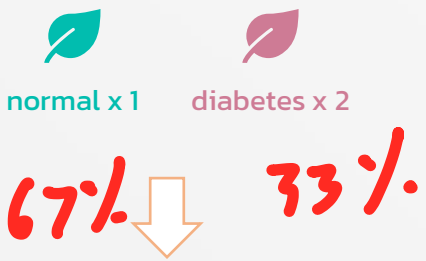
☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

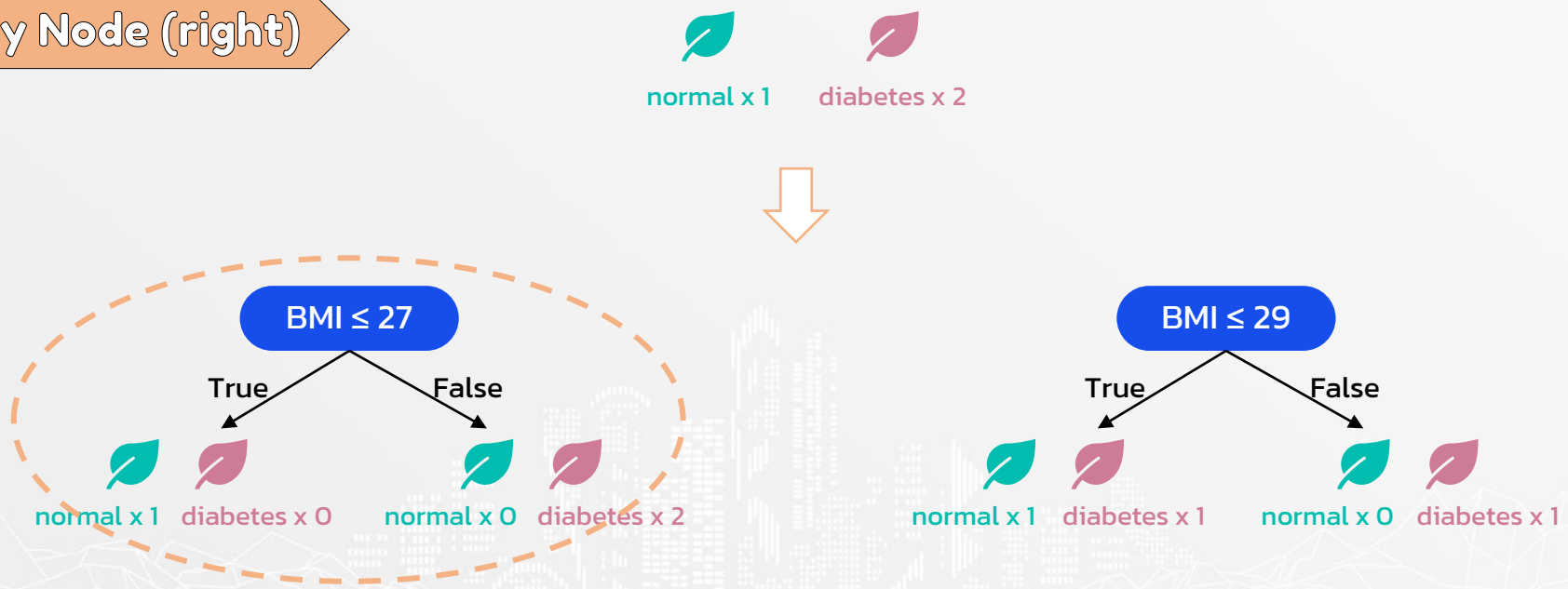
Secondary Node (right)



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

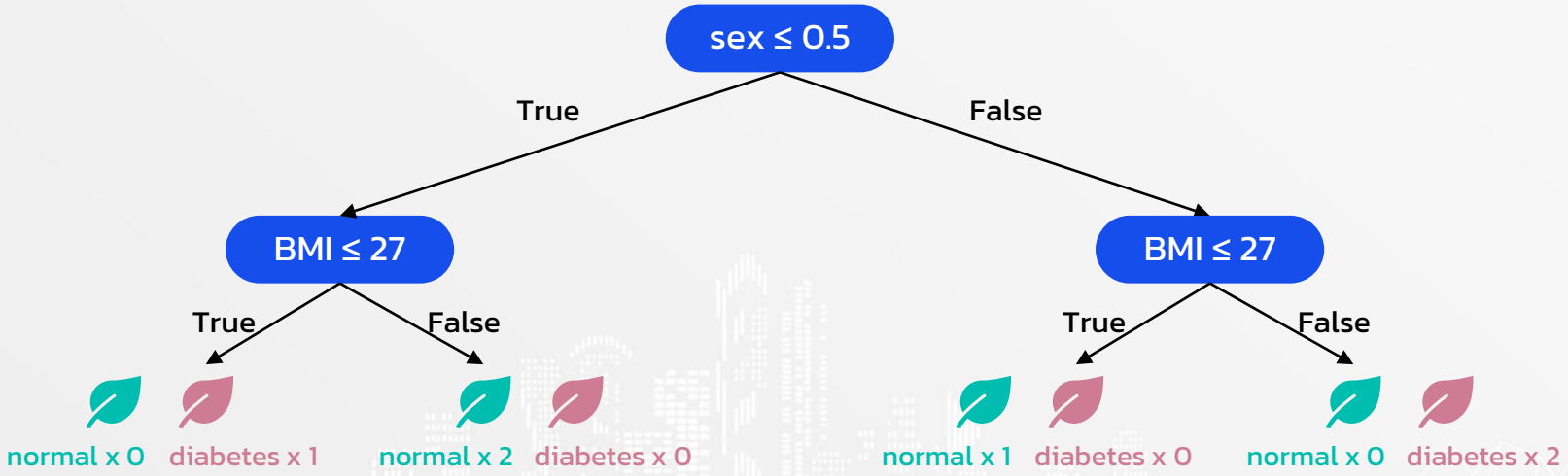
Secondary Node (right)



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

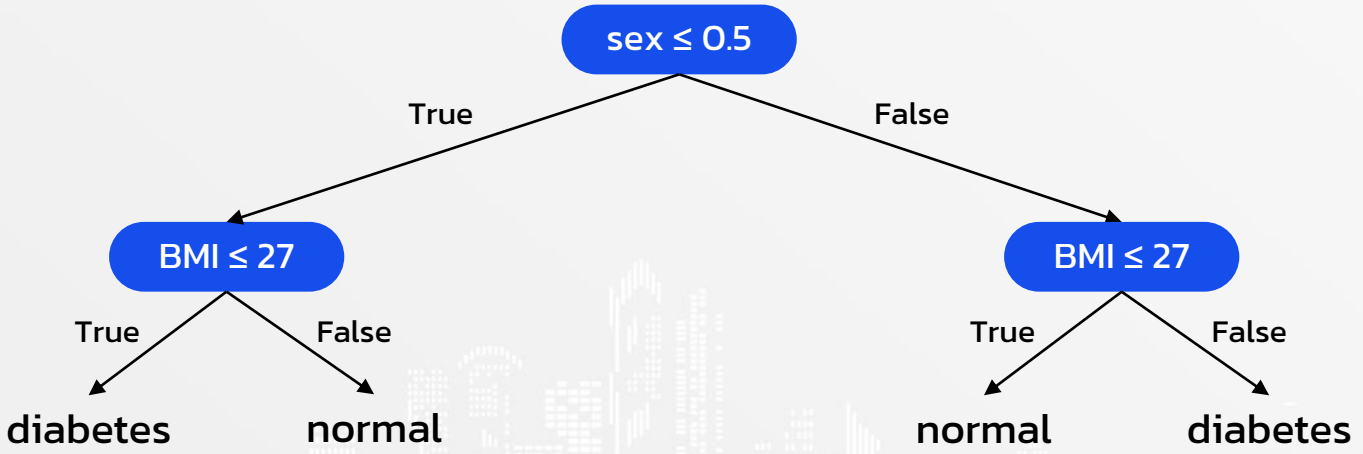
Full Tree



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

Full Tree



How to Create Model (Math)

☑ **Step 3** : ในแต่ละชั้น, ตั้งคำถามที่ทำให้ classification tree มีความสามารถในการพยากรณ์มากยิ่งขึ้น ภายใต้เงื่อนไขที่กำหนด

If sex <= 50 and BMI <= 27, then ‘diabetes’

If sex <= 0.5 and BMI > 27, then ‘normal’

If sex > 0.5 and BMI <= 27, then ‘normal’

If sex > 0.5 and BMI > 27, then ‘diabetes’

Model Creation

- Assumption
- Real Face of the Model
- How to Create Model (Math)
- How to Create Model (Code)
- Further Reading

How to Create Model (Code)

ตัวอย่าง Code สำหรับ Classification Tree

sex	BMI	target
0	26	diabetes
1	26	normal
1	28	diabetes
1	30	diabetes
0	28	normal
0	30	normal

ตารางแสดงข้อมูลผู้ป่วยที่เป็นโรคเบาหวาน

How to Create Model (Code)

- Code สำหรับสร้าง model จากข้อมูลของเราโดยที่

$$X = \begin{bmatrix} 0 & 26 \\ 1 & 26 \\ 1 & 28 \\ 1 & 30 \\ 0 & 28 \\ 0 & 30 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} diabetes \\ normal \\ diabetes \\ diabetes \\ normal \\ normal \end{bmatrix}$$

```
1  clf = DecisionTreeClassifier()
2  clf.fit(X, y)
```

```
DecisionTreeClassifier()
```

How to Create Model (Code)

```
1  r = export_text(clf, feature_names=list(X.columns))
```

```
1  print(r)
```

```
|--- sex <= 0.50
|    |--- BMI <= 27.00
|    |    |--- class: diabetes
|    |--- BMI >  27.00
|    |    |--- class: normal
|--- sex >  0.50
|    |--- BMI <= 27.00
|    |    |--- class: normal
|    |--- BMI >  27.00
|    |    |--- class: diabetes
```

How to Create Model (Code)

Code for this section

Open File

Model Creation.ipynb

Model Creation

- Assumption
- Real Face of the Model
- How to Create Model (Math)
- How to Create Model (Code)
- Further Reading

Further Reading

CART

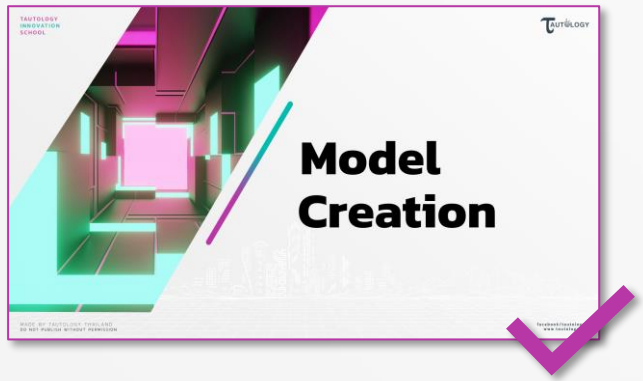
- Gini Impurity
- Entropy
- Information Gain
- ID3 Algorithm (algorithm Turing)

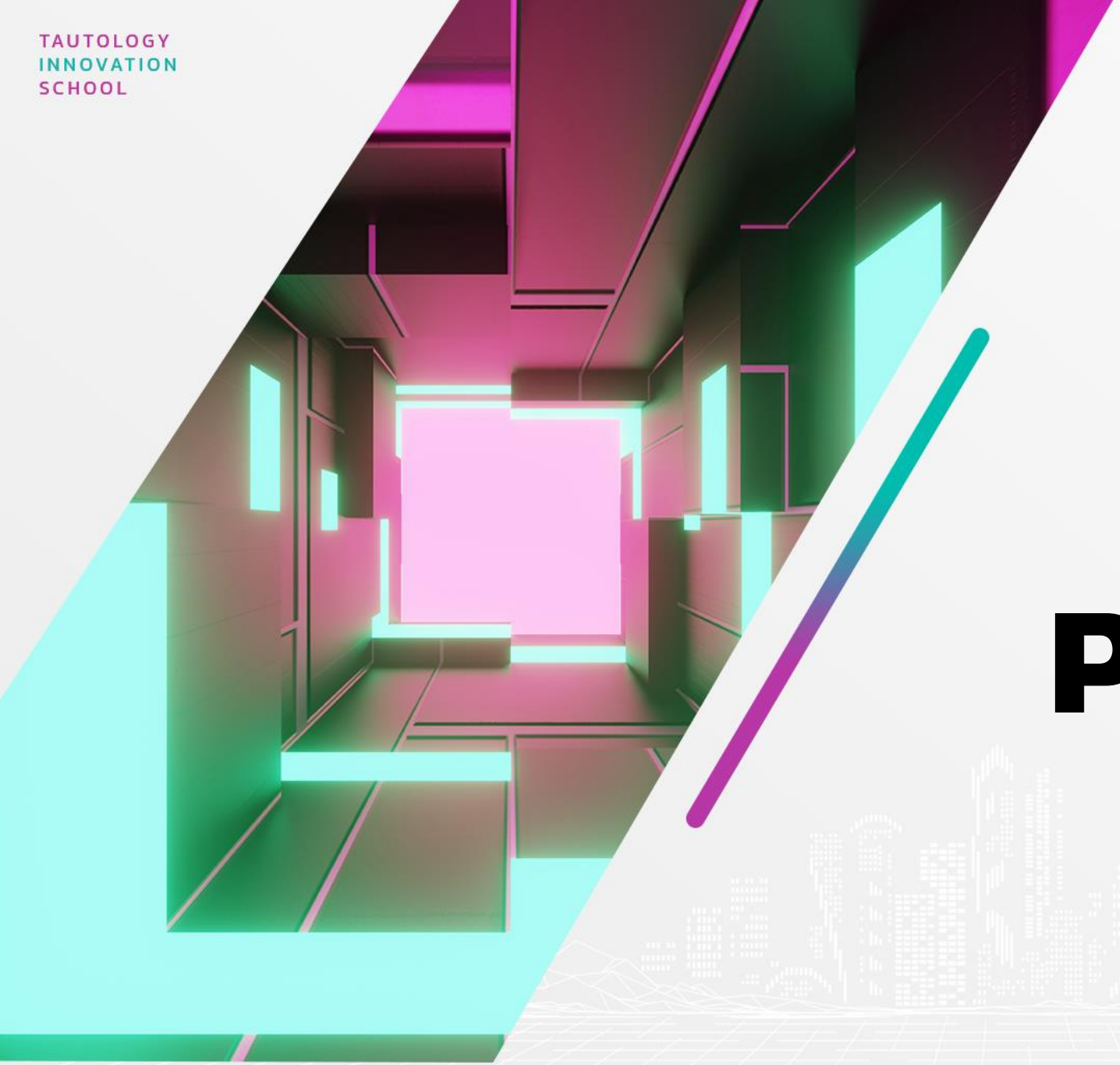
- KL Divergence
- Entropy
- Cross Entropy

Model Creation

- Assumption
- Real Face of the Model
- How to Create Model (Math)
- How to Create Model (Code)
- Further Reading

Classification Tree





Prediction

Prediction

Classification Tree คือ ชุดของกฎเพื่อจำแนกประเภทของข้อมูล

```
sex <= 0.5
        BMI <= 27
                predict = 'diabetes'
        BMI > 27
                predict = 'normal'
sex > 0.5
        BMI <= 27
                predict = 'normal'
        BMI > 27
                predict = 'diabetes'
```

=> If sex <= 0.5 and BMI <= 27, then 'diabetes'

=> If sex <= 0.5 and BMI > 27, then 'normal'

=> If sex > 0.5 and BMI <= 27, then 'normal'

=> If sex > 0.5 and BMI > 27, then 'diabetes'

Prediction

If sex <= 0.5 and BMI <= 27, then ‘diabetes’

If sex <= 0.5 and BMI > 27, then ‘normal’

If sex > 0.5 and BMI <= 27, then ‘normal’

If sex > 0.5 and BMI > 27, then ‘diabetes’

Prediction

1-Sample

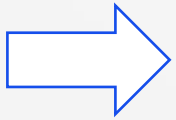
Multi-Sample

Code

1-Sample

ตัวอย่างการคำนวณ $\hat{y}$

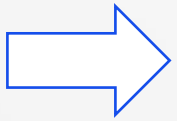
sex	BMI
0	29



$\hat{y}$
?

1-Sample

sex	BMI
0	29



$\hat{y}$
normal

If sex <= 0.5 and BMI <= 27, then ‘diabetes’

If sex <= 0.5 and BMI > 27, then ‘normal’

If sex > 0.5 and BMI <= 27, then ‘normal’

If sex > 0.5 and BMI > 27, then ‘diabetes’

Prediction

1-Sample

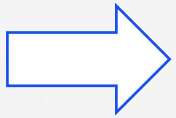
Multi-Sample

Code

Multi-Sample

ตัวอย่างการคำนวณ $\hat{y}$

sex	BMI
0	29
0	26
1	30
1	28



$\hat{y}$
?
?
?
?

Multi-Sample

If sex <= 0.5 and BMI <= 27, then ‘diabetes’

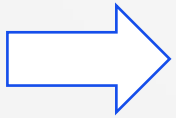
If sex <= 0.5 and BMI > 27, then ‘normal’

If sex > 0.5 and BMI <= 27, then ‘normal’

If sex > 0.5 and BMI > 27, then ‘diabetes’

Multi-Sample

sex	BMI
0	29
0	26
1	30
1	28



$\hat{y}$
normal
diabetes
diabetes
diabetes

Prediction

1-Sample

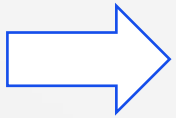
Multi-Sample

Code

Code

ตัวอย่าง code สำหรับการคำนวณ $\hat{y}$

sex	BMI
0	29
0	26
1	30
1	28



$\hat{y}$
?
?
?
?

Code

- Code สำหรับสร้าง model จากข้อมูลของเราโดยที่

$$X = \begin{bmatrix} 0 & 29 \\ 0 & 26 \\ 1 & 30 \\ 1 & 28 \end{bmatrix}$$

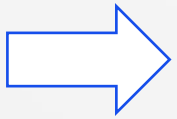
```
1  clf.predict(X)
```

```
array(['normal', 'diabetes', 'diabetes', 'diabetes'], dtype=object)
```

Code

ดังนั้น เราจะได้ $\hat{y}$ สำหรับข้อมูลชุดนี้คือ

sex	BMI
0	29
0	26
1	30
1	28



$\hat{y}$
normal
diabetes
diabetes
diabetes

Code

Code for this section

Open File

Model Creation.ipynb

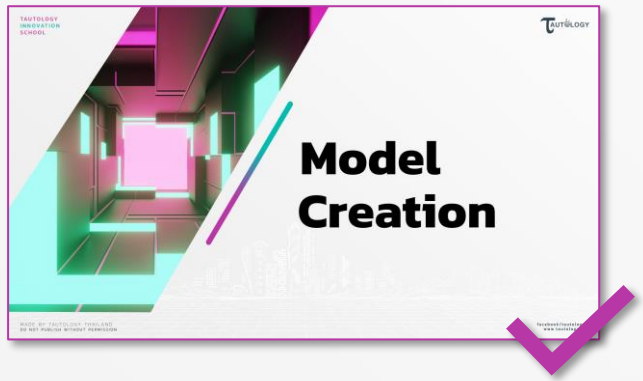
Prediction

1-Sample

Multi-Sample

Code

Classification Tree



TAUTOLOGY
INNOVATION
SCHOOL

Workshop

AI in Healthcare

- Abstract
- Why this project important?
- Who this project for?
- Hepatitis C Dataset
- What we learn from this project?

Abstract

สร้าง model เพื่อวินิจฉัยผู้ป่วยโรคไวรัสตับอักเสบ C (Hepatitis C) โดย feature ที่นำมาใช้ คือ ข้อมูลทั่วไปของผู้ป่วย และ ผลตรวจการทำงานของตับ

Why this project important?

- สามารถสร้างระบบสำหรับตรวจโรคไวรัสตับอักเสบ c ที่ทำงานได้ตลอด 24 ชั่วโมง
- สามารถนำไปต่อยอดกับการวินิจฉัยโรคอื่น ๆ
- สามารถใช้เป็นพื้นฐานสำหรับการแพทย์ทางไกล

Who this project is for?

- ผู้บริหารโรงพยาบาล
- บุคลากรทางการแพทย์
- นักวิเคราะห์ข้อมูล

Hepatitis C Dataset

https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

Hepatitis C Dataset

Feature

- Age : อายุ (ปี)
- Sex : เพศ (m = ชาย, f = หญิง)
- ALB : ปริมาณโปรตีน Albumin ในตับ
- ALP : ปริมาณเอนไซม์ Alkaline phosphatase ในตับ
- ALT : ปริมาณเอนไซม์ Alanine transaminase ในตับ
- AST : ปริมาณเอนไซม์ Aspartate transaminase ในตับ
- BIL : สาร Bilirubin ในตับ

Hepatitis C Dataset

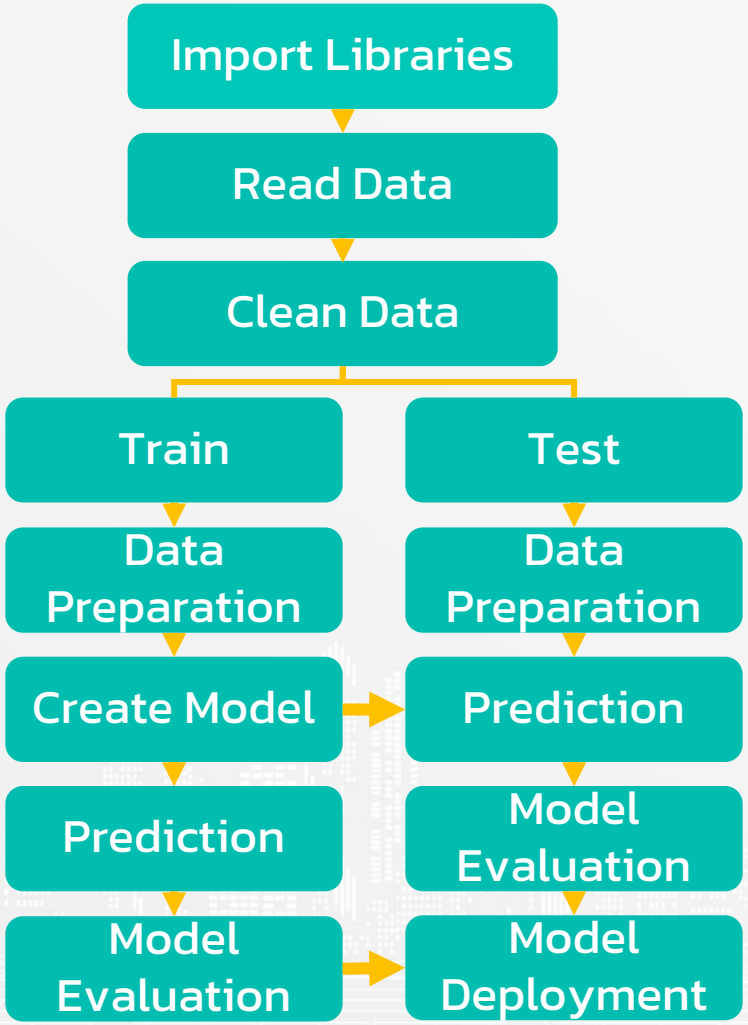
Feature

- CHE : ปริมาณเอนไซม์ Cholinesterase ใน serum
- CHOL : ปริมาณ Cholesterol ในตับ
- CREA : ปริมาณ Creatinine ในตับ
- GGT : ปริมาณโปรตีน Gamma glutamic transpeptidase ในตับ
- PROT : ปริมาณโปรตีน Prothrombin ในตับ

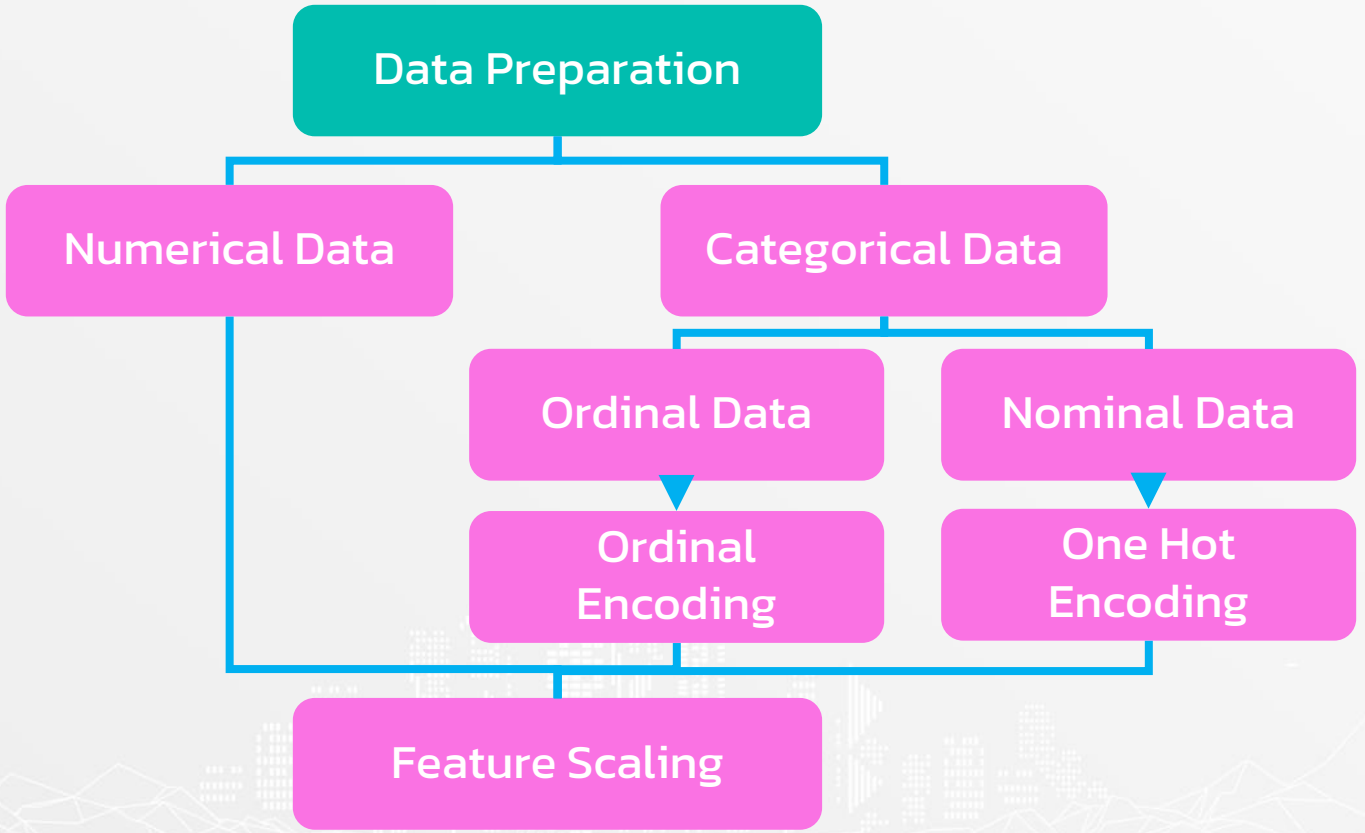
Target

- target : การเป็นโรคไวรัสตับอักเสบ c (0 = ไม่เป็น, 1 = เป็น)

What we learn from this project?



Data Preparation



File

02. HEPATITIS C

hepatitis_c_model.pickle

hepatitis_c_mc.ipynb

hepatitis_c_md.ipynb

hepatitis_c_dataset.csv

Classification Tree