Workshop

Workshop

2-class, table

AI in Healthcare

2-class, ข้อความ (อังกฤษ)

Fake News

multi-class, image

AI in Fruit Industry

Healthcare

AI in Insurance

- Abstract
- Why this project important?
- Who this project for?
- Heart Disease Dataset
- What we learn from this project?

Abstract

สร้าง model เพื่อวินิจฉัยผู้ป่วยโรคหัวใจ โดย feature ที่นำมาใช้ คือ ข้อมูลสภาวะร่างกาย เช่น

- ลักษณะการเจ็บหน้าอก
- ค่าความเข้มข้นน้ำตาลในเลือด
- ระดับคอลเลสเตอรอลในเลือด

Why this project important?

- สามารถสร้างระบบสำหรับตรวจโรคหัวใจที่ทำงานได้ตลอด 24 ชั่วโมง
- สามารถนำไปต่อยอดกับการวินิจฉัยโรคอื่น ๆ
- สามารถใช้เป็นพื้นฐานสำหรับการแพทย์ทางไกล

Who this project is for?

- ผู้บริหารโรงพยาบาล
- บุคลากรทางการแพทย์
- นักวิเคราะห์ข้อมูล

Heart Disease Dataset

https://www.kaggle.com/ronitf/heart-disease-uci

Heart Disease Dataset

Feature

- age : อายุ
- sex : เพศ (1 = ชาย, 0 = หญิง)
- cp : ลักษณะการเจ็บหน้าอก (0, 1, 2, 3)
- tresbph : ความดันโลหิตขณะพัก
- chol : ระดับคอลเลสเตอรอลในเลือด
- fbs : ค่าความเข้มข้นน้ำตาลในเลือด > 120 mg/dl (1 = จริง, 0 = เท็จ)
- restecg : ผลคลื่นไฟฟ้าหัวใจขณะพัก

ข้อดี
- ถามความรู้พื้นฐาน
- ช่วยงานง่าย ๆ เขียน code (ระบบที่ไม่ซับซ้อน)
- ช่วยคิดไอเดีย (สิ่งเร้าทางความคิด)

ข้อเสีย
- ข้อมูลที่มันให้มามีโอกาสผิดได้
- ข้อมูลลึก ๆ
- ระบบที่ต้องการความรัดกุม (เราควรเขียนเอง)

Heart Disease Dataset

Feature

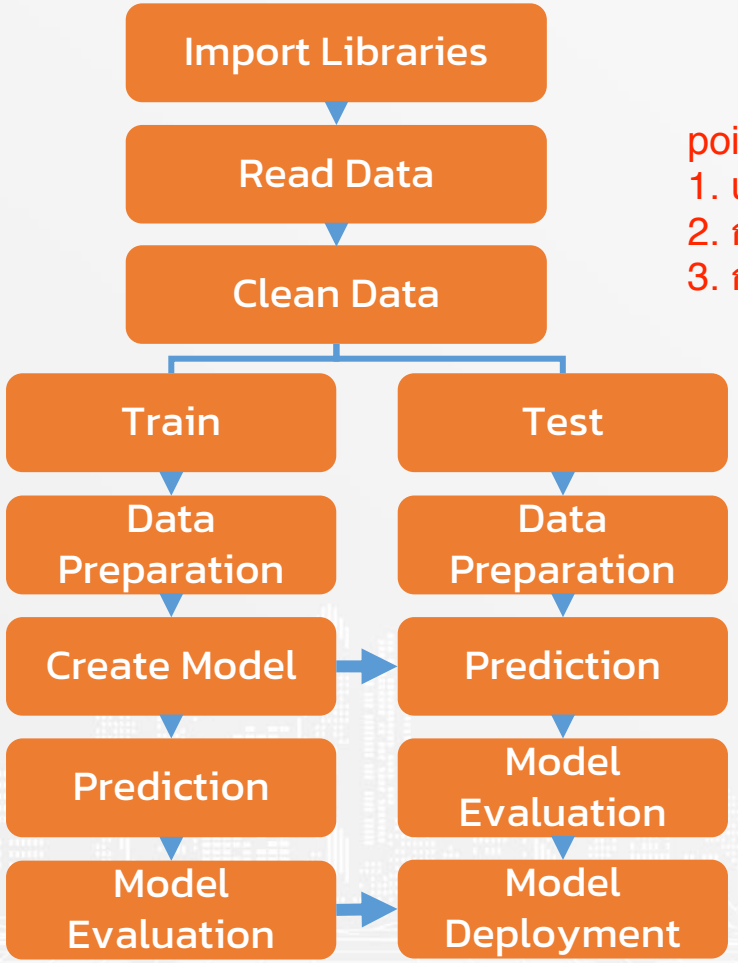
- thalach : อัตราการเต้นหัวใจสูงสุด
- exang : อาการเจ็บหน้าอกระหว่างออกกำลังกาย (1 = เจ็บ, 0 = ไม่เจ็บ)
- oldpeak : การเกิดกราฟ ST depression ในผลการตรวจคลื่นไฟฟ้าของหัวใจ
- slop : ลักษณะความชันของกราฟ ST segment (0 = ชันขึ้น, 1 = ราบ, 2 = ชันลง)
- ca : จำนวนเส้นเลือดตีบ
- thal : ลักษณะความเครียดของหัวใจ (0, 1, 2, 3)

Target

- target : การเป็นโรคหัวใจ (1 = เป็น, 0 = ไม่เป็น)

What we learn from this project?

key success ของการสร้างโมเดล
1. right algorithm
2. right feature

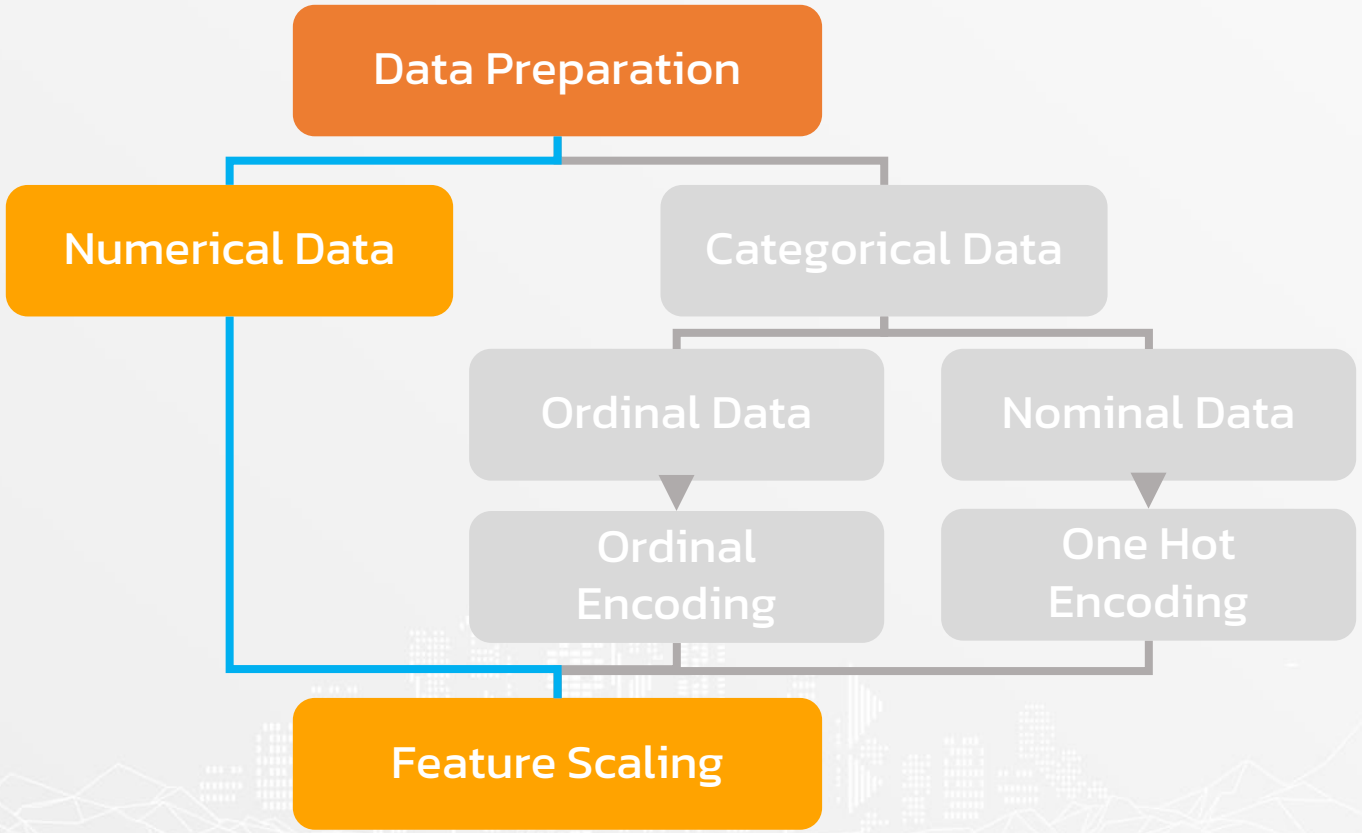


point
1. use case & inspiration
2. การสร้าง classification ตัวแรกของเรา
3. การใช้ chatGPT ช่วยอธิบายข้อมูล & ให้ไอเดีย

chatGPT

ควร ==> คิด feature
ไม่ควร ==> เขียน code (ไม่รัดกุม)

Data Preparation



01. HEART DISEASE

heart_disease_dataset.csv

heart_disease_mc.ipynb

heart_disease_md.ipynb

heart_disease_model.pickle

Workshop

AI in Healthcare

Fake News

AI in Fruit Industry

Fake News

- Abstract
- Why this project important?
- Who this project for?
- Fake News Dataset
- What we learn from this project?

Abstract

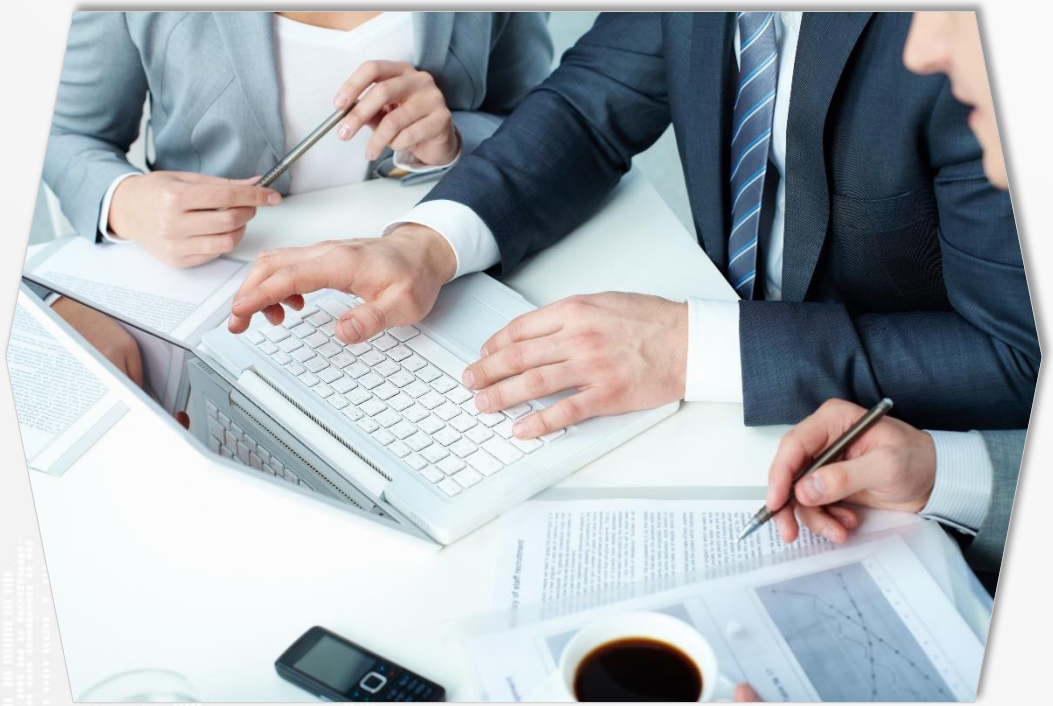
สร้าง model เพื่อตรวจสอบข่าวปลอมโดยพิจารณาจากหัวข้อข่าว, เนื้อหาข่าว และหมวดหมู่ข่าว

Why this project important?

- สามารถสร้างระบบตรวจสอบข่าวปลอมที่ทำงานได้ตลอด 24 ชั่วโมง
- สามารถนำไปต่อยอดเพื่อจัดอันดับความน่าเชื่อถือของสื่อมวลชน
- สามารถนำไปประยุกต์ใช้กับงานที่มีลักษณะใกล้เคียงได้ เช่น sentimental analysis

Who this project is for?

- บุคลากรด้านสื่อมวลชน
- นักลงทุน
- นักวิเคราะห์ข้อมูล



Fake News Dataset

https://www.kaggle.com/c/fake-news/data

Fake News Dataset

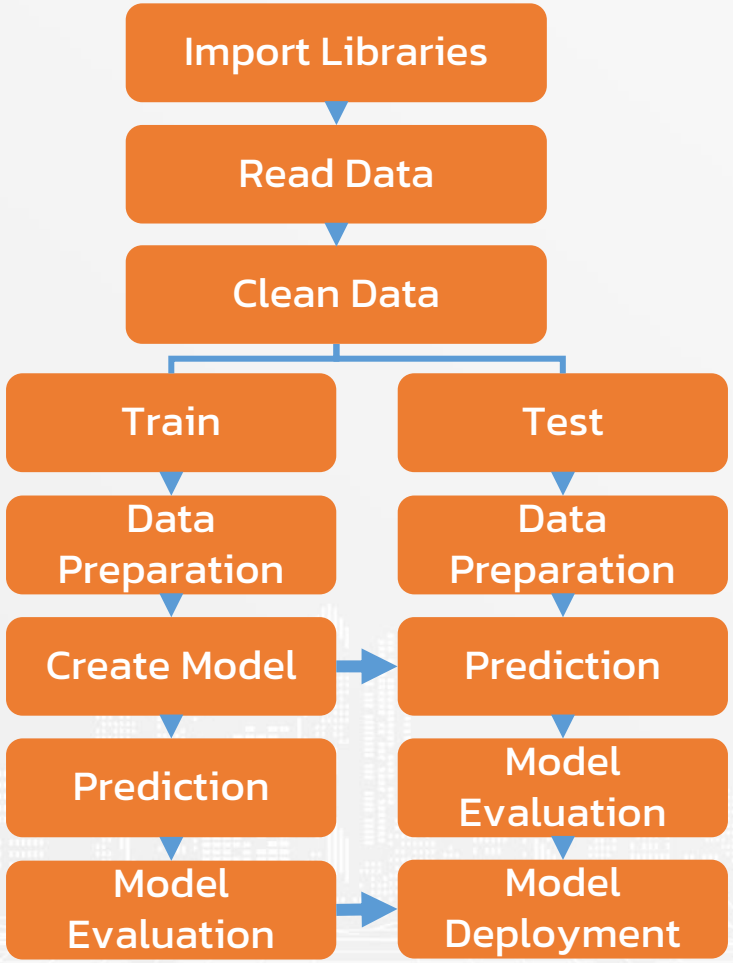
Feature

- title : หัวข้อข่าว
- text : เนื้อหาข่าว
- subject : หมวดหมู่ข่าว

Target

- class : ค่าความจริงของข่าว (fake, true)

What we learn from this project?

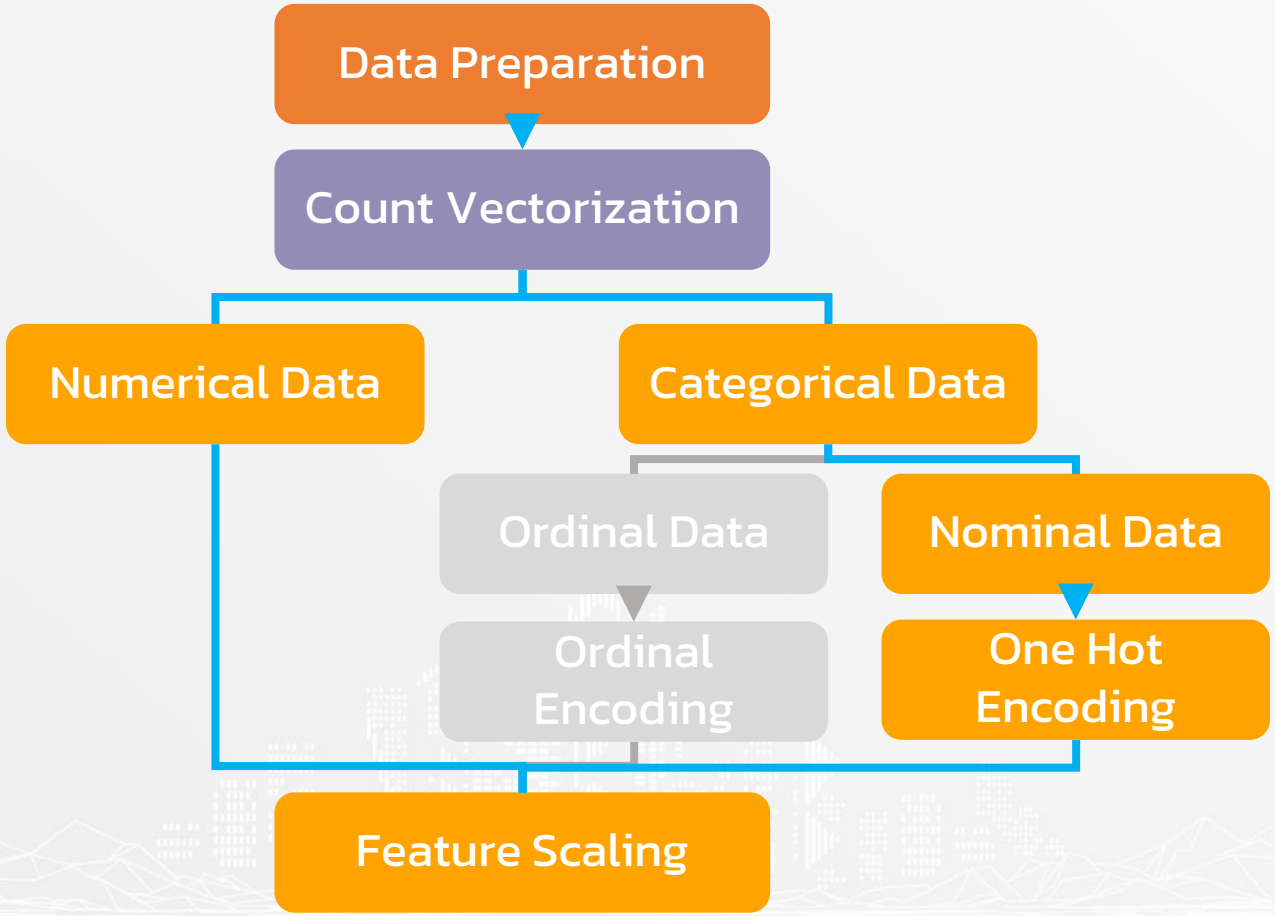


point
1. use case
2. วิธีการจัดการกับข้อมูลที่เป็น ข้อความ

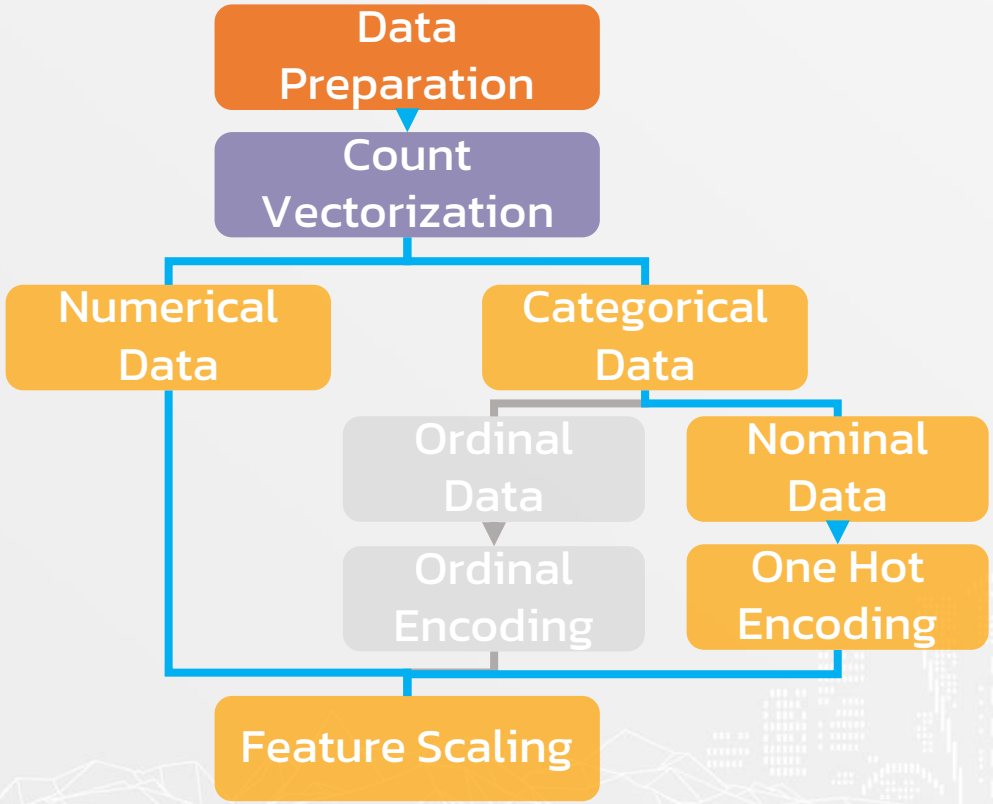
Import Libraries

```
1  import numpy as np
2  import pandas as pd
3  import matplotlib.pyplot as plt
4
5  from sklearn.model_selection import train_test_split
6  from sklearn.feature_extraction.text import CountVectorizer
7  from sklearn.preprocessing import OrdinalEncoder, OneHotEncoder, StandardScaler, MinMaxScaler
8  from sklearn.linear_model import LogisticRegression, LogisticRegressionCV
9  from sklearn.metrics import plot_confusion_matrix, classification_report
10
11 import warnings
12 warnings.filterwarnings('ignore')
13
14 np.random.seed(12345)
```

Data Preparation



Data Preparation



Count vectorization

สร้าง feature ใหม่ โดยการหา unique word จากข้อความทั้งหมดใน dataset จากนั้นให้พิจารณาว่าแต่ละข้อความประกอบด้วย unique word อะไรบ้าง และจำนวนกี่ครั้ง

	'apple'	'green'	'is'	'kiwi'	'orange'	'red'
'Apple is red'	1	0	1	0	0	1
'Kiwi is green'	0	1	1	1	0	0
'Orange is orange'	0	0	1	0	2	0

Count Vectorization

	cnt_title_000	cnt_title_10	cnt_title_100	...	cnt_title_year	cnt_title_years	...
BAGHDAD (Reuters) - A Russian Islamic State fi...	0	0	0	...	0	0	...
WASHINGTON (Reuters) - President Donald Trump ...	0	0	0	...	0	0	...
WASHINGTON (Reuters) - Russian President Vladi...	0	0	0	...	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

corpus = คลังคำศัพท์

Code

- Count vectorization for **training set**

```
1  corpus_train = X_train["title"].tolist()
2  title_vectorizer = CountVectorizer(max_features=1000)
3  title_vectorizer.fit(corpus_train)
4  title_cnt_vec_train = title_vectorizer.transform(corpus_train).toarray()
```

list unique word

count

```
1  title_cnt_vec_feature_name = [
2      "cnt_title_" + feature for feature in title_vectorizer.get_feature_names()
3  ]
```

สร้างชื่อ feature หลังจากทำ count vectorization

```
1  X_train[title_cnt_vec_feature_name] = title_cnt_vec_train
2  X_train.drop("title", axis=1, inplace=True)
```

เพิ่ม feature จากการทำ CVZ ลงใน training set

drop feature ที่เราทำ count vectorization ไปเสร็จแล้ว

Code

- Count vectorization for **test set**

```
1 corpus_test = X_test['title'].tolist()
2 title_cnt_vec_test = title_vectorizer.transform(corpus_test).toarray()
```

```
1 X_test[title_cnt_vec_feature_name] = title_cnt_vec_test
2 X_test.drop('title', axis=1, inplace=True)
```

03. FAKE NEWS

dataset

fake_news_mc.ipynb

fake_news_md.ipynb

fake_news_model.pickle

Workshop

AI in Healthcare

Fake News

AI in Fruit Industry

AI in Fruit Industry

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

Abstract

สร้าง model เพื่อจำแนกผลไม้สด และผลไม้เสีย สำหรับ apple, banana และ orange โดยพิจารณาจากรูปผลไม้

Why this project important?

- สามารถสร้างระบบคัดแยกผลผลิตที่ทำงานได้อย่างมีประสิทธิภาพ
- สามารถนำความรู้ไปต่อยอดเพื่อสร้าง smart farm
- สามารถนำไปต่อยอดเพื่อจำแนกผลไม้ หรือ สินค้าชนิดอื่น

Who this project is for?

- เกษตรกรที่สนใจ AI กับการเกษตร
- ผู้ควบคุมสายการผลิต
- นักวิเคราะห์ข้อมูล

Fruit Dataset

https://www.kaggle.com/sriramr/fruits-fresh-and-rotten-for-classification

Fruit Dataset

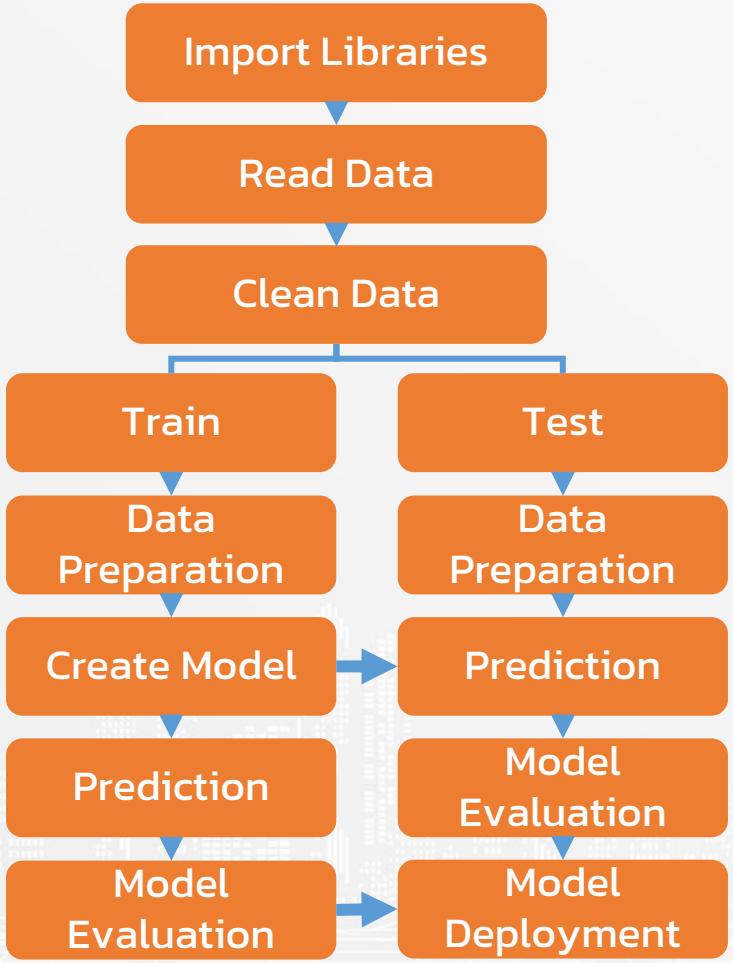
Feature

Q : ทำไมต้องมีทั้งรูปเอียง, รูปตรง, รูปถ่ายจากหลายมุม, และถ่ายผลไม้หลากหลายลูกด้วย

Target

- target : freshapples, freshbanana, freshoranges, rottenapples, rottenbanana, rottenoranges

What we learn from this project?



1. use case การใช้งาน
2. วิธีการดีลกับ image
3. วิธีการเตรียม dataset ที่เป็นรูปภาพที่ดี

Import Libraries

```
1  import numpy as np
2  import pandas as pd
3  import matplotlib.pyplot as plt
4
5  from glob import glob
6  from PIL import Image
7  import cv2
8  from tqdm.auto import tqdm
9  from sklearn.model_selection import train_test_split
10 from sklearn.preprocessing import OrdinalEncoder, OneHotEncoder, StandardScaler, MinMaxScaler
11 from sklearn.linear_model import LogisticRegression, LogisticRegressionCV
12 from sklearn.metrics import (
13     plot_confusion_matrix,
14     classification_report
15 )
16
17 import warnings
18 warnings.filterwarnings('ignore')
19
20 np.random.seed(12345)
```

ใช้ในการเข้าถึงชื่อไฟล์
ใช้ในการอ่านไฟล์ภาพ
ใช้ในการจัดการบางอย่างกับรูปภาพ
เป็นตัวโชว์ progress bar

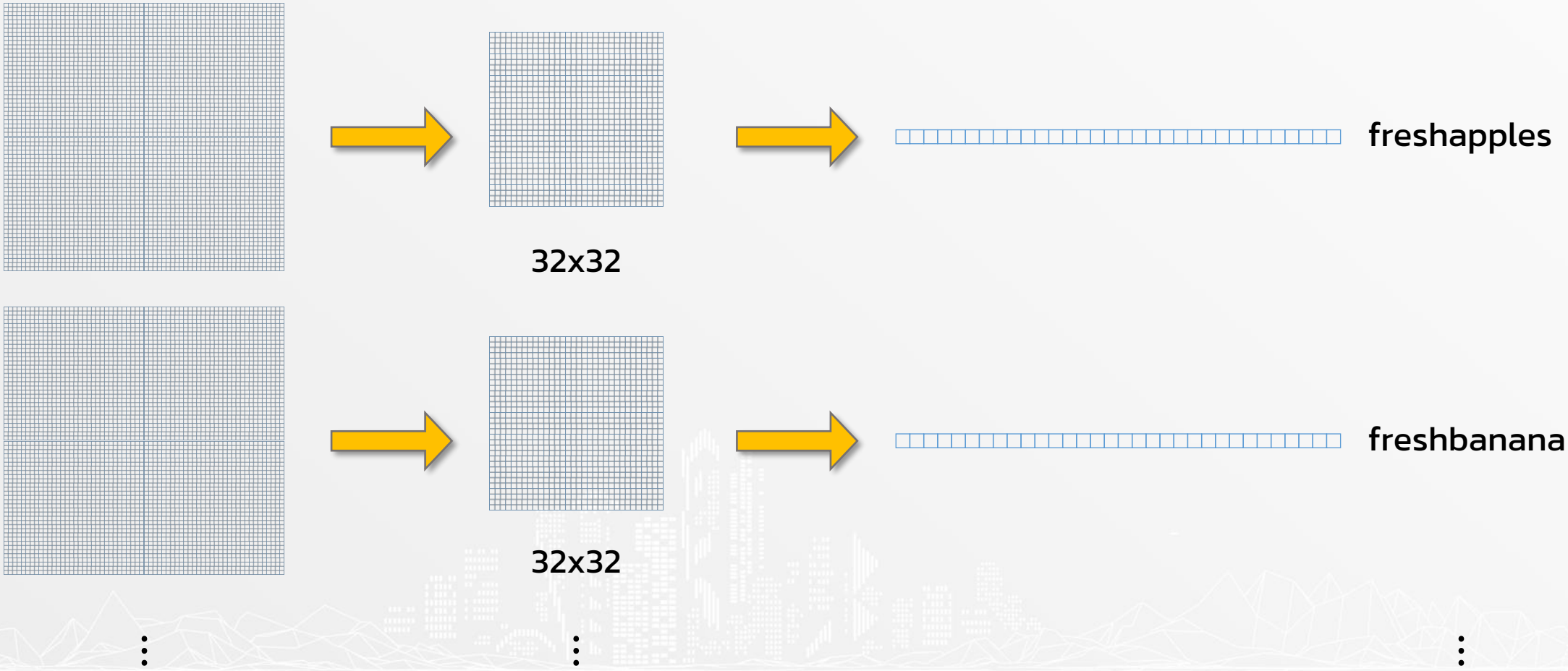
Read Data

```
classes = ['freshapples', 'freshbanana', 'freshoranges',
           'rottenapples', 'rottenbanana', 'rottenoranges']
```

```
X = np.empty([0, 32*32*3])   สร้าง matrix X ว่างเปล่าไว้รอ
y = np.empty([0, 1])   สร้าง vector y ว่างเปล่าไว้รอ

for _class in tqdm(classes):   ไล่อ่านข้อมูลทีละ folder
    img_path = glob('dataset/' + _class + '/*')   เข้าถึงชื่อไฟล์ทุกไฟล์ใน folder ที่กำลังพิจารณา
    for path in tqdm(img_path):   ไล่พิจารณาทีละชื่อไฟล์
        img = Image.open(path)   อ่านภาพ จากชื่อไฟล์
        img = img.resize((32, 32))   resize
        img = np.array(img)   แปลงให้เป็น matrix
        if img.shape[2] == 4:
            img = cv2.cvtColor(img, cv2.COLOR_BGRA2BGR)   การตรวจสอบและการแปลงไฟล์ภาพให้อยู่ใน format RGB
        img = img.reshape(1, -1)   ยืด dataset ให้เป็นเส้นตรง
        X = np.vstack([X, img])
        y = np.vstack([y, _class])   เอาข้อมูลแต่ละภาพมาต่อตึกกัน
```

Read Data



Read Data

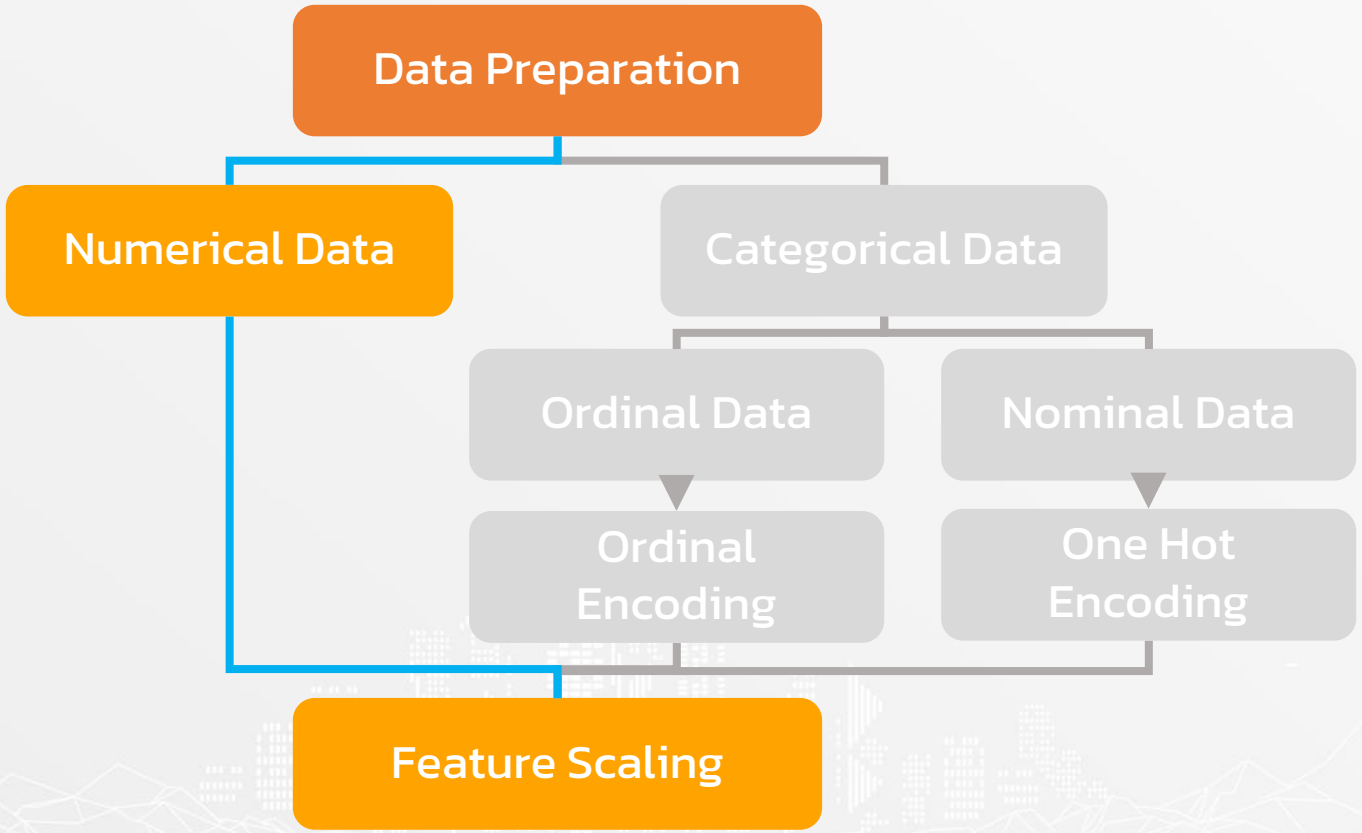
x_1	x_2	x_3	$\cdots$	x_{3072}
0.0	0.0	0.0	$\cdots$	0.0
0.0	0.0	0.0	$\cdots$	0.0
0.0	0.0	0.0	$\cdots$	0.0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
255.0	255.0	255.0	$\cdots$	255.0

X

y
freshapples
freshapples
freshapples
$\vdots$
rottenoranges

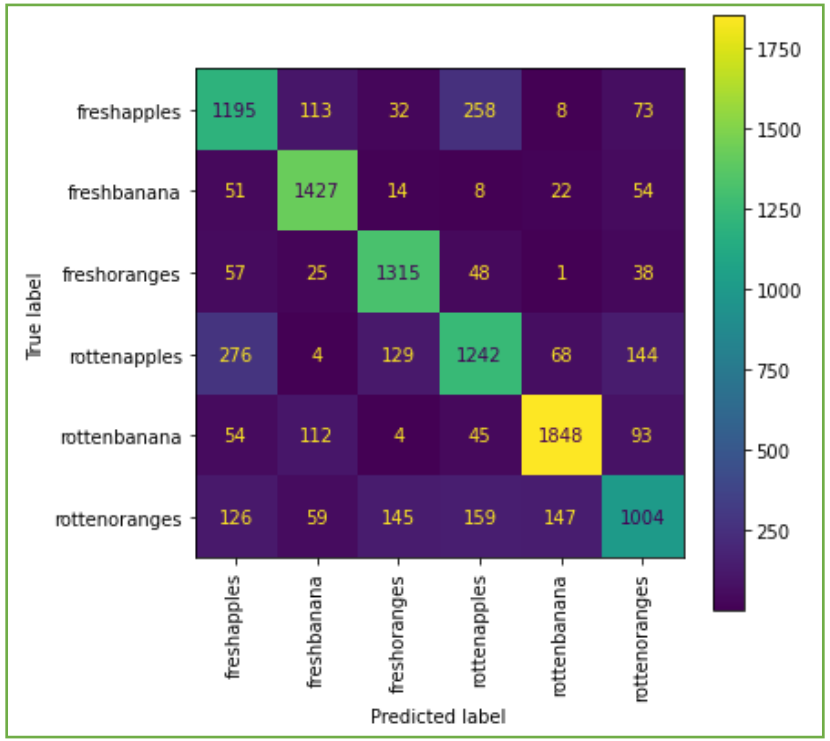
y

Data Preparation

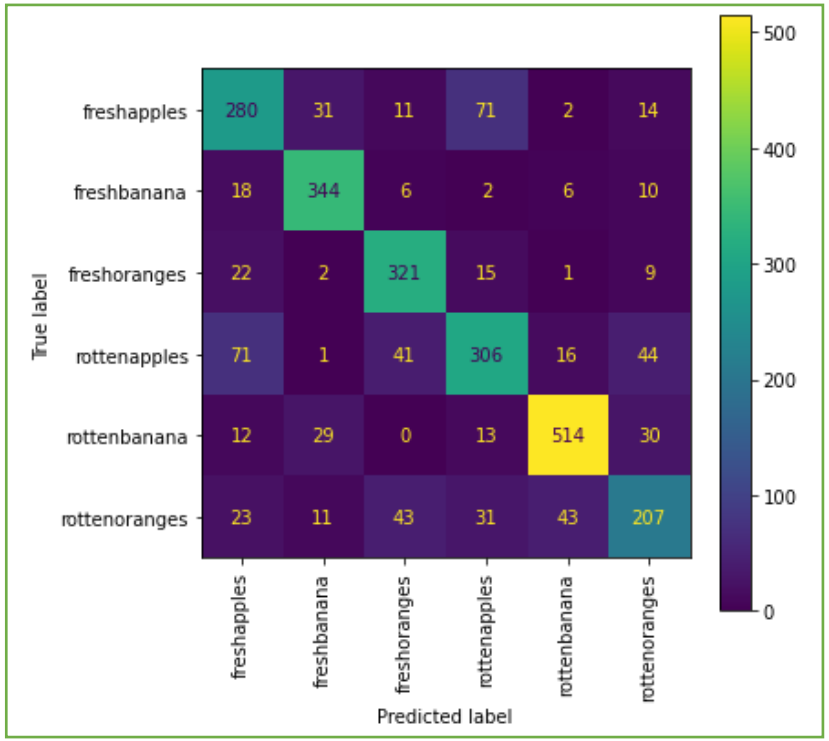


Model Evaluation

- Confusion Matrix for **training set**



- Confusion Matrix for **test set**



Code

- Confusion Matrix for **training set**

```
fig, ax = plt.subplots(figsize=(6, 6))
plot_confusion_matrix(clf, X_train_scaled, y_train, ax=ax)
plt.xticks(rotation=90)
plt.show()
```

- Confusion Matrix for **test set**

```
fig, ax = plt.subplots(figsize=(6, 6))
plot_confusion_matrix(clf, X_test_scaled, y_test, ax=ax)
plt.xticks(rotation=90)
plt.show()
```

Model Evaluation

- Scoring for **training set**

	precision	recall	f1-score	support
freshapples	0.679363	0.711733	0.695172	1679.00000
freshbanana	0.820115	0.905457	0.860676	1576.00000
freshoranges	0.802318	0.886119	0.842139	1484.00000
rottenapples	0.705682	0.666667	0.685620	1863.00000
rottenbanana	0.882521	0.857143	0.869647	2156.00000
rottenoranges	0.714083	0.612195	0.659225	1640.00000
accuracy	0.772360	0.772360	0.772360	0.77236
macro avg	0.767347	0.773219	0.768746	10398.00000
weighted avg	0.770561	0.772360	0.770028	10398.00000

Code

- Scoring for **training set**

```
1 report = classification_report(y_train, y_pred_train, output_dict=True)
```

```
1 print('accuracy =', report['accuracy'])
```

```
1 pd.DataFrame.from_dict(report).T
```

Model Evaluation

- Scoring for **test set**

	precision	recall	f1-score	support
freshapples	0.657277	0.684597	0.670659	409.000000
freshbanana	0.822967	0.891192	0.855721	386.000000
freshoranges	0.760664	0.867568	0.810606	370.000000
rottenapples	0.698630	0.638831	0.667394	479.000000
rottenbanana	0.883162	0.859532	0.871186	598.000000
rottenoranges	0.659236	0.578212	0.616071	358.000000
accuracy	0.758462	0.758462	0.758462	0.758462
macro avg	0.746989	0.753322	0.748606	2600.000000
weighted avg	0.756430	0.758462	0.756053	2600.000000

Code

- Scoring for **test set**

```
report = classification_report(y_test, y_pred_test, output_dict=True)
```

```
print('accuracy =', report['accuracy'])
```

```
pd.DataFrame.from_dict(report).T
```

05. FRUIT INDUSTRY

dataset

fruit_industry_mc.ipynb

fruit_industry_md.ipynb

fruit_industry_model.pickle

Workshop

AI in Healthcare

Fake News

AI in Fruit Industry

Logistic Regression

Logistic Regression (Binary)

Logistic Regression (Multi-Class)

Workshop