Naïve Bayes for Cancer Classification

1. Introduction

ในไฟล์ cancer.csv จะเป็นข้อมูลผู้ป่วยที่รักษาโรคมะเร็งตับ โดยจะแบ่งออกเป็น 4 ระยะ (CA Level) ซึ่งจะมีข้อมูลแต่ละคอลัมน์ดังนี้ Sex, Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Tryglicerides, Platelets, Prothrombin, Ca level ในไฟล์ cancer.csv จะเป็นข้อมูลผู้ป่วยที่รักษาโรคมะเร็งตับ โดยจะแบ่งออกเป็น 4 ระยะ (CA Level) ซึ่งจะมีข้อมูลแต่ละคอลัมน์ดังนี้ Sex, Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Tryglicerides, Platelets, Prothrombin, Ca level

2. Preprocessing

Sex	Ascites	Hepatomegaly	Spiders	Edema
F	Y	Y	Y	Y
F	N	Y	Y	N
M	N	N	N	S
F	N	Y	Y	S
F	N	Y	Y	N
F	N	Y	N	N

คอลัมน์เหล่านี้มีค่าข้อมูลเป็น String ซึ่งเราจะทำการเปลี่ยนให้อยู่ในรูปของตัวเลขก่อนทำการเทรน

```python
df['Sex'] = df['Sex'].map({'M': 1, 'F': 0})
df['Ascites'] = df['Ascites'].map({'Y': 1, 'N': 0, 'S': 2})
df['Hepatomegaly'] = df['Hepatomegaly'].map({'Y': 1, 'N': 0, 'S': 2})
df['Spiders'] = df['Spiders'].map({'Y': 1, 'N': 0, 'S': 2})
df['Edema'] = df['Edema'].map({'Y': 1, 'N': 0, 'S': 2})

X = df.drop(columns=['CA level']) # feature
y = df['CA level'] # class
```

✓ 0.0s Python

3. Training Model

3.1 train test split โดยกำหนด ข้อมูลที่ใช้ train กับ test อัตราส่วนคือ 7:3 และกำหนด random state = 3

```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=3, shuffle=True) # แบ่งข้อมูลเป็น train และ test
```

✓ 0.0s Python

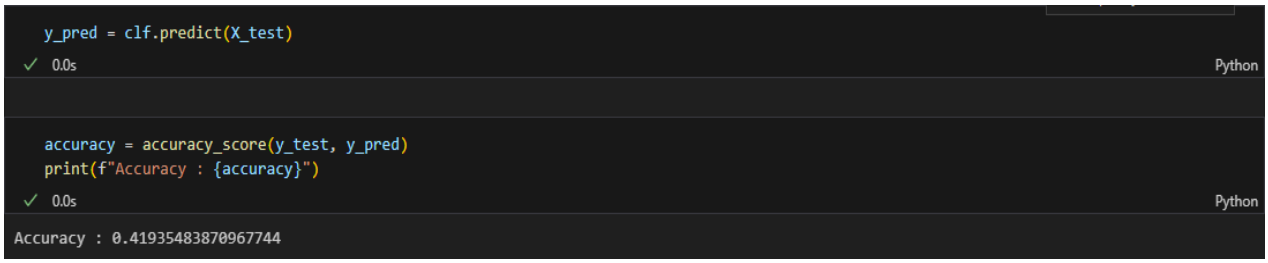
3.2 ใช้ MultinomialNB model ในการ train โดยค่า alpha = 1.5

```python
clf = MultinomialNB(alpha=1.5, class_prior=None, fit_prior=True)
clf.fit(X_train, y_train)
```

✓ 0.0s Python

4. Result

จาก model ที่เทรนด้วยวิธีที่นำเสนอมีค่าความแม่นยำประมาณ 41.9%



ภูมิระพี เสริญวณิชกุล

6510405750