K-Nearest Neighbors 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 ซึ่งเราจะทำการเปลี่ยนให้อยู่ในรูปของ Integer ก่อนทำการเทรน

```
df = df.apply(LabelEncoder().fit_transform)
df
```

3. Training Model

3.1 train test split โดยกำหนด ข้อมูลที่ใช้ train กับ test อัตราส่วนคือ 75:25 และกำหนด random state = 3

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=3) # แบ่งข้อมูลเป็น train และ test
```

3.2 ใช้ scaler เพื่อปรับข้อมูลในช่วง 0-1 และทำให้คำนวณเร็วขึ้น

```
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

3.3 สร้าง model และ เทรน

```
classifier = KNeighborsClassifier(n_neighbors=7)
classifier.fit(X_train, y_train)
```

4. Result

จาก model ที่เทรนด้วยวิธีที่นำเสนอมีค่าความแม่นยำประมาณ 58.97%

```
Accuracy: 0.5897435897435898
==================================================
Confusion Matrix:
[[ 0  0  1  0]
 [ 1  5  6  0]
 [ 0  6 19  5]
 [ 0  4  9 22]]
==================================================
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.33      0.42      0.37        12
           2       0.54      0.63      0.58        30
           3       0.81      0.63      0.71        35

    accuracy                           0.59        78
   macro avg       0.42      0.42      0.42        78
weighted avg       0.63      0.59      0.60        78
```

ภูมิระพี เสริญวณิชกุล

6510405750