# 데이터 베이스 시스템

건강 상태를 기반으로 적절한 약물 분류

- Decision Trees -

214683 я æ ΰ

# 목차

1. 모델링 배경 및 목적

2. 데이터 설명

3. 모델링 과정

4. 결과 및 성능

#### 1. 모델링 배경 및 목적





최근 일교차가 심해지며 감기로 고생하는 사람 이 많아졌습니다.

한 달 가까이 감기 증상에 감기약을 복용하였음 에도 큰 효과를 느끼지 못하면서, "현재 나에 게는 어떤 약이 더 효과적일까?" 하는 궁금증 이 생겼습니다.

현재 데이터셋에는 약의 종류가 정해지진 않았지만 비슷한 성향을 분류한다는 내용에 대해 집 중하였습니다.

#### 2. 데이터 설명

```
df = pd.read_csv("drug200.csv")
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
   Column Non-Null Count Dtype
            200 non-null int64
0
   Age
               200 non-null object
   Sex
           200 non-null object
   BP
   Cholesterol 200 non-null object
   Na to K 200 non-null
                            float64
    Drug 200 non-null
                             object
dtypes: float64(1), int64(1), object(4)
memory usage: 6.3+ KB
```

200개의 데이터셋 존재 5개의 독립변수, 1개의 종속 변수 종속 변수 : Drug

0. Age : 나이

1. Sex : 성별

2. BP : 혈압 상태

3. Cholesterol : 콜레스테롤 상태

4. Na\_to\_K 혈중 나트륨 대비 칼륨

# 3. 모델링 과정

```
from sklearn.preprocessing import LabelEncoder
                                        데이터셋이 문자로 되어있는 관계로 숫자로
                                        인식 될 수 있도록 인코딩 과정을 거침
# Label Encoding
le sex = LabelEncoder()
le bp = LabelEncoder()
                                        인코딩 이후 데이터셋 x,y분리 및 훈련
le chol = LabelEncoder()
le drug = LabelEncoder()
df['Sex'] = le sex.fit transform(df['Sex']) # 'F', 'M' + 0, 1
                                   # 'HIGH', 'NORMAL', 'LOW' →
df['BP'] = le bp.fit transform(df['BP'])
df['Cholesterol'] = le chol.fit transform(df['Cholesterol'])
df['Drug'] = le drug.fit transform(df['Drug']) # 'A'~'Y' → 0~4
x = df.drop(columns='Drug')
y = df['Drug']
x train, x test, y train, y test = train test split(X, y, stratify=y, random state:
```

#### 4. 결과 및 성능

```
export graphviz(
훈련 및 테스트 성능 검사
                                                                                                       dTreeAll,
                                                                                                       out file="decisionTree0.dot".
                                                                                                       class names=le drug.classes .
from sklearn.tree import DecisionTreeClassifier
                                                                                                       feature names-x.columns,
                                                                                                       impurity=False,
                                                                                                       filled=True
dTreeAll = DecisionTreeClassifier(random state=0)
                                                                                                    (graph,) = pydot.graph from dot file('decisionTree0.dot', encoding='utf8'
dTreeAll.fit(x_train, y_train)
                                                                                                    raph.write_png('decisionTree0.png')
                                                                                                                             Na_to_K <= 14.906
                                                                                                                               samples = 150
                                                                                                                          value = [17, 12, 12, 41, 68]
print("Train Score: {:.2f}".format(dTreeAll.score(x train, y train)))
                                                                                                                               class = drugY
print("Test Score : {:.2f}".format(dTreeAll.score(x test, y test)))
                                                                                                                           True
                                                                                                                                           False
                                                                                                                     BP <= 0.5
                                                                                                                                          samples = 68
                                                                                                                    samples = 82
                                                                                                                                       value = [0, 0, 0, 0, 68]
                                                                                                               value = [17, 12, 12, 41, 0]
                                                                                                                                          class = drugY
                                                            Pydot을 이용한 트리 구조
                                                                                                                    class = drugX
Train Score: 1.00
Test Score: 0.98
                                 Figure 1
                                                                                                          Age <= 52.0
                                                                                                                                BP <= 1.5
                                                   ROC Curve for Drug Y
                                                                                                          samples = 29
                                                                                                                               samples = 53
                                   1.0 -
                                                                                                      value = [17, 12, 0, 0, 0]
                                                                                                                           value = [0, 0, 12, 41, 0]
                                                                                                         class = drugA
                                                                                                                               class = drugX
                                   0.8
                                                                                                                             Cholesterol <= 0.5
                                 9.0 ge
                                                                                     samples = 17
                                                                                                          samples = 12
                                                                                                                                                    samples = 28
                                                                                                                               samples = 25
                                                                                  ∨alue = [17, 0, 0, 0, 0]
                                                                                                      value = [0, 12, 0, 0, 0]
                                                                                                                                                 value = [0, 0, 0, 28, 0]
                                                                                                                           value = [0, 0, 12, 13, 0]
                                                                                     class = drugA
                                                                                                          class = drugB
                                                                                                                                                   class = drugX
                                                                                                                               class = drugX
                                                                                                                     samples = 12
```

# 4. 결과 및 성능

```
Pydot을 이용한 트리 구조 및 Curve
```

value = [17, 0, 0, 0, 0]

class = drugA

value = [0, 12, 0, 0, 0]

class = drugB

```
# 3. PR Curve
                                                                                                           precision, recall, = precision recall curve(y test binary, y proba)
export graphviz(
   dTreeAll,
                                                                                                           plt.figure(figsize=(6, 5))
   out file="decisionTree0.dot",
                                                                                                           plt.plot(recall, precision, label="PR Curve")
   class names=le_drug.classes_
   feature names=x.columns,
                                                                                                           plt.xlabel("Recall")
   impurity=False.
                                                                                                           plt.ylabel("Precision")
   filled True
                                                                                                           plt.title("Precision-Recall Curve for Drug Y")
                                                                                                           plt.legend()
                                                                             Na to K <= 14.906
                                                                                                           plt.grid()
(graph,) = pydot.graph from dot file('decisionTree0.dot', encoding='utf8')
                                                                               samples = 150
graph.write png('decisionTree0.png')
                                                                           value = [17, 12, 12, 41, 68]
                                                                                                           plt.tight layout()
                                                                                class = drugY
                                                                                                           plt.show()
                                                                           True
                                                                                            False
                                                                                                           # 4. AUC 값 출력
                                                                     BP <= 0.5
                                                                                           samples = 68
                                                                                                           roc auc val = roc auc score(y test binary, y proba)
                                                                    samples = 82
                                                                                        value = [0, 0, 0, 0, 68]
                                                               value = [17, 12, 12, 41, 0]
                                                                                                           print("ROC AUC Score (Drug Y vs others):", round(roc_auc_val, 3))
                                                                                           class = drugY
                                                                    class = drugX
                                                                                                                                                                                      - 0 X
                                                                                                                                                 Figure 1
                                                                                                                                                                   ROC Curve for Drug Y
                                                                                 BP <= 1.5
                                                          Age <= 52.0
                                                                                                                                                   1.0 -
                                                                               samples = 53
                                                         samples = 29
                                                     value = [17, 12, 0, 0, 0]
                                                                            value = [0, 0, 12, 41, 0]
                                                         class = drugA
                                                                                                                                                   0.8
                                                                               class = drugX
                                                                             Cholesterol <= 0.5
                                    samples = 17
                                                         samples = 12
                                                                                                     samples = 28
                                                                               samples = 25
```

samples = 13

value = [0, 0, 0, 13, 0]

class = drugX

value = [0, 0, 12, 13, 0]

class = drugX

samples = 12

value = [0, 0, 12, 0, 0]

class = drugC

value = [0, 0, 0, 28, 0]

class = drugX

9 0.4

0.2

0.0

ROC Curve (AUC = 1.00)

1.0

--- Random

False Positive Rate

# 출처

데이터 셋:

https://www.kaggle.com/datasets/pablomgomez21/drugs-a-b-c-x-y-for-decision-trees