**R – Py 컴퓨팅: Homework 1**

**Due: 2023-5-15 23:59:00, LMS 제출**

**파일 형식은 html, 워드, 한글, 혹은 pdf로, 코드는 부록으로 첨부**

**Part 1 보스턴 주택가격 데이터 분석하기: boston\_csv.csv 파일**

* **보스턴 주택 가격 데이터의 변수명은 다음과 같다.**

|  |  |
| --- | --- |
| CRIM | 자치시(town) 별 1인당 범죄율 |
| ZN | 25,000 평방피트를 초과하는 거주지역의 비율 |
| INDUS | 비소매상업지역이 점유하고 있는 토지의 비율 |
| CHAS | 찰스강에 대한 더미변수  (강의 경계에 위치한 경우는 1, 아니면 0) |
| NOX | 10ppm 당 농축 일산화질소 |
| RM | 주택 1가구당 평균 방의 개수 |
| AGE | 1940년 이전에 건축된 소유주택의 비율 |
| DIS | 5개의 보스턴 직업센터까지의 접근성 지수 |
| RAD | 방사형 도로까지의 접근성 지수 |
| TAX | 10,000 달러 당 재산세율 |
| PTRATIO | 자치시(town)별 학생/교사 비율 |
| B | 1000(Bk-0.63)^2,   여기서 Bk는 자치시별 흑인의 비율을 말함. |
| LSTAT | 모집단의 하위계층의 비율(%) |
| MEDV | 본인 소유의 주택가격(중앙값) (단위: $1,000) |
| CAT.MEDV | MEDV 가 $30,000 을 넘는지에 대한 변수 (넘는 경우 1, 아닌 경우 0) |

* **보스턴 주택 가격 데이터에서 결측치는 na와 NaN으로 표시되어 있다.**
* **다음과 같은 라이브러리를 불러와야 프로그램이 구현될 것이다.**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.metrics import mean\_squared\_error**

**import numpy as np**

**import matplotlib.pyplot as plt**

**import pandas as pd**

**import seaborn as sns**

**Q1 다음과 같은 데이터 전처리를 시행하라.**

1. **제공된 boston\_csv.csv 파일을 사용하여 pandas 데이터 프레임 객체를 만든다, 결측치 코드인 na와 NaN이 모두 실제 결측치로 되도록 한다.**
2. **이와 같은 결측치가 있는 관측치를 모두 제거한다.**

**Q2 다음과 같은 요약 통계를 구하라.**

1. **describe 메소드를 적용해서 각 변수병 요약 통계를 구한다.**
2. **상관관계를 구한후 seaborn 라이브러리의 heatmap 을 구현한다.**

**Q3 다음과 같은 단순회귀분석 모형을 Training Set과 Test Set을 통해 구현하라.**

* 모집단의 하위계층의 비율(LSTAT)이 독립변수
* 본인 소유의 주택가격(중앙값)인 MEDV가 종속변수
* Training set이 표본의 75%를 차지한다.
* Training Set에대해서는 회귀 분석 추정 계수 및 R2 값 mean squared error 값을 보고한다.
* Training Set에대해서는 회귀 분석 추정 계수 값을 바탕으로 Test Set에서 예측한 후 mean squared error 값을 보고한다.

**Q4 다음과 같은 다중회귀분석 모형을 Training Set과 Test Set을 통해 구현하라.**

* 모집단의 하위계층의 비율(LSTAT)과 10,000 달러 당 재산세율(TAX)가 독립변수
* 본인 소유의 주택가격(중앙값)인 MEDV가 종속변수
* Training set이 표본의 75%를 차지한다.
* Training Set에대해서는 회귀 분석 추정 계수 및 R2 값 mean squared error 값을 보고한다.
* Training Set에대해서는 회귀 분석 추정 계수 값을 바탕으로 Test Set에서 예측한 후 mean squared error 값을 보고한다.

PART2. 와인 클래스에 대한 kNN 알고리즘 적용

와인 데이터 구성

The data is the results of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in the three types of wine.

이 와인의 화학적 분석 결과는 다음과 같이 정리된다.

Attribute Information:

- Alcohol

- Malic acid

- Ash

- Alcalinity of ash

- Magnesium

- Total phenols

- Flavanoids

- Nonflavanoid phenols

- Proanthocyanins

- Color intensity

- Hue

- OD280/OD315 of diluted wines

- Proline

와인 종류는 class 변수로 구분되어 있으며, 1, 2, 3을 각 코드값으로 지니고 있다.

이 와인데이터 (wine\_data.csv) 파일에 대해서 다음과 같은 분석을 행하라.

1. Pandas를 사용해 wine\_data.csv파일을 wine 데이터프레임을 만든다.

2. Wine 데이터에 describe 메소드를 사용하여 요약통계량을 구한다.

3. 주어진 데이터에 첫 변수, class를 우리가 예측하는 타겟 변수, y, 나머지 모든 변수를 예측 변수 X로 활용하고, Train/Test 표본을 70% 와 30%가 되도록 데이터를 구성한다. 각각의 명칭은 X\_train/X\_test, y\_train/y\_test로 칭한다.

4. Scikit-learn의 KNeighborsClassifier를 사용하여 70%인 X\_train과 y\_train을 바탕으로 모형을 트레이닝 시킨다.

5. 이 Training 된 모형을 바탕으로 Training Set에서 예측해보고 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력하고, k=5로 결정하고 다수결 및 유클리디안 거리를 사용한다. .

6. 이 Training 된 모형을 바탕으로 Test Set에서 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력한다. .

7. 5-6번을 k=3일 경우에 대해서도 시행한다.

8. 3-6번에서 X 데이터를 Alchol, Malic acid, Ash, Alcalinity of ash 네 변수만 활용하여 다시 수행한다.