2장 – 머신러닝 프로젝트 처음부터 끝까지

<4조 의견>

- 처음 설정 부분의 코드는 파이썬, 사이킷런의 버전을 설정하거나 그래프를 출력할 때 더욱 깔끔하게 하기 위해 초기 설정 단계라고 생각한다. 또한 그림의 저장 위치 또한 설정한다.
- 필요한 라이브러리를 설치하고 데이터를 다운로드 하기 위해 파일의 위치 접근 코드를 실행하여 데이터를 불러온다. 데이터를 불러와 구조를 간단하게 살펴보기 위해 head 함수를 사용한다.
- Info()를 통해 해당 컬럼의 정보와 dtype 형태를 살펴볼 수 있다. Value_counts()를 통해 컬럼에 해당하는 정보의 개수를 파악할 수도 있다.
- Describe()를 통해 각 컬럼별 개수, 평균, 최소값, 최대값 등을 살펴볼 수 있다.
- 히스토그램에서 잘못된 이상치 데이터가 보이는데 이것을 제거하고 머신러닝을 돌려야 될 것 같다. 만약 제거를 하지 않고 머신러닝을 돌린다면 이상치 데이터 때문에 잘못된 학습이 이루어질 것이라고 생각된다. 그렇기 때문에 전처리 과정이 정말 중요한 것 같다고 생각한다.

<4조 의견>

※ 바로 머신러닝을 돌리지 않고 왜 데이터를 살펴볼까?

- 데이터의 정규분포를 확인하기 위해
- 변수의 결측값 여부를 확인하기 위해
- train 데이터와 test 데이터의 비율을 정확히 나누기 위해
- 데이터의 누락값이나 결측값으로 이상치 데이터가 존재하는 상황에서 머신러닝을 해버리면 잘못된 학습 결과가 나오기 때문
- 편향된 데이터로 인해 과대적합, 과소적합이 발생할 수 있기 때문

<4조 의견>

※ 훈련세트와 데이터 세트를 구분하는 방식

- 무작위 샘플링은 소규모보다는 대규모 데이터 세트에 사용될 것으로 생각한다. 또한 간단하고 편리하고 특별한 경우가 아닌 이상 균등한 분포를 유지할 것이라고 생각한다.

- 계층 샘플링은 각 계층에서 무작위 샘플링을 수행하는 방법이므로 여러 클래스나 범주로 나뉘나 있을 때 유용할 것이라고 생각이 든다. 그렇기 때문에 불균형 문제를 다룰 때 유용할 것 같다.