Pair Programing

3주차

1조
201984014
성도윤
201904126
허준혁

Q1. matplotlib는 무엇일까?

A1. 기본 그래프를 그리는 라이브러리
 -> 뒤에 붙은 mpl, plt는 무엇일까?
 -> 그냥 별칭으로 설정하는 코드

Q2. pandas는 무엇일까?

A2. 파이썬 라이브러리일 것 같다
 -> 데이터 조작과 분석을 위한 라이브러리

Q3. 데이터를 포함한 파일은 무엇일까?

A3. housing.csv
 -> 어떠한 형태를 띄고 있나요?
 ->엑셀 형태로 다양한 데이터를 저장하고있다

Q4. 히스토그램에서 비정상적인 값이 나오는 이유는?

A4. 잘못된 데이터이다
 -> 특정 값을 넘어가면 기존 패턴과 다른 패턴이 나온다
 -> 머신러닝이 이 값에 대해서도 패턴으로 인식한다
 -> 이 부분에 대해서는 개발자가 관여를 해야한다
 -> 특정 값의 범위를 지정해놔야한다

Q5. train_set, test_set 비율을 8대2로 설정하였는데 과대적합을 피할 수 있는가?

A5. 왜 그럴까?
-> 잘 모르겠다

Q6. 머신러닝을 통해서 학습하면 되는데 왜 계속 기존 데이터를 보고 결과물을 도출하는지?

A6. 학습을 할 필요가 없어서...
-> 최신 값에 대한 데이터는 적용이 안되는데?
-> 우리가 배우기에는 너무 어려워서
-> 그 후의 데이터는 지금까지 기록된 값을 토대로 유추
-> 정확한 값이 나오지않는다
-> 정확한 값을 향해 점점 나아가기 위한 과정이다

Q7. 무작위 샘플링과 계층적 샘플링의 장단점

A7.
* 무작위 샘플링
-> 간단하다, 범위가 무한하다
-> 편향이 발생할 수 있다
-> 무슨 값이 나올지 모른다
-> 값이 랜덤이다

* 계층적 샘플링
 -> 편향이 발생하지 않는다
 -> 복잡하다
 -> 더욱 정확한 값을 가져올 수 있다
 -> 전체 계층을 대표하도록 할 수 있다

Q8. 데이터 전처리기란 무엇인가?

A8. 모델 학습을 효율적으로 진해행하기 위해 주어진 데이터를 변환시키는 것
 -> 머신러닝에서 매우 중요하다

Q9. 튜닝을 위한 방법은 어떤 것이 있나요?

A9. 그리드 탐색, 랜덤 선택, 앙상블 방법