Machine learning Final

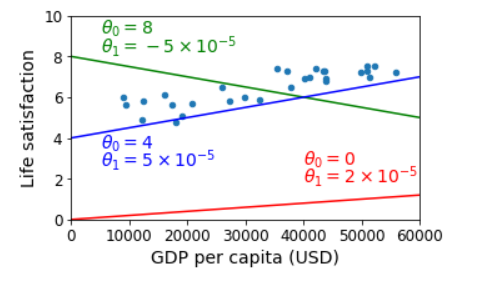
이름: \_\_\_최윤영 \_\_\_\_\_\_\_

**해당 과목은 오픈북입니다**

* 이 시험은 오픈북 테스트 이며 Colab 파일에서 문제를 풀거나(Question 1, Question 3) 인터넷 서칭(Question 2) 후 이 문서에 정리후 제출하시면 됩니다.
* 설명에 필요하다면 화면캡쳐 하셔도 좋습니다.
* 질문에 해당하는 코드 위치는 Colab 파일에서 Question 1, Question 2, Question 3으로 마킹되어 있습니다. Ctrl + F를 활용하셔서 쉽게 그 위치를 찾을 수 있습니다

Question 1. 1인당 GDP와 삶의 만족도가 어느 정도 선형 관계에 있는 것처럼 보인다. 그런데 어떤 선형 관계가 가장 적절한가 판단해야 한다. 어느 선형관계가 가장 적절한지 파악하고 그 이유를 적으시오

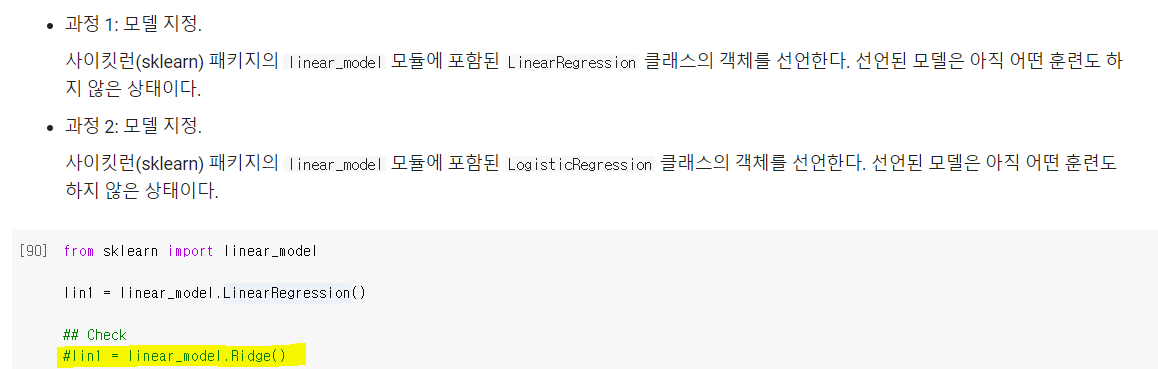
파란색 직선이 가장 적절하다. 데이터들의 분포를 적절하게 선형으로 잘 나타내고 있기 때문이다. 초록색과 빨간색 직선은 일부 데이터들의 값을 대표하지 못한다.



Question 2. 대표적인 linear model에는 LInearRegression() 뿐만 아니라 Ridge()도 포함된다. 선형회귀 모델과 Ridge 모델이 어떤 점이 개념적으로 다른지 설명하시오

Ridge()모델은 선형 회귀의 연장선으로, 손실함수가 모델의 복잡도를 최소화하기 위해 수정된 것을 말한다. 손실함수의 수정은 penalty parameter를 설정함으로써 이루어진다.

Question 3. 대표적인 linear model에는 LInearRegression() 뿐만 아니라 Ridge()에 대해서 1인당 GDP에 대한 삶의 만족도 데이터를 훈련시켜보고 키프러스(Cyprus)의 삶의 만족도를 예측해본다. 둘의 훈련 결과와 예측 결과는 어떤지 언급하고, 설명해보시오

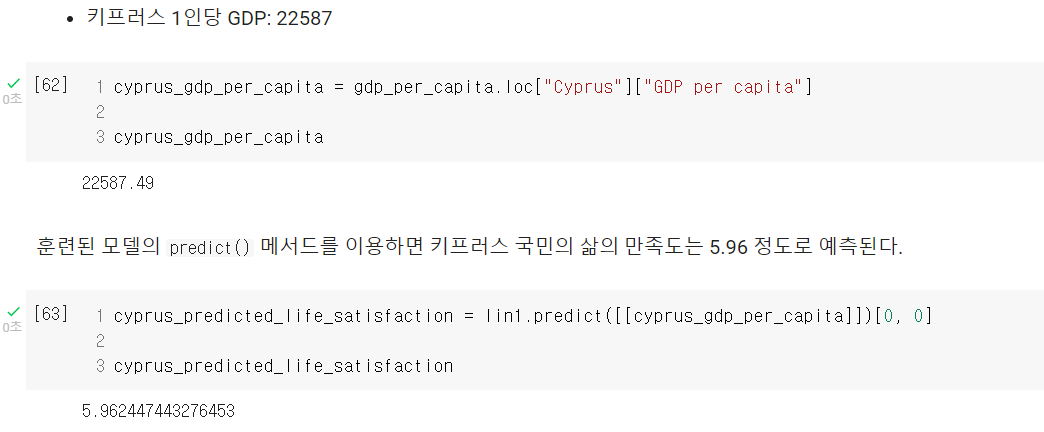


※ 모델 선택은 둘 중하나의 모델 지정 코드라인을 주석처리 하는 방식으로 진행하면 간편할 것이다

1. linear regression



1. ridge



linear regression의 예측값은 5.96244744318815,

ridge 의 예측값은 5.962447443276453

으로 거의 차이가 없다.

ridge regression은 linear regression을 할 때 파라미터 여러개로 모델을 만들 때, 과대적합을 방지하기 위해 사용하는데, 여기에서는 하나의 파라미터로 진행하기 때문에 두 모델간의 차이가 크게 나지 않는다.