Assignment 4. **Due date**: 08/10

Titanic data는 classification 에서 기본으로 다루는 예제입니다. Kaggle에 많은 분석들이 올라와 있습니다. 자료분석에서 가장 시간을 많이 차지하는 부분이 전처리입니다. missing값에 대한 처리가 평균을 대체하는 방법 혹은 크게 영향이 없으면 아예 자료를 없애는 방법도 있지만 추천하지는 않습니다. 저는 성별, 나이, 좌석등급을 변수로 보았지만 이 부분도 다른 기준으로 얼마든지 분석가능합니다. 두번째 자료가 breast cancer입니다. 이 자료도 classfication예제로 많이 사용됩니다.

Knn example파일은 knn의 원리에 대해서 설명한 자료입니다. Knn은 훈련데이터로 모델을 생성하는 것이 아니라 훈련데이터를 기반으로 테스트 데이터가 들어오면서 정확도를 살펴보게 되는 원리라는 점이 다른 방법과 다른 점입니다. 모형의 원리를 정확하게 알고 공부하는 것이 이 수업의 목적입니다.

코드를 보면 알겠지만 코드는 누구나 돌릴 수 있고 결과도 얻습니다. 그러나 그것이 무엇을 의미하는지 모른다면 잘못된 분석을 할 가능성이 크겠지요. 이론을 아는 것이 그래서 중요합니다.

1. Knn example 파일의 첫번째 코드에서는 각 단계별로 무슨 작업을 하고 있는지 설명을 적어줍니다. (knn의 원리를 정확하게 이해시키기 위함이니 순서도를 그리듯이 정확하게 이해해서 써주세요.)

2. Knn example 의 두번째 breast cancer에 대한 결과를 iris나 titanic에 대해서도 해봅니다. 그리고 무엇을 뜻하는 코드인지도 설명을 합니다.

3. Predict 파일에는 titanic에 대한 다양한 classification 방법에 대한 결과가 담겨 있습니다. 자료를 breast cancer로 바꾸어서 결과를 도출합니다. Breast cancer가 어려운 분은 iris 의 종이 3개인데 두 개로만 분류되도록 자료를 변경하셔서 label이 2개만 있는 자료로 분석하시면 되겠습니다.

4. Data rescaling에서는 breast cancer에 대해 정규화를 하기 전과 후에 대한 분석 결과의 차이를 보여주고 있습니다. 그래서 분석을 하기 전에는 꼭 box plot을 그려서 자료가 얼마나 단위가 다르게 분포하고 있는지 살펴본 후 classification을 진행합니다. 정규화 혹은 표준화를 해서 큰 차이가 없는 경우도 있지만 지금처럼 이렇게 차이가 큰 경우도 있기 때문에 머신러닝에서 자료의 정규화 또는 표준화는 아주 중요한 부분을 차지합니다.

   3번 과제에 4번과 같이 정규화(또는 표준화)를 해야 하는지 확인하는 작업을 하시고(제가 한 것을 그대로 따라하지 마시고 그 중 필요한 부분만 확인) 정규화(또는 표준화) 이전과 이후의 결과 또한 함께 제시합니다.