DSC3011 Assignment 1: Kaggle InClass Competition

1. **My Kaggle Account:** User Name - yujinkim26
2. **How many submissions have you tried to improve the performance?**

   38 Submissions.
3. **What methods have you tried?**

   먼저 Outlier에 민감한 MinMax Scaling 등을 다루기 위해, IQR 방식을 이용해 Outlier 제거 과정을 거쳤다. 단순한 categorial 값을 갖는 컬럼을 제외한 나머지 컬럼 중, trestbps, chol, thalach, oldpeak 컬럼에서 Outlier가 확인되어 제거해주었다.

   이후 여러가지 Classification 알고리즘을 적용시키는 과정에서, GridSearchCV 함수를 이용해 상황에 맞는 파라미터 값을 찾기 위해 노력했다. KFold에서 n_split 값을 조절하면서 best score가 높은 파라미터 값을 확인했고, 결과로 나온 n, C, 혹은 gamma값 등을 각각의 알고리즘에 적용시키며 성능을 비교했다. 위와 같은 과정으로 모든 Scaling 방법과 Classification method를 조합해보며 sample을 생성하고 점수를 확인했다.
4. **Did the methods improve the performance? Why or Why not?**

   1) Outlier 제거 성능 확인: MinMax Scaling의 성능 개선 확인을 위해 outlier 제거 전과 후 데이터를 각각 똑같이 MinMax Scaling과 Naïve Bayes 알고리즘을 적용시켜보았다. 그 결과, Outlier제거 전에는 0.72857, 제거 후에는 0.74285로 성능이 개선됨을 확인할 수 있었다. 이처럼 MinMax Scaling과 Standaradization은 확실히 outlier의 영향을 받아 Outlier를 제거해주는 것이 성능 개선에 도움이 되었으나, MaxAbs scaling의 성능 개선에는 도움이 되지 않았다.

   2) GridSearchCV 함수 사용 성능 확인: 처음에는 단순히 Classification 알고리즘의 파라미터들을 default로 실행시켰다. 하지만 GridSearchCV를 이용해 찾은 파라미터 값을 이용해 학습을 시키니 더 높은 점수가 나왔다.
5. **Please explain your best solution with the highest score.**

   1) sample 5- MaxAbs Scaling + Linear Support Vector Machine

   2) sample 15- MinMax Scaling + RBF Support Vector Machine(C = 25, gamma = 0.1)

   3) sample 31- MaxAbs Scaling + RBF Support Vector Machine(C = 50, gamma = 0.025)

   세가지 solution을 살펴보았을 때, Classification은 전부 SVM이었으며 대체로 MaxAbs Scaling이 성능이 좋았다. Sample 15와 31은 GridSearchCV함수를 이용해 파라미터 값을 찾았는데, Scaling 방식에 따라 값이 상이하게 나와 위와 같이 적용시켰다.
6. **What have you learned from the competition?**

   아직 데이터 분포에 따라 적절한 Scaling 방식과 Classification 알고리즘을 선택하는 능력이 부족함을 느꼈다. 또한 수업이나 구글링을 통해 알게 된 이론들이 실제 데이터와 맞지 않는 경우가 있어서 헷갈렸다. 예를 들어 SVM같은 경우, 데이터가 가우시안 분포를 가지고 있다고 가정하고 구현되어서 Standardization이 바람직하다는 설명을 보았는데, 실제 데이터에 적용 및 학습을 시켜보니 그다지 성능이 좋지 않았다. 앞으로 있을 프로젝트에서는 체계적으로 주어진 데이터를 분석하고 전처리하며 적절한 알고리즘을 선택할 수 있도록 많은 연습을 해야겠다는 생각이 들었다.