>서문.

여러 저널이나 기사로 우리는 대학교에 대한 순위를 매기려고 한다.

대학 평가에 대한 지표를 사용하는데에 있어 우리는 어느 부분에 가중치를 두냐에 따라 순위가 바뀌고는 한다.

따라서, 여러 저널들의 평가를 바탕으로 비지도 학습 ML모델을 활용하여 대학교들의 군집을 파악하여 각 학교들의 군집을 파악해 나아갈 방향을 제시한다.

>본문.

데이터는 kaggle에서 제공하는 'World University Rankings'라는 데이터 셋을 사용했음.

각 데이터를 설명하자면

timesData.csv | 영국에서 설립된 회사에서 측정한 ‘Times Higher Education World University Ranking’이다.

ShanghaiData.csv | 중국에서 설립된 회사에서 측정한 Shanghai Ranking 이라 알려진 ‘The Academic Ranking of World Universities’ 이다.

cwurData.csv | 사우디 아라비아에서 설립된 회사에서 측정한 ‘The Center for World University Rankings’이다.

해당 데이터들을 가지고 비지도 학습인 cluster 분석을 사용하여 대학교들의 군집 분석을 진행하려 한다.

데이터 전처리 이전 데이터 자체가 고차원이라 시각화와 차원 축소를 위해 PCA기법을 사용할 예정이며, 주성분개수로 Scree plot의 elbow기법을 활용하여 차원을 축소하고 저차원 평면에 scatter plot으로 시각화를 진행할 예정이다.

해당 군집 분석을 통해 나뉘어진 군집들을 세곳에서 측정한 데이터들의 평균 랭킹을 활용하여 각 군집을 대학교의 수준에 따라 나누고, 해당 군집에 대한 분석과 타당성을 검증을 진행하였다.

>결과.

우선 ML모델 진행에 앞서 전처리 이전에 PCA방법을 적용해 scatter plot을 그려보았고, 해당 plot에서 가시적으로 5개의 군집으로 나뉘는 것을 확인해 볼 수 있었다.

이를 통해 Kmeans 분석을 해보았으나 각 주성분의 단위가 달라 거리기반의 cluster인 Kmeans는 별 소용이 없었다. 따라서 밀도 기반 군집분석인 DBSCAN을 활용해보았으나 이또한 마찬가지로 각 주성분의 단위가 달라 아래와 같이 잘 나뉘어 지지않았다.

각 주성분의 단위가 다르기 때문에 PC1의 단위를 조금 축소(1/10)해서 DBSCAN을 활용했다.

해당 분석을 해본 결과 가시적으로 5개에 근접한 군집을 형성하였고 각 군집의 평균 world\_rank를 보았으나 해당 결과에서는 유의미한 군집의 world\_rank결과가 눈에 보이지 않았다.

이후 Standard preprocessing을 진행하였고 위와 같이 PCA방법을 적용하여 Scree plot을 통해 2개의 주성분으로 적용하였다.

해당 scatter plot에서는 가시적으로 눈에 보이는 군집은 없었고, Kmeans 방법을 적용하여 군집 분석을 진행했다.

Kmeans의 최적의 군집 개수를 찾는데 intertia라고 부르는 클러스터 중심과 각 표본점들의 제곱합을 계산하는 방식의 plot을 elbow기법을 적용하여 찾았다.

또한, 실루엣 계수를 활용하여 최적 군집의 개수 k를 찾는 데 보충하여 사용했다.

위 그림과 같이 군집 개수를 2로 설정해 학습시킨 데이터의 scatter plot을 그려보았다.

각 군집의 world\_rank의 평균을 확인해보면 눈에띄게 확인이 가능했고, 각 군집의 distplot을 그려보았다.

해당 결과를 바탕으로 두 군집의 평균차이가 유의미 한지 파악하기 위해 정규성, 등분산성 검정을 실시하였다.

두집단의 정규성 검정을 위해 QQ-plot을 그려보았고, 아래 그림과 같이 어느정도 정규성을 만족한다고 가정할 수 있었다.

등분산성 검정은 두 표본이 정규성을 만족하므로 bartlett 등분산 검정을 진행하였고, 해당 결과처럼 p-value가 매우 작으므로 두 표본의 분산은 같다는 귀무가설을 기각하므로 비등분산을 한다는 결과를 확인 할 수 있었다.

등분산성이 만족하지 않으므로 대중적으로 알려진 독립표본 t-test를 사용하지 못하고, 등분산성 가정이 충족하지 못했을 때 사용하는 Games Howell’s test를 진행하였다.

GH HOC test를 통해 두 집단의 평균은 유의미한 차이를 보인다고 파악할 수 있었다.

원 데이터 전체에서 절반에 가까운 설명력을 가진 PC1을 기준으로 hici, ns, teaching column들이 높은 절댓값을 지니고 있는것으로 보아 world\_rank가 낮은 군집에 속한 대학들은 해당 부분에 대해 집중적으로 관리를 할 필요가 있다고 생각한다.