**Assignment #4**: **Data Visualization**

**2016314726 정영준**

Glass Identification dataset을 이번 Data Visualization 과제를 위해 사용할 데이터셋으로 정하였습니다. 데이터 파일을 다운받아보니 data파일 형식으로 되어있어 read\_csv를 통해 로드하였습니다. 데이터는 id를 포함한 총 11개의 column과 214개의 row로 이루어져 있었습니다. Id columns을 제거하고 마지막 class columns을 제외한 데이터를 x로 지정하였고 class column을 y로 지정하였습니다. 9개의 feature는 각 유리의 RI (refractive index), Na, Mg, Al, Si, K, Ca, Ba, Fe의 함유량을 나타냅니다. Class는 총 7개의 종류가 있었습니다. 1에서 7까지 각각 building\_windows\_float\_processed, building\_windows\_non\_float\_proce-ssed, vehicle\_windows\_float\_processed, vehicle\_win-dows\_non\_float\_processed, containers, tableware, headlamps를 나타냅니다. 4에 해당하는 vehicle\_windows\_non\_float\_processed는 주어진 데이터셋에 존재하지 않았습니다. 데이터 시각화를 하는 것이 목표이기 때문에 train\_test\_split을 이용한 train set과 test set으로 데이터를 분리하지 않았습니다. 먼저 PCA를 통해 차원축소를 하였습니다. Sklearn.decomposition의 PCA를 사용하였으며 시각화를 하기 위해 feature가 2개인 데이터를 얻기 위해 n\_components를 2로 설정하였습니다. fit과 transform을 사용하여 label 데이터를 사용하지 않은 비지도 학습으로 x를 학습시켰습니다. Seaborn의 scatterplot을 사용하여 변환된 feature가 2개인 데이터를 시각화 하였습니다. 다음으로 StandardScaler를 적용하고 같은 과정을 한 번 더 시행하고 이를 matplotlib의 subplots을 사용하여 나란히 2개의 scatterplot을 비교하였습니다. Label의 종류가 총 6개인 데이터고 완벽하게 class별로 영역의 분리가 이루어지지 않았습니다. Class 7의 경우 scaler를 사용과 상관없이 모두 다른 class의 데이터 들과 거리를 두고 군집을 형성하고 있었습니다. Class 1과 Class 3은 비슷한 분포를 보였으며 Class 2는 전체적으로 퍼져 있는 형태였습니다. 다음으로 sklearn.manifold의 TSNE를 사용하여 시각화를 하였습니다. PCA에 비해 시간이 오래걸렸지만 PCA에서 불가능한 비선형적 분포에 대한 차원 축소 덕분에 PCA보다 더 class간 구분이 확실해졌습니다. Class 1과 Class 3의 분포가 겹치고 Class 2가 전체적으로 퍼져있는 것은 동일하였으나 나머지 Class의 구분이 어느정도 가능해졌습니다. StandardSclaer가 데이터 시각화에 큰 도움을 주지는 못하였습니다. 시각화 결과로부터 9개의 feature를 가진 고차원 데이터를 2차원으로 mapping하여 각 Class에 대한 대략적인 분포를 파악할 수 있었습니다. 시각화를 통해 해석한 결과는 Float processed가 이루어진 유리의 분포가 비슷하고 headlamps의 경우 다른 유리와 다른 점이 많다는 것입니다.











