**PCA\_1210\_데이터마이닝 팀 프로젝트**

**#1 성능 비교**: 훈련 데이터 70%로 했을 때

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명 PCA 하기 전 성능

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명 PCA 한 후 성능(주성분 2개)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명 (주성분 16개)

**#2 PCA 실행**

기울기가 급격하게 낮아지는 부분의 수치로 주성분의 개수를 정한다 = 2

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명 Scree plot

위에 따르면 주성분 2개지만, 설명력 99%일 때의 주성분을 보겠음

붉은색 선분은 주성분의 각 개수 마다의 변동의 비율값이고, 초록색 선분은 주성분의 개수가 증가할 때의 변동의 누적 비율이다. X 축은 주성분들의 개수이고 y축은 변동의 비율 값이다.  이 도형에서 y축은 계산된 PCA의 값이 갖는 원 데이터세트에 대한 설명력이라고 보면 되며 값이 0.8, 즉 80% 이상의 설명력을 가진 때의 x 축 값을 보면 된다. 위의 그림에서는 PC(주성분)의 개수가 1일 때 0.3(충분한 설명력 보장X)이고 주성분의 개수가 16일 때 99%로 거의 100%의 설명력을 가짐을 알 수 있다.

텍스트, 스크린샷, 번호, 디스플레이이(가) 표시된 사진

자동 생성된 설명 19개 feature에서 16개의 PC 생성

여기서는 PC1 즉 첫번째 주성분이 분산을 0.31을 보존했으므로 설명력이 31%라고 볼 수 있다. PC2는 두번째 주성분으로써 분산이 0.13 정도로써 PC1보다는 설명력이 높지 않음을 알 수 있다.

텍스트, 스크린샷, 소프트웨어, 번호이(가) 표시된 사진

자동 생성된 설명 informative projections 순위

텍스트, 스크린샷, 도표, 지도이(가) 표시된 사진

자동 생성된 설명 PC1과 PC3을 x,y축

IF 주성분의 개수를 2개로 하였을 때

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 소프트웨어이(가) 표시된 사진

자동 생성된 설명 PC1, PC2 설명력 변화 없음

텍스트, 스크린샷, 지도이(가) 표시된 사진

자동 생성된 설명 t-SNE

t-distributed stochastic neighbor embedding: 높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방법, 낮은 차원 공간의 시각화에 주로 사용하며 차원 축소할 때는 비슷한 구조끼리 데이터를 정리한 상태이므로 데이터 구조를 이해하는 데 도움을 준다. 매니폴드(Manifold) 학습의 하나로 복잡한 데이터의 시각화가 목적이다. 높은 차원의 데이터를 2차원 또는 3차원으로 축소시켜 시각화 합니다. 높은 차원 공간에서 비슷한 데이터 구조는 낮은 차원 공간에서 가깝게 대응하며, 비슷하지 않은 데이터 구조는 멀리 떨어져 대응된다.