다변량및시계열데이터분석 – 다변량 과제

AI융합학부 2022113170 장정의

**1. wine dataset** : 와인 데이터셋을 선정한 이유는, 13개의 화학적 특성들이 3개의 와인의 종류에 따라 어떻게 다른지 분석하는 데 적합하기 때문이다. 각 특성이 와인의 종류에 미치는 영향을 동시에 고려할 수 있기에, 여러 종속 변수를 동시에 분석할 수 있는 다변량 모델이 필요하다. 이에 적합한 모델로는 MANOVA를 선택했으며, 와인의 화학적 특성들이 와인 종류에 미치는 종합적인 영향을 분석하였다. 분석 결과, 모든 테스트에서 p-value가 유의수준 0.05보다 훨씬 낮으므로 와인의 종류와 절편이 다변량 종속 변수들에 영향을 미치지 않는다는 귀무 가설을 기각할 수 있다. 따라서 와인의 종류와 절편이 화학적 특성들에 실제로 유의미한 영향을 미친다는 결론을 내릴 수 있다. pairplot을 통해서도 클래스간 군집과 변수간 관계를 알 수 있다.

**2. breast cancer dataset** : 유방암 데이터셋은 다변량 데이터를 포함하며 여러 특성을 동시에 분석하여 보다 정확한 분석이 가능하기에 선정하였다. 다변량 모델이 필요한 이유는 데이터셋의 여러 특징들이 상호 작용하며 종양의 악성과 양성을 구별하는데 중요한 역할을 하기 때문이다. 위 다변량 데이터를 차원을 축소하고 주요 정보를 유지하면서 패턴을 파악하기 위해 PCA를 사용하였다. 첫 두 개의 주성분으로 데이터를 투영함으로써 악성과 양성 종양 사이의 명확한 분리를 관찰할 수 있었다. PC1과 PC2에 의해 설명되는 분산은 약 63%이며, 누적 설명 분산 그래프를 통해서도 몇 개의 주성분만을 가지고 거의 모든 분산을 설명할 수 있음을 알 수 있다. 이는 차원 축소의 효율성을 뒷받침한다.

**3. Shoulder Implant X-Ray Manufacturer Classification dataset** : 위 데이터셋은 어깨 X-ray 이미지를 포함하고 있으며, 4개의 제조업체에서 이식된 어깨보철물을 나타낸다. 이미지 데이터에서 여러 특징을 추출해 다변량 분석을 통해 제조업체 분류가 가능하다. 이미지 픽셀 정보가 고차원적이며, 특징이 상호작용하므로 다변량 모델이 필요하다. 따라서 PCA 활용이 가능한데, 누적 설명 분산 그래프를 통해 약 100개의 주성분을 사용하면 데이터 대부분의 분산을 설명할 수 있음을 알 수 있다. PCA는 노이즈가 많은 데이터에 필터링 기법으로도 사용이 가능한데, 직접 노이즈를 추가한 후 주성분 100개를 사용해 노이즈가 제거된 이미지를 재구성한 결과, 원본 이미지와 유사한 모습을 확인할 수 있다. 즉, 신호는 유지하고 노이즈는 제거한다.

**4. Wholesale customers dataset** : 위 데이터셋은 변수가 다양하고 연속형이며, 변수간 뚜렷한 잠재적 관계가 존재할 것이라 판단하여 선택하게 되었다. 여러 변수 간 상호작용을 고려하기에 다변량 모델을 사용해야 하며, 그 중 숨겨진 관계를 찾기 위해 요인분석을 사용하였다. 요인 적재치 출력을 통해 요인 0은 유제품, 식료품, 세제 및 종이제품에 대한 소비와 강한 관련이 있고, 요인1은 냉동 식품, 조제 식품에 대한 소비와 강한 관련이 있으며, 요인2는 신선 식품과 조제 식품에 대한 소비에 어느 정도 관련이 있음을 알 수 있었다. 또한, 공통성 그래프를 통해, 식료품과 세제 및 종이제품에 대한 소비는 요인 구조에 적합한 반면 냉동 식품은 그렇지 않다는 결과를 얻었으며 각 변수의 소비 패턴이 요인들에 의해 얼마나 설명되는지를 알 수 있다.