과제 참고 자료

데이터 소개

과제에 사용된 데이터는 성인 소득수준 데이터셋으로 알려진 데이터셋입니다. 원 데이터는 14개의 수치형/범주형 특징을 가진 48,842개의 소득 데이터로 구성되어 있으며 데이터는 50K 소득을 기준으로 상위 및 하위의 2개 유형으로 구분됩니다. 원 데이터와 기초 분석 내용은 https://www.kaggle.com/datasets/wenruliu/adult-income-dataset에서 확인하실 수 있습니다.

과제에서는 학생들이 데이터셋을 추측할 수 없도록 원 데이터를 변형하여 사용합니다. 변형 내용은 다음과 같습니다.

- 미국 태생 백인 데이터만 사용 (48,842명 중 38,493명 사용, 78.8%)
- 무직자와 무급 노동자 데이터 폐기 (38,493명 중 36,328명 사용, 94.4%)
- 기혼자와 미혼자, 이혼자 데이터만 사용 (36,328명 중 34,164명 사용, 94.0%)
- 남성 데이터만 사용 (34,164명 중 24,495명 사용, 71.7%)
- 6개 특징만 사용 (Age, Workclass, Fnlwgt, Educational-num, Marital-status, Hours-per-week)
- 범주형 특징들에 대해 정수값 부여 (Workclass, Marital-status)
- 24,495개의 데이터 중 20,000개 데이터 무작위 선별
    - ■ 남은 4,495개 데이터로 모델 성능 평가 예정
- 데이터 레코드 및 컬럼 순서 셔플
- 컬럼명 폐기
- 임금에 대한 언급을 만족도로 대체

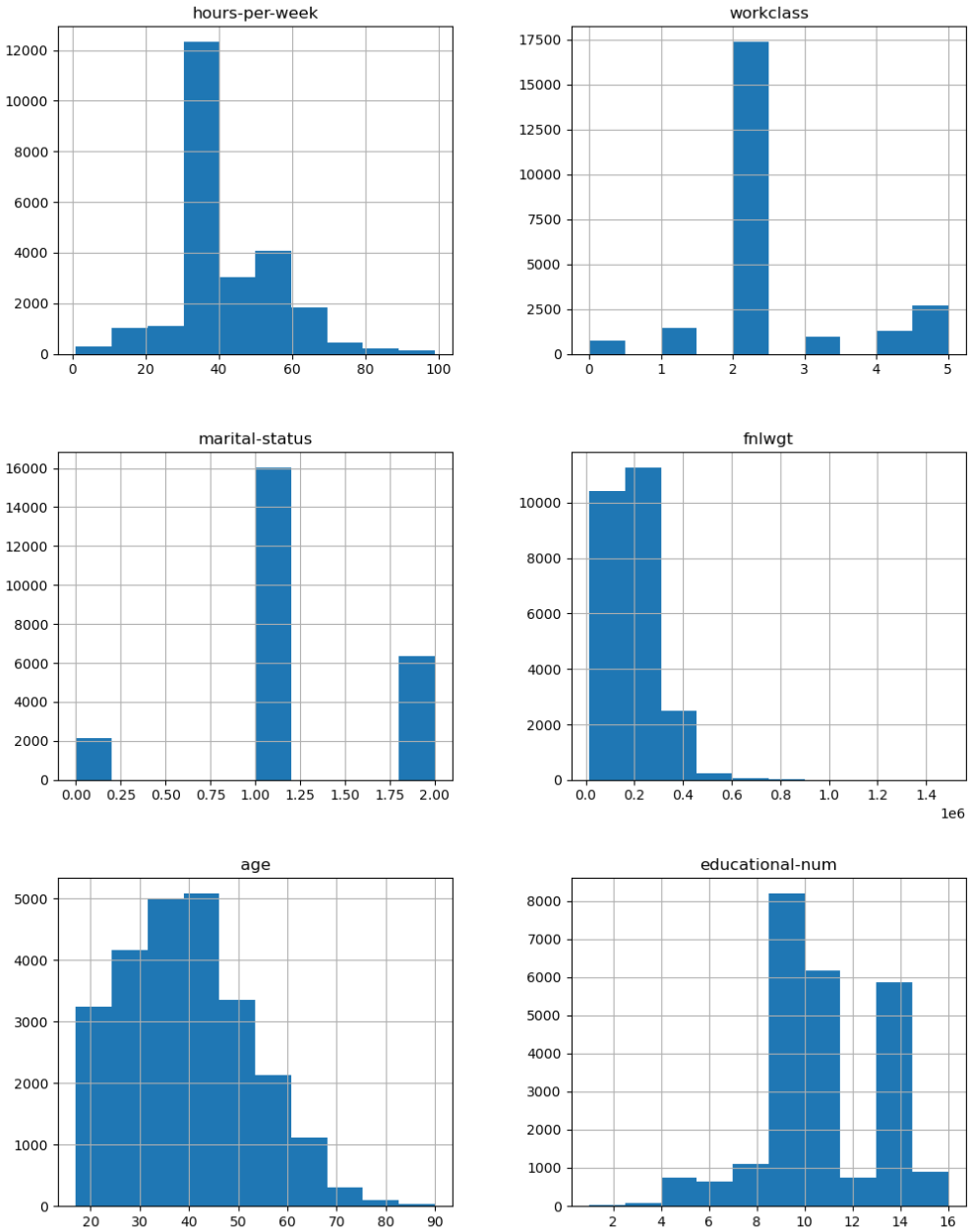
데이터 상세

분류	개수	비율
50K 이하 소득	24,637	72.1%
50K 초과 소득	9,527	27.9%

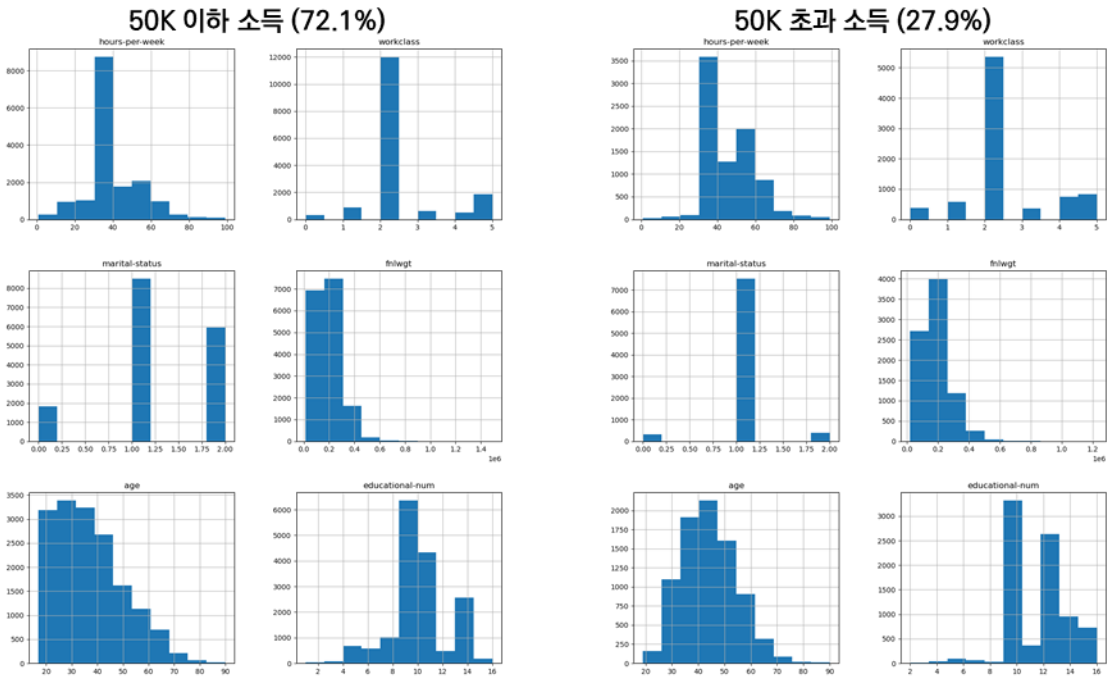
[데이터 비율]

특징명	활용여부	설명
Age	활용	나이
Workclass	활용	근무지유형 (민간, 자영업, 공공, 무급, 무직 등)
Fnlwgt	활용	통계청에서 지정한 가중치 (비슷할수록 비슷한 소득정보)
Education	비활용	교육수준
Education-num	활용	교육기간
Marital-status	활용	결혼상태
Occupation	비활용	직업
Relationship	비활용	가족구성 (남편, 아내, 편부모, 미혼 등)
Race	비활용	인종
Gender	비활용	성별
Capital-gain	비활용	자본 증가액
Capital-loss	비활용	자본 감소액
Hours-per-week	활용	주당 근무시간
Native-country	비활용	출생국가

[데이터 컬럼 설명]

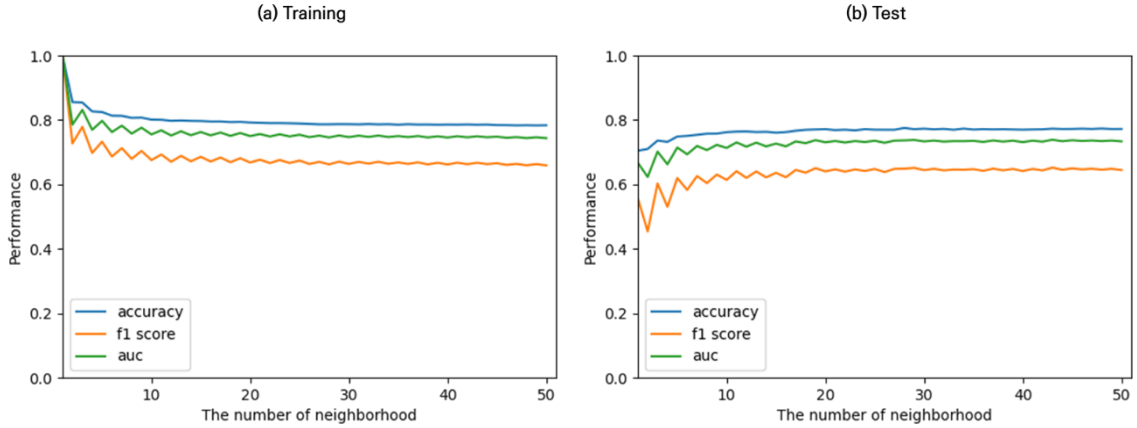


[전체 클래스 컬럼별 데이터 분포]



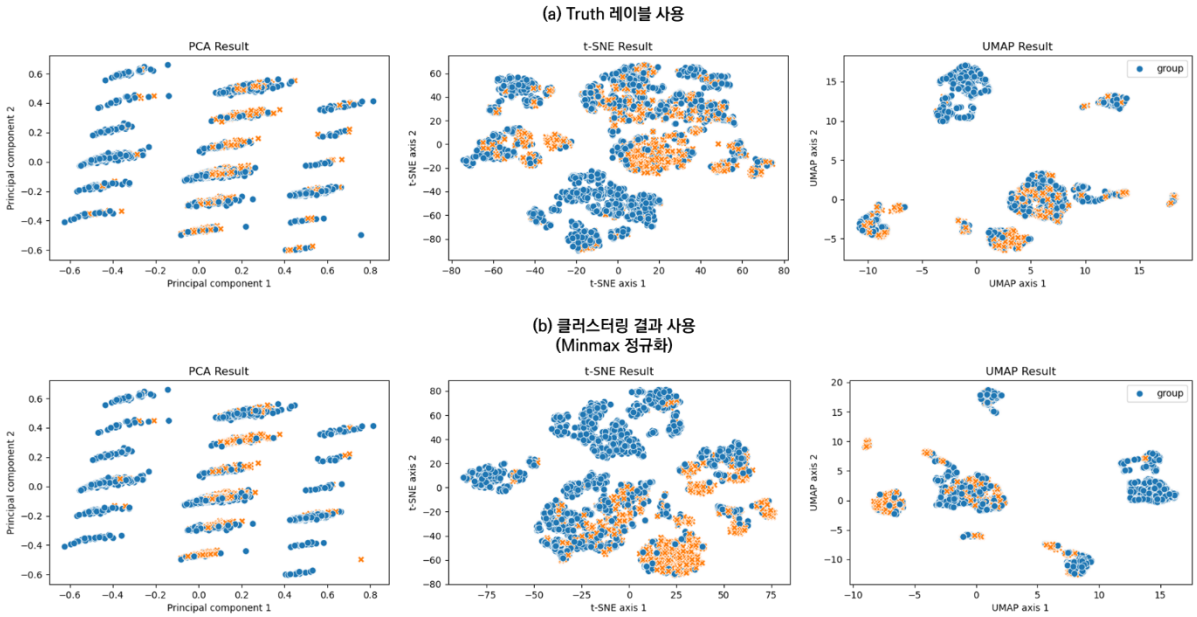
[클래스별 컬럼별 데이터 분포]

분석 결과



[평가 결과]

*** k=5정도에서 0.7정도의 AUC 성능을 보이며 유클리드 거리와 해밍 거리를 조합해 사용하면 결과가 더 좋아질 것으로 기대 (현재 분석에서는 범주형 데이터도 유클리드 거리로 비교하였음)**



[원 데이터 및 클러스터링 결과 시각화 자료]

*** 2차원에서 분류하기 어려운 데이터 분포일 것으로 추정**

PCA: Zero-centered 데이터의 공분산 행렬로부터 분산이 가장 큰 축을 순서대로 찾아 원 데이터를 각 축으로 사영하는 기법

t-SNE: t분포와 KL-divergence를 바탕으로 고차원 데이터 분포와 유사한 저차원 데이터 분포를 형성하고 해당 저차원에서 시각화하는 기법

UMAP: 데이터의 주요 성분이 리만 다양체 위에 놓여있음을 가정하고 데이터를 리만 다양체로 사영하는 기법