1. Deepfake Detection Challenge
https://www.kaggle.com/c/deepfake-statml2021s-smwu/data

인간의 안면을 실제로 촬영한 사진과 인공지능이 생성한 안면 사진이 혼합되어 있는 데이터셋. 실제 사진은 Nvidia가 모은 Flickr dataset에서 추출되었으며 가짜 사진은 Bojan이 StyleGAN을 이용하여 생성한 데이터임.

Attributes:
- 사진 파일
- 진짜/조작 여부 (binary classification 레이블)

Instances:
- fake image 10,000장
- real image 10,000장

모델 시나리오
Classification:
이 데이터셋은 진짜 사진과 인공지능이 생성한 가짜 사진을 구분하는 이진 분류 문제이다. CNN(Convolutional Neural Network) 기반의 모델을 사용하여 사진을 입력받고, 진짜(real)인지 가짜(fake)인지 분류할 수 있다.
- 데이터 전처리: 각 사진 파일을 동일한 해상도로 리사이징하고, 이미지 정규화를 수행하여 학습에 적합하게 만든다.
- 모델 구조: CNN 기반의 ResNet이나 EfficientNet 같은 대규모 이미지 분류 모델을 사용할 수 있으며, 데이터셋 크기에 맞춰 여러 층의 합성곱 층과 풀링 층을 결합해 특징 추출을 수행한다.
- 전이 학습: 사전 학습된 모델을 사용하여 빠른 학습을 구현하고, Deepfake 특성에 맞게 파인튜닝한다.
- 평가 방법: Accuracy, Precision, Recall, F1-score 등을 이용하여 모델 성능을 평가할 수 있다. 특히, false positive(진짜 사진을 가짜로 분류)와 false negative(가짜 사진을 진짜로 분류)에 민감한 응용 환경에서 이들을 구분하는 것이 중요할 것이다.

Numerical Prediction:
사진이 얼마나 조작되었는지 그 정도를 수치적으로 예측하는 모델을 설계할 수도 있다.
- 모델 개념: 각 사진의 조작된 픽셀 비율을 수치로 나타내고, 이 수치를 예측하는 회귀 모델을 만들 수 있다. 예를 들어, 0(완전히 진짜)에서 1(완전히 가짜) 사이의 값을 예측하는 방식이다.
- 활용 가능성: 완전히 가짜인 사진뿐 아니라, 일부만 조작된 사진의 경우에도 이를 구별할 수 있게 되어, 조작 정도에 따라 가짜를 감지하는 세부적인 평가가 가능할 것이다.

2. Shop Customer Segmentation
https://www.kaggle.com/code/utkarshsaxenadn/shop-customer-clustering

Attributes:
이 데이터셋은 고객의 인구 통계 및 소비 습관에 관한 데이터를 포함하고 있다. 주요 속성(attribute)들은 다음과 같다:
- 고객ID
- 성별
- 나이
- 연간 소득 ($)
- 지출 점수(1-100)
- 직업
- 근무 경험
- 가족 크기

Instances:
- 2000개의 인스턴스로 구성되어 있으며, 각 인스턴스는 한 명의 고객에 대한 정보를 나타냄

모델 시나리오
Clustering:
이 모델은 고객들의 소비 성향을 바탕으로 비슷한 성향을 가진 그룹을 나누는 군집화 모델이다.
- 데이터 전처리: 성별과 같은 범주형 변수는 원-핫 인코딩을 하고, 소득 및 나이와 같은 수치형 변수는 스케일링을 통해 정규화하여 모델에 입력한다.
모델 선택: K-means 클러스터링을 사용할 경우, 군집 수(k)를 선택하는데 엘보우 방법을 사용할 수 있다. 고객 데이터를 여러 군집으로 나누어 비슷한 성향을 가진 고객들을 그룹화한다. 또한, Hierarchical Clustering(계층적 군집화) 방식으로 고객들 간의 관계를 트리 구조로 표현할 수 있다.
- 군집 활용: 각 군집에 속한 고객들의 특성을 분석하여, 특정 군집(예: 고소득, 높은 지출 성향을 가진 고객층)에 대한 맞춤형 마케팅 캠페인을 계획할 수 있다. 이를 통해 기업의 마케팅 자원을 더 효율적으로 배분하고, 타깃 마케팅을 진행할 수 있다.
Association Rule Mining:
고객들의 성별, 나이, 연소득, 지출 점수 등의 속성을 바탕으로 연관 규칙을 찾는 모델이다. 이를 통해 소비 습관과 인구통계학적 특성 사이의 관계를 발견할 수 있다.
- Apriori 알고리즘: Apriori 또는 FP-Growth 알고리즘을 사용하여 고객 데이터에서 연관 규칙을 찾을 수 있다. 예를 들어, "연간 소득이 $100,000 이상이고, 나이가 40대 이상인 고객은 지출 점수가 80 이상일 가능성이 높다"라는 규칙을 도출할 수 있다.
- 응용 시나리오: 이러한 규칙을 바탕으로 특정 고객 세그먼트에 대한 맞춤형 제품 추천 또는 프로모션을 진행할 수 있다. 또한, 고객의 성향을 파악하여 잠재 고객의 구매 가능성을 예측하는 데에도 활용할 수 있다.

참고문헌: OpenAI. (2024). ChatGPT (Sep 18 GPT-4-o version) [Large language model]. https://chat.openai.com.