<YDMS 2022-1> 1주차 과제

Subject : 데이터 마이닝 프로세스 (Data Minig Process)

주하연

1. 데이터마이닝 프로세스란?

데이터마이닝이란 인공 지능, 기계 학습, 통계 및 데이터 베이스 시스템의 교차점에 있는 방법을 포함하는 대규모 data set에서 patterns, insights, knowledge, valuable information을 찾아내는 프로세스이다. 여기서 찾아내는 정보는 사소하지 않고 암시적, 이전에 알려지지 않았거나 잠재적으로 유용한 것들이다. 데이터마이닝은 다양한 분야에서 더 깊이 이해하고 진행하면서 얻은 통찰력을 기반으로 우리가 찾아낸 정보를 기반으로 한 의사결정을 내리기 위해 사용된다.

그리고 데이터마이닝의 목표는 비즈니스의 성과를 개선하고 효율성을 높이며 비용을 절감하는 데 사용할 수 있는 패턴과 통찰력을 발견하는 것이다. 기업은 기업 운영에 도움이 되는 정보를 알아낼 수 있다. (예를 들어, 수익성이 높은 고객 식별, 공급망을 적절하게 설정 등) 따라서 데이터마이닝은 기업 운영에 있어서 아주 중요한 부분 중 하나이다. 이 과정은 기업과 조직이 경쟁 우위를 확보하고 데이터에서 얻은 통찰력을 기반으로 우리가 찾은 정보를 기반으로 한 의사 결정을 내리는 데에 필수적인 과정이다. 또한 기업이 앞으로 어떻게 나아가야 할지 미래의 추세에 대해 예측할 수도 있다.

데이터 마이닝은 데이터 수집, 처리 및 분석을 위한 데이터 과학 방법론인 KDD(knowledge Discovery in Database) 프로세스의 한 단계이다.

데이터 마이닝 프로세스는 일반적으로 여러 단계로 이루어져 있다.

- Business understanding

데이터마이닝 프로젝트의 문제 및 목표를 정의한다. (ex. 회사의 요구사항 : 수익개선)
여기서 데이터를 분석하는 데이터사이언티스트와 회사관계자는 주어진 프로젝트에 대한 문제와 목표를 정의하기 위해 협력해야 한다.

- Data understanding

문제를 해결하는 데 우리는 어떤 데이터가 필요한지 알아간다. 데이터를 수집하고 탐색하여 데이터의 속성과 특성을 이해한다. 다양한 source에서 데이터를 수집하고 분석을 위해 적절한 형식으로 저장한다.

- Data preparation

데이터를 정리, 변환 및 전처리하여 분석할 수 있도록 준비한다.
이 단계가 제대로 되어야 분석에 더 도움이 됨.

■ 데이터 탐색 (EDA) **

이 단계에서는 데이터의 특성과 관계를 더 깊이 이해하기 위해 데이터를 시각화하고 분석한다. 탐색적 데이터 분석(EDA) 기법을 사용하여 데이터의 패턴, 추세 또는 특이치를 식별할 수 있다.

■ 데이터 품질 관리

■ 데이터 변환

데이터 분석에 적합한 형식으로 변환한다. ex) 수치 데이터를 표준화하거나 정규화

■관련이 없거나 중복된 데이터 제거, 결측값 처리 및 불일치 수정

■크기 조정 및 차원 축소 (데이터 세트에 따라 피처가 너무 많으면 계산속도 저하 우려)

--> 모델 내에서 최적의 정확도를 보장하기 위해 가장 중요한 예측 변수를 유지

- Modeling(데이터마이닝)

데이터의 패턴과 관계를 알아내는 모델을 구축하기 위해 적합한 데이터 마이닝 기술과 알고리즘을 선택하고 적용한다. 문제의 특성에 따라 분류, 회귀, 클러스터링 또는 연결 규칙 마이닝과 같은 다양한 기법이 사용될 수 있다. 여기서 데이터 사이언티스트는 순차적 패턴, 연관 규칙, 상관 관계와 같은 흥미로운 데이터 관계를 조사할 수 있다.

- Evaluation

정확도, 정밀도, 리콜 또는 F1 score을 사용하여 발견된 패턴 및 모델의 품질과 성능, 유용성을 평가한다. Cross-validation과 holdout validation 기법을 사용하여 모델의 일반화 능력을 테스트할 수 있다.

- Deployment

모델을 사용하여 새 데이터를 기반으로 예측이나 결정을 내리는 과정을 수행한다. 이 모델은 응용되어 다른 더 큰 프로그램에도 통합될 수 있다. 이를 통해 비즈니스 문제를 해결한다.

2. 지도학습 & 비지도학습 개념 및 사용 기법 조사하기 - (타켓변수란?)

머신러닝 문제는 두 종류가 있다.

1. 타겟 변수가 하나 또는 여러 개가 존재하는 지도 학습
2. 타켓 변수가 존재하지 않는 비지도 학습

이 둘은 각각 다른 학습 방법을 사용하고, 다른 유형의 데이터에 적용된다. 그리고 타겟변수 사용에 차이가 있다.

- 타겟변수 : 타겟 변수는 지도학습에서 학습을 진행할 때 예측하려는 대상 변수이다. 즉, 모델이 학습할 때 입력 데이터와 함께 학습되는 라벨(Label)이나 정답(Target)으로 사용되는 변수이다. 타겟 변수는 모델의 예측 성능과 정확도를 결정하는 중요한 역할을 한다. 모델은 입력 데이터와 타겟 변수 사이의 관계를 학습하고, 이를 바탕으로 새로운 입력 데이터에 대한 예측값을 출력한다. 따라서 타겟 변수는 학습에 사용되는 데이터와 함께 주어져야 하며, 모델이 학습하는 동안 정확한 라벨이 제공되어야 한다.

● **지도학습** : 정답을 학습시킨다. = 타겟변수가 있는 것

데이터는 각각의 입력 데이터와 그것에 대응하는 정답(label) 쌍으로 이루어져 있다. 모델은 이러한 레이블 된 데이터를 사용하여 입력 데이터와 출력 데이터 간의 관계를 학습하고, 새로운 입력 데이터에 대한 출력 데이터를 예측할 수 있다. 지도 학습의 목표는 입력 기능과 대상 변수 사이의 관계를 학습하여 모델이 보이지 않는 새로운 예에 대한 예측을 할 수 있도록 하는 것이다.

- 사용기법

■분류(Classification)

범주 값을 예측한다.

입력 데이터와 이산적인 출력 값(카테고리, 클래스) 간의 관계를 학습하는 작업이다.

이 값은 예/아니오 로 답이 나올 수도 있고, 여러 가지 값이 될 수도 있다.
ex) 사기, 이탈, 재방문, 재구매, 부실

■ 회귀(Regression)

수치 값을 예측한다. 즉, 양적 데이터
입력 데이터와 연속적인 출력 값(숫자) 간의 관계를 학습하는 작업이다. (선형 회귀, 다항 회귀, 결정 트리 회귀 등)
ex) 구매 금액, 구매 빈도 등

레이블된 데이터가 많은 분야에서 사용된다. (음성 인식, 이미지 인식, 자연어 처리 등)

● **비지도학습** : 정답을 알려주지 않는다. = 맞춰야하는 타겟변수가 없는 것

입력 데이터만 제공되며, 모델은 입력 데이터 간의 패턴, 관계, 혹은 구조를 학습하게 된다.
- 사용기법(기법들은 데이터의 특성과 목적에 따라 선택되어 적용되어야 함)

■ 군집화(클러스터링)

기준과 정답이 없는 데이터들을 분석하여 정답이 없는 데이터들을 분석하여 유사한 데이터들끼리 그룹을 묶어주는 작업이다. = 서로 비슷한 특징을 가진 그룹으로 나누는 작업(K-means, DBSCAN 등)

■ 차원 축소

고차원의 입력 데이터를 저차원의 데이터로 변환하는 작업이다. 이를 통해 입력 데이터의 특징을 보존하면서도 계산 복잡도를 줄일 수 있다. (PCA, t-SNE 등)

■ 이상치 탐지

정상적인 데이터와 다른 패턴을 가지고 있는 이상치 데이터를 찾는 작업이다. 이상치 탐지는 보안, 금융 등 다양한 분야에서 사용된다. (One-class SVM, Isolation Forest 등)

■ 연관 규칙 학습

력 데이터에서 상호 연관성이 높은 데이터들을 찾아내는 작업이다. 이를 통해 마케팅, 추천 시스템 등 다양한 분야에서 사용된다. (Apriori, FP-Growth 등)

■ 생성 모델

입력 데이터와 비슷한 새로운 데이터를 생성하는 작업이다. 생성 모델은 음성, 이미지, 자연어 등 다양한 분야에서 사용된다. (GAN, VAE 등)

이러한 기법들은 데이터를 자동으로 분류하거나, 데이터의 특징을 추출하는 작업에서 유용하다. 레이블이 없거나 적은 경우에 사용된다. (사진 분류, 신용카드 부정거래 탐지 등)

3. 데이터 셋을 TRAIN, VALIDATION, TEST로 나누는 이유에 대하여 조사 (데이터 분할)

데이터 분할은 머신 러닝 모델을 구축하고 평가하는 데 필수적인 단계이다. 왜 데이터 분할을 해야 하는지 이유를 조사해 보았다.

● TRAIN, VALIDATION, TEST DATA

■ TRAIN DATA

학습을 하기 위한 용도 (트레이닝 과정에서 사용)

데이터셋의 70프로 정도를 할당함.

- VALIDATION DATA

학습을 하는 과정에서 중간평가를 하기 위한 용도이다. TRAIN DATA에서 일부를 떼내서 가져온다. (트레이닝 과정에서 사용)
모델 설계를 지속적으로 반복하여 VALIDATION DATASET에서 모델의 성능을 높일 수 있다.
데이터셋의 20프로 정도를 할당함.

- TEST DATA

훈련한 모델을 한번도 보지 못한 데이터를 이용해서 평가를 하기 위한 용도이다. (트레이닝 과정이 끝난 후 성능을 측정하기 위해 사용)
데이터셋의 10프로 정도를 할당함.

● WHY DO WE NEED TO SPLIT?

- 모델 성능을 측정

데이터 분할은 모델의 성능을 평가하는 데 필요하다. 데이터 분할을 하여 각각의 데이터셋을 사용해 모델을 학습시키고 검증하며, 모델의 성능을 측정한다. 기계 학습 모델을 교육할 때, 단일 데이터 세트에서 교육하면 모델의 성능을 평가할 수 없다.

- 일반화 성능을 평가

모델이 학습용 데이터셋에 대해 잘 동작하더라도, 실제로 새로운 데이터에 대해서도 잘 동작하는지는 보장할 수 없다. 따라서, 모델의 일반화 성능을 평가하기 위해서는 TARIN 데이터셋과 다른 데이터셋을 사용해야 한다. 따라서, TEST 데이터셋을 사용하여 모델의 일반화 성능을 평가합니다.
--> 여기서 모델이 TEST 데이터셋에 지나치게 적합되지 않았는지 확인 할 수 있다. (과적합 문제 방지)

- 데이터 분석에 용이

데이터 분할은 데이터셋을 살펴보고 분석하는 데 도움이 된다. 예를 들어, TRAIN 데이터셋에서 레이블(label) 분포를 확인하여 클래스 불균형(class imbalance) 문제가 발생하는지 확인할 수 있다.

- 모델의 하이퍼파라미터 조정

모델의 성능을 최적화하기 위해 하이퍼파라미터를 조정해야 하는데, 이 때 VALIDATION 데이터셋을 사용한다. VALIDATION 데이터셋에서 모델의 성능을 평가하면서 하이퍼파라미터를 조정하여 모델의 성능을 개선할 수 있다.

4. 과적합(과대적합)에 대하여 조사하기

과적합(overfitting)은 TRAIN 데이터셋에 대해서는 모델의 성능이 좋지만, 새로운 데이터에 대해서는 성능이 나쁘게 나타나는 문제이다. 이는 모델이 TRAIN 데이터셋에 지나치게 적합되어 새로운 데이터에 대해서는 일반화되지 못하기 때문이다. = TRAIN DATA를 과하게 학습
-> 예측의 정확도에 영향을 미침

- 과적합이 발생되는 경우
    - 모델의 분산이 높을 때
    - 잘 정리되지 않은 데이터를 사용할 때
    - TRAIN DATA의 크기가 충분하지 않아 제한된 데이터를 학습할 때
    - 모델에는 신경망이 쌓여 있는데, 신경망이 복잡하고 훈련하는데 상당한 시간이 필요할 때

    + 3번에서 언급하였던 것처럼 데이터를 제대로 일반화하지 않으면 과적합문제가 있을 수 있다.

- 과적합을 방지/해결하는 방법
    - 3번에서 알아보았던 데이터 분할에서 일반화 성능을 평가
    - 더 많은 데이터 수집 : 모델이 일반화하는데 도움이 될 만한 데이터를 추가하여 모델이 더 일반적인 패턴을 학습하도록 합니다. 이는 알고리즘이 신호를 더 잘 감지하게 하여 오류를 최소화할 수 있다.
    - 데이터 증강 : 더 많은 데이터 수집이 어려운 경우 이 방법을 사용할 수 있다. 이는 샘플 데이터가 모델에서 처리될 때마다 약간 다르게 보이게 한다. 각 데이터 세트가 모델에 고유하게 나타나도록 하고 모델이 데이터 세트의 특성을 학습하지 못하도록 한다. 그리고 비교적 비용이 저렴하다.
    - 모델의 복잡도 감소(단순화) : 모델의 파라미터 수, 매개변수 수 등을 줄여 모델이 간단한 패턴을 학습할 수 있도록 한다.
    - 규제화(Regularization) : 가중치 감소(weight decay)나 드롭아웃(dropout) 등의 규제화 방법을 적용하여 모델이 학습 데이터에 너무 맞추지 않도록 한다.
    - 조기 종료(Early Stopping) : 학습 데이터와 검증 데이터의 손실(loss) 값을 모니터링하여 검증 데이터의 손실 값이 증가하기 시작할 때 학습을 중단한다.

        그러나, 너무 빨리 학습을 중지하면 과소적합이 발생할 위험이 있다.
    - 앙상블 : 여러 모델의 예측 결과를 결합하여 더 좋은 성능을 얻는 방법이다.

        - 부스팅 : 간단한 기본 모델을 사용하여 시퀀스의 각 학습자가 이전 학습자의 실수로부터 학습하도록 한다. 약간 학습자를 결합 -> 강한 학습자
        - 배깅 : 병렬 패턴으로 배열된 많은 수의 강력한 학습자를 훈련한 다음 결합하여 예측을 최적화하는 방식

과적합은 머신러닝에서 오차를 증가시키는 원인이다. BUT, 실제로는 해결이 불가능한 수준인 경우가 대부분이다.

- ㄴ WHY? (실제로는 왜 해결이 어려운가)
    - 일반적으로 학습 데이터는 실제 데이터의 부분집합이며, 실제 데이터를 모두 수집하는 것은 불가능하다.
    - 만약 실제 데이터를 모두 수집하여도 모든 데이터를 학습시키기 위한 시간이 측정 불가능한 수준으로 증가할 수 있다.
    - 학습 데이터만 가지고 실제 데이터의 오차가 증가하는 지점을 예측하는 것은 매우 어렵거나 불가능하다.

5. 변수 TYPE에 대해 정리하기 (Ex. 범주형, 수치형 / 명목형, 순서형, 연속형, 이산형 등)

데이터 분석에서 변수 TYPE을 파악하는 것은 분석의 첫 단계 중 하나이고 적절한 분석 방법을 선택하는 데에 중요한 역할을 한다. 이는 결과를 해석하는 방법이 달라지게 한다.

- 범주형 : 질적자료 = 몇 개의 범주로 나누어진 자료, 수치로 측정 불가
    - 명목형

    범주 간에는 동등한지 여부만을 나타낸다.

    ex) 성별, 혈액형, mbti, 종교
    - 순서형

    각 범주 간에 상대적인 순서를 가진다.

    ex) 학점, 만족도, 순위, 직급, 온도

- 범주형 단독 : 파이차트, 막대그래프
- 범주형 자료의 표현 : 도수 분포표, 막대그래프로 나타낼 수 있다.
    - 막대 그래프 – 각 카테고리 값에 대한 빈도수 또는 비율을 막대 형태로 나타냄
    - 원 그래프 – 각 카테고리 값의 비율을 부채꼴 모양으로 나타냄
    - 히트맵 – 카테고리 값 간의 관계를 색상으로 나타냄. 두 개의 카테고리 값 간의 상관 관계를 표시할 때 사용
    - 테이블 – 각 카테고리 값에 대한 빈도수 또는 비율을 표로 나타냄
    - 도넛 차트 – 중앙에 구멍이 뚫려 있고 원 그래프보다 비율을 더 직관적으로 파악

- 수치형 : 양적자료, 수치로 측정 가능. 수치형 자료를 구간으로 나누어 준다면 범주형 자료로 변환하였다고 할 수 있다.
    - 연속형

    무한한 값을 가지며, 정확한 값을 측정하는 데에는 무한대의 가능한 값이 존재한다.

    ex) 키, 무게, 온도

    - 연속형 단독 : 히스토그램, 상자그림
    - 이산형

    정수 값을 가지며, 값이 일정한 간격으로 측정되지 않고 한 값이 다음 값과 어떤 관계를 가지는지 명확하지 않다.

    ex) 나이, 개수, 시간

- 수치형 자료의 표현
    - 히스토그램 – 자료의 분포를 파악
    - 상자그림 – 자료의 분포와 이상치 파악
    - 산점도 – 변수 간의 관계 파악
    - 밀도 그림 – 자료의 분포를 부드러운 곡선으로 파악
    - 라인 차트 – 시간 또는 다른 연속 변수에 따른 수치형 자료의 변화 파악

--> 이러한 시각화 방법들은 자료들을 효과적으로 분석하고 시각화하여 정보를 얻을 수 있도록 해준다.

6. EDA(데이터 시각화) 정리하기

데이터를 분석할 때 어떠한 가설들을 세우고, 그에 맞는 데이터들을 찾아서 그 데이터들을 이용해 유용한 정보를 추출한다. 그런데 본격적인 데이터를 분석함에 있어서 그 전에 꼭 해야 하는 것이 EDA이다.

EDA란 데이터 분석 과정에서 데이터의 특성과 구조를 파악하기 위한 과정이다. 데이터 분석을 시작하기 전에 데이터를 이해하기 위한 첫 번째 단계로, 데이터셋의 크기, 변수의 종류, 변수 간의 관계, 이상치나 결측치 등을 확인할 수 있다. EDA를 통해 얻은 정보는 데이터 분석 전략을 수립하는 데 도움이 되며, 데이터의 품질을 평가하고 전처리 방법을 결정하는 데도 중요한 역할을 한다.

그리고나서 산점도/히스토그램/Bosplot 등을 이용한 데이터 시각화로 보다 쉽게 변수들 사이의 다양한 패턴과 상관관계를 발견할 수 있고, 변수들의 변화양상을 보고 각 변수들 사이의 특징을 확인할 수 있다.

즉, EDA는 방대한 양의 데이터를 분석하기 전에 데이터의 구조를 파악하고, 기초통계량을 파악, 오류를 식별, 그리고 시각화를 통해 자료의 분포를 알고 문제점을 발견하면서 데이터 분석의 질을 높여주는 필수적인 과정이다.

- EDA 단계
    - ■ 데이터 수집

    데이터를 수집하고, 데이터의 출처와 형식, 크기 등을 파악
        - 데이터를 임포트하고 데이터를 불러온다
        - 데이터의 모양 확인
        - 데이터의 타입 확인
    - ■ 데이터 정제

    데이터에 결측치, 이상치, 중복 등의 문제가 있는지 파악하고 이를 처리
        - NULL값 체크 등 데이터 정제(NULL값 개수, 비율 확인 후 정제)
        - 합리적 접근법, 완전제거법, 다중 대체법, 대표값대체 등으로 처리
    - ■ 데이터 탐색(시각화)

    변수들의 통계치 계산

    시각화 기법을 사용하여 데이터를 탐색 -> 변수 간의 관계 파악
        - 범주형 변수의 분포를 살펴본다
        - 연속형 변수의 분포를 살펴본다
    - ■ 데이터 분석

    위의 결과를 바탕으로 변수들 간의 관계 분석
        - 변수간의 관계 분석

- EDA 장점
    - ■ 데이터의 특성 파악

    데이터의 분포, 이상치, 결측치 등을 시각화하여 쉽게 확인

    기초통계량을 파악하고 변수에 대한 이해도 향상
    - ■ 시간 절약을 위한 더 나은 전처리 데이터

    기존 데이터셋에서 실수, 이상, 누락된 값을 식별하여 데이터의 올바른 전처리를 보

장하고 나중에 모델을 적용할 때 실수를 방지하여 상당한 시간을 절약
불필요한 비용과 시간 절약

- 분석 방법 결정

EDA를 수행하여 데이터의 특성을 파악하면, 이를 바탕으로 추가적인 분석 방법을 결정할 수 있다. 분석 방법을 더욱 효율적으로 결정할 수 있으며, 분석 결과의 신뢰성을 높일 수 있다.

7. 변수변환의 필요성과 방법 조사하기

변수변환은 데이터 분석 과정에서 데이터의 분포를 바꾸거나 데이터의 특성을 강조하기 위해 변수의 값을 변경하여 모델의 성능을 향상시키거나 분석 결과를 더욱 해석하기 쉽도록 만드는 작업이다. = 데이터를 분석하기 좋은 형태로 바꾸는 작업
변수들이 선형 관계가 아닌 로그, 제곱, 지수 등의 모습을 보일 때 변수 변환을 통해 선형 관계를 만들면 분석하기 쉽다.

- 변수 변환의 필요성

  변수 변환을 하지 않으면 데이터 분석 결과가 왜곡될 수 있다. 변수의 분포가 비대칭이거나 변수 간 스케일 차이가 크다면, 이를 고려하지 않은 분석 결과는 부정확하거나 잘못된 결론을 내릴 수 있다. 이를 고려하지 않고 분석을 진행한다면 예측력이 떨어지거나 오류를 발생할 가능성이 높다.
  변수변환을 통해 데이터 분포의 비대칭성을 보정, 이상치의 처리, 변수간 스케일 차이 보정을 할 수 있다. 이를 통해 데이터 분석 결과를 더 정확하고 유용하게 만든다.

- 변수변환 방법

  - Z-Score 변환

  데이터를 평균과 표준 편차를 이용하여 표준화하는 방법이다. 변수의 값이 평균과 일치하면 0으로 정규화하고, 평균보다 작으면 음수, 평균보다 크면 양수로 변환
  Outlier은 잘 처리하지만, 정확히 같은 척도로 정규화된 데이터를 생성하지 못한다는 단점이 있다.

  - 로그변환

  로그 변환은 자연 로그나 상용 로그를 사용하여 변수를 변환하는 것이다. 로그 변환은 데이터가 치우친 경우, 즉 한 쪽으로 치우친 분포를 가진 경우에 많이 사용된다. 로그 변환은 이러한 분포를 보다 정규 분포에 가깝게 만들어줄 수 있다.

  - 제곱근 변환

  제곱근 변환은 변수의 값에 제곱근 함수를 취하는 것이다. 이 방법은 데이터의 분포가 비대칭이며, 값의 차이가 큰 경우에 사용된다.

  - Box Cox 변환

  정규성을 만족하지 않는 데이터에 대해, 데이터를 정규분포에 가깝게 만들거나 데이터의 분산을 안정화하는 것이다. 이 방법은 최적의 파라미터를 선택하도록 구현된다.

  - Min-Max 정규화

  모든 변수에 대해 최솟값은 0, 최댓값은 1로, 최솟값 및 최댓값을 제외한 다른 값들은 0과 1 사이의 값으로 변환
  변수의 값이 일정한 범위 내에 있어야 하는 경우에 사용된다.
  Outlier에 영향을 많이 받는 단점이 있다.

- Binning

데이터 값을 몇 개의 Bin으로 분할하여 계산하는 방법
데이터 평활화에서도 사용되는 기술이며, 기존 데이터를 범주화하기도 한다.
(*데이터 평활화 : 데이터 추세에 벗어나는 값들을 변환하는 기법)
연속형, 이산형 데이터를 범주형으로 변환

- Encoding

범주형 데이터를 연속형, 이산형 데이터로 변환(last encoding, One-hot encoding)

8. 계급 불균형 자료(Imbalanced data) 및 해결법에 대하여 조사 - (정확도의 역설)

계급 불균형자료(Imbalanced Data)는 목표 변수(target/output variable) 가 범주형 데이터일 때, 범주 별로 관측치의 개수, 비율의 차이가 많이 나는 데이터이다. 실제 데이터셋에서 일반적인 문제이다.

- 불균형 데이터의 문제점

계급 불균형 자료에서는 모델이 예측을 잘하기 위해 다수 계급에 집중할 가능성이 높아진다. 이 경우, 소수 계급의 예측 정확도는 저하될 수 있지만 다수 계급의 정확도는 높게 나온다. 그러나 이 모델은 정확도가 높아도 모델의 성능이 좋다고 할 수 없다. 소수 계급은 감지하지 못하지만 정확도는 높은 모델인 것이다. = 실제 문제를 해결할 수 없는 모델
--> 이것을 "정확도의 역설" 이라고 한다.
또한 특정 계급에 대푯값이 집중되어서 대푯값이 왜곡될 수 있고 이는 모델의 일반화 성능을 저하 시킬 수 있다.

- 불균형 데이터 해결법
  - Resampling
    - 언더샘플링 : 다수 계급의 크기를 줄여 데이터셋의 균형을 맞춘다.
    - 오버샘플링 : 소수 계급의 크기를 늘려 데이터셋의 균형을 맞춘다.
      -> 과적합위험?
    - 오버&언더샘플링 : 두 샘플링을 반반씩 섞은 기법이다. 각 계급의 크기를 두 계급의 크기의 평균만큼 되도록 샘플링을 하면 된다.
  - SMOTE

이것 또한 소수 계급의 크기를 늘리는 방법이다. Sampling과의 차이점은 기존의 데이터를 랜덤으로 복원추출하는 것이 아니라 기존의 데이터들을 적절하게 조합하여 새로운 데이터를 만드는 것이다.

  - 앙상블

데이터의 서로 다른 하위 집합에 대해 훈련된 여러 모델을 결합하거나 분류 성능을 향상 시키기 위해 서로 다른 알고리즘을 학습시킨다.

  - 알고리즘 수정

기존 알고리즘을 수정하거나 불균형 데이터에 덜 민감한 새 알고리즘을 설계한다.

9. 주어진 framingham data set을 이용하여 데이터 탐색을 진행하고 탐색에 따른 자신의 견해 작성하기

https://colab.research.google.com/drive/1u8XnG1sCWyyZystB5CPRze14DKm0BGFi#scrollTo=06sIiC4Sh_uK
colab으로 탐색 진행하였습니다.

*discussion
데이터타입을 확인할 때, 범주형 변수는 object, string / 수치형 변수는 int64, float64로 맞춰줘야 한다고 공부했습니다. 이번에 진행한 과제에서 education, BPMeds, prevalentStroke, prevalentHyp, TenYearCHD은 범주형인데 수치형 변수타입이 나왔습니다. 이를 string타입으로 변경하였는데 이런 경우에는 object 타입과 string 타입 중 어떤 것으로 변경을 해주어도 분석에는 차이가 없는지, 혹은 어떤 타입에는 어떤 것이 더 나은지 궁금합니다. 그리고 str로 변경하였는데 타입을 확인하니 object라고 떴습니다. 이 이유도 궁금합니다.

결측치가 있는 범주형 자료는 평균값으로 대체하면 소수점의 값이 나오기 때문에 대체하지 못하고 아예 삭제하는 방법으로 진행하였습니다. --> 최빈값으로 대체하는게 더 나을 듯

그러나 총 150개가 넘는 값들이 삭제되었습니다. 이 개수는 데이터 분석에 있어서 생각보다 큰 값이 제거 된 것인데 이럴 땐 어떤 다른 방법으로 하는 것이 더 좋나요?