데이터 전처리의 이해

데이터 전처리	데이터 분석을 위한 필수 과정으로 데이터를 정제한 뒤, 데이터 가공, 통합, 정리, 변환을 통해 데이터 분석 변수를 처리하는 등의 작업으로 데이터 분석 결과의 신뢰도를 높이기 위한 과정

데이터 정제

데이터 정제	결측값, 잡음, 이상값 등 데이터 오류의 원인을 분석 작업 전에 처리하는 것을 의미
결측값	분석대상에서 제외 또는 보완하여 처리 가능
이상값	삭제, 대체, 스케일링, 정규화 등의 방법으로 처리 가능

데이터 결측값

결측값 유형	①비무작위(NMAR): 결측값에 영향 미친다. ②무작위(MAR): 연관은 있지만 결과에는 영향 미치지 않는다. ③완전 무작위(MCAR): 연관 없이 완전히 무관한 결측
결측값 대체 방법	①평균 대체: 대푯값으로 대체 ②단순 확률 대체: 단순 확률값으로 대체 ③보삽법: 비슷한 시기, 다른 해의 데이터를 참고한 평균값으로 대체 ④평가치 추정법: 맥락적/행렬식 자료를 고려하여 원래의 값 추정 ⑤다중 대치법: 결측치 추정을 통해 완성한 데이터셋을 이용하여 결측치 추정 ⑥완전 정보 최대우도법: 최대우도 바탕으로 가중평균 구성하여 대체

데이터 이상값

이상값 검출	①분산: 정규분포 97.5% 외의 값 ②우도함수: 우도확률값 외의 값 ③근접 이웃 기반 이상치 탐지: 정상값 거리와 거리가 먼 값 ④밀도 기반: 상대적 밀도 값이 먼 값 ⑤군집: 특정 군집에 속하지 않는 값 ⑥사분위수: 양쪽 말단에 1.5분위수를 벗어나는 값

■ 변수 선택

변수 선택	종속변수에 영향을 미칠 독립변수를 선택하는 과정, 선택적으로 변수를 적용하여 모델 성능 향상 가능
선택적 변수 선택의 이점	① 머신러닝 알고리즘 학습속도 향상 ② 모델 해석 용이 ③ 모델 정확도 향상 ④ 과적합 감소, 성능 향상

■ 단계적 변수 선택 방법

전진 선택법	가장 많은 영향을 줄 것 같은 변수부터 하나씩 추가(AIC 작은 것부터 추가)
후진 제거법	가장 적은 영향을 주는 변수부터 하나씩 제거(AIC 큰 것부터 제거)
단계적 방법	전진 선택법에 의한 유의한 변수 추가, 후진 선택법에 의한 유의성 낮은 변수 제거 작업 반복(변수 연속적 추가와 제거를 통해 AIC가 낮아지는 모델을 완성) ※AIC: 작을수록 좋은 데이터 모델이라고 할 수 있다.

■ 차원축소

차원축소	▪ 변수의 수 증가로 인해 차원이 커지면서 데이터 모델링 성능 저하 문제가 발생하는 것을 '차원의 저주'라고 한다. ▪ 공간은 증가하는 데 비해 데이터 수의 변화가 없는 경우 불필요한 정보와 공간으로 인해 모델링 성능의 저하를 유발할 수 있기 때문에 차원축소가 필요하다.
PCA **(주성분 분석)**	▪ 변수 간 상관관계를 파악하고 선형 연관성이 없는 저차원으로 축소하는 방법 ▪ 데이터 분산을 최대로 보존하는 축(PC1)을 찾고 PC1과 직교하면서, 두 번째로 분산이 최대인 축(PC2)을 찾는다. 이를 n회 반복하여 n만큼의 축을 찾는다.
LDA **(선형판별분석)**	▪ 지도학습을 통해 데이터 결정경계를 만들어 데이터 분류하는 것 ▪ LDA의 2가지 가정 ①다변량정규분포를 따르는 데이터 분포여야 한다. ②파라미터는 평균(벡터)과 공분산(행렬)이어야 한다.
t-SNE **(t-분포 확률적 임베딩)**	▪ 고차원의 데이터 거리를 보존하며 그 관계를 저차원으로 축소하는 방법 ▪ 하나의 점(t1)을 선택하여 다른 점들 간의 거리를 측정한 뒤 이를 T 분포 그래프에 표현한 다음 t1을 중앙에 위치시키고 친밀도가 가까운 값끼리 그룹화
SVD **(특잇값 분해)**	행렬의 크기와 모양에 상관없이 적용할 수 있는 방법이다. $(M = U \Sigma V^T)$

■ 변수 변환

범주형 데이터 변환	범주형 변수를 숫자로 변환 (남자:1, 여자:2)
연속형→범주형으로	연속형 데이터를 범주형으로 (10~19세: 10대)
비정형 데이터 변환	단어의 빈도수 등을 이용해서 정형화
더미 변수화	어떤 특징의 존재 여부를 1 또는 0으로 변환
스케일링	최소-최대 표준화, 정규화

■ 클래스 불균형

여러 클래스 중 데이터 양에 큰 차이가 있는 경우 클래스 불균형이 있다고 한다.

과소표집	소수 클래스의 데이터 수만큼 감소시킨다. 데이터 손실 우려.
과대표집	다수 클래스의 데이터 수만큼 증가시킨다. 과적합 문제 발생 가능.
SMOTE	주변값을 기준으로 소수 클래스의 데이터 수를 증가시켜 다수 클래스의 수와 동일해지게 한다.