〈 머신러닝으로 예측 모형 생성 〉

아래에 제시된 데이터셋을 사용하여 예측 모형을 생성해보자.
수업시간에 학습한 다양한 머신러닝 모형을 사용하여 다양한 예측 모형을 생성할 수 있다.
다양한 데이터 전처리 방법, 변수 선택, 하이퍼파라미터 튜닝 등을 사용해 예측 성능을 향상할 수 있는 방법을 시도해보자.

1. Medical Insurance Costs 예측 모형

Medical Insurance Costs 데이터셋은 개인의 의료비 청구액을 예측하는 데 사용될 수 있는 데이터셋이다.

데이터 소스 :
https://raw.githubusercontent.com/aonekoda/reference/refs/heads/main/data/insurance.csv

Medical Insurance Costs 데이터셋의 주요 속성:

- **age**: 보험 수혜자의 나이
- **sex**: 보험 계약자의 성별 (여성, 남성)
- **bmi**: 체질량 지수 (Body Mass Index), 키와 몸무게를 이용하여 비만도를 측정하는 지수
- **children**: 건강 보험에 가입된 자녀 또는 부양 가족 수
- **smoker**: 흡연 여부 (흡연, 비흡연)
- **region**: 수혜자의 거주 지역 (미국 내 4개 지역: northeast, southeast, southwest, northwest)
- **charges**: 건강 보험으로 청구된 개별 의료비 (예측해야 할 종속 변수)

데이터셋의 특징:

- 해당 데이터는 개인의 의료 보험 관련 정보를 나타낸다.
- 범주형 데이터(성별, 흡연 여부, 지역)와 수치형 데이터(나이, BMI, 자녀 수, 의료비)가 혼합되어 있다.
- 선형 회귀 모델을 포함한 다양한 머신러닝 모델을 훈련하여 의료비(charge) 예측 분석을 수행하는 데 적합하다.

2. Titanic 생존여부 예측 모형

이 데이터셋은 타이타닉 승객의 생존여부를 예측하는데 사용될 수 있다.

데이터 소스 :
https://raw.githubusercontent.com/aonekoda/reference/main/data/titanic.csv

titanic 데이터셋의 주요 속성 :

- Survived - 생존 여부 (0 = 사망, 1 = 생존), 종속변수수
- Pclass - 티켓 클래스 (1 = 1등석, 2 = 2등석, 3 = 3등석)
- Sex - 성별
- Age - 나이
- SibSp - 함께 탑승한 형제 / 배우자의 수
- Parch - 함께 탑승한 부모 / 자녀의 수
- Ticket - 티켓 번호
- Fare - 탑승 요금
- Cabin - 선실 번호
- Embarked - 선착장 (C = Cherbourg, Q = Queenstown, S = Southampton)

데이터셋의 특징 :

- 해당 데이터는 타이타닉 승객에 대한 정보이다.
- 선실등급, 성별, 선착장 정보는 범주형이다.
- 머신러닝 모형을 생성하여 승객의 생존여부(Survived)를 예측하는 분류모형을 생성할 수 있다.