**Ⅱ. 단변량 데이터 탐색**

**2.1 머신러닝 프로세스**

(1) Data Set을 분할한다.

- 학습 데이터를 랜덤으로 학습데이터셋 / 검증데이터셋을 분할하고, 테스트 셋도 준비한다.

- 학습 데이터는 전체 데이터 중 70~90%, 검증 데이터는 10~20%, 테스트 데이터는 10~20%

(2) 데이터 전처리

- 특성치(X) 변수의 정규화와 범주자료 원핫인코딩 과정이 핵심이다.

- 다만 특성치에 범주자료가 없을 경우, 원핫인코딩은 생략한다.

- 정규화는 표준화(평균 0 / 표준편차 1) 또는 Min-max 정규화가 대표적이다.

(3) 모델 적용

(4) 다양한 Hyper Parameter 적용하고 변경하면서 최적의 Hyper Parameter 및 최종 모델을 결정

- 최종 분류기에서 테스트 셋으로 성능을 평가한다. 또한 검증 셋은 사용하지 않는 것이 좋다.

- 즉, 오로지 테스트 데이터의 결과에 집중한다.

**2.1.1 DATA SET 분할**

1) 학습데이터(train data) : 전체 분석데이터 중 머신러닝 모델 알고리즘을 학습하기 위한 데이터

전체 데이터의 70~80%

2) 테스트데이터(test data) : 학습된 모델이 다른 데이터에도 맞는지를 확인하기 위한 데이터

전체 데이터의 20~30%

- 전체 데이터가 아니라 테스트 데이터를 따로 떼어내는 이유는 모델 결과가 다른 데이터에도 적용 가능한지, 즉 일반화가 가능한지를 검증하고 일반화에 적합한 모델을 확인하기 위함이다.

**2.1.2 DATA 전처리**

- 데이터셋을 나눈 후, 목적변수인 레이블(y)은 건드리지 않고, 특성치(X)는 전처리 중 정규화를 반드시 진행한다.

- 정규화는 X 변수들의 단위가 다르기 때문에 생기는 문제를 해결하기 위해 모든 단위를 동일하게 만드는 과정

- 정규화는 표준화(Standardization)와 민맥스(Min-Max) 두 가지 방법이 주로 사용된다.

- 특성치에 범주형 자료가 있을 경우, 그대로 학습시키면 안된다.

(1) 예를 들어, 성별 1=남자, 2=여자를 구분한 1과 2는 크고 작음의 의미가 없고 단지 집단을 분류하는 것이다.

(2) 따라서 1, 2라는 숫자를 없애고 모두 0과 1로 변경해야한다.

(3) 성별이라면 남자인지 아닌지(0,1), 여자인지 아닌지(0,1)라는 2개 변수를 만들어서 무조건 0과 1로 만드는 과정을 **원핫인코딩** 이라고 한다.

(4) 다만 **원핫인코딩**은 데이터셋을 나누기 전에 하는 것이 더 좋다.

**2.1.3 모델 학습**

- 목적변수인 y(레이블이 존재한다면 지도학습(supervised learning)을 적용하고, 그렇지 않다면 군집이나 연관분석을 수행한다.

- **회귀문제**(레이블이 **연속형**), **분류문제**(레이블이 **범주형**)

- 반면 **로지스틱회귀모델**처럼 **분류문제**만 적합하거나 **선형회귀·릿지(Ridge)·라소(Lasso)·엘라스틱넷(ElasticNet)**과 같이 **회귀문제**만 적합한 알고리즘이 있다.

[머신러닝 알고리즘의 분류]

1. 지도 – 분류

* 로지스틱회귀모델

1. 지도 - 분류+회귀

* 최근접이웃법
* 나이브 베이즈
* 인공신경망
* 서포트 벡터 머신
* 의사결정나무
* 랜덤 포레스트

1. 지도 – 회귀

* 선형 회귀모델
* 릿지
* 라소
* 엘라스틱넷

1. 지도 -앙상블

* 투표기반 앙상블
* 배깅
* 부스팅
* 스태킹

1. 비지도

* 군집분석
* DBSCAN
* 연관규칙분석

**2.1.4 하이퍼파라미터 탐색과 모델 튜닝**

- 머신러닝은 다양한 하이퍼파라미터를 조정하며 데이터에 적합한 최적의 알고리즘을 찾아야한다.

- 이러한 과정을 하이퍼파라미터 탐색(Search) 또는 모델튜닝(Model Tuning)이라고 한다.

- 사이킷런에서는 그리드 탐색(Grid Search) 또는 랜덤 탐색(Random Search)을 통해 여러 조합의 결과를 파악할 수 있는 기능을 제공한다.

1) 그리드 탐색 : 분석자가 몇 가지 임의의 파라미터를 설정하여 그 결과 중 가장 좋은 정확도 (모델평가지표)를 선정하는 방법

2) 랜덤 탐색 : 분석자가 설정한 일정 범위 이내에서 무작위로 하이퍼파라미터를 뽑아 최적 결과를 도출하는 방식

- 하이퍼파라미터는 규제를 의미하는 일종의 **모델 조정계수**라고 할 수 있다.

- 서포트벡터머신에서 C와 gamma값을 조정하면서 얼마나 분류의 기준을 완화할 것인가를 결정

- KNN에서 이웃수를 조정하면서 정교한 분류를 할 것인가(k값이 작을수록), 아니면 더 일반화에 중점을 둘 것인가(k값이 커질수록)를 결정

**2.1.5 모델 성능 평가**

- 모델 성능이란 ‘얼마나 정확하게 맞추었는가’를 지표화한 것이다.

- **분류의 문제**라면 **실제분류와 예측분류가 얼마나 일치하였는가**이고, **회귀의 문제**라면 **실제 레이블값과 예측 레이블값이 얼마나 맞는지**를 봐야 한다.

- 분류 알고리즘에서 보는 대표적인 모델 결과는 오차행렬(confusion matrix)이다.

- 오차행렬 : 실제분류를 행(row)에, 모델의 예측분류를 열(columns)에 배치한 교차표로 정리하며, 정분류와 오분류를 쉽게 알 수 있도록 한 결과표

- 회귀에서 대표적인 모델 성능 결과는 R²(설명력)과 RMSE가 있다.

- R² : 0과 1 사이의 값으로 1에 가까울수록 실제값과 예측값이 일치하는 정도가 높다

- RMSE : 오차로서 실제값과 예측값의 차이를 평균적으로 계산한 값