프로젝트 논의 및 연습문제

프로젝트를 하면서 도움이 되는 팁과 성능 평가 방법

프로젝트 진행 팁

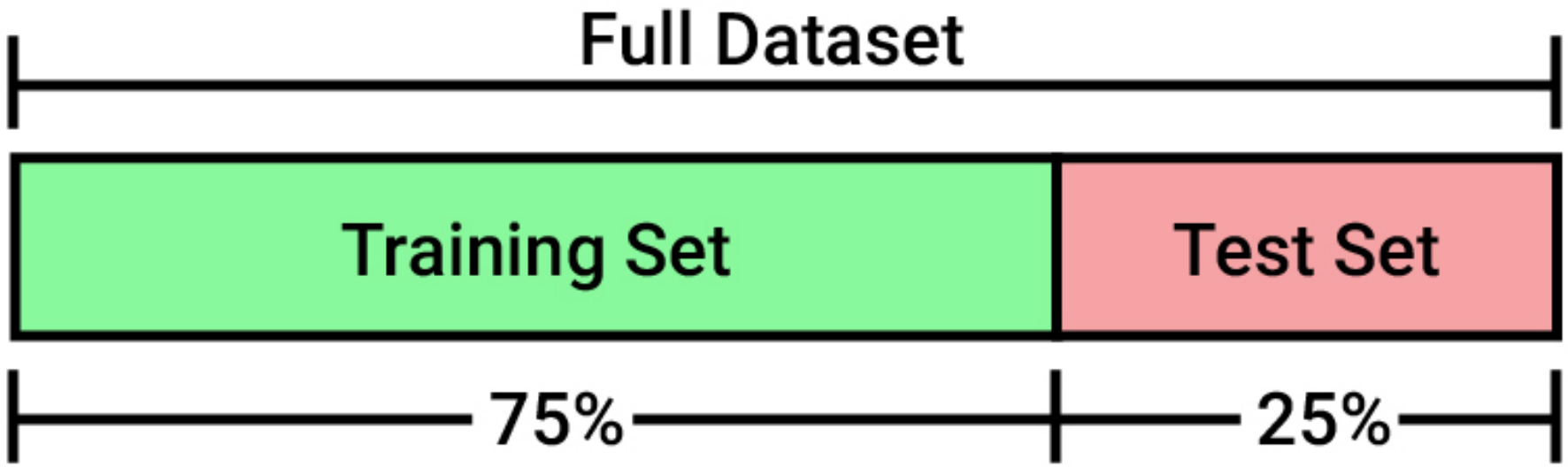
해결하려는 문제에 대한 명확한 정의 필요

- 어떤 **데이터**를 사용할 것인지: **학습에 사용할 변수**, **데이터 전처리** 방법
  - 정규화, 이산화, 결측값 제거 및 채우기 작업
  - 변수가 너무 많다면, **상관관계** 분석을 통해 영향을 주는 변수만 사용 !
- 어떠한 **머신러닝** 방법을 사용할 것인지: 데이터 분석 방법 및 예측 방법
  - SVM, 회귀 (선형/로지스틱), 신경망, 클러스터
- 예측의 **입력**과 **출력**은 어떤 것인지: 입력과 결과의 대상
- 데이터의 특성을 고려했을 때 **어떤 학습 방법**이 적합한지: 지도학습과 비지도학습
- 어떠한 방식으로 **예측의 성능을 평가**할 것인지: **정확도**를 어떻게 계산하고 측정할 것인지 기준
  - **교차 검증 (Cross validation), 혼동 행렬 (Confusion matrix)**

교차 검증 (Cross validation)

교차 검증의 필요성

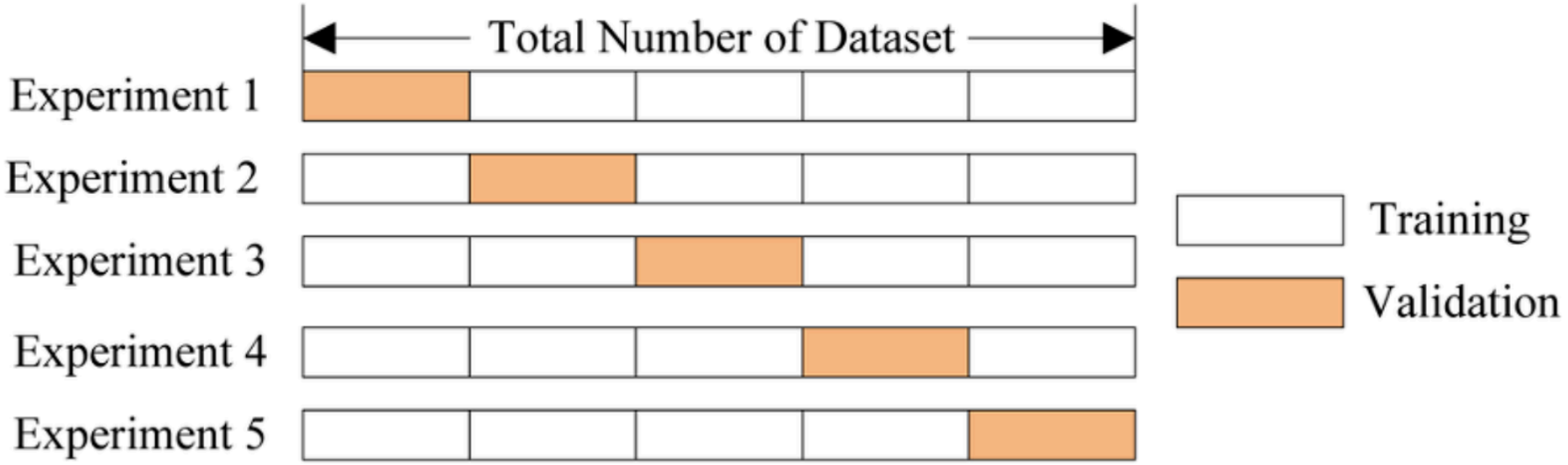
- 예측 모델을 만들기 위한 데이터가 있고, 이때 데이터는 **학습**과 **평가**에 해당되는 부분만 있다면
  - 모델을 학습하면서 **검증과정에서 평가에 해당되는 데이터 셋을 사용**
  - 고정된 평가 데이터를 사용하여 성능을 평가하면, 평가 데이터에만 성능이 좋은 현상 발생 (Overfitting)



교차 검증 (Cross validation)

교차 검증의 방법

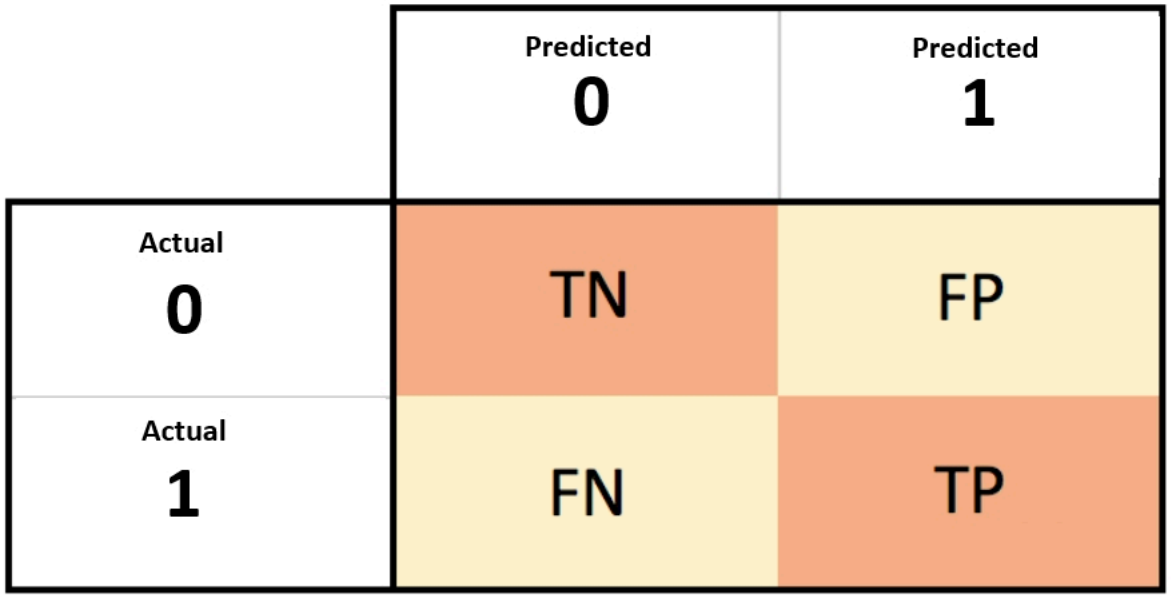
- 데이터를 **학습**과 **평가**에 해당되는 부분으로 구분하고
  - 학습에 해당되는 데이터를 **K개의 Fold**로 나눔
  - 하나는 **검증 데이터**, 나머지는 **훈련 데이터**로 사용
  - **또 다른 부분을 검증 데이터, 나머지를 훈련 데이터로 사용**
  - **K번 반복하여 K번 성능의 평균을 구함**



혼동 행렬 (Confusion matrix)

훈련된 모델의 성능을 측정하기 위한 Matrix

- 모델을 **평가하는 지표** (정밀함, 실용적인 분류, 정확한 분류)
  - **레이블 0, 1을 가진 데이터를 분류**한다고 할 때, 관심 범주를 1이라고 가정
  - True Positives (TP): 1인 레이블을 1이라고 함. (관심 범주를 정확하게 분류)
  - False Negatives (FN): 1인 레이블을 0이라고 함. (관심 범주가 아닌 것으로 잘못 분류)
  - False Positives (FP): 0인 레이블을 1이라고 함. (관심 범주라고 잘못 분류)
  - True Negatives (TN): 0인 레이블을 0이라고 함. (관심 범주가 아닌 것을 정확하게 분류)



혼동 행렬 (Confusion matrix)

4가지 정보를 바탕으로 3가지 척도를 계산

- **정확도 (Accuracy):** 정확도는 1을 1로, 0을 0로 **정확하게 분류**한 것을 의미
- **정밀도 (Precision):** 모델을 1이라고 분류한 그룹 A가 있을 때, **믿을 만한 정도**로 A를 만들어 냈는지 평가
  - 예) 어부가 그물을 던져 물고기를 잡을 때, 그물안에 1이라는 물고기가 얼마나 있을지에 대한 척도
- **재현도 (Recall):** 정밀도와 비교되는 척도, **전체 예측 중에 TP가 얼마나 많은 것인가에 대한 정보**
  - 관심있는 영역만을 추출했는지를 의미하는 것으로 모형의 실용성과 관련된 척도

- **Accuracy(정확도)** $= \dfrac{TP + TN}{TP + TN + FP + FN}$

- **Precision(정밀도)** $= \dfrac{TP}{TP + FP}$

- **Recall(재현도)** $= \dfrac{TP}{TP + FN}$

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

프로젝트 추천 주제

번호	논문 제목	내용	구현방법
1	재입원 예측 모형 개발에 관한 연구	불필요한 재입원 방지	로지스틱 회귀
2	다중모델을 이용한 자동차 보험 고객의 이탈예측	자동차 보험 고객의 이탈 예측	다중 모델
3	빅데이터 프로세싱이 보험 상품에 미칠수 있는 영향	보험 상품을 만들기 위한 요인 분석	회귀 분석
4	제3 보험의 해약 결정요인에 관한 연구	제3 보험 해약 결정에 영향을 미치는 요인 분석	로지스틱 회귀
5	데이터 마이닝 기법을 이용한 건강보험 사기 탐지	다양한 머신러닝 기법을 활용한 사기 탐지	다양한 방법
6	뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정 모형	감성분석을 활용한 주가 예측	로지스틱 회귀
7	SVM 기반 재무 정보를 이용한 주가 예측	회사 재무 정보를 SVM에 적용	SVM
8	딥러닝 분서고가 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측	SVM 기반의 코스피 예측	SVM
8	데이터 마이닝을 이용한 코스닥 시장의 상장폐지 예측모형 구축에 관한 연구	상장폐지 주식 예측	인공신경망
9	한국 성인 암 수검 관련 요인에 대한 선행 연구 고찰	요인들과 암 수검의 관계 분석	-
10	DEA와 선형 회귀분석을 활용한 콜센터 상담원의 성과 상대효율성 분석	콜센터 품질을 높이는 요인 분석 (상담원 성별)	선형 회귀
11	국민건강보험 빅데이터 기반의 질병트렌드에 따른 지역 군집화 방법론 개발	특정 지역의 질병 발생 추이 분석	클러스터
12	데이터 마이닝을 이용한 차량 사고자 사망확률 모형	교통사고 사망확률 예측 및 요인 분석	군집화, 트리
13	생명보험회사의 파생상품사용 결정요인에 관한 연구	파생상품사용 결정요인 분석	로지스틱 회귀
14	생명보험회사에 대한 만족도, 신뢰도, 충성도에 영향을 미치는 요인 분석	생멍보험 계약자의 보험회사에 대한 분석 (만족도, 신뢰도, 충성도)	회귀 분석

연습문제

타이타닉 생존자 데이터 셋에서 결측치 제거

- DataFrame.info (): 각 속성의 정보 확인
- DataFrame.dropna (): Nan에 해당되는 값 제거
- DataFrame.isnull (): 결측치 값인지 확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId     891 non-null int64
Survived        891 non-null int64
Pclass          891 non-null int64
Name            891 non-n
Sex             891 non-n
Age             714 non-n
SibSp           891 non-n
Parch           891 non-n
Ticket          891 non-n
Fare            891 non-n
Cabin           204 non-n
Embarked        889 non-n
dtypes: float64(2), int6
memory usage: 83.7+ KB
```

```python
import numpy as np
import pandas as pd

df = pd.read_csv("train.csv")
df.info()

# 결측치 함수를 제거합니다.
# 결측치는 df.isnull()을 사용하여 확인합니다.

```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 183 entries, 1 to 889
Data columns (total 12 columns):
PassengerId     183 non-null int64
Survived        183 non-null int64
Pclass          183 non-null int64
                183 non-null object
                183 non-null object
                183 non-null float64
                183 non-null int64
                183 non-null int64
                183 non-null object
                183 non-null float64
                183 non-null object
                183 non-null object
at64(2), int64(5), object(5)
e: 18.6+ KB
```