230519 B팀 주간발표

Heading To the Ground

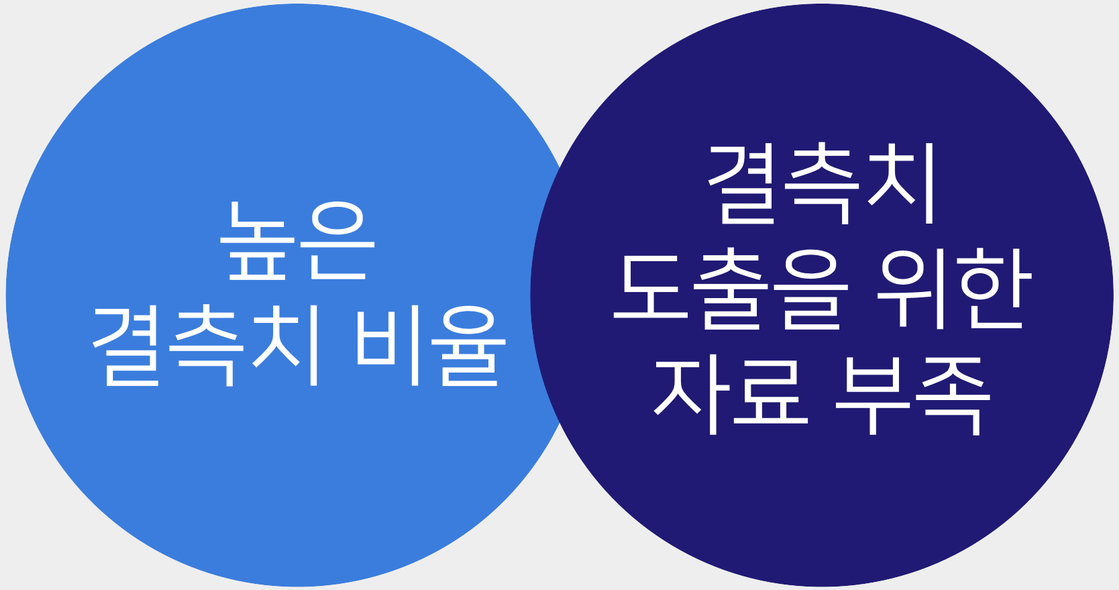
5월 8일 데이콘 마감

월간 데이콘 - 항공편 지연 예측 모델

Semi-Supervised Learning

결측치 처리

결측치 처리가 중요했던 이유



**결측치 처리가
성능을 결정짓는
중요한 요인!**

결측치 시각화

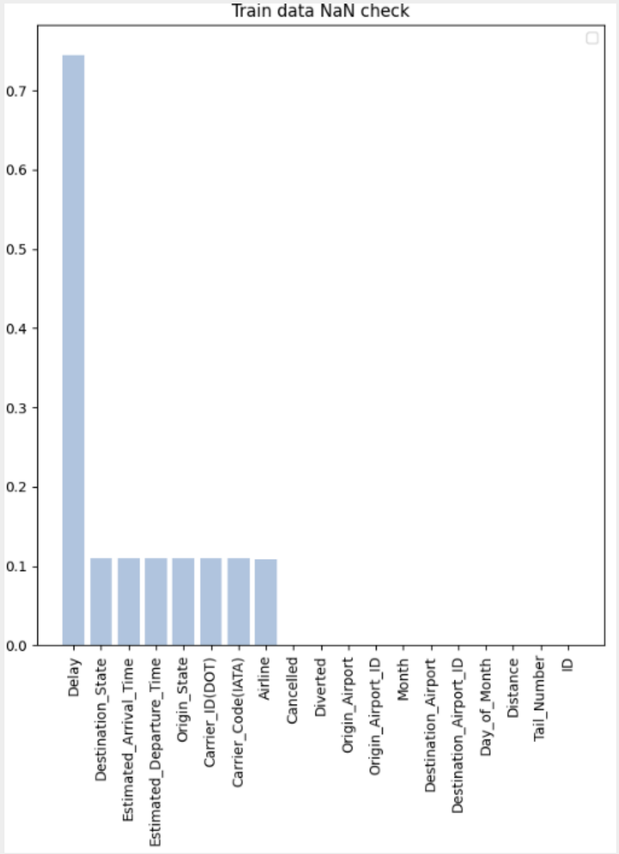
train

```
ID                          0
Month                       0
Day_of_Month                0
Estimated_Departure_Time    109019
Estimated_Arrival_Time      109040
Cancelled                   0
Diverted                    0
Origin_Airport              0
Origin_Airport_ID           0
Origin_State                109015
Destination_Airport         0
Destination_Airport_ID      0
Destination_State           109079
Distance                    0
Airline                     108920
Carrier_Code(IATA)          108990
Carrier_ID(DOT)             108997
Tail_Number                 0
Delay                       744999
dtype: int64
```

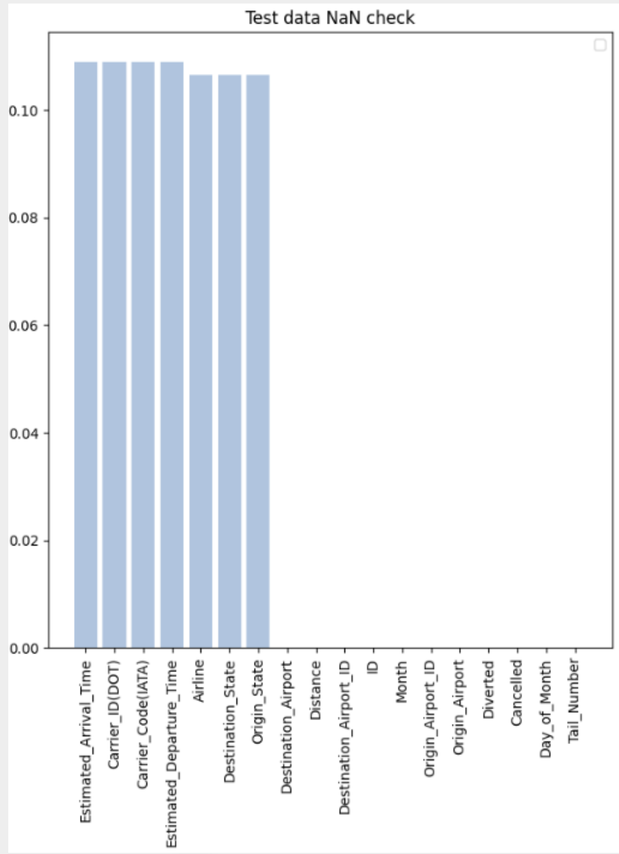
test

```
ID                          0
Month                       0
Day_of_Month                0
Estimated_Departure_Time    108984
Estimated_Arrival_Time      109048
Cancelled                   0
Diverted                    0
Origin_Airport              0
Origin_Airport_ID           0
Origin_State                106505
Destination_Airport         0
Destination_Airport_ID      0
Destination_State           106523
Distance                    0
Airline                     106527
Carrier_Code(IATA)          108993
Carrier_ID(DOT)             109006
Tail_Number                 0
dtype: int64
```

train



test



Pseudo-labeling

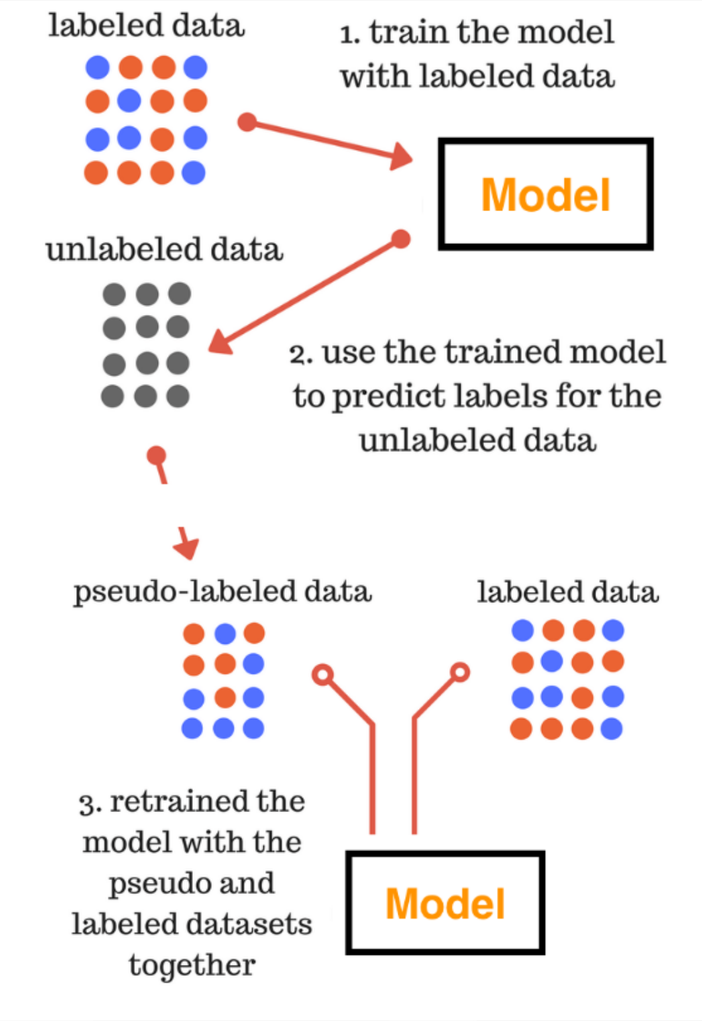
Pseudo-labeling이란?

기존에 정답라벨이 있는 데이터에 **정답라벨이 없는 데이터를 학습**시켜

도출된 결과를 접목시켜서 기존 가지고 있는 정답 Label 데이터에 기반하여

확률적인(대략적인) 정답 라벨을 부여하는 기법

Pseudo-labeling



정답라벨이 있는 데이터를 모델 학습 진행

학습한 모델을 활용하여
라벨이 없는 데이터(test data)를 예측하고
그 결과를 사용하는 pseudo labeled data 생성

2번에서 만든 pseudo-labeled data와 기존에 정답
라벨이 있는 데이터를 모두 사용하여 다시 모델 학습

Self-Training Module

어려웠던 부분

Airline Self_training

```
[16]  1 original_col = ['Month', 'Origin_Airport', 'Destination_Airport', 'Distance', 'Tail_Number']
```

```
1 # Airline Self_training
2
3 X = train_st[original_col]
4 y = labeled_data[(labeled_data['Airline'].astype(str) != 'None')][original_col + ['Airline']]['Airline']
5
6 stclf = SelfTrainingClassifier(
7     base_estimator = RandomForestClassifier(n_estimators = 100),
8     verbose = True)
9
10
11 stclf.fit(X, y)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-17-69b90ae30913> in <cell line: 11>()
      9
     10
---> 11 stclf.fit(X, y)

                                    4 frames
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py in _assert_all_finite(X, allow_nan, msg_dtype, estimator_name, input_name)
    109     if X.dtype == np.dtype("object") and not allow_nan:
    110         if _object_dtype_isnan(X).any():
--> 111             raise ValueError("Input contains NaN")
    112
    113     # We need only consider float arrays, hence can early return for all else.

ValueError: Input contains NaN
```

Self-Training

sklearn.semi_supervised.
SelfTrainingClassifier

ValueError :
Input contains NaN

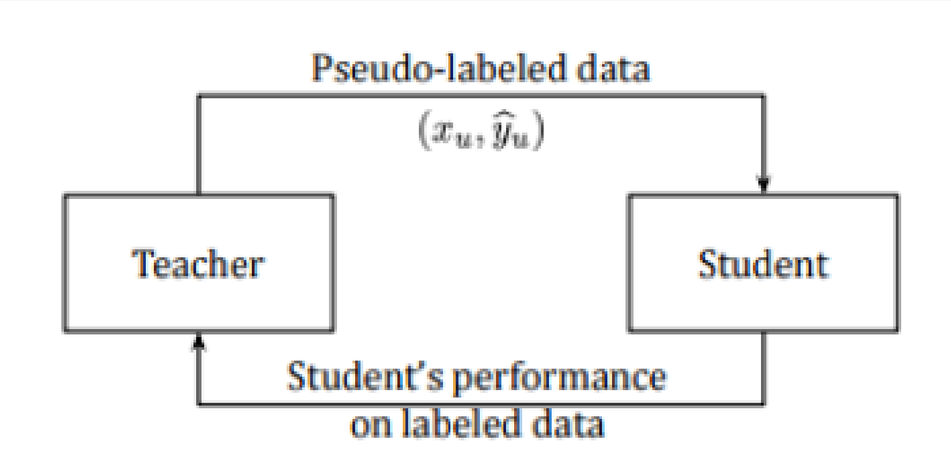
fit (x)

Meta-Pseudo-labeling

Meta-Pseudo-labeling이란?

기존에 정답라벨이 있는 데이터(Teacher)에 정답라벨이 없는 데이터(Student)를 동시에 학습시켜 Student의 성능이 Teacher에게 Reward로 전달

Meta-Pseudo-labeling



student는 teacher에서 생성된 pseudo labeled data로 학습

↓

teacher은 student가 labeled data에 얼마나 잘 작동하는지의 reward signal로 학습

↓

teacher은 더 나은 pseudo label을 생성 가능

Pseudo-labeling

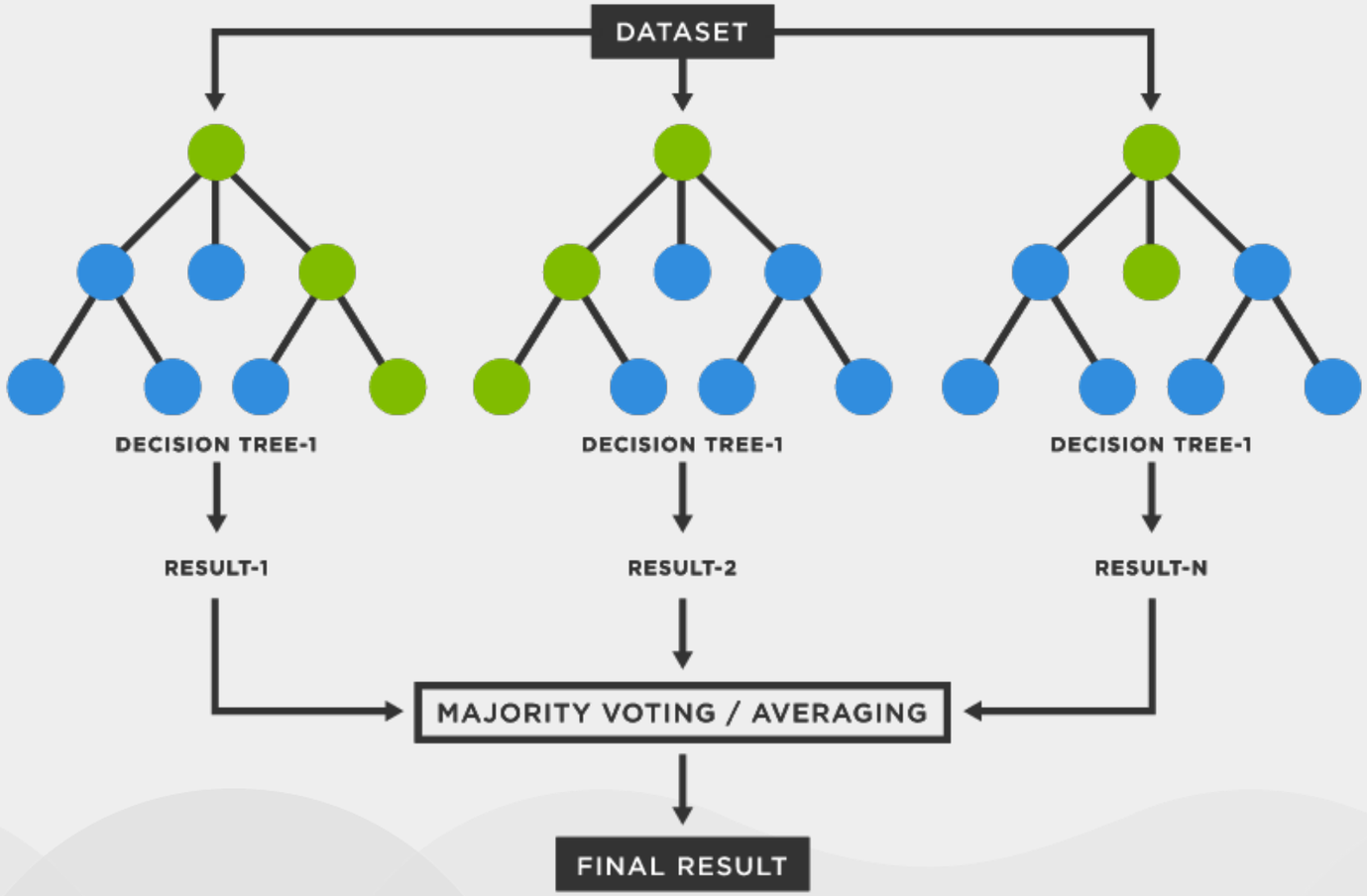
Meta-Pseudo-labeling을 이용한 성과

Logloss

0.9147043451

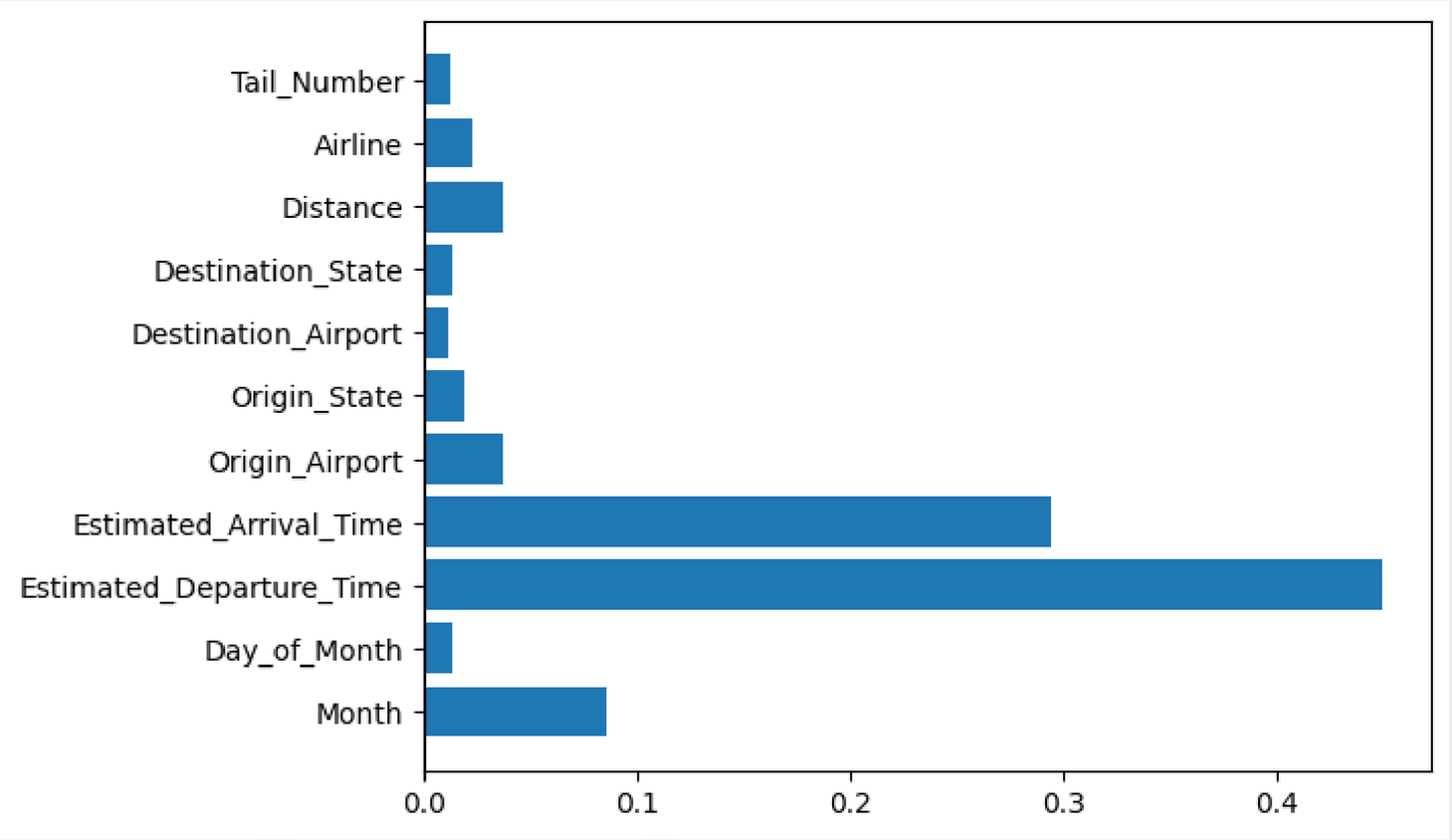
Random Forest

준지도 학습 X



시간 데이터의 함정

중요도 분석

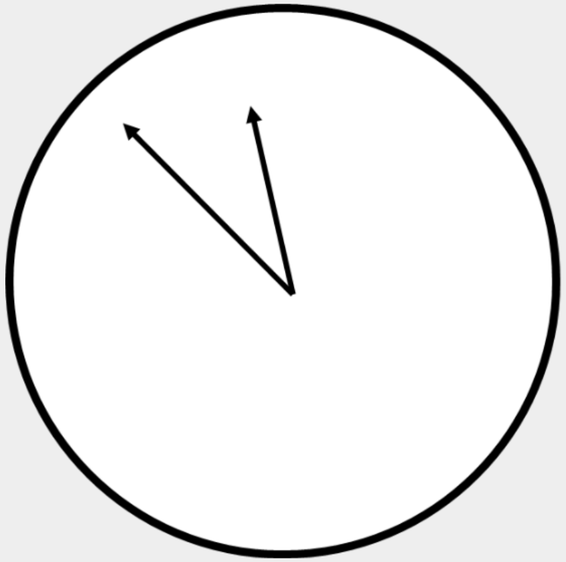


Estimated_Arrival_Time

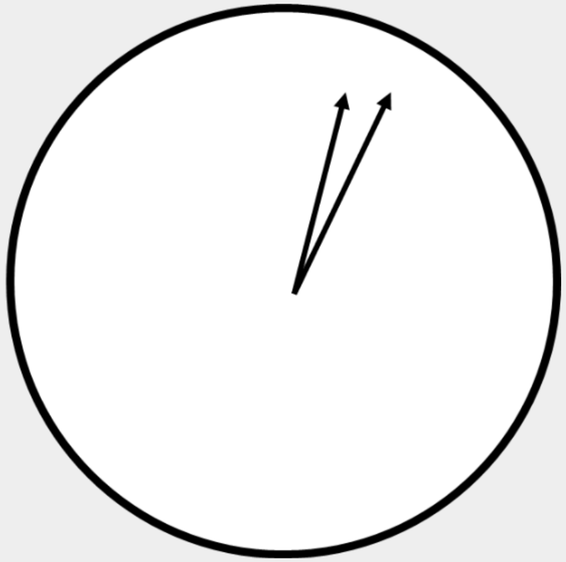
Estimated_Departure_Time

시간 데이터의 함정

다음 두 시간의 평균은 얼마일까요?



23:50

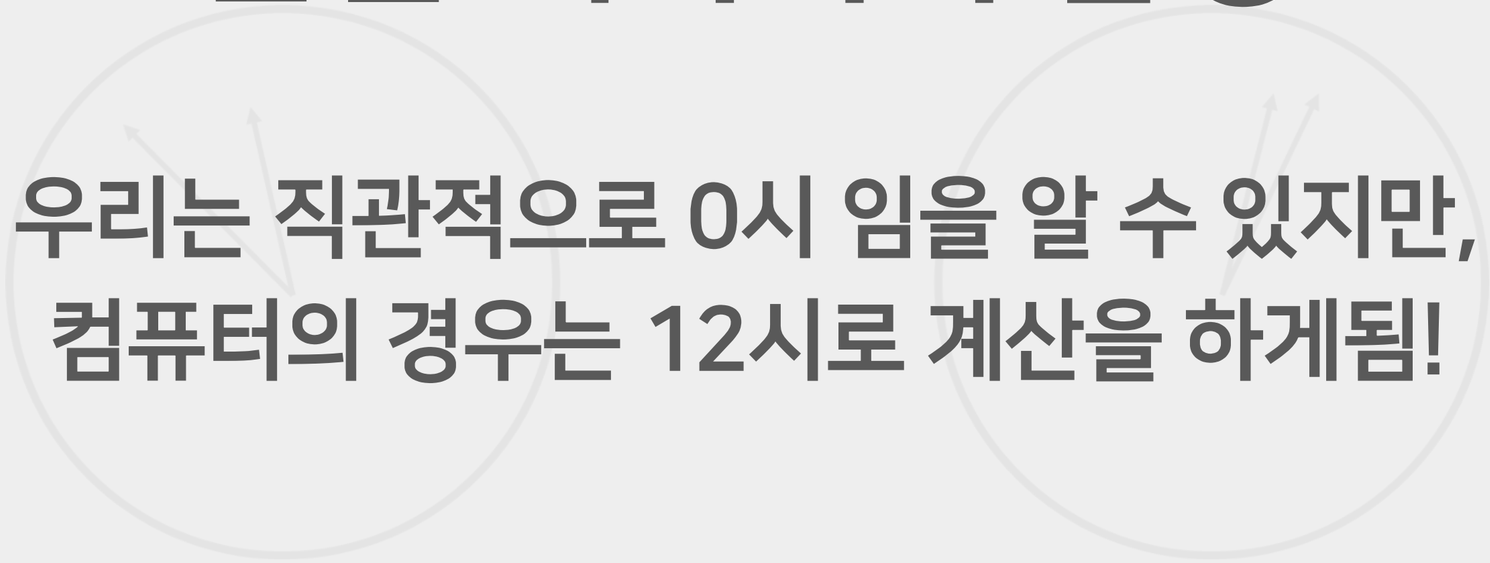


00:10

시간 데이터의 함정

다음 두 시간의 평균은 얼마일까요?

순환 데이터의 함정



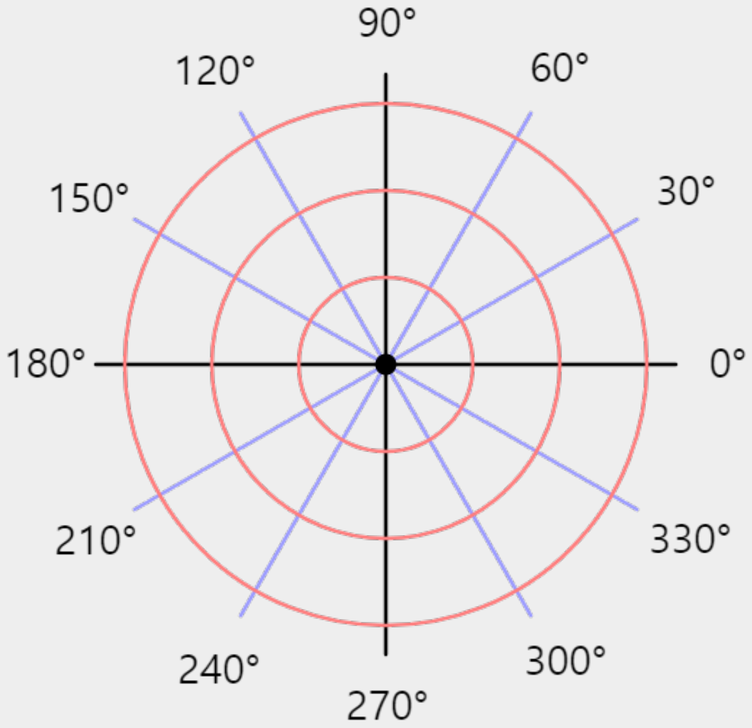
우리는 직관적으로 0시 임을 알 수 있지만,
컴퓨터의 경우는 12시로 계산을 하게됨!

시간 데이터의 함정

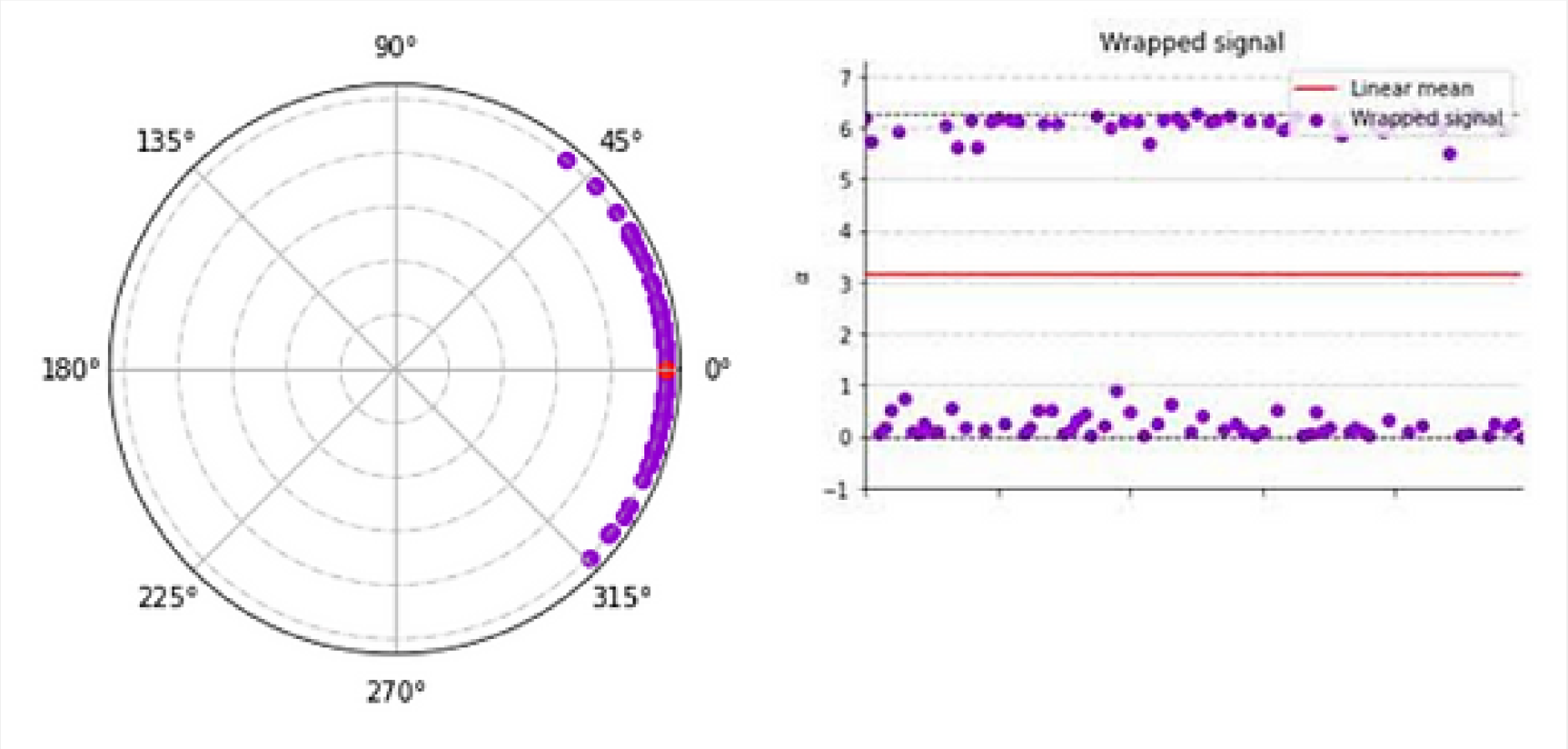
극좌표 변환

극좌표란?

한 점을 원점까지의 거리 r과
x축의 양의 부분에서
반시계방향으로 이루어지는
각도 θ를 이용하여
(r, θ)를 표현하는 방식



시간 데이터의 함정



좌표값의 중간지점을 찾아 평균값을 수월하게 얻을 수 있다

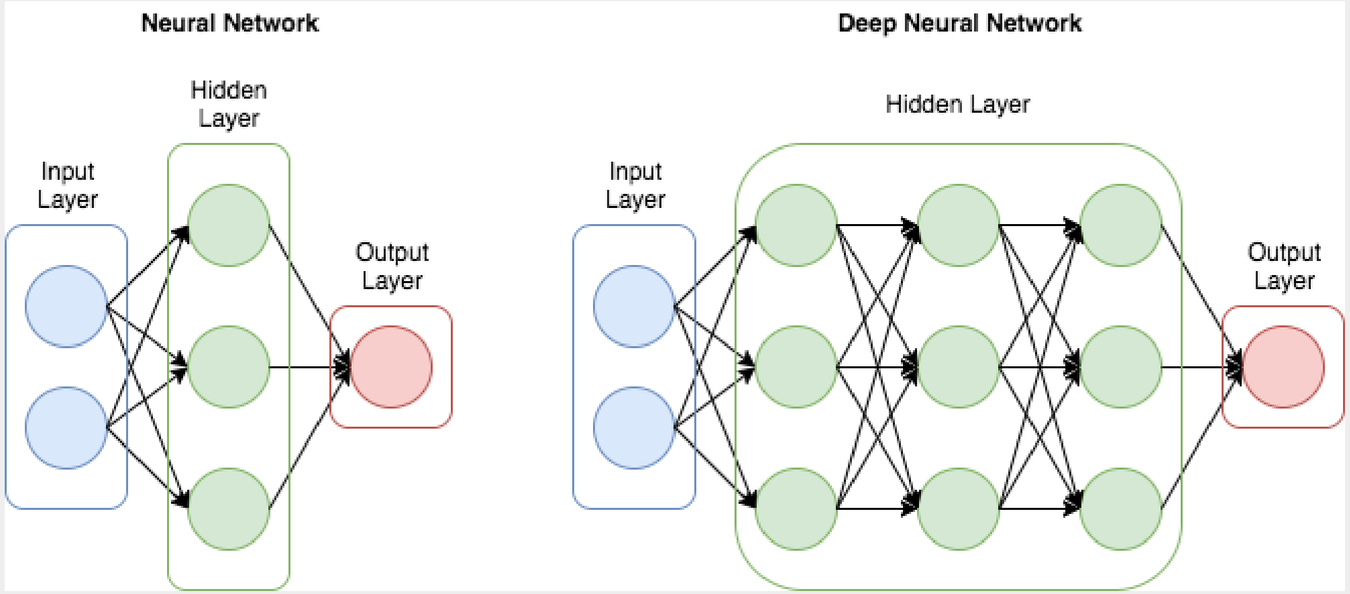
Model

DNN

DNN?

간단한 딥러닝 모델

입력층과 출력층 사이에
여러개의 은닉층들로 이루어진
인공신경망



DNN

DNN을 이용한 성과

Logloss

0.6132553363

추후 목표

코드리뷰

스터디

데이터 시각화

QnA