LG Aimers
1,Feb

Sunjun Hwang
RAISE Lab

목차

데이터 형식

```
test_data = pd.read_csv('./dataset/test.csv')
test_df = pd.DataFrame(data=test_data)
train_data = pd.read_csv('./dataset/train.csv')
train_df = pd.DataFrame(data=train_data)
print(test_df)
print(train_df)
```

-> 먼저 간단하게 csv 파일을 읽어와 DataFrame으로 만들었다.

데이터 형식

```
              ID 시술 시기 코드 시술 당시 나이  ...  난자 혼합 경과일 배아 이식 경과일 배아 해동 경과일
0      TEST_00000   TRYBLT  만35-37세  ...        0.0       NaN       NaN
1      TEST_00001   TRDQAZ  만18-34세  ...        0.0       NaN       NaN
2      TEST_00002   TRCMWS  만40-42세  ...        0.0       3.0       NaN
3      TEST_00003   TRJXFG  만40-42세  ...        0.0       2.0       NaN
4      TEST_00004   TRJXFG  만35-37세  ...        0.0       5.0       NaN
...           ...      ...      ...  ...        ...       ...       ...
90062  TEST_90062   TRDQAZ  만18-34세  ...        0.0       NaN       NaN
90063  TEST_90063   TRYBLT  만43-44세  ...        NaN       0.0       0.0
90064  TEST_90064   TRVNRY  만18-34세  ...        0.0       5.0       NaN
90065  TEST_90065   TRCMWS  만43-44세  ...        NaN       4.0       0.0
90066  TEST_90066   TRDQAZ  만18-34세  ...        0.0       NaN       NaN

[90067 rows x 68 columns]
                ID 시술 시기 코드 시술 당시 나이  ...  배아 이식 경과일 배아 해동 경과일 임신 성공 여부
0       TRAIN_000000   TRZKPL  만18-34세  ...        3.0       NaN        0
1       TRAIN_000001   TRYBLT  만45-50세  ...        NaN       NaN        0
2       TRAIN_000002   TRVNRY  만18-34세  ...        2.0       NaN        0
3       TRAIN_000003   TRJXFG  만35-37세  ...        NaN       NaN        0
4       TRAIN_000004   TRVNRY  만18-34세  ...        3.0       NaN        0
...              ...      ...      ...  ...        ...       ...      ...
256346  TRAIN_256346   TRYBLT  만18-34세  ...        5.0       NaN        0
256347  TRAIN_256347   TRYBLT  만38-39세  ...        3.0       NaN        1
256348  TRAIN_256348   TRVNRY  만35-37세  ...        3.0       NaN        0
256349  TRAIN_256349   TRZKPL  만38-39세  ...        1.0       NaN        1
256350  TRAIN_256350   TRXQMD  만35-37세  ...        0.0       0.0        1

[256351 rows x 69 columns]
```

데이터 형식

일단, 주어진 명세서를 통해 각 데이터변수의 정보를 정리했음.

번호	컬럼명	설명	범주형 여부	범주 정보
1	시술 시기 코드	난임 시술을 받은 시기를 기준으로 코드 부여	1	['TRCMWS', 'TRDQAZ', 'TRJXFG', ...]
2	시술 당시 나이	환자의 시술 당시 나이(연령대)	1	['만18-34세', '만35-37세', ...]
3	임신 시도 또는 마지막 임신 경과 연수	환자가 처음 임신을 시도한 시점 또는 마지막 임신 이후 현재까지의 경과 연수 (년 단위)	0	없음
4	시술 유형	IVF 또는 DI 시술 여부	1	['DI', 'IVF']
5	특정 시술 유형	IVF – 체외 수정, ICSI – 세포질 내 정자 주입 등 시술 유형에 대한 설명	-	없음
...	...	...	...	...
34	다른 변수 34번	설명 34번	1	['Category1', 'Category2', ...]

데이터의 수가 68개 이므로 직접 보시는걸 권장함.

모델 선정

각자 알고 있는 모델을 이용해서 분석해보시는걸 권장드립니다.

시계열 데이터로 분석할 수 있으므로 시계열 데이터를 분석하기에 적합한 모델인
RNN, LSTM, Tranformer, Arima, Sarima 등
최대한 많은 모델을 이용해서 각자 분석하는 시간을 가졌으면 합니다.

Thanks!

Do you have any questions?

sunjun7559012@yonsei.ac.kr
010 -8240-7559 | https://sites.google.com/view/seonjunhwang