

# Bigdata Analytics

Final Report

INE5015 – 22057

DO NOT LEAK THIS DOCUMENT

**빅데이터 애널리틱스 8조**

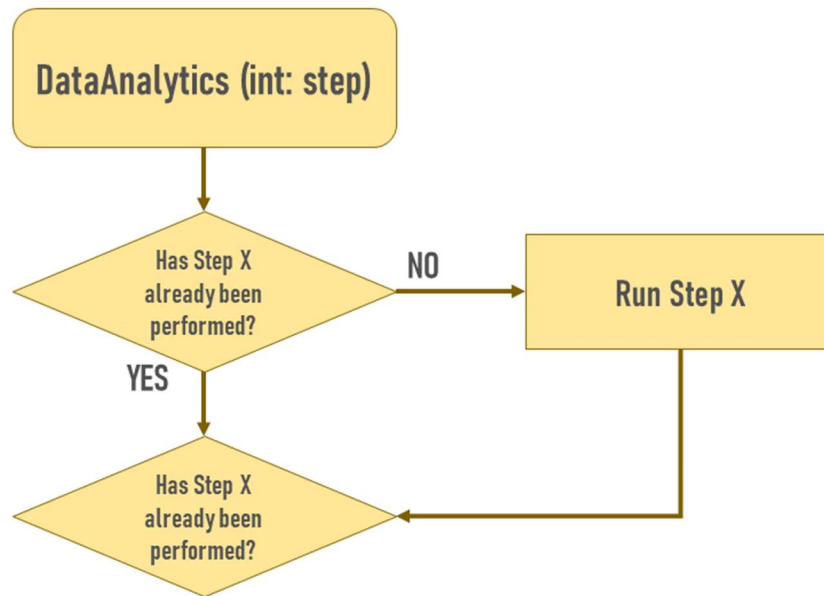
| 최준희 | 이강산 | 장혜연 | 황태영 |

# List of Contents

- Logic of Data Preprocessing
- step-by-step process
  - Step 0 ~ Step 2 : raw file refining
  - Data Cleaning Overview (3 Steps)
    - ◆ Step 3 : correlation check and correction
    - ◆ Step 4 : Missing Value Imputation
    - ◆ Step 5 : Outlier corrections
  - Step 6 : Data Scaling
- Feature Selection
- Data over/down Sampling overview
- Performance
  - The performance of two samplings
  - Prediction accuracy according to each algorithm
    - ◆ Logistic Regression
    - ◆ Decision Tree
    - ◆ Random Forest
    - ◆ Boosting

## Logic of our Data Processing

전반적인 Preprocessing 과정은 기능, 목적에 따라 Step으로 구별했고, 매 Step마다 시각적, 통계적 데이터를 확인하면서 진행하였다. 아래는 Step에 대한 구성 Logic이다.



Proceed with 8 steps as above

DataAnalytics 함수에서 필요한 함수를 호출하여 사용하는 구조이며, 각 단계마다 csv 파일을 저장하도록 하여 변화를 직관적으로 확인하기 용이하도록 디자인하였다.

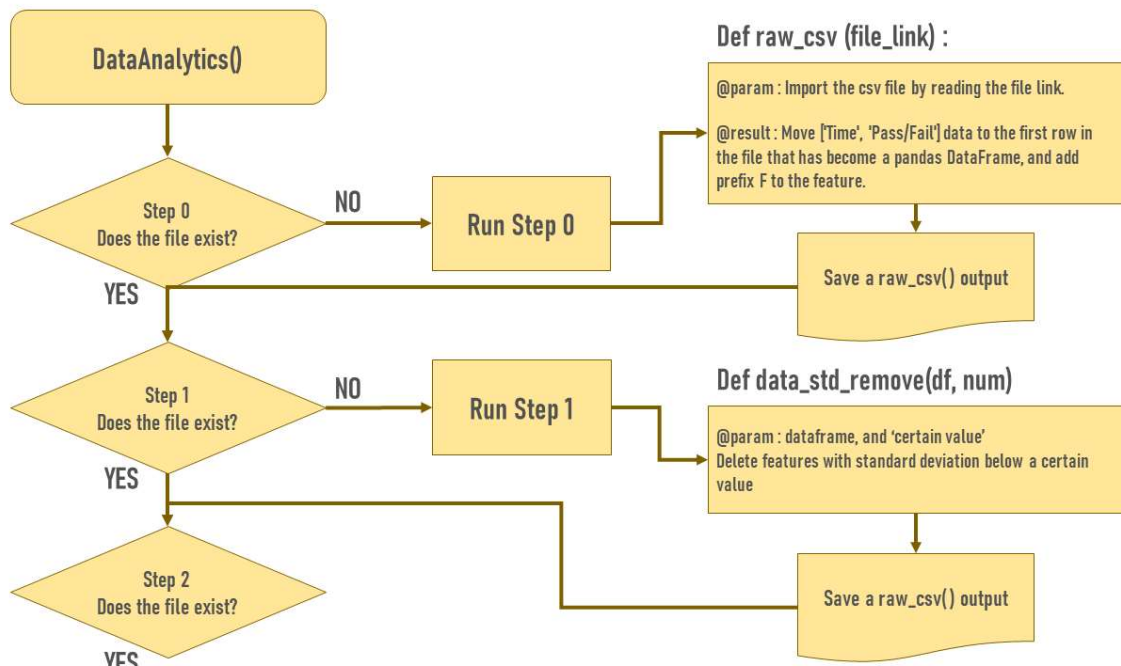
각 Step을 크게 나누어 보면 Raw data를 가공에 용이하도록 조정하는 부분, Pass/Fail 데이터를 분리해 총 3가지 데이터셋으로 나누어 진행할 수 있도록 하는 부분이 있다.

이후 각 Feature (독립변수)간 종속성을 확인하고, 제거하기 위해 진행하는 Correlation 과, 결측치 처리/보정 과정, 이상치 처리/보정 과정. 즉, 데이터 클리닝 과정이 있다.

마지막으로 Feature Selection과 샘플링을 통해 최종 데이터 셋 후보를 선정하고, 퍼포먼스를 확인하는 과정으로 마무리한다.

## Step 0 ~ Step 2 : Ready to Preprocess

이 단계는 데이터 파일을 불러오고, 이후 보정에 있어 편의성을 높이기 위해 보정한다.



### Explanation of Steps

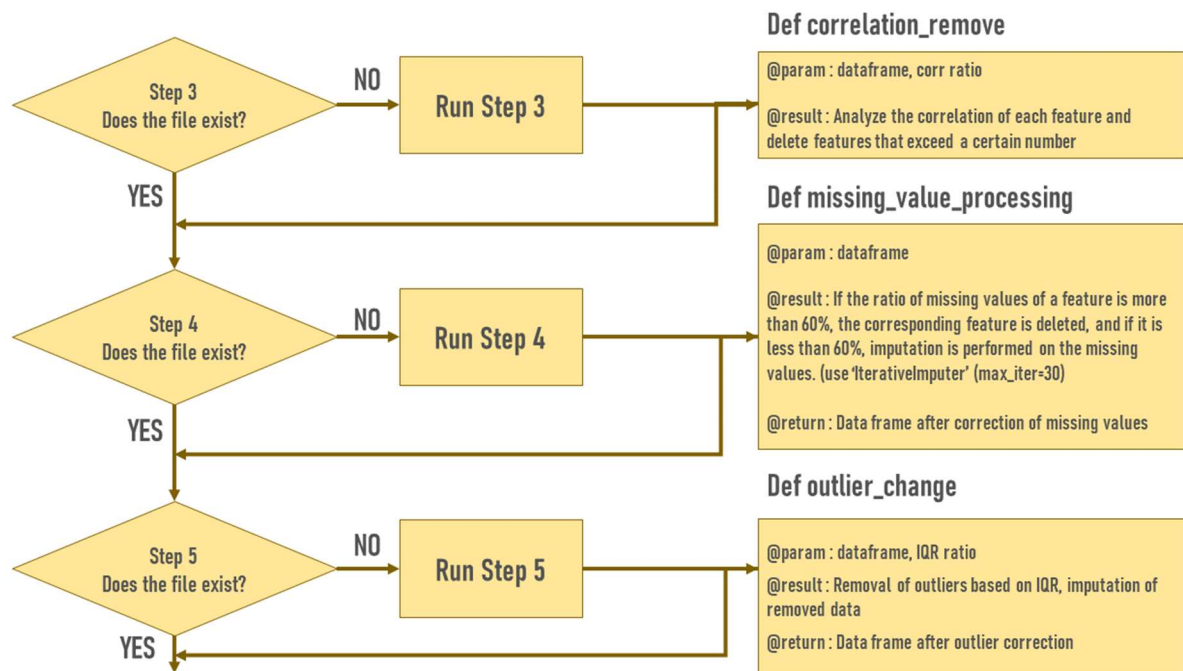
STEP	STEP DESCRIPTION
STEP 0	raw_csv 함수를 통해, Pandas Read_CSV를 실행하여 데이터 셋을 받아온다. 이후 작업의 편의를 위해 행 위치 변경, Prefix 추가 등의 과정을 거친 Dataframe 파일을 받고, 저장한다.
STEP 1	일정 계수 미만의 표준편차를 가진, 특징 선택의 필요도가 낮은 Feature들을 제거하고, Dataframe 파일을 저장한다. 그리고 데이터 셋을 3개로 나누는 과정을 거친다.
STEP 2	데이터 셋은 Pass Data, Fail Data, Both Data로 나뉘며, 이렇게 데이터 셋을 나누는 것은 프로젝트 초기의 아이디어 중 유용한, 유의미한 결과를 가져오는 계기가 된다.

Step 0 부터 Step 2까지의 3 과정은 데이터 전 처리를 위한 준비 단계에 불과하다.

> 위 과정에서 1566개 (1463+104)의 테스트 케이스, 247개의 Feature Data가 남았다.

## Step 3 ~ Step 5 : Data Cleaning

이 단계에서는 데이터셋의 노이즈를 제거하기 위해 데이터 클리닝을 진행하는 과정이다. 크게 독립변수 간 종속성 (비율), 결측치 제거 및 보정, 이상치 제거 및 보정으로 나뉜다.



### Explanation of Steps

STEP	STEP DESCRIPTION
STEP 3	각 독립변수 (Feature)간 상관성이 높다는 것은 다중공선성 발생 소지가 있고, 정확한 예측을 방해하는 요인으로 작용할 수 있다. 상관관계가 높은 Feature들을 제거하는 과정이다.
STEP 4	데이터에 결측치가 많아도 정확한 예측을 방해한다. 독립변수의 수를 최대한 유지하기 위해, 결측치가 60% 이상인 Feature를 제거하고, 60% 미만은 IterativeImputer를 통해 보정한다.
STEP 5	각 Feature의 사분위 값을 확인하고, IQR 방식을 통해 양 극단에 있는 이상치들을 제거한다. 이 때, Fail 데이터셋은 오버샘플링을 진행하지 않았기 때문에 가중치를 달리 설정한다.

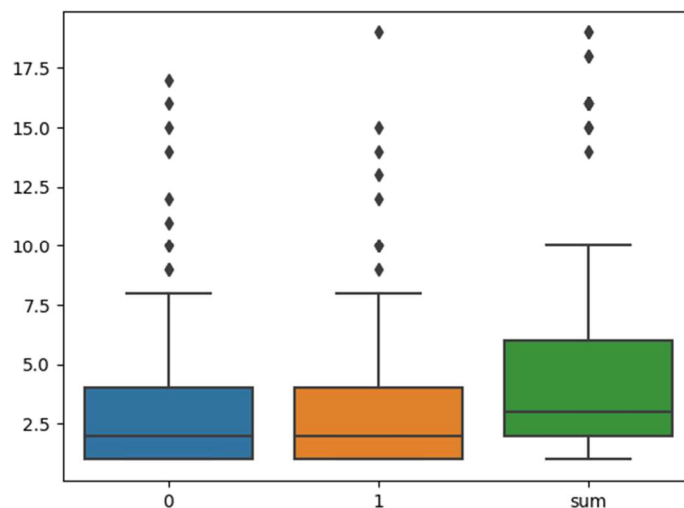
다음 페이지에는 Step3, 4, 5의 구체적인 내용과 편차 제거의 효용성을 비교한다.

## Data Cleaning - Overview

데이터 전처리 과정에 있어, 데이터 클리닝 과정은 매우 중요하다. 특히 결측치와 이상치에 대해 처리하는 것은 데이터 분석 및 모델링 결과를 크게 변화할 수 있기 때문에, 데이터의 불안전성과 잡음, 불일치 등을 최대한 효과적으로 처리해야 한다.

### Step 3 - Correlation

상관관계를 분석 (Correlation Analysis)을 한다는 것은 통계학적으로 두 변수간 선형적 관계를 분석하는 것이다. 독립 변수 (=Feature)간 상관관계가 높다면 두 변수의 연관성이 높다는 것이고, 필요치 않은 연산을 할 뿐만 아니라 Clustering 에도 방해가 된다.



우리는 이 프로젝트 과정에서 상관관계가 높은 Feature들을 계산했고 아래 예시처럼 선택된 Feature들을 제거했다. (여기서, Fail 데이터와 All/Pass 데이터간 상관관계 분석 내용이 달라 약 20여개의 Feature가 Fail 데이터에는 남아있게 되었다.

(시각화 자료 들어갈 위치)

## Step 4 - Missing Value

이 프로젝트에 있는 결측치는 패턴이 없는, Random Missing Feature 라 가정하고 진행하였다. 결측치를 처리하는 방법에는 삭제, 대치, 예측이 있는데, 결측치들의 특성이 패턴을 가지고 있다는 가정 하에 예측 모델을 구현해야 하기 때문에 이 프로젝트에서는 Deletion, Imputation 두 가지 방법을 사용하였다.

Scikit Learn에서 제공하고 있는 impute 중 Iterative Imputer를 사용하였다. 다른 모든 특성에서 개별 특성을 추정하는 다변량 대치 방식이며, Round Robin 알고리즘으로 각 Feature를 모델링 하여 결측값을 대치하는 기능을 한다.

또한 Iterative Imputer는 KNN 알고리즘으로 결측치를 예측하여 채워 넣는 방식인데, Max-Iter를 30으로 조정함으로써 최종 라운드 동안 계산된 결과를 반환하기전에, 수행할 Round의 수를 늘림으로써 데이터의 완전성이 높아지기를 기대하였다.

```

133 # 결측치 보충 단계
134 s3_c30_all = pd.read_csv('./[step 3] - correlation/s3_c30_all.csv')
135 s3_c30_pass = pd.read_csv('./[step 3] - correlation/s3_c30_pass.csv')
136 s3_c30_fail = pd.read_csv('./[step 3] - correlation/s3_c30_fail.csv')
137
138 #결측치 처리 -> 0.6 이상 비율인 경우 삭제
139 #결측치 처리 -> 0.6 미만 비율인 경우 Imputation (max-iter=30)
140
141
142 s4_MVP_all = missing_value_processing(s3_c30_all)
143 s4_MVP_pass = missing_value_processing(s3_c30_pass)
144 s4_MVP_fail = missing_value_processing(s3_c30_fail)
145
146 s4_MVP_all.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_all')
147 s4_MVP_pass.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_pa')
148 s4_MVP_fail.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_fa')
149
150
2008-07-19 13:17    1  2932.61 ...  9.2721  3.1745  82.860200
2008-07-19 14:43   -1  2988.72 ...  8.5831  2.0544  73.843200
2008-07-19 15:22   -1  3032.24 ... 10.9698  99.3032  73.843200
...
2008-10-16 15:13   -1  2899.41 ... 11.7256  2.8669  203.172000
2008-10-16 20:49   -1  3052.31 ... 17.8379  2.6238  203.172000
2008-10-17 5:26    -1  2978.81 ... 17.7267  3.0590  43.523100
2008-10-17 6:01    -1  2894.92 ... 19.2104  3.5662  93.494100
2008-10-17 6:07    -1  2944.92 ... 22.9183  3.6275 137.784400
1567 rows x 181 columns

```

ALL

```

...
1458  2008-10-16 15:13    -1  2899.41 ... 11.7256  2.8669  203.172000
1459  2008-10-16 20:49   -1  3052.31 ... 17.8379  2.6238  203.172000
1460  2008-10-17 5:26    -1  2978.81 ... 17.7267  3.0590  43.523100
1461  2008-10-17 6:01    -1  2894.92 ... 19.2104  3.5662  93.494100
1462  2008-10-17 6:07    -1  2944.92 ... 22.9183  3.6275 137.784400
[1463 rows x 181 columns]

```

PASS

```

..
99   2008-06-10 15:00    1  2988.39 ...  0.000000  3.0992  0.0000
100  2008-07-10 13:10    1  3052.98 ... 52.701400  3.1106  52.7014
101  2008-09-10 4:34    1  2951.84 ... 93.125224  2.3773  67.7994
102  2008-09-10 15:55    1  3173.18 ... 88.152800  3.8289  88.1528
103  2008-10-15 2:42    1  2903.34 ... 85.312200  3.5250  85.3122
[104 rows x 201 columns]

```

FAIL

Normal (ALL) 데이터와 Fail 데이터가 다른 이유는, STEP3 에서 약 20여개의 Feature가 덜 제거된 것으로 보인다. (이 페이지는 수정될 예정입니다)

## Step 5 - Outlier Value

무야호~

여기서는 Step 3, Step 4와 다르게 편차가 달라지기 때문에, 편차 제거를 한번 더 진행하였다.

### Step 3 / 4

```

119 # step 3 / 4.0은 의미 없음 / step 3만
120 s3_c30_all = correlation_remove(std30_all,
121 print("----- 필터링 -----")
122 print(s3_c30_all)
123
124 s3_c30_all_2 = data_std_remove(s3_c30_all,
125 print("----- 필터링 -----")
126 print(s3_c30_all_2)
127 print("----- 필터링 -----")
128 return
129 df.to_csv('./[step 0] - rawfile_low_refine/
130 s3_c30_all.to_csv('./[step 3] - correlation
131
1567 rows x 202 columns]
----- 필터링 -----
Time Pass/Fail F0 ...
2008-07-19 11:55 -1 3030.93 ...
2008-07-19 12:32 -1 3095.78 ...
2008-07-19 13:17 1 2932.61 ...
2008-07-19 14:43 -1 2988.72 ...
2008-07-19 15:22 -1 3032.24 ...
...
562 2008-10-16 15:13 -1 2899.41 ...
563 2008-10-16 20:49 -1 3052.31 ...
564 2008-10-17 5:26 -1 2978.81 ...
565 2008-10-17 6:01 -1 2894.92 ...
566 2008-10-17 6:07 -1 2944.92 ...
1567 rows x 202 columns]

158 s4_all = pd.read_csv('./[step 4] - DC
159 s4_pass = pd.read_csv('./[step 4] - DC
160 s4_fail = pd.read_csv('./[step 4] - DC
161
162 print("----- 필터링 -----")
163 print(s4_all)
164
165 s4_all_2 = data_std_remove(s4_all, 1)
166 print("----- 필터링 -----")
167 print(s4_all_2)
168 print("----- 필터링 -----")
169 return
170
171 #이상치 처리 -> fail 데이터가 너무 적어
172
173 s5_w15_p50_pass = outlier_change(s4_p
174
1567 rows x 181 columns]
----- 필터링 -----
Time Pass/Fail F0 ...
2008-07-19 11:55 -1 3030.93 ...
2008-07-19 12:32 -1 3095.78 ...
2008-07-19 13:17 1 2932.61 ...
2008-07-19 14:43 -1 2988.72 ...
2008-07-19 15:22 -1 3032.24 ...
...
562 2008-10-16 15:13 -1 2899.41 ...
563 2008-10-16 20:49 -1 3052.31 ...
564 2008-10-17 5:26 -1 2978.81 ...
565 2008-10-17 6:01 -1 2894.92 ...
566 2008-10-17 6:07 -1 2944.92 ...
1567 rows x 181 columns]

```

### STEP 5 (유의미!)

```

155 # (step - 5)
156 #이상치 필터링 단계
157
158 s4_all = pd.read_csv('./[step 4] - DC
159 s4_pass = pd.read_csv('./[step 4] - DC
160 s4_fail = pd.read_csv('./[step 4] - DC
161
162 print("----- 필터링 -----")
163 print(s4_all)
164
165 s4_all_2 = data_std_remove(s4_all, 0.1)
166 print("----- 필터링 -----")
167 print(s4_all_2)
168 print("----- 필터링 -----")
169 return
170
171 #이상치 처리 -> fail 데이터가
172
173 s5_w15_p50_pass = outlier_change(s
174 s5_w15_p50_fail = outlier_ch
175 s5_w15_p50_all = outlier_ch
176
177 print("----- 필터링 -----")
178 print(s5_w15_p50_all)
179
180 s4_all_3 = data_std_remove(s
181 print("----- 필터링 -----")
182 print(s4_all_3)
183 print("----- 필터링 -----")
184 return
185
186 # 이상치 제거 과정에서 의미있는
187 # 2000번의 데이터는 제거 (기존)
188
1567 rows x 181 columns]
----- 필터링 -----
Time Pass/Fail F0 ...
2008-07-19 11:55 -1 30
2008-07-19 12:32 -1 30
2008-07-19 13:17 1 29
2008-07-19 14:43 -1 29
2008-07-19 15:22 -1 30
...
562 2008-10-16 15:13 -1 28
563 2008-10-16 20:49 -1 30
564 2008-10-17 5:26 -1 29
565 2008-10-17 6:01 -1 28
566 2008-10-17 6:07 -1 29
1567 rows x 153 columns]

189 print("----- 필터링 -----")
190 print(s4_all_3)
191
192 s4_all_4 = data_std_remove(s4_all_3, 0.1)
193 print("----- 필터링 -----")
194 print(s4_all_4)
195 print("----- 필터링 -----")
196 return
197
198 #이상치 처리 -> fail 데이터가 너무 적어
199
200 s5_w15_p50_pass = outlier_change(s4_p
201 s5_w15_p50_fail = outlier_ch
202 s5_w15_p50_all = outlier_ch
203
204 print("----- 필터링 -----")
205 print(s5_w15_p50_all)
206
207 s4_all_5 = data_std_remove(s4_all_4, 0.1)
208 print("----- 필터링 -----")
209 print(s4_all_5)
210 print("----- 필터링 -----")
211 return
212
213 #이상치 처리 -> fail 데이터가 너무 적어
214
215 s5_w15_p50_pass = outlier_change(s4_p
216 s5_w15_p50_fail = outlier_ch
217 s5_w15_p50_all = outlier_ch
218
219 print("----- 필터링 -----")
220 print(s5_w15_p50_all)
221
222 s4_all_6 = data_std_remove(s4_all_5, 0.1)
223 print("----- 필터링 -----")
224 print(s4_all_6)
225 print("----- 필터링 -----")
226 return
227
228 #이상치 처리 -> fail 데이터가 너무 적어
229
230 s5_w15_p50_pass = outlier_change(s4_p
231 s5_w15_p50_fail = outlier_ch
232 s5_w15_p50_all = outlier_ch
233
234 print("----- 필터링 -----")
235 print(s5_w15_p50_all)
236
237 s4_all_7 = data_std_remove(s4_all_6, 0.1)
238 print("----- 필터링 -----")
239 print(s4_all_7)
240 print("----- 필터링 -----")
241 return
242
243 #이상치 처리 -> fail 데이터가 너무 적어
244
245 s5_w15_p50_pass = outlier_change(s4_p
246 s5_w15_p50_fail = outlier_ch
247 s5_w15_p50_all = outlier_ch
248
249 print("----- 필터링 -----")
250 print(s5_w15_p50_all)
251
252 s4_all_8 = data_std_remove(s4_all_7, 0.1)
253 print("----- 필터링 -----")
254 print(s4_all_8)
255 print("----- 필터링 -----")
256 return
257
258 #이상치 처리 -> fail 데이터가 너무 적어
259
260 s5_w15_p50_pass = outlier_change(s4_p
261 s5_w15_p50_fail = outlier_ch
262 s5_w15_p50_all = outlier_ch
263
264 print("----- 필터링 -----")
265 print(s5_w15_p50_all)
266
267 s4_all_9 = data_std_remove(s4_all_8, 0.1)
268 print("----- 필터링 -----")
269 print(s4_all_9)
270 print("----- 필터링 -----")
271 return
272
273 #이상치 처리 -> fail 데이터가 너무 적어
274
275 s5_w15_p50_pass = outlier_change(s4_p
276 s5_w15_p50_fail = outlier_ch
277 s5_w15_p50_all = outlier_ch
278
279 print("----- 필터링 -----")
280 print(s5_w15_p50_all)
281
282 s4_all_10 = data_std_remove(s4_all_9, 0.1)
283 print("----- 필터링 -----")
284 print(s4_all_10)
285 print("----- 필터링 -----")
286 return
287
288 #이상치 처리 -> fail 데이터가 너무 적어
289
290 s5_w15_p50_pass = outlier_change(s4_p
291 s5_w15_p50_fail = outlier_ch
292 s5_w15_p50_all = outlier_ch
293
294 print("----- 필터링 -----")
295 print(s5_w15_p50_all)
296
297 s4_all_11 = data_std_remove(s4_all_10, 0.1)
298 print("----- 필터링 -----")
299 print(s4_all_11)
300 print("----- 필터링 -----")
301 return
302
303 #이상치 처리 -> fail 데이터가 너무 적어
304
305 s5_w15_p50_pass = outlier_change(s4_p
306 s5_w15_p50_fail = outlier_ch
307 s5_w15_p50_all = outlier_ch
308
309 print("----- 필터링 -----")
310 print(s5_w15_p50_all)
311
312 s4_all_12 = data_std_remove(s4_all_11, 0.1)
313 print("----- 필터링 -----")
314 print(s4_all_12)
315 print("----- 필터링 -----")
316 return
317
318 #이상치 처리 -> fail 데이터가 너무 적어
319
320 s5_w15_p50_pass = outlier_change(s4_p
321 s5_w15_p50_fail = outlier_ch
322 s5_w15_p50_all = outlier_ch
323
324 print("----- 필터링 -----")
325 print(s5_w15_p50_all)
326
327 s4_all_13 = data_std_remove(s4_all_12, 0.1)
328 print("----- 필터링 -----")
329 print(s4_all_13)
330 print("----- 필터링 -----")
331 return
332
333 #이상치 처리 -> fail 데이터가 너무 적어
334
335 s5_w15_p50_pass = outlier_change(s4_p
336 s5_w15_p50_fail = outlier_ch
337 s5_w15_p50_all = outlier_ch
338
339 print("----- 필터링 -----")
340 print(s5_w15_p50_all)
341
342 s4_all_14 = data_std_remove(s4_all_13, 0.1)
343 print("----- 필터링 -----")
344 print(s4_all_14)
345 print("----- 필터링 -----")
346 return
347
348 #이상치 처리 -> fail 데이터가 너무 적어
349
350 s5_w15_p50_pass = outlier_change(s4_p
351 s5_w15_p50_fail = outlier_ch
352 s5_w15_p50_all = outlier_ch
353
354 print("----- 필터링 -----")
355 print(s5_w15_p50_all)
356
357 s4_all_15 = data_std_remove(s4_all_14, 0.1)
358 print("----- 필터링 -----")
359 print(s4_all_15)
360 print("----- 필터링 -----")
361 return
362
363 #이상치 처리 -> fail 데이터가 너무 적어
364
365 s5_w15_p50_pass = outlier_change(s4_p
366 s5_w15_p50_fail = outlier_ch
367 s5_w15_p50_all = outlier_ch
368
369 print("----- 필터링 -----")
370 print(s5_w15_p50_all)
371
372 s4_all_16 = data_std_remove(s4_all_15, 0.1)
373 print("----- 필터링 -----")
374 print(s4_all_16)
375 print("----- 필터링 -----")
376 return
377
378 #이상치 처리 -> fail 데이터가 너무 적어
379
380 s5_w15_p50_pass = outlier_change(s4_p
381 s5_w15_p50_fail = outlier_ch
382 s5_w15_p50_all = outlier_ch
383
384 print("----- 필터링 -----")
385 print(s5_w15_p50_all)
386
387 s4_all_17 = data_std_remove(s4_all_16, 0.1)
388 print("----- 필터링 -----")
389 print(s4_all_17)
390 print("----- 필터링 -----")
391 return
392
393 #이상치 처리 -> fail 데이터가 너무 적어
394
395 s5_w15_p50_pass = outlier_change(s4_p
396 s5_w15_p50_fail = outlier_ch
397 s5_w15_p50_all = outlier_ch
398
399 print("----- 필터링 -----")
400 print(s5_w15_p50_all)
401
402 s4_all_18 = data_std_remove(s4_all_17, 0.1)
403 print("----- 필터링 -----")
404 print(s4_all_18)
405 print("----- 필터링 -----")
406 return
407
408 #이상치 처리 -> fail 데이터가 너무 적어
409
410 s5_w15_p50_pass = outlier_change(s4_p
411 s5_w15_p50_fail = outlier_ch
412 s5_w15_p50_all = outlier_ch
413
414 print("----- 필터링 -----")
415 print(s5_w15_p50_all)
416
417 s4_all_19 = data_std_remove(s4_all_18, 0.1)
418 print("----- 필터링 -----")
419 print(s4_all_19)
420 print("----- 필터링 -----")
421 return
422
423 #이상치 처리 -> fail 데이터가 너무 적어
424
425 s5_w15_p50_pass = outlier_change(s4_p
426 s5_w15_p50_fail = outlier_ch
427 s5_w15_p50_all = outlier_ch
428
429 print("----- 필터링 -----")
430 print(s5_w15_p50_all)
431
432 s4_all_20 = data_std_remove(s4_all_19, 0.1)
433 print("----- 필터링 -----")
434 print(s4_all_20)
435 print("----- 필터링 -----")
436 return
437
438 #이상치 처리 -> fail 데이터가 너무 적어
439
440 s5_w15_p50_pass = outlier_change(s4_p
441 s5_w15_p50_fail = outlier_ch
442 s5_w15_p50_all = outlier_ch
443
444 print("----- 필터링 -----")
445 print(s5_w15_p50_all)
446
447 s4_all_21 = data_std_remove(s4_all_20, 0.1)
448 print("----- 필터링 -----")
449 print(s4_all_21)
450 print("----- 필터링 -----")
451 return
452
453 #이상치 처리 -> fail 데이터가 너무 적어
454
455 s5_w15_p50_pass = outlier_change(s4_p
456 s5_w15_p50_fail = outlier_ch
457 s5_w15_p50_all = outlier_ch
458
459 print("----- 필터링 -----")
460 print(s5_w15_p50_all)
461
462 s4_all_22 = data_std_remove(s4_all_21, 0.1)
463 print("----- 필터링 -----")
464 print(s4_all_22)
465 print("----- 필터링 -----")
466 return
467
468 #이상치 처리 -> fail 데이터가 너무 적어
469
470 s5_w15_p50_pass = outlier_change(s4_p
471 s5_w15_p50_fail = outlier_ch
472 s5_w15_p50_all = outlier_ch
473
474 print("----- 필터링 -----")
475 print(s5_w15_p50_all)
476
477 s4_all_23 = data_std_remove(s4_all_22, 0.1)
478 print("----- 필터링 -----")
479 print(s4_all_23)
480 print("----- 필터링 -----")
481 return
482
483 #이상치 처리 -> fail 데이터가 너무 적어
484
485 s5_w15_p50_pass = outlier_change(s4_p
486 s5_w15_p50_fail = outlier_ch
487 s5_w15_p50_all = outlier_ch
488
489 print("----- 필터링 -----")
490 print(s5_w15_p50_all)
491
492 s4_all_24 = data_std_remove(s4_all_23, 0.1)
493 print("----- 필터링 -----")
494 print(s4_all_24)
495 print("----- 필터링 -----")
496 return
497
498 #이상치 처리 -> fail 데이터가 너무 적어
499
500 s5_w15_p50_pass = outlier_change(s4_p
501 s5_w15_p50_fail = outlier_ch
502 s5_w15_p50_all = outlier_ch
503
504 print("----- 필터링 -----")
505 print(s5_w15_p50_all)
506
507 s4_all_25 = data_std_remove(s4_all_24, 0.1)
508 print("----- 필터링 -----")
509 print(s4_all_25)
510 print("----- 필터링 -----")
511 return
512
513 #이상치 처리 -> fail 데이터가 너무 적어
514
515 s5_w15_p50_pass = outlier_change(s4_p
516 s5_w15_p50_fail = outlier_ch
517 s5_w15_p50_all = outlier_ch
518
519 print("----- 필터링 -----")
520 print(s5_w15_p50_all)
521
522 s4_all_26 = data_std_remove(s4_all_25, 0.1)
523 print("----- 필터링 -----")
524 print(s4_all_26)
525 print("----- 필터링 -----")
526 return
527
528 #이상치 처리 -> fail 데이터가 너무 적어
529
530 s5_w15_p50_pass = outlier_change(s4_p
531 s5_w15_p50_fail = outlier_ch
532 s5_w15_p50_all = outlier_ch
533
534 print("----- 필터링 -----")
535 print(s5_w15_p50_all)
536
537 s4_all_27 = data_std_remove(s4_all_26, 0.1)
538 print("----- 필터링 -----")
539 print(s4_all_27)
540 print("----- 필터링 -----")
541 return
542
543 #이상치 처리 -> fail 데이터가 너무 적어
544
545 s5_w15_p50_pass = outlier_change(s4_p
546 s5_w15_p50_fail = outlier_ch
547 s5_w15_p50_all = outlier_ch
548
549 print("----- 필터링 -----")
550 print(s5_w15_p50_all)
551
552 s4_all_28 = data_std_remove(s4_all_27, 0.1)
553 print("----- 필터링 -----")
554 print(s4_all_28)
555 print("----- 필터링 -----")
556 return
557
558 #이상치 처리 -> fail 데이터가 너무 적어
559
560 s5_w15_p50_pass = outlier_change(s4_p
561 s5_w15_p50_fail = outlier_ch
562 s5_w15_p50_all = outlier_ch
563
564 print("----- 필터링 -----")
565 print(s5_w15_p50_all)
566
567 s4_all_29 = data_std_remove(s4_all_28, 0.1)
568 print("----- 필터링 -----")
569 print(s4_all_29)
570 print("----- 필터링 -----")
571 return
572
573 #이상치 처리 -> fail 데이터가 너무 적어
574
575 s5_w15_p50_pass = outlier_change(s4_p
576 s5_w15_p50_fail = outlier_ch
577 s5_w15_p50_all = outlier_ch
578
579 print("----- 필터링 -----")
580 print(s5_w15_p50_all)
581
582 s4_all_30 = data_std_remove(s4_all_29, 0.1)
583 print("----- 필터링 -----")
584 print(s4_all_30)
585 print("----- 필터링 -----")
586 return
587
588 #이상치 처리 -> fail 데이터가 너무 적어
589
590 s5_w15_p50_pass = outlier_change(s4_p
591 s5_w15_p50_fail = outlier_ch
592 s5_w15_p50_all = outlier_ch
593
594 print("----- 필터링 -----")
595 print(s5_w15_p50_all)
596
597 s4_all_31 = data_std_remove(s4_all_30, 0.1)
598 print("----- 필터링 -----")
599 print(s4_all_31)
600 print("----- 필터링 -----")
601 return
602
603 #이상치 처리 -> fail 데이터가 너무 적어
604
605 s5_w15_p50_pass = outlier_change(s4_p
606 s5_w15_p50_fail = outlier_ch
607 s5_w15_p50_all = outlier_ch
608
609 print("----- 필터링 -----")
610 print(s5_w15_p50_all)
611
612 s4_all_32 = data_std_remove(s4_all_31, 0.1)
613 print("----- 필터링 -----")
614 print(s4_all_32)
615 print("----- 필터링 -----")
616 return
617
618 #이상치 처리 -> fail 데이터가 너무 적어
619
620 s5_w15_p50_pass = outlier_change(s4_p
621 s5_w15_p50_fail = outlier_ch
622 s5_w15_p50_all = outlier_ch
623
624 print("----- 필터링 -----")
625 print(s5_w15_p50_all)
626
627 s4_all_33 = data_std_remove(s4_all_32, 0.1)
628 print("----- 필터링 -----")
629 print(s4_all_33)
630 print("----- 필터링 -----")
631 return
632
633 #이상치 처리 -> fail 데이터가 너무 적어
634
635 s5_w15_p50_pass = outlier_change(s4_p
636 s5_w15_p50_fail = outlier_ch
637 s5_w15_p50_all = outlier_ch
638
639 print("----- 필터링 -----")
640 print(s5_w15_p50_all)
641
642 s4_all_34 = data_std_remove(s4_all_33, 0.1)
643 print("----- 필터링 -----")
644 print(s4_all_34)
645 print("----- 필터링 -----")
646 return
647
648 #이상치 처리 -> fail 데이터가 너무 적어
649
650 s5_w15_p50_pass = outlier_change(s4_p
651 s5_w15_p50_fail = outlier_ch
652 s5_w15_p50_all = outlier_ch
653
654 print("----- 필터링 -----")
655 print(s5_w15_p50_all)
656
657 s4_all_35 = data_std_remove(s4_all_34, 0.1)
658 print("----- 필터링 -----")
659 print(s4_all_35)
660 print("----- 필터링 -----")
661 return
662
663 #이상치 처리 -> fail 데이터가 너무 적어
664
665 s5_w15_p50_pass = outlier_change(s4_p
666 s5_w15_p50_fail = outlier_ch
667 s5_w15_p50_all = outlier_ch
668
669 print("----- 필터링 -----")
670 print(s5_w15_p50_all)
671
672 s4_all_36 = data_std_remove(s4_all_35, 0.1)
673 print("----- 필터링 -----")
674 print(s4_all_36)
675 print("----- 필터링 -----")
676 return
677
678 #이상치 처리 -> fail 데이터가 너무 적어
679
680 s5_w15_p50_pass = outlier_change(s4_p
681 s5_w15_p50_fail = outlier_ch
682 s5_w15_p50_all = outlier_ch
683
684 print("----- 필터링 -----")
685 print(s5_w15_p50_all)
686
687 s4_all_37 = data_std_remove(s4_all_36, 0.1)
688 print("----- 필터링 -----")
689 print(s4_all_37)
690 print("----- 필터링 -----")
691 return
692
693 #이상치 처리 -> fail 데이터가 너무 적어
694
695 s5_w15_p50_pass = outlier_change(s4_p
696 s5_w15_p50_fail = outlier_ch
697 s5_w15_p50_all = outlier_ch
698
699 print("----- 필터링 -----")
700 print(s5_w15_p50_all)
701
702 s4_all_38 = data_std_remove(s4_all_37, 0.1)
703 print("----- 필터링 -----")
704 print(s4_all_38)
705 print("----- 필터링 -----")
706 return
707
708 #이상치 처리 -> fail 데이터가 너무 적어
709
710 s5_w15_p50_pass = outlier_change(s4_p
711 s5_w15_p50_fail = outlier_ch
712 s5_w15_p50_all = outlier_ch
713
714 print("----- 필터링 -----")
715 print(s5_w15_p50_all)
716
717 s4_all_39 = data_std_remove(s4_all_38, 0.1)
718 print("----- 필터링 -----")
719 print(s4_all_39)
720 print("----- 필터링 -----")
721 return
722
723 #이상치 처리 -> fail 데이터가 너무 적어
724
725 s5_w15_p50_pass = outlier_change(s4_p
726 s5_w15_p50_fail = outlier_ch
727 s5_w15_p50_all = outlier_ch
728
729 print("----- 필터링 -----")
730 print(s5_w15_p50_all)
731
732 s4_all_40 = data_std_remove(s4_all_39, 0.1)
733 print("----- 필터링 -----")
734 print(s4_all_40)
735 print("----- 필터링 -----")
736 return
737
738 #이상치 처리 -> fail 데이터가 너무 적어
739
740 s5_w15_p50_pass = outlier_change(s4_p
741 s5_w15_p50_fail = outlier_ch
742 s5_w15_p50_all = outlier_ch
743
744 print("----- 필터링 -----")
745 print(s5_w15_p50_all)
746
747 s4_all_41 = data_std_remove(s4_all_40, 0.1)
748 print("----- 필터링 -----")
749 print(s4_all_41)
750 print("----- 필터링 -----")
751 return
752
753 #이상치 처리 -> fail 데이터가 너무 적어
754
755 s5_w15_p50_pass = outlier_change(s4_p
756 s5_w15_p50_fail = outlier_ch
757 s5_w15_p50_all = outlier_ch
758
759 print("----- 필터링 -----")
760 print(s5_w15_p50_all)
761
762 s4_all_42 = data_std_remove(s4_all_41, 0.1)
763 print("----- 필터링 -----")
764 print(s4_all_42)
765 print("----- 필터링 -----")
766 return
767
768 #이상치 처리 -> fail 데이터가 너무 적어
769
770 s5_w15_p50_pass = outlier_change(s4_p
771 s5_w15_p50_fail = outlier_ch
772 s5_w15_p50_all = outlier_ch
773
774 print("----- 필터링 -----")
775 print(s5_w15_p50_all)
776
777 s4_all_43 = data_std_remove(s4_all_42, 0.1)
778 print("----- 필터링 -----")
779 print(s4_all_43)
780 print("----- 필터링 -----")
781 return
782
783 #이상치 처리 -> fail 데이터가 너무 적어
784
785 s5_w15_p50_pass = outlier_change(s4_p
786 s5_w15_p50_fail = outlier_ch
787 s5_w15_p50_all = outlier_ch
788
789 print("----- 필터링 -----")
790 print(s5_w15_p50_all)
791
792 s4_all_44 = data_std_remove(s4_all_43, 0.1)
793 print("----- 필터링 -----")
794 print(s4_all_44)
795 print("----- 필터링 -----")
796 return
797
798 #이상치 처리 -> fail 데이터가 너무 적어
799
800 s5_w15_p50_pass = outlier_change(s4_p
801 s5_w15_p50_fail = outlier_ch
802 s5_w15_p50_all = outlier_ch
803
804 print("----- 필터링 -----")
805 print(s5_w15_p50_all)
806
807 s4_all_45 = data_std_remove(s4_all_44, 0.1)
808 print("----- 필터링 -----")
809 print(s4_all_45)
810 print("----- 필터링 -----")
811 return
812
813 #이상치 처리 -> fail 데이터가 너무 적어
814
815 s5_w15_p50_pass = outlier_change(s4_p
816 s5_w15_p50_fail = outlier_ch
817 s5_w15_p50_all = outlier_ch
818
819 print("----- 필터링 -----")
820 print(s5_w15_p50_all)
821
822 s4_all_46 = data_std_remove(s4_all_45, 0.1)
823 print("----- 필터링 -----")
824 print(s4_all_46)
825 print("----- 필터링 -----")
826 return
827
828 #이상치 처리 -> fail 데이터가 너무 적어
829
830 s5_w15_p50_pass = outlier_change(s4_p
831 s5_w15_p50_fail = outlier_ch
832 s5_w15_p50_all = outlier_ch
833
834 print("----- 필터링 -----")
835 print(s5_w15_p50_all)
836
837 s4_all_47 = data_std_remove(s4_all_46, 0.1)
838 print("----- 필터링 -----")
839 print(s4_all_47)
840 print("----- 필터링 -----")
841 return
842
843 #이상치 처리 -> fail 데이터가 너무 적어
844
845 s5_w15_p50_pass = outlier_change(s4_p
846 s5_w15_p50_fail = outlier_ch
847 s5_w15_p50_all = outlier_ch
848
849 print("----- 필터링 -----")
850 print(s5_w15_p50_all)
851
852 s4_all_48 = data_std_remove(s4_all_47, 0.1)
853 print("----- 필터링 -----")
854 print(s4_all_48)
855 print("----- 필터링 -----")
856 return
857
858 #이상치 처리 -> fail 데이터가 너무 적어
859
860 s5_w15_p50_pass = outlier_change(s4_p
861 s5_w15_p50_fail = outlier_ch
862 s5_w15_p50_all = outlier_ch
863
864 print("----- 필터링 -----")
865 print(s5_w15_p50_all)
866
867 s4_all_49 = data_std_remove(s4_all_48, 0.1)
868 print("----- 필터링 -----")
869 print(s4_all_49)
870 print("----- 필터링 -----")
871 return
872
873 #이상치 처리 -> fail 데이터가 너무 적어
874
875 s5_w15_p50_pass = outlier_change(s4_p
876 s5_w15_p50_fail = outlier_ch
877 s5_w15_p50_all = outlier_ch
878
879 print("----- 필터링 -----")
880 print(s5_w15_p50_all)
881
882 s4_all_50 = data_std_remove(s4_all_49, 0.1)
883 print("----- 필터링 -----")
884 print(s4_all_50)
885 print("----- 필터링 -----")
886 return
887
888 #이상치 처리 -> fail 데이터가 너무 적어
889
890 s5_w15_p50_pass = outlier_change(s4_p
891 s5_w15_p50_fail = outlier_ch
892 s5_w15_p50_all = outlier_ch
893
894 print("----- 필터링 -----")
895 print(s5_w15_p50_all)
896
897 s4_all_51 = data_std_remove(s4_all_50, 0.1)
898 print("----- 필터링 -----")
899 print(s4_all_51)
900 print("----- 필터링 -----")
901 return
902
903 #이상치 처리 -> fail 데이터가 너무 적어
904
905 s5_w15_p50_pass = outlier_change(s4_p
906 s5_w15_p50_fail = outlier_ch
907 s5_w15_p50_all = outlier_ch
908
909 print("----- 필터링 -----")
910 print(s5_w15_p50_all)
911
912 s4_all_52 = data_std_remove(s4_all_51, 0.1)
913 print("----- 필터링 -----")
914 print(s4_all_52)
915 print("----- 필터링 -----")
916 return
917
918 #이상치 처리 -> fail 데이터가 너무 적어
919
920 s5_w15_p50_pass = outlier_change(s4_p
921 s5_w15_p50_fail = outlier_ch
922 s5_w15_p50_all = outlier_ch
923
924 print("----- 필터링 -----")
925 print(s5_w15_p50_all)
926
927 s4_all_53 = data_std_remove(s4_all_52, 0.1)
928 print("----- 필터링 -----")
929 print(s4_all_53)
930 print("----- 필터링 -----")
931 return
932
933 #이상치 처리 -> fail 데이터가 너무 적어
934
935 s5_w15_p50_pass = outlier_change(s4_p
936 s5_w15_p50_fail = outlier_ch
937 s5_w15_p50_all = outlier_ch
938
939 print("----- 필터링 -----")
940 print(s5_w15_p50_all)
941
942 s4_all_54 = data_std_remove(s4_all_53, 0.1)
943 print("----- 필터링 -----")
944 print(s4_all_54)
945 print("----- 필터링 -----")
946 return
947
948 #이상치 처리 -> fail 데이터가 너무 적어
949
950 s5_w15_p50_pass = outlier_change(s4_p
951 s5_w15_p50_fail = outlier_ch
952 s5_w15_p50_all = outlier_ch
953
954 print("----- 필터링 -----")
955 print(s5_w15_p50_all)
956
957 s4_all_55 = data_std_remove(s4_all_54, 0.1)
958 print("----- 필터링 -----")
959 print(s4_all_55)
960 print("----- 필터링 -----")
961 return
962
963 #이상치 처리 -> fail 데이터가 너무 적어
964
965 s5_w15_p50_pass = outlier_change(s4_p
966 s5_w15_p50_fail = outlier_ch
967 s5_w15_p50_all = outlier_ch
968
969 print("----- 필터링 -----")
970 print(s5_w15_p50_all)
971
972 s4_all_56 = data_std_remove(s4_all_55, 0.1)
973 print("----- 필터링 -----")
974 print(s4_all_56)
975 print("----- 필터링 -----")
976 return
977
978 #이상치 처리 -> fail 데이터가 너무 적어
979
980 s5_w15_p50_pass = outlier_change(s4_p
981 s5_w15_p50_fail = outlier_ch
982 s5_w15_p50_all = outlier_ch
983
984 print("----- 필터링 -----")
985 print(s5_w15_p50_all)
986
987 s4_all_57 = data_std_remove(s4_all_56, 0.1)
988 print("----- 필터링 -----")
989 print(s4_all_57)
990 print("----- 필터링 -----")
991 return
992
993 #이상치 처리 -> fail 데이터가 너무 적어
994
995 s5_w15_p50_pass = outlier_change(s4_p
996 s5_w15_p50_fail = outlier_ch
997 s5_w15_p50_all = outlier_ch
998
999 print("----- 필터링 -----")
1000 print(s5_w15_p50_all)
1001
1002 s4_all_58 = data_std_remove(s4_all_57, 0.1)
1003 print("----- 필터링 -----")
1004 print(s4_all_58)
1005 print("----- 필터링 -----")
1006 return
1007
1008 #이상치 처리 -> fail 데이터가 너무 적어
1009
1010 s5_w15_p50_pass = outlier_change(s4_p
1011 s5_w15_p50_fail = outlier_ch
1012 s5_w15_p50_all = outlier_ch
1013
1014 print("----- 필터링 -----")
1015 print(s5_w15_p50_all)
1016
1017 s4_all_59 = data_std_remove(s4_all_58, 0.1)
1018 print("----- 필터링 -----")
1019 print(s4_all_59)
1020 print("----- 필터링 -----")
1021 return
1022
1023 #이상치 처리 -> fail 데이터가 너무 적어
1024
1025 s5_w15_p50_pass = outlier_change(s4_p
1026 s5_w15_p50_fail = outlier_ch
1027 s5_w15_p50_all = outlier_ch
1028
1029 print("----- 필터링 -----")
1030 print(s5_w15_p50_all)
1031
1032 s4_all_60 = data_std_remove(s4_all_59, 0.1)
1033 print("----- 필터링 -----")
1034 print(s4_all_60)
1035 print("----- 필터링 -----")
1036 return
1037
1038 #이상치 처리 -> fail 데이터가 너무 적어
1039
1040 s5_w15_p50_pass = outlier_change(s4_p
1041 s5_w15_p50_fail = outlier_ch
1042 s5_w15_p50_all = outlier_ch
1043
1044 print("----- 필터링 -----")
1045 print(s5_w15_p50_all)
1046
1047 s4_all_61 = data_std_remove(s4_all_60, 0.1)
1048 print("----- 필터링 -----")
1049 print(s4_all_61)
1050 print("----- 필터링 -----")
1051 return
1052
1053 #이상치 처리 -> fail 데이터가 너무 적어
1054
1055 s5_w15_p50_pass = outlier_change(s4_p
1056 s5_w15_p50_fail = outlier_ch
1057 s5_w15_p50_all = outlier_ch
1058
1059 print("----- 필터링 -----")
1060 print(s5_w15_p50_all)
1061
1062 s4_all_62 = data_std_remove(s4_all_61, 0.1)
1063 print("----- 필터링 -----")
1064 print(s4_all_62)
1065 print("----- 필터링 -----")
106
```



## Feature Selection

여기서는 RFE, KBS 내용 설명

DO NOT LEAK THIS DOCUMENT

## 오버샘플링

STD와 MMS 간 차이가 없음

오버샘플링 SMOTE 랜덤 100회 평균 (test\_size = 0.2)

acc, pre, rec, f1, roc\_curve

```
RFE_MMS_ALL  
0.8061 0.7864 0.8414 0.8128 0.8807
```

```
RFE_STD_ALL  
0.8022 0.7848 0.8338 0.8084 0.8793
```

```
KBS_MMS_ALL  
0.7999 0.7774 0.8414 0.808 0.873
```

```
KBS_STD_ALL  
0.7997 0.7774 0.8408 0.8077 0.8728
```

RFE > KBS, MMS와 STD는 큰 차이 없음.

성능 평가

각 모델에 맞추어 다시 성능을 평가 해야 함.

DO NOT LEAK THIS DOCUMENT