Feature Importance Using Logistic Regression:
(Semiconductor Manufacturing Process Data)

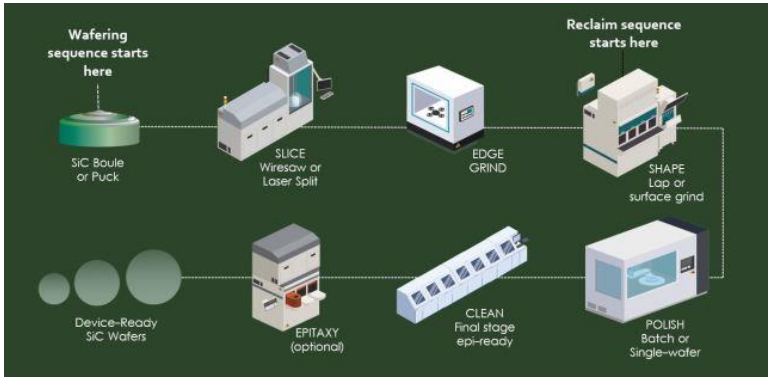
주건재

OVERVIEW

1. INTRODUCTION
2. DATA
3. MODEL
4. RESULT

1. INTRODUCTION

1.1 TOPIC



Research Question

Wafer 제조 공정 중 수집한 센서데이터로 Wafer결함을 예측하고 Logistic Regression 모델을 통해 중요한 Feature를 선택한다.

Goals

1st goal: Highly Imbalance 데이터를 적절한 방법으로 처리한다.

2nd goal: 적절한 Feature의 개수를 정의하고 선택한다.

1.2 DATA OVERVIEW

Data Description

독립변수: Wafer 공정 중 수집한 590개 Sensor Data와 Time

종속변수: Wafer 불량 여부 (-1:양품, 1: 불량)

(각 Sensor가 무엇을 측정하는지는 모름)

Data Resource: uci ML Repository (https://archive.ics.uci.edu/ml/datasets/SECOM)

	Time	0	1	2	3	4	5	6	7	8	...	581	582	583	584	585	586	587	588	589	Pass/Fail
0	2008-07-19 11:55:00	3030.93	2564.00	2187.7333	1411.1265	1.3602	100.0	97.6133	0.1242	1.5005	...	NaN	0.5005	0.0118	0.0035	2.3630	NaN	NaN	NaN	NaN	-1
1	2008-07-19 12:32:00	3095.78	2465.14	2230.4222	1463.6606	0.8294	100.0	102.3433	0.1247	1.4966	...	208.2045	0.5019	0.0223	0.0055	4.4447	0.0096	0.0201	0.0060	208.2045	-1

Data Summary

```
Data shape: (1567, 592)
----------------------------
Data Types:
float64    590
object       1
int64        1
Name: Types, dtype: int64
----------------------------
```

	Types	Counts	Uniques	Nulls	Min	Max
0	float64	1561	1521	6	2743.24	3356.35
1	float64	1560	1505	7	2158.75	2846.44
10	float64	1565	393	2	-0.0349	0.053
100	float64	1561	36	6	-0.003	0.0023
101	float64	1561	30	6	-0.0024	0.0017
...	...	...	...	...	...	...
97	float64	1561	2	6	0	0
98	float64	1561	1421	6	-5.2717	2.5698
99	float64	1561	273	6	-0.5283	0.8854
Pass/Fail	int64	1567	2	0	-1	1
Time	object	1567	1534	0	2008-01-08 02:02:00	2008-12-10 18:47:00

1.3 OUT LINE

작업	3월 22일	3월 29일	4월 5일	4월 12일	4월 19일
데이터 찾기					
문제 정의					
1차 발표					
전처리					
EDA					
Regression 모델링					
Feature간 상관성 확인 및 선택					
후행연구 제안					
발표자료					
최종 발표					
리포트 작성					