회귀분석을 이용한 반도체 웨이퍼 기울기 조절의 영향 인자 도출 및 개선

Deriving and improving Factors Affecting the Adjustment of Semiconductor Wafer Slope Using Regression Analysis

길준혁, 이나영, 신윤경, 배장원
한국기술교육대학교 산업경영학부
{gjhkms00929, 4yeong0318, rt0326, jangwon_bae}@koreatech.ac.kr

Abstract

사회가 발전함에 따라 반도체의 수요가 많아지고 있다. 많은 기업에서 반도체 공정에서 사용하는 웨이퍼의 기울기에 대한 수요가 증가하고 있으며, 기업의 다양한 수요에 맞는 반도체 제조 능력이 중요해지고 있다. 본 연구는 원하는 반도체 웨이퍼 기울기를 제작하는데 요구되는 반도체 공정의 영향 인자 도출 및 개선을 목적으로 하며, 이를 위해 실제 기업의 반도체 공정 데이터를 회귀분석을 통해 웨이퍼의 기울기에 영향을 주는 주요한 요인들을 분석하였다. 분석 결과 확인한 인자와 웨이퍼 기울기 사이의 유의미한 관계를 볼 수 있었으며, 이를 응용하여 원하는 기울기를 맞추는 데 필요한 적절한 대응 방법을 제시할 수 있었다.

1. 서론

1.1 연구 배경

전 세계적인 반도체 시장의 규모는 장기적으로 보았을 때 지속적으로 성장하고 있다. 앞으로의 반도체 시장은 5G, 인공지능(AI) 등의 첨단 산업 분야가 성장하는 만큼 반도체 수요는 증가할 것으로 전망된다[4]. 반도체 산업이 성장하는 만큼 웨이퍼의 품질은 중요해지고 있다. 최근 반도체 공정에서 중요한 부분 중 하나는 더욱 섬세한 반도체 생산을 위해 기업이 원하는 웨이퍼의 기울기를 정교하게 제작할 수 있는 능력이다. 따라서, 웨이퍼의 기울기 대한 정밀도는 점점 더 많은 수요 기업들의 요구사항으로 자리매김이 되고 있다. 하지만 공정에서 사용하는 웨이퍼가 더욱 넓어지고 더욱 얇아지는 상황[3]이기 때문에 웨이퍼의 기울기를 정확하게 제작하는 것은 더욱 어려워지고 있다. 따라서, 시장의 변화하는 수요를 따라가기 위해서 웨이퍼의 기울기에 영향을 주는 주요한 요인이 무엇인지 파악할 필요가 있으며, 수요 기업의 다양한 요구사항에 능동적으로 대응하기 위해서 빅데이터 기술을 활용하는 것이 필요하다.

1.2 연구 동향

실리콘 웨이퍼 연삭의 평탄도 특성 평가 연구 중 Sangchul Kim 등[2]은 웨이퍼 평탄도가 악화되는 것을 보완하는 방법으로 틸팅각을 부여하였고 평탄도를 향상시키기 위해 틸팅각의 변화에 따라 형상을 임의로 제어하는 형상 시뮬레이션을 제작하였다.

Wafer Backside Grinding 공정의 총두께 변화 제어 및 최적화 연구 중 Yuanhang Liu 등[5]은 간결한 스핀들 자세 조정 작업을 통해 원하는 웨이퍼 두께의 균일성을 달성하기 위

해서 전체 두께 변동(TTV)을 위한 효과적인 전략을 제안하였다. TTV 최적화 실험을 수행하였고 TTV 제어 전략이 웨이퍼 두께 균일성 향상에 중요한 역할을 하는 것을 입증하였다.

1.3 반도체 제조 공정 및 측정 데이터

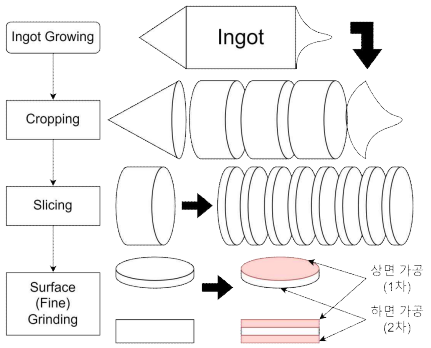

그림 1 두께 가공 공정

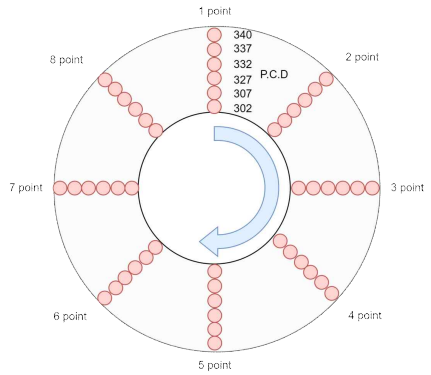

그림 2 웨이퍼 P.C.D 및 point

반도체 8대 공정 중 첫 번째 단계인 웨이퍼 제조 공정 데이터를 사용하였다. 해당 과정은 그림 1과 같으며 잉곳을 생성하는 Ingot Grwoing, 잉곳을 두꺼운 원기둥 형태로 자르는 Cropping, 두꺼운 원기둥을 얇은 원판 형태로 자르는 Slicing, 얇은 원판 형태인 웨이퍼의 표면을 매끄럽고 평평하게 가공하는 Surface Grinding 순서로 진행된다. 마지막 과정은 Fine Grinding 과정이라고도 하며, 해당 과정의 데이터를 H사로부터 받아 사용하였다. Fine Grinding 데이터는 크게 상면 가공(1차)과 하면 가공(2차)으로 나누어져 있으며, 그림 2와 같이 모두 각 P.C.D의 여러 포인트에 대한 측정값 데이터를 가지고 있다. P.C.D란 웨이퍼의 중앙에서 외곽으로 이동하면서 두께 값을 측정하는데 이때, 측정하는 구간의 명칭을 P.C.D라고 한다. 또한, 각 P.C.D 값을 시계방향으로 총 8번 측정하는데 각각을 포인트라고 한다. 즉, 하나의 P.C.D에 대해서 8번 측정하게 된다.

1차 가공과 2차 가공의 차이점은 2차에서 다루는 P.C.D 값 중 처음 값과 마지막 값을 1차에서 사용하므로 사용하는 P.C.D의 개수가 다르다. 추가적으로, H사에서 제공받은 데이터는 제품 특징에 따라 총 4개의 제품 코드에 대한 데이터를 받았는데 제품 코드마다 2차에서 사용하는 P.C.D의 종류와 개수가 다르게 구성되어 있다.

1차와 2차 파일은 이러한 측정값 외에도 여러 계산 과정을 거친 값들을 가지고 있다. 1차와 2차 데이터 모두 계산값(difference)과 기울기값(Slope) 그리고 각 P.C.D가 범위 안에 들어오는지 판단하는 spec_judgment를 가지고 있으며, 1차의 경우 2차 공정에서 기울기 조정이 필요한 Tilt_judgment를 포함하고 있다. 웨이퍼의 최종 기울기를 정확하게 제작하는데 필요한 중요한 값이다. H사에서 판단한 Tilt_judgment값에 영향을 주는 요소는 1차 기울기(Slope_1), 2차 기울기(Slope_2), Slope_2 - Slope_1 값인 1차와 2차 기울기 변화량(Slope_change)이 있다.

1.4 연구 목적 및 방법

본 연구는 빅데이터 기반 원하는 기울기를 제작하는데 요구되는 반도체 공정의 영향 인자 도출 및 개선을 목적으로 한다. H사로부터 제공받은 데이터를 python을 통해 제품 코드별로 구분할 수 있도록 한 후, 최종적으로 데이터프레임 형태로 나타내었다. 이후 EDA를 통해 영향 인자가 될 수 있는 데이터에 대해 파악하였고, 원하는 기울기를 정확하게 제작하기 위한 Tilt_judgment값에 영향을 주는 인자들을 Linear Regression을 통해 분석하였다. 또한 최적의 영향 인자를 찾기 위해 변수 선택법을 사용하였다. 모델의 성능을 파악하기 위한 지표로 MAPE를 사용하였고, 계산 결과

MAPE가 낮은 요인들에 대해서 유의미한 결과로 판단하였다.

2. 공정 데이터 전처리 모듈 개발

2.1 데이터 전처리

H사로부터 받은 raw 데이터는 4개의 제품 코드 파일 안에 1차와 2차로 구분된 후 날짜별로 정리되어 있다. 제품마다 P.C.D 측정값, 계산값, 기울기값을 가지며 1차의 경우 Tilt_judgment를 추가로 가진다. 제품 코드마다 파일 경로가 다르고 모든 파일이 1차와 2차 가공 데이터의 연결성을 쉽게 알 수 있도록 정렬되어 있지 않았다. 따라서 가장 먼저 필요한 데이터에 쉽게 접근할 수 있도록 구성하는 것이 필요하였다.

먼저 최종 파일까지의 경로와 최종 파일명을 추출할 수 있도록 구현하였으며 1차와 2차를 분리한 후 해당 파일명과 파일까지의 경로를 4개의 리스트에 분리하여 저장하였다. 데이터프레임으로 구현하기 위해서 1차와 2차 데이터를 함께 사용해야 하기에 1차와 2차 연결되는 데이터를 쌍으로 묶어 리스트에 저장할 수 있도록 하였다. 최종적으로 중복을 제거한 후 1차 2차가 쌍으로 존재하는 파일 경로 리스트와 파일명 리스트를 각각 pair, pair_file 변수에 할당하였다. 결과는 그림 3과 같다.

```
['A3R0VT-SM/1차 (상면가공 후)/11호기/1월/11일/210T8HA16021.xls'
'A3R0VT-SM/2차 (하면가공 후)/11호기/1월/11일/214D1HA16021.xls'],
['A3R0VT-SM/1차 (상면가공 후)/11호기/1월/16일/210T5HA16021.xls'
```

```
'214D3HA16021.xls', '21E13HA16021.xls',
'21DVUHA16021.xls', '21FKNHA16021.xls',
'21CDRHA16021.xls', '21DCDHA16021.xls',
```

그림 3 pair, pair_file 실행 결과 일부

2.2 데이터프레임 구성 및 결과

데이터 분석을 진행할 때 raw 데이터를 쉽고 빠르게 가져오기 위해 데이터프레임으로 구성하고 csv 파일로 저장하여 각 제품 코드마다 필요한 csv 파일을 불러와 데이터를 사용하였다.

전체적인 코드는 pandas와 numpy를 불러온 후 하나의 class 안에 여러 method로 구성하였다. class를 실행시킨 후 method에 제품 코드를 입력하여 특정 데이터에 대해 method가 독립적으로 실행될 수 있도록 하였다.

전체적으로 적용되는 원리가 있다. 첫째, 제품 코드마다 사용하는 P.C.D가 다르며, 같은 제품 코드 내 1차 가공과 2차 가공에서도 사용하는 P.C.D가 다르다. 따라서 제품 코드, 1차 가공, 2차 가공을 구분하여 사용하였다. 둘째, 각 함수의 return값을 이용해 최종적으로 하나의 데이터프레임을 구성하였다. 이때, 칼럼마다 계산 방식이 다르기에 값 자체가 존재하지 않는 부분이 있다. 이러한 부분에 대해서 'x' 혹은 'not_counting'으로 처리하여 결측치 없이 합쳐질 수 있도록 하였다. 셋째, 데이터프레임을 구성할 때, pair와 pair_file을 반복문을 사용하여 제품(파일)을 하나씩 사용하였다. 이때, 각 반복문의 값을 하나로 합치기 위해서 반복문의 마지막에 pd.concat을 사용하였다. 넷째, 데이터프레임의 인덱스는 각 제품 코드의 P.C.D와 'total'을 사용하였다. 'total'은 P.C.D 전체에 대한 값을 의미한다.

3. EDA

3.1 목적

EDA를 진행하는 목적은 다음과 같다. 첫째, 기존 데이터들의 특징 및 관계를 확인하기 위함이다. 데이터의 타입, 데이터의 분포, 이상치 혹은 결측치 유무 등 각 칼럼 값의 특징을 확인할 수 있다. 더 나아가 2개 이상의 칼럼들에 대한 값을 서로 비교하여 수치적인 관계를 확인할 수 있다. 이러한 활동을 통해 데이터를 분석하기 위해 알아야 하는 정보들을 수집할 수 있으며, 데이터 자체를 이해할 때 도움이 될 수 있고 앞으로의 분석 과정에서 유용하게 사용될 수 있다. 둘째, 최종적인 웨이퍼 기울기에 대해 Pass 혹은 Unpass를 구분하는 주요한 요인을 추가로 발견하기 위함이다. 입력으로 사용되는 x와 출력으로 사용되는 y 혹은 x와 x에 대한 유의미한 관계를 찾아 분석의 x 값으로 적당한 칼럼을 사용한다면 더욱 효율

이 좋은 분석기를 만들 수 있다. 또한, 데이터의 관계, 특징에 대해서 시각적인 자료를 사용한다면 수치적으로만 볼 때 알기 어려웠던 부분에 대해서 직관적으로 데이터를 이해할 수 있다. 이는 여러 가설을 설정하고 실험을 진행할 때 아이디어가 되어 주요한 요인들을 추가로 탐색해 볼 수 있다.

3.2 그래프 구현 및 결과

최종 데이터프레임을 sum_1에 할당하였고 가장 먼저 그림 4와 같이 info()를 통해 확인한 결과 float 형태가 아닌 칼럼이 대부분이었다. 따라서 sum_1의 특정 값을 추출 후 재구성하여 각각 sum_1_0240, sum_1_total, s_0240 변수에 할당하여 사용였다.

다음으로 다양한 기초통계를 확인하였다. 4개의 데이터프레임에 대해서 모두 info(), describe()을 사용하였고 sum_1을 대상으로 isnull().sum()과 value_counts()를 실행하였다.

이후 boxplot, distplot, heatmap, scatterplot을 통해 그래프로 표현하였다. boxplot과 distplot은 3가지의 같은 기준으로 구분하였다. 첫째는 raw 데이터를 이용하여 각 P.C.D 마다 그래프 생성, 둘째는 두께 및 계산값 칼럼이 있는 부분이 종합되어 있는 그래프 생성, 셋째는 데이터프레임의 인덱스 'total'에 데이터가 저장되어 있는 칼럼만 존재하는 그래프 생성이다. 3가지 부분의 각 목적은 다음과 같다. 첫 번째는 raw 데이터의 P.C.D 분포를 spec 판정값에 따라 분리하여 수치적인 특징을 확인하는 것과 P.C.D 별로 전체적인 비교를 진행하기 위함이다. 두 번째는 두께와 계산값이 각 칼럼, P.C.D, 가공 전후에 따라 어떻게 변화하는지 시각적으로 파악하여 어떠한 변화가 있는지 판단하기 위함이다. 세 번째는 'total'이 아닌 인덱스에는 값이 저장되어 있지 않아 다른 칼럼과 함께 비교할 수 없는 칼럼들의 분포를 보기 위해 구성하였다.

```
sum_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6120 entries, 0 to 6119
Data columns (total 49 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   file_name             6120 non-null   object
 1   PCD_name              6120 non-null   object
 2   Thickness_min_1       6120 non-null   object
 3   Thickness_max_1       6120 non-null   object
 4   Thickness_avg_1       6120 non-null   object
 5   Undulation_1          6120 non-null   object
 6   PCD_difference_Min_1  6120 non-null   object
 7   PCD_difference_Max_1  6120 non-null   object
 8   PCD_difference_Avg_1  6120 non-null   object
 9   one_UCL               6120 non-null   object
10   one_LCL               6120 non-null   object
11   One_spec_judgment     6120 non-null   object
12   Slope_1               6120 non-null   object
13   Tilt_judgment         6120 non-null   object
14   Thickness_min_2       6120 non-null   float64
15   Thickness_max_2       6120 non-null   float64
16   Thickness_avg_2       6120 non-null   float64
17   Undulation_2          6120 non-null   object
18   PCD_difference_Min_2  6120 non-null   object
19   PCD_difference_Max_2  6120 non-null   object
20   PCD_difference_Avg_2  6120 non-null   object
21   two_UCL               6120 non-null   object
22   two_LCL               6120 non-null   object
23   Two_spec_judgment     6120 non-null   object
24   Slope_2               6120 non-null   object
25   one_1                 6120 non-null   object
26   one_2                 6120 non-null   object
27   one_3                 6120 non-null   object
28   one_4                 6120 non-null   object
29   one_5                 6120 non-null   object
30   one_6                 6120 non-null   object
31   one_7                 6120 non-null   object
32   one_8                 6120 non-null   object
33   two_1                 6120 non-null   object
34   two_2                 6120 non-null   object
35   two_3                 6120 non-null   object
36   two_4                 6120 non-null   object
37   two_5                 6120 non-null   object
38   two_6                 6120 non-null   object
39   two_7                 6120 non-null   object
40   two_8                 6120 non-null   object
41   two_1_j               6120 non-null   object
42   two_2_j               6120 non-null   object
43   two_3_j               6120 non-null   object
44   two_4_j               6120 non-null   object
45   two_5_j               6120 non-null   object
46   two_6_j               6120 non-null   object
47   two_7_j               6120 non-null   object
48   two_8_j               6120 non-null   object
dtypes: float64(3), object(46)
memory usage: 2.3+ MB
```

그림 4 sum_1.info() 결과

boxplot과 distplot의 결과 중 일부는 그림 5 그림 6과 같다. boxplot과 distplot의 결과 많은 칼럼에서 극단적인 이상치가 등장함을 알

수 있었다. boxplot의 경우 이상치로 인해 꼬리부터 꼬리까지의 부분이 잘 보이지 않아 일정 수준 제외하였다. 또한 대체로 1차 공정에 비해 2차 공정에서 같은 칼럼 내의 다른 P.C.D에 대해 값의 분포가 거의 유사해짐을 알 수 있었다. 또한, 2차 공정 값이 1차 공정 값보다 두께가 얇아졌고 두께의 균일도가 상승한 것을 알 수 있었다.

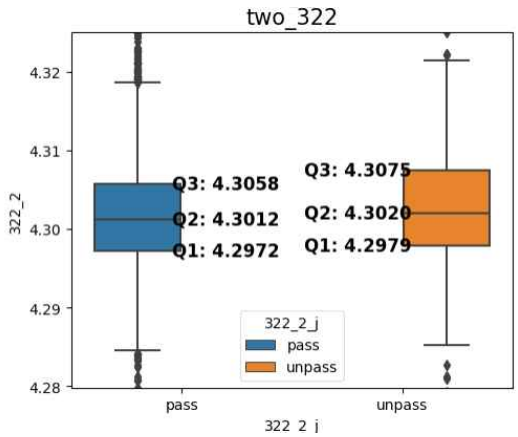

그림 5 boxplot 결과 일부

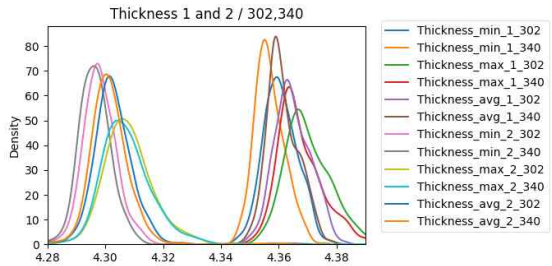

그림 6 distplot 결과 일부

scatterplot의 경우는 x와 x와의 관계와 x와 y와의 관계 두 가지 경우로 나누어 진행하였다. 변수들이 서로 어떤 연관성을 가지는지 알아보기 위함이기 때문에 모든 칼럼을 변수로 사용하지 않고 heatmap의 결과 상관계수가 0.5 이상인 칼럼들을 대상으로 진행하였다.

scatterplot을 실행시키기 위해 3개의 함수를 생성하였다. 함수의 입력값으로 x와 x 혹은 x와 y의 칼럼을 입력하도록 구성하였다. 추가로, 어떤 칼럼에 의해 구분될지 입력하도록 하여 다양한 변수들이 다양한 기준으로 표현될 수 있도록 하였다.

그림 7은 scatterplot의 결과 일부이다. pass와 unpass가 분명히 구분되는 여러 유의미한 모습을 볼 수 있었다. 유의미한 그래프의 경우 주요한 요인이 될 가능성이 높다고 판단하였다. Tilt_judgment과 가장 유의한 칼럼은 Slope_1과 계산값(difference) 칼럼이었으며, spec_judgment와 가장 유의한 칼럼은 1차와 2차 계산값(difference) 칼럼이었다. 이러한 결과는 예측 모델의 변수를 선택할 때 기준이 될 수 있으며 최적의 해를 찾을 때 도움이 될 수 있다.

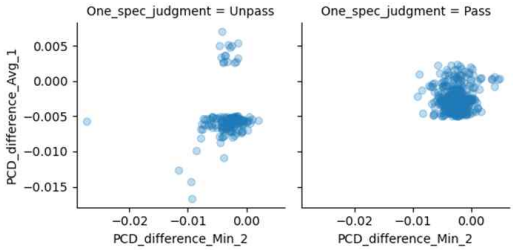

그림 7 scatterplot 결과 일부

그림 8은 클래스 함수들의 데이터프레임에서 정의한 칼럼들의 관계도이다. 이 관계도는 EDA의 결과를 종합하여 도출한 것으로, 모든 데이터가 관계성과 연결성을 가지고 있음을 확인할 수 있다.

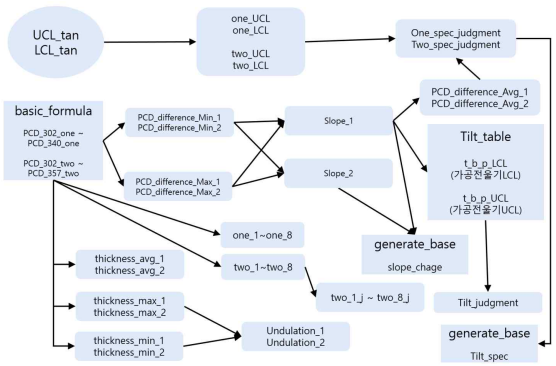

그림 8 데이터프레임 변수들 관계도

4. Linear Regression

4.1 Linear Regression 모델 개발

Linear Regression 모델을 개발하기 위해 학습 데이터와 테스트 데이터의 비율을 8:2로

설정해 구분하였다. 학습 데이터와 테스트 데이터를 비교하여 모델을 학습시키고, 성능 지표를 활용해 수치를 확인하는 과정을 거쳤다. 데이터셋 구분은 표 1과 같다.

Product_code	Train : Test = 8 : 2	
A (3,098)	Train	Test
	620	2,478
B (2,577)	Train	Test
	516	2,061
C (509)	Train	Test
	102	407
D (2,755)	Train	test
	551	2,204

표 1 제품 코드별 train, test 비율

원하는 웨이퍼의 기울기를 정확하게 제작하기 위해서 필요한 것은 Tilt_judgment이다. 따라서 회귀식에 필요한 X와 Y의 값에서 X는 주요한 요인들이 되고 Y는 Tilt_judgment가 된다. Tilt_judgment에 영향을 주는 요소로 H사에서 Slope_1, Slope_2, Slope_change이라고 판단하였다. 따라서 3가지 변수로 가능한 6가지 조합을 표 2와 같이 설정하였다.

X	Y
Slope_change	Tilt_judgment
Slope_1	
Slope_2	
Slope_1, Slope_change	
Slope_2, Slope_change	
Slope_1, Slope_2, Slope_change	

표 2 X와 Y의 6가지 조합

선형회귀식을 만들기 위해서 sklearn.linear_model의 LinearRegression를 사용하였다. matplotlib 라이브러리를 이용하여 그래프로 표현하였다. 모델의 성능을 평가하기 위한 지표로는 RMSE와 MAPE를 사용을 계획하였다. RMSE는 실제값과 예측값에 대한 오차 제곱 평균의 제곱근을 의미한다. MAPE는 실제값과 예측값 간의 백분율 오차의 평균을 나타낸다. 두 지표 모두 값이 낮을수록 모델의 예측 성능이 좋다고 할 수 있다. 또한, R-squared를 통해 모델이 예측하는 값과 실제 값의 유사도를 구하여 모델의 성능에 대한 보조 지표로써 사용하였다. R-squared는 1에 가까울수록 유사도가 높다는 것을 의미한다.

4.2 회귀식 결과

제품코드	X = slope_ change	X = slope_1	X = slope_2	X = slope_1, slope_ change	X = slope_2, slope_ change	X= slope_1, slope_2, slope_ change
A	0.4277 0.841	0.3843 0.984	0.3641 0.020	**0.3456** **0.985**	0.3705 0.985	0.4087 0.985
B	0.3839 0.825	0.3867 0.988	0.4148 0.007	0.4105 0.988	0.3811 0.988	**0.3119** **0.988**
C	0.5016 0.714	0.4086 0.969	0.4564 0.007	0.3998 0.971	0.4039 0.971	**0.3858** **0.971**
D	0.3561 0.841	0.3564 0.986	0.3853 0.009	**0.3511** **0.986**	0.4719 0.986	0.3665 0.986

표 3 MAPE와 R-squared 결과표

회귀식 결과는 표 3과 같다. 위의 값이 MAPE이며 아래의 값이 R-squared이다. 제품 코드마다 6가지의 X 조합에 대해 MAPE와 R-squared를 계산하였다. EDA를 통해 이상치가 적지 않게 존재하며 극단적인 값도 있음을 확인했기 때문에 RMSE는 사용하지 않았다[1]. 회귀식 결과 MAPE가 가장 낮고 R-squared가 가장 큰 X 조합은 A와 D는 Slope_1, Slope_change이며 B와 C는 Slope_1, Slope_2, Slope_change이라는 결과를 얻을 수 있었다. 따라서 적절한 Tilt_judgment를 위해 각 제품 코드마다 회귀식 결과와 같이 적용 및 참고한다면 더 좋은 웨이퍼의 기울기 제작이 가능하다고 예측할 수 있다.

4.3 변수 선택법 개발

X가 될 수 있는 변수는 Slope_1, Slope_2, Slope_change의 조합만이 아닌 다양한 변수가 있을 수 있다. 따라서, 더욱 다양하게 주요한 변수를 탐색할 수 있는 변수 선택법을 사용하였다. 변수 선택법을 통해서 적절한 변수를 선택한다면 모델의 성능을 향상시킬 수 있다. 변수 선택법의 방법으로 전진 선택법, 후진 소거법, 전진 단계적 선택법을 사용하였다. Y는 기

제품 코드	전진 선택법			후진 소거법			전진 단계별 선택법		
	X	MAPE	Ad-R	X	MAPE	Ad-R	X	MAPE	Ad-R
A	slope_1 slope_2	0.379	0.985	slope_1 slope_2 slope_change thickness_avg_1 thickness_min_1 thickness_max_2 undulaion_2	0.302	0.985	slope_1 slope_2	0.405	0.985
B	slope_1	0.315	0.988	slope_1 slope_2 slope_change thickness_min_1 thickness_max_1 thickness_avg_1	0.454	0.988	slope_1	0.315	0.988
C	slope_1 slope_2 slope_change undulaion_1	0.435	0.971	slope_1 slope_2 slope_change undulaion_1 thickness_max_2 undulaion_2	0.362	0.971	slope_1 slope_2 slope_change undulaion_1	0.411	0.971
D	slope_1	0.31	0.986	slope_1 slope_2 slope_change thickness_min_1 thickness_avg_1 thickness_max_2 thickness_avg_2	0.395	0.986	slope_1	0.315	0.986

표 4 변수 선택법에 따른 MAPE 및 Adjusted R-squared 결과표

울기 조절에 필요한 Tilt_judgment를 사용하였다. 변수 선택 및 소거의 기준이 되는 값은 p-value로 설정하였다. 전진 선택법은 칼럼을 순서대로 X값에 대입하며 최적의 해가 나오면 더 이상 대입하지 않고 멈춘다. 후진 소거법은 X에 모든 칼럼을 대입한 상태에서 시작하여 변수를 하나씩 소거해가는 과정을 거치며 최적의 해가 나오면 멈추게 된다. 전진 단계별 선택법은 전진 선택법의 코드를 사용하여 변수를 하나씩 추가하며, 후진 소거법의 코드를 사용하여 불필요한 변수를 제거한다. 이러한 방식을 반복하여 더 이상 추가하거나 소거할 변수가 없으면 멈추게 된다. 각 방법을 통해 최종 해가 결정되었다면 MAPE와 Adjusted R-squared를 사용하여, 모델의 성능을 평가하였다. 이전과 달리 Adjusted R-squared를 사용한 이유는 변수의 수가 이전에 비해 상대적으로 많아짐에 따라 오버피팅 문제가 발생할 확률이 높아지는 문제가 발생할 수 있으므로, 이를 조절하기 위해서 사용하였다.

4.4 변수 선택법 결과

변수 선택법 결과는 표 4와 같다. 대부분 가장 낮은 MAPE와 가장 높은 Adjusted-R-squared가 일치하지만 제품 코드 B의 경우 다르다는 것을 알 수 있었다. 따라서 B의 경우 다른 성능 지표와 함께 참고하는 것이 필요하다고 판단할 수 있다. 변수 선택법에

따라 최적의 변수는 다음과 같다. A는 Slope_1, Slope_2, Slope_change, Tthickness_avg_1, Thickness_min_1, Thickness_max_2, Undulaion_2이고, C는 Slope_1, Slope_2, Slope_change, Undulaion_1, Thickness_max_2, Undulaion_2이고, D는 Slope_1이다.

최종적으로, 이전 회귀분석과 함께 비교하여 X에 대입했을 때 최적인 변수는 다음과 같다. A는 Slope_1, Slope_2, Slope_change, Tthickness_avg_1, Thickness_min_1, Thickness_max_2, Undulaion_2이다. B는 Slope_1, Slope_2, Slope_change이다. C는 Slope_1, Slope_2, Slope_change, Undulaion_1, Thickness_max_2, Undulaion_2이다. D는 Slope_1이다.

5. 결론

본 연구는 적절한 웨이퍼 기울기를 제작하기 위한 영향 인자를 도출하기 위해 H사의 데이터를 python을 이용하여 데이터프레임으로 정리한 후 EDA를 통해 여러 변수에 대해서 관계를 살펴보았고 유의미한 변수들을 탐색했으며, Linear Regression을 통해 주요한 요인들을 분석하였다. EDA 결과 칼럼들의 유의미한 상관성을 토대로 예측 모델의 변수를 선택할 때 주요한 변수로써 새로운 최적 해를 찾아볼 수 있다. 또한, Linear Regression과 변수 선택법을 통해 발견한 주요 요인들을 실제 현장에서 사용 가능한 형태로 적용한다면 이전보다 더욱 정확한 기울기를 제작할 수 있을 것으로 기대된다.

이러한 측면에서 본 연구는 웨이퍼 기울기에 영향을 줄 수 있는 여러 가지 요인들과 다양한 가능성을 확인하였다는 점과 수요 기업의 다양한 요구사항을 효과적으로 대응할 수 있다는 가능성을 빅데이터 기술을 통해 발견할 수 있었다는 점에서 의의가 있다.

본 연구는 다음과 같은 한계점을 가지고 있다. 기울기 및 두께와 관련된 칼럼에 집중하였기 때문에 EDA의 결과 중 일부를 Linear Regression에 변수로 사용하지 못하였고, 그 결과 다양하고 폭넓은 결과를 도출하지 못하였다.

향후 연구에서는 본 연구의 결과물이 실제로 효과가 있는지 확인한 후, 시뮬레이션 및 생성 모델을 통해 더욱 다양한 데이터를 생성하여 적절한 변수 탐색 및 정밀한 예측 모델을 만들 예정이다.

참고문헌

[1] 김문과의 데이터. (게시일 불명). 왜 직관적인 MAE 말고 RMSE를 쓰시나요. *Notion*. https://data101.oopy.io/mae-vs-rmse.

[2] 김상철, 이상직, 정해도, 이석우, 최헌종.(2003).실리콘 웨이퍼 연삭의 평탄도 특성 평가.대한기계학회 춘추학술대회,(),151-156.

[3] 양대규. (2020년 06월 20일). “두께를 줄여라”...얇아지는 반도체, 박형(Thin) 웨이퍼 시장도 성장세. IT BizNews. https://www.itbiznews.com/news/articleView.html?idxno=17425.

[4] 한지연. (2023년 02월 06일). 반도체 혹한기에도...글로벌 매출, 사상 최고치 찍었다. 머니투데이. https://news.mt.co.kr/mtview.php?no=2023020615104041488.

[5] Liu, Y., Tao, H., Zhao, D., & Lu, X. (2022). An Investigation on the Total Thickness Variation Control and Optimization in the Wafer Backside Grinding Process. Materials (1996-1944), 15(12), 4230–N.PAG. https://doi-org.libproxy.koreatech.ac.kr/10.3390/ma15124230