센서 신호를 활용한 반도체 제조공정 머신러닝 이상 탐지: SECOM 데이터셋 기반 SMOTE, PCA 및 통계 기반 피처 선택 통합 파이프라인 접근

박성진
University of Wisconsin-Madison
Computer Science & Data Science
(seongjinpark99@gmail.com)

https://github.com/Seongjin74/Efficient-Semiconductor-Anomaly-Detection

1.서론

현제 제조공정에서는 한정된 데이터와 다양한 품질의 데이터가 혼재되어 있어, 이를 효과적으로 활용한 분석 및 예측의 중요성이 점점 커지고 있다. 특히, 제조공정에서의 데이터는 센서로부터 수집되는 방대한 정보와 함께 결측치, 노이즈, 다중공선성 등 다양한 문제점을 내포하고 있으며, 데이터의 전처리 및 변환 과정이 최종 예측 모델의 성능에 지대한 영향을 미친다. 이에 따라, 주어진 데이터를 얼마나 효율적으로 전처리하고, 적절한 분석 기법을 적용하느냐가 제품의 품질 관리와 이상 탐지에 있어 핵심적인 역할을 수행하게 된다.

반도체 제조 공정은 이러한 관점에서 특히 주목할 만하다. 반도체 제조 과정에서는 수많은 센서 데이터를 기반으로 제품 품질을 관리하고, 불량 및 이상 징후를 조기에 탐지하는 것이 필수적이다. 그러나 센서 데이터는 클래스 불균형, 고차원성, 결측치 및 노이즈, 그리고 다중공선성 문제 등 여러 한계점을 가지고 있어, 이를 그대로 활용할 경우 모델의 성능 저하와 해석의 어려움이 발생할 수 있다.

2.관련연구

제조공정 데이터의 불균형, 노이즈, 고차원성 등 다양한 문제를 해결하기 위해 기존 연구에서는 여러 전처리 및 모델링 기법을 제안해왔다. 예를 들어, 데이터 불균형 문제를 완화하기 위해 Chawla et al. [2]가 제안한 SMOTE(Synthetic Minority Over-sampling Technique)는 소수 클래스 데이터를 인위적으로 증강하여 모델이 소수 클래스에 대해 충분히 학습할 수 있도록 돕는 대표적인 방법론이다. SMOTE 는 이후 다양한 변형 기법과 함께 실제 제조공정 데이터에 적용되어 효과를 검증한 연구들이 보고되었으며, He & Garcia [5]는 불균형 데이터 문제의 심각성과 이를 해결하기 위한 다양한 접근법을 종합적으로 논의하였다.

한편, 제조공정에서 수집된 센서 데이터는 다수의 피처를 포함하고 있어 '차원의 저주'와 노이즈 문제를 동반한다. 이에 따라, 데이터의 주요 정보를 보존하면서 불필요한 노이즈를 제거하기 위한 차원 축소 기법으로 PCA(Principal Component Analysis)가 널리 활용되고 있으며, Jolliffe [3]는 PCA 의 이론적 배경과 응용 사례를 상세히 정리하였다. 최근에는 단순 PCA 의 한계를

극복하기 위해 계층적 PCA 와 같은 비선형 및 군집 기반 차원 축소 기법이 제안되었으며, 이러한 기법은 데이터의 복잡한 상호작용을 보다 효과적으로 반영할 수 있다는 점에서 주목받고 있다.

또한, 단일 전처리 기법만으로는 제조공정 데이터에 내재한 복잡한 문제들을 완전히 해결하기 어려움이 확인되면서, 여러 전처리 기법과 분류기를 결합하는 하이브리드 및 앙상블 방법론이 활발히 연구되고 있다. Breiman [4]는 Random Forest 와 같은 앙상블 학습 기법을 통해 개별 모델의 약점을 보완함으로써 예측 성능을 향상시킬 수 있음을 입증하였고, Van Hulse et al. [6]와 Blagus & Lusa [7]는 통계 기반 피처 선택과 차원 축소 기법의 결합이 복잡한 제조공정 데이터에서 의미 있는 특성을 효과적으로 추출하는 데 기여함을 제시하였다.

본 연구는 이와 같은 선행 연구들을 토대로, SECOM 데이터셋을 대상으로 결측치 보완, SMOTE 를 통한 클래스 불균형 해결, PCA 및 통계 기반 피처 선택, 그리고 GridSearchCV 를 활용한 모델 최적화 과정을 통합한 파이프라인을 제안한다. GridSearchCV 는 다양한 모델의 하이퍼파라미터 조합에 대해 교차 검증을 수행하여 최적의 설정을 도출하는 기법으로, 이를 통해 모델 성능을 극대화할 수 있다. 이를 통해 제조공정 이상 탐지 문제에서 기존 기법들의 한계를 극복하고, 보다 신뢰할 수 있는 예측 모델을 구축하는 데 기여하고자 한다.

3.방법 제안론

본 연구는 위와 같은 선행 연구들을 종합하여, 신호 데이터를 활용하여 이상 탐지를 예측하는 머신러닝 모델의 개선을 목표로 한다. 구체적으로, secom 데이터셋을 대상으로 결측치 보완, 단일 값 및 노이즈 제거, 다중공선성 완화 등의 전처리 기법을 적용하고, SMOTE 를 통한 클래스 불균형 문제를 해결한다. 이후, 계층적 PCA 와 통계 기반 피처 선택을 통해 데이터의 효율성을 높인 후, 6 개의 분류기를 개별적으로 학습 및 평가하여 기본 성능을 확인하고, GridSearchCV 를 활용한 최적 분류기 선정 과정을 수행한다.

이와 같은 접근법은 제조공정의 품질 관리 및 이상 탐지에 있어 데이터 활용의 효율성을 극대화하고, 보다 신뢰할 수 있는 예측 모델을 개발하는 데 기여할 것으로 기대된다. 본 연구에서는 제안하는 통합 파이프라인의 각 단계별 성능을 단계 1(원본 데이터)부터 단계 2(SMOTE 적용), 단계 3(SMOTE + PCA), 단계 4(SMOTE + 통계 기반 피처 선택 + PCA), 그리고 단계 5(SMOTE + 통계 기반 피처 선택 + PCA + GridSearchCV)까지 비교 분석함으로써, 전처리 및 최적화 과정이 최종 모델의 성능에 미치는 영향을 종합적으로 평가하고자 한다.

SECOM 데이터셋은 2008 년 11 월 18 일에 제공된 반도체 제조 공정 데이터로, 1567 개의 예제와 591 개의 피처로 구성되어 있으며, 각 예제는 생산 단위에 해당하는 센서 신호와 함께 간단한 Pass/Fail 라벨(실패 104 건 포함)을 갖고 있다. 이 데이터는 실제 공정에서 수집된 다양한 신호가 포함되어 있어, 유의미한 특성 선택과 노이즈 제거가 중요한 역할을 한다.[1]

4.실험

4.1 데이터 전처리

데이터 전처리는 모델 성능에 직결되는 중요한 단계로, 원천 데이터에 존재하는 결측치, 노이즈, 불필요한 피처 및 클래스 불균형 문제를 해결하기 위해 다양한 기법을 적용하였다.

```
[8 rows x 591 columns]
Dataset shape after removing 28 columns with >50% missing values: (1567, 564)
Dataset shape after removing 116 columns with a single unique value: (1567, 448)
Dataset shape after deleting 'Time' column: (1567, 447)
Dataset shape after removing highly collinear features: (1567, 205)
Final dataset shape: (1567, 205)
```

<그림 1> 데이터 전처리 후 데이터 지표

먼저, 제조공정 데이터는 센서 측정 과정에서 누락된 값이 빈번하게 발생하는 특성이 있어, 이러한 결측치가 데이터 분석 및 모델 학습 시 노이즈로 작용할 수 있다는 문제를 인식하였다. 이에 전체 데이터 중 50% 이상의 결측치를 포함하는 컬럼은 정보의 신뢰도가 낮다고

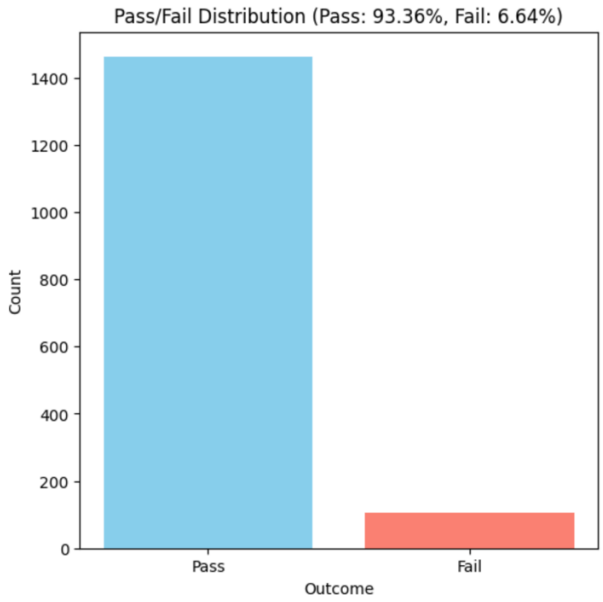
판단하여 제거하였으며, 남아 있는 결측치는 데이터의 순서를 고려한 forward fill 및 backward fill 기법을 통해 보완함으로써 데이터의 연속성을 유지하였다. 이러한 처리 방식은 결측치가 많은 피처를 제거하여 불필요한 노이즈를 줄이고, 채움 기법을 통해 데이터 손실을 최소화하는 효과를 가져왔으나, 순차적 채움 방식이 데이터의 시간적 특성을 가정하기 때문에 데이터의 순서나 계절성이 반영되지 않는 경우 다른 보간법과 비교하여 한계가 있을 수 있다.

또한, 모든 샘플에서 동일한 값을 가지는 컬럼은 모델에 어떠한 구별 정보도 제공하지 않고, 오히려 계산량을 늘리는 요인으로 작용하기 때문에, 이러한 단일 값 피처를 제거하였다. 단일 값 피처의 제거는 모델의 학습 효율성을 높이고 불필요한 차원을 줄여 계산 비용을 절감하는 효과가 있다.

분석 목적과 직접적으로 관련이 없는 'Time' 컬럼은 모델의 성능에 영향을 주지 않으므로 제거하였으며, 다수의 센서 데이터 간에는 상호 간에 높은 상관관계가 나타날 수 있다. 상관계수가 0.7 이상인 피처들 사이에는 중복된 정보가 존재할 가능성이 크기 때문에, 하나의 대표 피처만 남기고 나머지를 제거하는 방식으로 다중공선성을 완화하였다. 이러한 처리를 통해 불필요한 중복 정보를 제거함으로써 모델의 해석력을 향상시키고 과적합의 위험을 줄일 수 있으나, 지나친 피처 제거는 오히려 유의미한 정보를 잃을 수 있으므로 상관계수 기준의 설정에 신중할 필요가 있다.

각 피처의 스케일이 상이할 경우, 거리 기반 또는 확률 기반 알고리즘에서 특정 피처가 과도한 영향을 미칠 수 있다는 점을 고려하여, 모든 피처를 평균 0, 표준편차 1 의 분포로 변환하는 StandardScaler 를 사용해 정규화를 수행하였다. 이를 통해 각 피처가 동일한 스케일에서 비교될 수 있게 되어 모델 학습의 안정성과 수렴 속도를 개선하는 효과를 얻었으나, 일부 트리 기반 알고리즘에서는 정규화가 큰 영향을 미치지 않을 수 있음을 고려해야 한다.

마지막으로, 원천 데이터에서 정상(Pass)과 결함(Fail)의 비율이 약 93.36% 대 6.64%로 크게 불균형되어 있는 상황을 시각화를 통해 확인하였다.



<그림 2> Pass/Fail 분포 그래프 시각화

타깃 변수의 분포를 시각화함으로써 불균형 문제를 직관적으로 파악할 수 있었으며, 이를 바탕으로 이후 SMOTE 와 같은 오버샘플링 기법의 필요성을 인식하였다.

4.2 모델 학습 및 평가

데이터 전처리 후 전체 데이터를 학습용과 테스트용으로 분할하여 모델의 일반화 성능을 평가하였다. 본 연구에서는 Decision Tree, Naive Bayes, Logistic Regression, K-NN, SVM, Neural Network 등 6 가지 분류기를 대상으로 각 단계별로 성능 변화를 면밀히 비교 분석하였다.

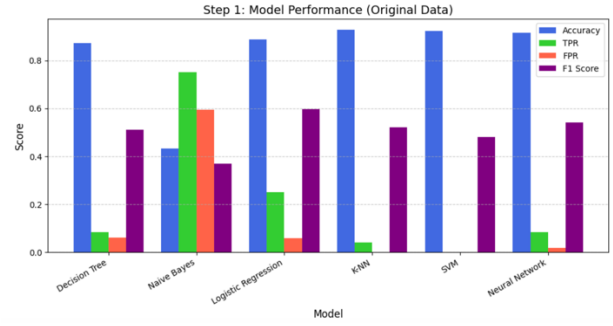
먼저, **Step 1** 에서는 전처리만을 거친 원본 불균형 데이터를 사용하여 모델을 학습시켰다.

```
----- Step 1: Original Data (Imbalanced) -----
                Model  Accuracy       TPR       FPR  F1 Score
0       Decision Tree  0.872611  0.083333  0.062069  0.511208
1         Naive Bayes  0.433121  0.750000  0.593103  0.369136
2  Logistic Regression  0.888535  0.250000  0.058621  0.597539
3                 K-NN  0.926752  0.041667  0.000000  0.520929
4                  SVM  0.923567  0.000000  0.000000  0.480132
5       Neural Network  0.914013  0.083333  0.017241  0.541903

Step 1 Average Results:
Accuracy    0.826433
TPR         0.201389
FPR         0.121839
F1 Score    0.503475
dtype: float64
```

<그림 3> Step 1: 모델 평균 성능 지표

이 단계에서 평균 Accuracy 는 약 0.8264 로 상대적으로 높은 편이었으나, 불균형 데이터 특성으로 인해 소수 클래스(결함)에 대한 예측이 미흡하여 평균 TPR(참 양성률)이 약 0.2014, 평균 F1 Score 는 0.5035 로 낮게 나타났다.

<그림 4> Step 1: Confusion Matrices



<그림 5> Step 1: Model Performance

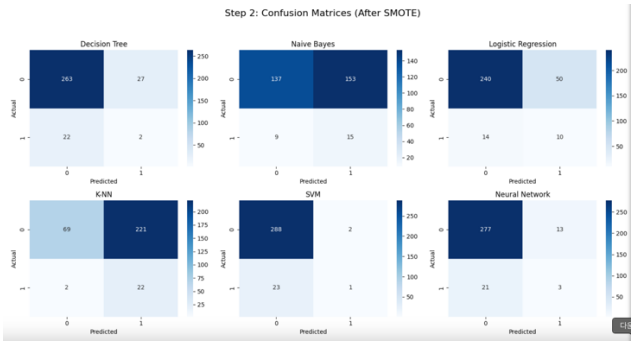
특히 Naive Bayes 모델의 경우 TPR 이 75%에 달하는 반면 FPR(거짓 양성률)이 59%에 이르는 등, 일부 모델은 정상 클래스를 과도하게 오탐하는 경향을 보였으며, K-NN 과 SVM 은 소수 클래스에 대한 인식이 거의 이루어지지 않는 문제를 확인할 수 있었다. 이 결과는 단순 Accuracy 지표만으로 모델의 실제 결함 탐지 성능을 평가하기 어렵다는 점을 명확하게 보여준다.

이후 **Step 2** 에서는 SMOTE 를 적용하여 소수 클래스 데이터를 인위적으로 증강하였다.

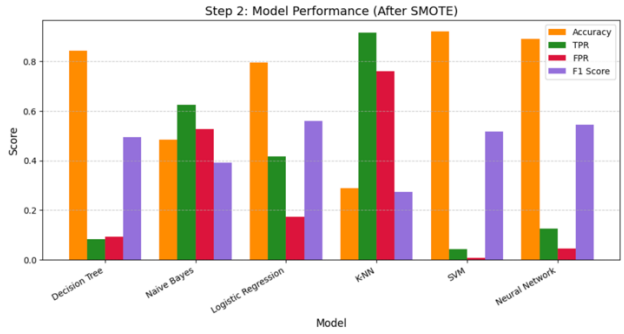
```
----- Step 2: SMOTE Applied -----
Shape after rebalancing training data: (2346, 204)
                Model  Accuracy       TPR       FPR  F1 Score
0       Decision Tree  0.843949  0.083333  0.093103  0.495127
1         Naive Bayes  0.484076  0.625000  0.527586  0.392345
2  Logistic Regression  0.796178  0.416667  0.172414  0.560224
3                 K-NN  0.289809  0.916667  0.762069  0.273533
4                  SVM  0.920382  0.041667  0.006897  0.516238
5       Neural Network  0.891720  0.125000  0.044828  0.546088

Step 2 Average Results:
Accuracy    0.704352
TPR         0.368056
FPR         0.267816
F1 Score    0.463926
dtype: float64
```

<그림 6> Step 2: 모델 평균 성능 지표

이 과정에서 평균 Accuracy 는 약 0.7044 로 다소 하락하였으나, 평균 TPR 은 약 0.3681 로 개선되어 결함 탐지 능력이 강화된 것을 확인할 수 있었다. 그러나 SMOTE 적용으로 인해 정상 클래스를 오탐하는 비율인 평균 FPR 이 약 0.2678 로 상승하면서 전체 F1 Score 는 0.4639 로 약간 낮아지는 부작용이 발생하였다.



<그림 7> Step 2: Confusion Matrices



<그림 8> Step 2: Model performance

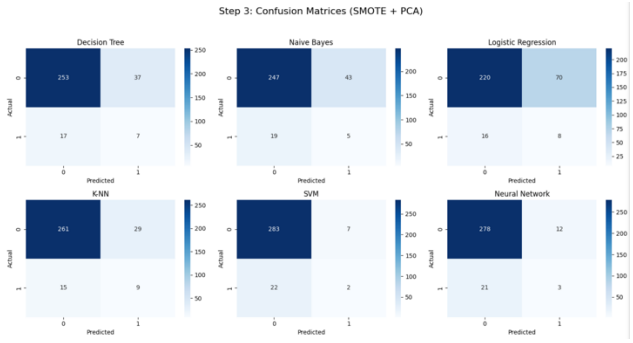
특히 K-NN 모델의 경우 TPR 은 극단적으로 높아 91.67%에 달하였으나, 동시에 FPR 이 76.21%로 지나치게 상승하는 양상을 보이며, 오히려 소수 클래스에 치우친 예측을 나타내어 SMOTE 의 증강 효과와 함께 과도한 노이즈 학습 위험을 내포하고 있음을 시사하였다.

Step 3 에서는 SMOTE 로 증강된 데이터를 대상으로 PCA 를 적용하여 차원을 50 개로 축소하였다. 이 과정을 통해 불필요한 노이즈를 줄이고 계산 비용을 절감할 수 있었다.

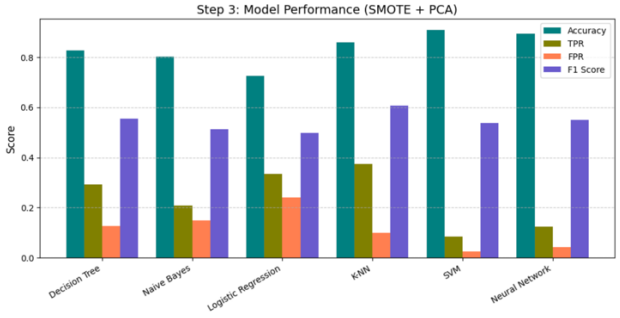
```
----- Step 3: SMOTE + PCA (Dimensionality Reduction) -----
Shape after rebalancing training data (SMOTE applied): (2346, 204)
X_train shape after PCA: (2346, 50)
                 Model  Accuracy       TPR       FPR  F1 Score
0        Decision Tree  0.828025  0.291667  0.127586  0.554727
1          Naive Bayes  0.802548  0.208333  0.148276  0.513689
2  Logistic Regression  0.726115  0.333333  0.241379  0.496682
3                  K-NN  0.859873  0.375000  0.100000  0.606292
4                  SVM  0.907643  0.083333  0.024138  0.536236
5       Neural Network  0.894904  0.125000  0.041379  0.548909

Step 3 Average Results:
Accuracy    0.836518
TPR         0.236111
FPR         0.113793
F1 Score    0.542756
dtype: float64
```

<그림 9> Step 3: 모델 평균 성능 지표

평균 Accuracy 는 약 0.8365 로 회복되었고, 평균 FPR 은 0.1138 로 크게 낮아진 반면, 평균 TPR 은 SMOTE 단계보다는 다소 낮은 0.2361 을 기록하였다. 그 결과 F1 Score 는 0.5428 로 상승하였는데, 이는 모델이 Precision 과 Recall 간의 균형을 어느 정도 회복하였음을 의미한다. PCA 적용으로 인해 일부 정보 손실이 발생할 수 있는 한계가 있으나, 전체적으로 모델 성능을 안정화시키는 긍정적인 효과가 있음을 확인할 수 있다.



<그림 10> Step 3: Confusion Matrices



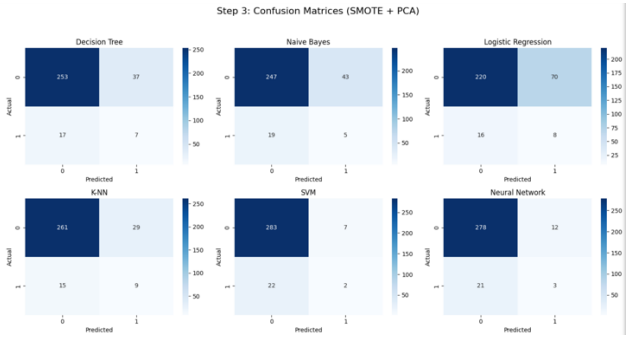
<그림 11> Step 3: Model performance

Step 4 에서는 통계 기반 피처 선택과 계층적 PCA 를 도입한 후, GridSearchCV 를 적용하지 않은 상태에서 모델을 평가하였다. 이 단계에서는 통계적 기준에 따라 유의미한 15 개 피처를 선택하고, 계층적 PCA 를 통해 각 피처 그룹별 주요 정보를 추출하였다.

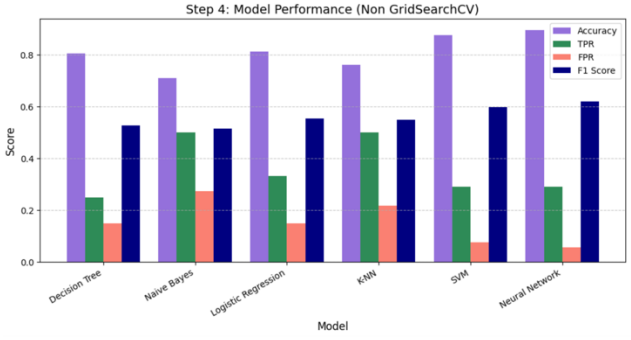
```
----- Step 4: SMOTE + Statistical Feature Selection + PCA (Non GridSearchCV) -----
Shape after rebalancing training data: (1759, 204)
Number of selected features: 15
                 Model  Accuracy       TPR       FPR  F1 Score
0        Decision Tree  0.805732  0.250000  0.148276  0.527237
1          Naive Bayes  0.710191  0.500000  0.272414  0.515654
2  Logistic Regression  0.812102  0.333333  0.148276  0.553321
3                  K-NN  0.761146  0.500000  0.217241  0.550324
4                  SVM  0.875796  0.291667  0.075862  0.598162
5       Neural Network  0.894904  0.291667  0.055172  0.620537

Step 4 Average Results:
Accuracy    0.809979
TPR         0.361111
FPR         0.152874
F1 Score    0.560872
dtype: float64
```

<그림 12> Step 4: 모델 평균 성능 지표

평균 TPR 은 약 0.3611, 평균 FPR 은 0.1529, 평균 Accuracy 는 0.8100, 평균 F1 Score 는 0.5609 로 나타났다. 피처 선택과 차원 축소 기법의 결합이 소수 클래스 탐지 성능을 개선하는 동시에 F1 Score 를 상승시키는 효과를 보였으나, 약간의 FPR 상승이 관찰되어 정상 제품에 대한 오탐률이 다소 증가하는 트레이드오프가 존재함을 시사한다.



<그림 13> Step 4: Confusion Matrices



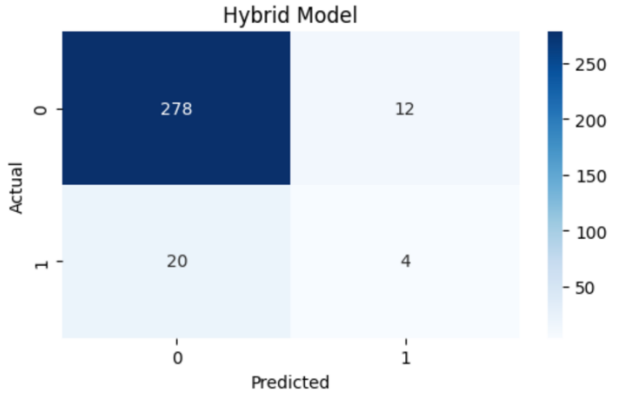
<그림 14> Step 4: Model performance

마지막 **Step 5** 에서는 GridSearchCV 를 활용하여 다양한 분류기와 계층적 PCA 의 클러스터 수 등 주요 파라미터를 최적화함으로써 최종 하이브리드 모델을 도출하였다.

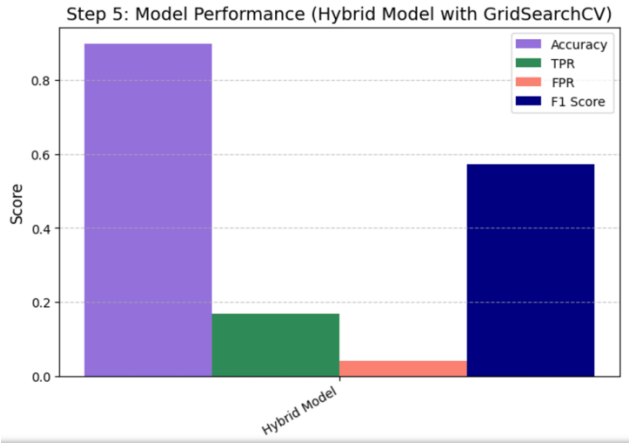
```
----- Step 5: SMOTE + Statistical Feature Selection + PCA (GridSearchCV) -----
         Model  Accuracy       TPR       FPR  F1 Score
0  Hybrid Model  0.898089  0.166667  0.041379  0.572789

Best Estimator:
Pipeline(steps=[('feature_selection',
                 ColumnTransformer(transformers=[('selector', 'passthrough',
                                                  ['28', '59', '63', '64', '79',
                                                   '103', '121', '122', '129',
                                                   '133', '144', '170', '183',
                                                   '468', '510'])])),
                ('hpca', HierarchicalPCA()),
                ('classifier', MLPClassifier(max_iter=1000))])

Best Parameters:
{'classifier': MLPClassifier(max_iter=1000), 'hpca__n_clusters': 15}

Step 5 Average Results:
Accuracy    0.898089
TPR         0.166667
FPR         0.041379
F1 Score    0.572789
dtype: float64
```

<그림 15> Step 5: 모델 평균 성능 지표

이 과정에서 후보 모델들(Decision Tree, Naive Bayes, Logistic Regression, K-NN, SVM, Neural Network 등)을 비교 평가한 결과, 최종적으로 선택된 모델은 통계 기반 피처 선택과 계층적 PCA 를 전처리 단계로 결합하고, 분류기로는 MLPClassifier(max_iter=1000)를 사용한 하이브리드 파이프라인이었다. GridSearchCV 를 통해 도출된 최적 파라미터는 계층적 PCA 의 클러스터 수를 15 로 설정하였으며, MLPClassifier 는 최대 반복 횟수를 1000 으로 설정하여 학습을 진행하도록 하였다. 이 최적화 결과, 최종 모델은 평균 Accuracy 0.8981, 평균 FPR 0.0414, 평균 TPR 0.1667, 평균 F1 Score 0.5728 를 달성하였으며, 특히 낮은 FPR 은 제조공정 품질 관리 측면에서 매우 중요한 성과로 평가된다. 다만, 최적화 과정에서 결함 탐지율(TPR)이 다소 낮게 나타난 점은 정상 제품에 대한 오탐을 최소화하기 위한 비용 균형을 반영한 결과로 해석할 수 있으며, 이 부분은 향후 추가 연구를 통해 개선할 필요가 있다.



<그림 16> Step 5: Confusion Matrices



<그림 17> Step 5: Model performance

5. 실험 평가

본 연구에서는 원본 불균형 데이터(Step 1)를 기준으로 SMOTE, 차원 축소(PCA), 통계 기반 피처 선택 및 계층적 PCA, 그리고 GridSearchCV 를 적용한 최적화 과정을 순차적으로 도입하여 모델의 성능 변화를 정량적으로 평가하였다.

1. 불균형 데이터 극복(SMOTE)의 효과

Step 1 (원본 데이터): 평균 Accuracy 0.8264, 평균 TPR 0.2014, 평균 FPR 0.1218, 평균 F1 Score 0.5035 로 나타났다.

Step 2 (SMOTE 적용): 평균 TPR 이 0.3681 로 약 82.7% 상승하였으나, 평균 FPR 은 0.2678 로 약 119.7%

증가하였다. Accuracy 는 0.7044 로 약 14.8% 하락하였고, F1 Score 는 0.4639 로 약 7.9% 감소하였다.

이 결과는 SMOTE 가 소수 클래스(결함) 탐지 능력(TPR)을 크게 개선하였으나, 동시에 정상 클래스에 대한 오탐률(FPR)이 크게 증가하는 부작용을 나타낸다.

2. 차원 축소(PCA)의 기여

Step 3 (SMOTE + PCA): PCA 를 통해 차원을 50 개로 축소한 결과, Accuracy 가 0.8365 로 1.2% 상승하였고, FPR 이 0.1138 로 약 6.6% 개선되었다. 다만, TPR 은 0.2361 로 약 17.2% 상승에 그쳤으며, F1 Score 는 0.5428 로 약 7.8% 개선되었다.

PCA 도입은 노이즈를 줄이고 계산 비용을 낮추어 모델의 전반적인 안정성을 높이는 효과가 있음을 수치로 확인할 수 있다.

3. 피처 선택 및 계층적 PCA 의 효과

Step 4 (SMOTE + 통계 기반 피처 선택 + PCA): 15 개의 유의미한 피처를 선택하고 계층적 PCA 를 적용한 결과, Accuracy 는 0.8100, TPR 은 0.3611, FPR 은 0.1529, F1 Score 는 0.5609 로 나타났다.

Step 1 대비 TPR 은 약 79.2% 상승하였으며, F1 Score 는 약 11.4% 개선되었다. 다만, FPR 은 약 25.5% 상승하는 결과를 보였다.

이로써 불필요한 피처 제거와 각 그룹별 주요 정보 추출이 모델이 보다 집중된 정보를 학습하는 데 기여하였음을 알 수 있다.

4. GridSearchCV 를 통한 최적화 효과

Step 5 (SMOTE + 통계 기반 피처 선택 + PCA + GridSearchCV): 최적화된 하이브리드 모델은 평균 Accuracy 0.8981, 평균 TPR 0.1667, 평균 FPR 0.0414, 평균 F1 Score 0.5728 를 달성하였다.

Accuracy 는 Step 1 대비 약 8.7% 상승하였으며, FPR 은 약 66.0% 감소하여 제조공정 품질 관리 측면에서 매우 긍정적인 결과를 보였다. 반면, TPR 은 기준치에 비해 약 17.2% 감소하였는데, 이는 오탐률(FPR) 최소화를 위한 비용 균형으로 해석할 수 있다.

F1 Score 는 약 13.8% 향상되어 전체적인 Precision 과 Recall 의 균형이 개선되었음을 보여준다.

5. 평가의 적합성:

a. F1 Score 는 Precision 과 Recall 의 균형을 평가하는 지표로, 특히 불균형 데이터에서 결함 탐지 성능을 종합적으로 파악할 수 있는 장점이 있습니다.

b. 제조공정 이상 탐지와 같이 오탐과 미탐의 비용이 중요한 상황에서는, F1 Score 와 함께 Accuracy, TPR, FPR 등 다각도의 평가 지표를 함께 고려하는 것이 바람직합니다.

6. 결론 및 향후 연구

본 연구에서는 SECOM 데이터셋을 기반으로 반도체 제조공정에서 수집된 센서 신호를 활용하여 머신러닝 이상 탐지 모델의 성능을 개선하기 위한 통합 파이프라인을 제안하였다. 제안된 파이프라인은 데이터 전처리, SMOTE 를 통한 소수 클래스 증강, PCA 및 계층적 PCA 를 이용한 차원 축소, 통계 기반 피처 선택과 GridSearchCV 를 통한 최적화 과정을 순차적으로 적용하였다. 그 결과, 제조공정 품질 관리의 핵심 지표인 Accuracy 와 F1 Score 는 개선되었으며, 특히 정상 제품에 대한 오탐률(FPR)을 현저하게 낮추는 성과를 달성하였다. 이는 반도체 제조공정에서 발생하는 데이터 품질 문제와 불균형 문제를 극복하는 데 기여하며, 실무 적용 가능성을 높이는 중요한 기반을 마련하였음을 시사한다.

그러나 본 연구에는 몇 가지 한계점이 존재한다.

첫째, SMOTE 를 통한 소수 클래스 증강 과정에서는 인위적인 데이터 확장이 실제 데이터 분포와 차이를 발생시켜 일부 모델에서 FPR 이 급증하고 과도한 오탐 현상이 나타났다.

둘째, PCA 및 계층적 PCA 적용 과정에서 일부 유의미한 정보가 손실되어 결함 탐지 능력(TPR)이 기대에 미치지 못하는 결과를 초래하였다.

셋째, 현재의 피처 선택은 통계 기반 방법론에 의존하고 있어 변수 간의 비선형 관계나 복잡한 상호작용을 충분히 반영하지 못하는 한계가 있다.

향후 연구에서는 이러한 한계점을 보완하기 위해 다음과 같은 방향을 고려할 필요가 있다.

첫째, SMOTE 외에도 ADASYN, 언더샘플링 등 다양한 불균형 처리 기법을 종합적으로 비교 분석하여 보다 정교한 데이터 증강 방법을 도입할 필요가 있다.

둘째, PCA 대신 오토인코더와 같은 비선형 차원 축소 기법을 적용하여 중요한 비선형 특성을 효과적으로 보존하면서 차원을 축소하는 방안을 모색할 수 있다.

셋째, 피처 선택 단계에서는 Mutual Information 과 같은 비선형 지표를 도입하여 변수 간 복잡한 관계를 보다 정밀하게 반영하는 특성 선택 기법을 개발할 필요가 있다.

마지막으로, 본 연구는 오프라인 데이터를 대상으로 하였으나, 제조공정 특성상 실시간 데이터 스트리밍 환경에서의 적용이 요구된다. 따라서, 온라인 학습 기법 및 실시간 데이터 처리 시스템과의 통합을 통해 현장 환경에서 자동화된 이상 탐지 시스템을 구현하는 방향으로 확장하는 것이 중요하다.

이와 같이, 본 연구는 제조공정 품질 관리 및 이상 탐지에 있어서 데이터 전처리와 최적화 기법의 결합이 효과적인 접근임을 확인하였으며, 향후 다양한 데이터 증강, 비선형 특성 보존, 정밀 피처 선택, 그리고 실시간 처리 시스템 구축 등 추가 연구를 통해 더욱 정교한 자동화 이상 탐지 시스템을 구축할 수 있을 것으로 기대된다.

참고문헌(Reference)

[1] McCann, M. & Johnston, A. (2008). SECOM [Dataset]. UCI Machine Learning Repository.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[3] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[5] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[6] Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In Proceedings of the 2007 ACM symposium on Applied computing (pp. 984-988).

[7] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14(1), 106.