석사학위논문

불균형 데이터셋에서의 다이캐스팅 공정의 불량예측 성능 연구 : SMOTE와 랜덤포레스트의 활용

2024년

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

이 상 민

석 사 학 위 논 문
지도교수 박인채

불균형 데이터셋에서의 다이캐스팅 공정의 불량예측 성능 연구 : SMOTE와 랜덤포레스트의 활용

Study on Defect Prediction Performance in Die Casting Process with Imbalanced Datasets: Utilization of SMOTE and Random Forest

2024년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

이 상 민

석 사 학 위 논 문
지도교수 박인채

불균형 데이터셋에서의 다이캐스팅 공정의 불량예측 성능 연구 : SMOTE와 랜덤포레스트의 활용

Study on Defect Prediction Performance in Die Casting Process with Imbalanced Datasets: Utilization of SMOTE and Random Forest

위 논문을 컨설팅학 석사학위 논문으로 제출함

2024년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

이 상 민

이상민의 컨설팅학 석사학위 논문을 인준함

2024년 6월 일

심사위원장 홍정완 (인)

심 사 위 원 윤주일 (인)

심 사 위 원 박인채 (인)

국 문 초 록

불균형 데이터셋에서의 다이캐스팅 공정의 불량예측 성능 연구: SMOTE와 랜덤포레스트의 활용

한성대학교 지식서비스&컨설팅대학원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 기 술 컨 설 팅 전 공
이 상 민

본 연구는 고압 다이캐스팅 공정에서 발생하는 불량품을 예측하기 위하여 랜덤 포레스트와 SMOTE(Synthetic Minority Over-sampling Technique)를 통해 성능지표를 높이는 샘플링 기법의 가이드를 제공 하는것을 목표로 한다. 다이캐스팅 공정은 공정 변수의 작은 변화로도 결함이 발생할 수 있어, 실시간 데이터 분석과 예측이 필요하다. 본 연구에서는 7개월간 수집된 105,748개의 데이터를 대상으로, 다양한 공정변수들을 분석하였다. 데이터 전처리 과정에서는 결측치와 이상치를 처리하였으며, 데이터 불균형 문제를 해결하기 위해 SMOTE와 언더샘플링 기법을 적용하였다. 랜덤 포레스트 모델을 사용하여 학습과 평가를 진행한 결과, SMOTE 기법을 적용한 모델이 높은 예측 성능을 보였다. 특히, SMOTE 10%~30%의 샘플링 비율을 적용한 모델에서 높은 F1-score를 유지하는 것이 확인되었다. 본 연구는 다이캐스팅 공정의 불량 예측 정확성을 높이고, 제조 공정의 품질 관리에 기여할 수 있는 데이터 기반 접근 방안을 제시하였다. 앞으로의 연구에서는 다양한 변수와 충

분한 데이터를 추가하여 모델 성능을 더욱 개선하고, 실제 산업 현장에서의 적용 가능성을 높이는 것이 필요하다.

【주요어】 머신러닝, 랜덤 포레스트, 다이캐스팅, SMOTE, 불량 예측,

목 차

표 목 차

그 림 목 차

제 1 장 서 론

제 1 절 연구의 배경 및 필요성

국내 다이캐스팅 산업은 다양한 공정 관리 문제로 인해 제조 현장에서 상당한 어려움을 겪고 있다. 이러한 문제들은 성형 과정에서 투입되는 소재량의 불균형, 금형 온도의 변동, 용탕의 청정도 부족, 주조 압력 및 냉각 조건의 변화 등 주조 조건 관리의 복잡성에서 기인한다. 특히, 다이캐스팅 공정의 특성상 최종 결함 여부는 공정이 완료된 후에만 확인할 수 있어 생산성 향상과 제품 납기 준수를 위한 관리가 매우 중요하다. 이에 따라 이러한 문제들에 대한 효과적인 해결책이 절실히 요구된다(김준, 2020). 이 산업의 노동집약적 특성과 낮은 영업이익률은 첨단 기술의 도입을 어렵게 만들고 있으며, 그 결과 다이캐스팅 공정은 다른 일반 제조 산업에 비해 상대적으로 높은 5~10%의 불량률을 경험하고 있다(이정수, 2022).

스마트 팩토리 기술은 4차 산업 혁명의 핵심 기술 중 하나로, 제조업 분야에서 생산성과 경쟁력을 향상시키는 데 필수적이다. 다이캐스팅 산업에서도 품질 혁신과 경쟁력 강화를 위해 생산 공정의 고도화 및 스마트 팩토리 기술 도입이 점점 더 중요해지고 있다. 이러한 기술을 활용하여 다이캐스팅 제조 데이터를 수집하고 관리함으로써 제품 품질을 향상시키는 연구가 요구되고 있다(이정수, 2022).

기존의 품질 관리 방법은 주로 경험에 의존하거나 공정 후 검사에 집중되어 있어 비효율적이며, 실시간 데이터를 통한 예측이나 적극적인 개입이 어려운 실정이다. 그러나 최근 인공지능(AI) 기술의 발전으로 대량의 공정 데이터를

활용하여 공정 중 발생 가능한 문제를 사전에 예측하고 개입하는 연구가 활발히 진행되고 있다.

본 연구에서는 고압 다이캐스팅 공정에서 발생하는 다양한 공정변수를 기반으로 랜덤 포레스트와 SMOTE 기법을 활용하여 불량 예측의 정확도를 향상시키는 방법을 탐구하였다. 이를 통해 데이터 특성이 예측 성능에 미치는 영향을 분석하여 보다 효과적인 불량 예측 방법을 제시하고자 한다. 본 연구의 결과는 제조업체들이 불량률 감소, 비용절감, 생산효율 증가 등의 직접적인 효과를 얻는 데 기여할 수 있을 것이다.

제 2 절 연구의 목적 및 내용

본 연구의 목적은 다이캐스팅 공정에서 발생하는 불량을 효과적으로 예측하고 이를 통해 공정 효율성을 높이는 것이다. 다이캐스팅은 주조 과정에서 여러 변수에 의해 불량이 발생할 가능성이 높기 때문에, 이러한 불량을 사전에 예측하고 관리하는 것은 매우 중요하다. 본 연구는 랜덤 포레스트(Random Forest) 알고리즘과 SMOTE(Synthetic Minority Oversampling Technique)를 활용하여 다이캐스팅 공정의 불량 예측 모델을 개발하고, 이를 통해 불량 예측의 정확성을 향상시키는 것을 목표로 한다.

본 연구에서는 지도학습 기반의 머신러닝 모델을 적용하여 불량 예측의 정확성을 높이는 방법을 탐구하였다. 이를 위해 다이캐스팅 공정에서 수집된 다양한 공정 데이터를 활용하였으며, 랜덤 포레스트 알고리즘을 사용하여 모델을 구축하였다. 또한, 데이터 불균형 문제를 해결하기 위해 SMOTE 기법을 적용하여 학습 데이터셋의 품질을 향상시켰다.

실험 및 분석 방법은 다음과 같다.

(1) 다이캐스팅 공정에서 수집된 데이터를 바탕으로 랜덤 포레스트 모델을 구축하고, 이를 통해 불량 예측의 정확성을 평가한다.

(2) SMOTE 기법을 적용하여 데이터 불균형 문제를 해결하고, 이를 통해 모델의 성능을 향상시킨다.

(3) 다양한 성능 평가 지표(정확도, 정밀도, 재현율, F1-score)를 활용하여 모델의 성능을 종합적으로 분석한다.

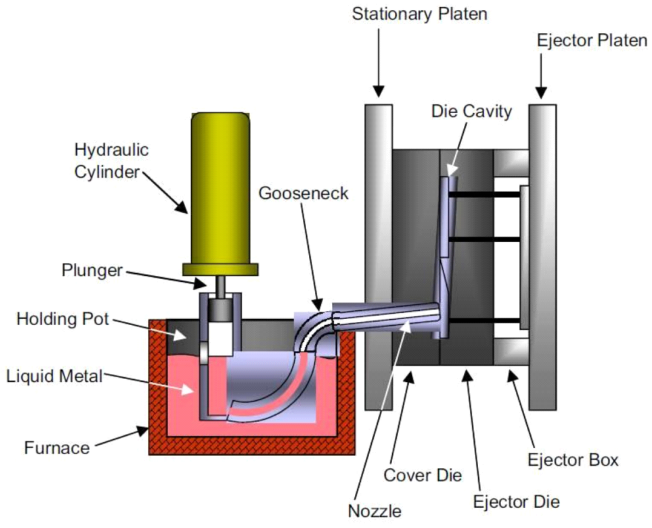
제 2 장 이론적 배경 및 기존연구

제 1 절 다이캐스팅에 대한 이론적 배경

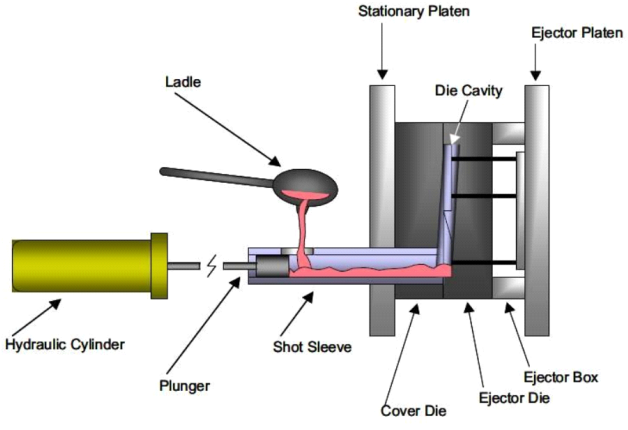
다이캐스팅은 금속 주조 방식 중 하나로, 가압 방식에 따라 고압 주조와 저압 주조로 구분된다. 고압 주조는 다시 핫 챔버 다이캐스팅(Hot Chamber Die Casting)과 콜드 챔버 다이캐스팅(Cold Chamber Die Casting)으로 나뉜다(최정훈, 2015). 핫 챔버 다이캐스팅 장비의 모식도를 [그림 2-1] (a)에 나타내었다. 방식은 사출 실린더가 용해로 내부에 설치되어, 플런저를 사용해 고온의 용탕을 금형에 주입하는 방법이다. 이 방식은 공기 유입이 적고 생산 속도가 빨라, 소형 및 얇은 벽 제품 생산에 적합하다. 용탕 온도 제어가 가능하며, 주로 아연 합금, 주석 합금, 납 합금 등 융점이 낮은 금속을 재료로 사용한다(최정훈, 2015). 반면 콜드 챔버 다이캐스팅 방식은 래들을 이용해 용탕을 샷 슬리브에 주입한 후 플런저로 가압하는 방법이다. 콜드 챔버 다이캐스팅 장비의 모식도가 [그림 2-1] (b)에 나와 있다. 이 방식은 높은 압력을 가할 수 있어, 두꺼운 벽 제품 제조에 주로 사용된다. 핫 챔버 방식보다 더 큰 압력을 가할 수 있으며, 최근에는 대용량 장비도 제작되어 사용된다. 주로 알루미늄 합금, 마그네슘 합금, 구리 합금 등 고융점 금속을 재료로 사용한다(최정훈, 2015). 핫 챔버 다이캐스팅은 고속 생산과 정밀한 부품 제작에 유리한 반면, 콜드 챔버 다이캐스팅은 고융점 금속을 사용하여 더 큰 부품이나 구조적으로 강한 제품 제조에 적합하다. 각 다이캐스팅 방법은 선택된 재료와 제품의 요구 사항에 따라 결정되며, 이러한 기술의 발전과 함께 다이캐스팅 산업은 지속적으로 효율성과 제품 품질을 향상시키고 있다(최정훈, 2015).

콜드 챔버 다이캐스팅 공정에서는 플런저의 직경과 저속/고속 속도와 같은 조건의 제어가 플런저 이동 중 내부 기공을 감소시키는 중요한 요소이다. 비

록 플런저 조건이 최적화되었더라도, 용융 금속이 충진되는 과정에서 기공이 형성될 수 있으며, 이는 기계적 성질의 저하를 초래할 수 있다. 최근에는 내부 기공을 제거하기 위해 진공 보조 장치를 갖춘 고압 다이캐스팅이 활용되고 있다(X. P. Niu, B. H. Hu, I. Pinwill and H. Li., 2000).



(a)



(b)

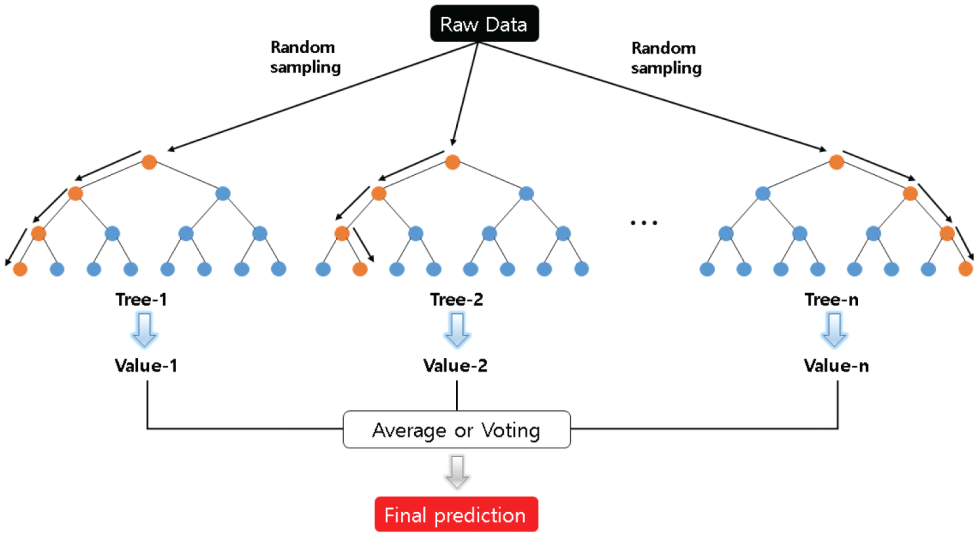
[그림 2-1] 고압 다이캐스팅 기계의 구성도; (a) hot chamber type and (b) cold chamber type (E.J. Vinarcik, 2003).

제 2 절 인공지능에 대한 이론적 배경

1) 랜덤 포레스트

의사결정나무(Decision Tree)는 이진 분류를 수행할 때 자주 사용되는 대표적인 모델 중 하나이다. 이 기법은 데이터의 속성을 기반으로 분할 기준을 설정하고, 그 기준에 따라 트리 형태로 데이터를 분기한다. 이 과정에서 첫 번째로 분기되는 최상위 노드를 뿌리 노드(Root Node)라 하고, 마지막으로 분기되는 지점을 최하위 노드(Terminal Node)라 한다. 뿌리 노드와 최하위 노드 사이에는 중간 노드(Internal Node)들이 존재한다. 의사결정나무는 시각적으로 영향을 미치는 주요 인자를 쉽게 식별할 수 있는 장점이 있지만, 과적합(Overfitting) 문제를 야기할 가능성이 크다. 또한, 상위 노드에서 분기점 설정이 잘못될 경우, 이는 하위 노드에 큰 영향을 미칠 수 있다(Mariana Belgiu, Lucian Dragut, 2016).

의사결정나무 모델의 신뢰성 문제를 해결하기 위해, 본 연구에서는 랜덤 포레스트(Random Forest)라는 앙상블 모델을 사용하였다. [그림 2-2]와 같이, 랜덤 포레스트는 여러 개의 의사결정트리(Decision Tree)를 무작위로 학습시키는 앙상블 방식의 분류기이다(Breiman, 2001). 앙상블 모델은 기본 모델을 여러 번 실행하여 평균이나 다수결과 같은 방법을 통해 예측 정확도를 향상시키는 접근 방식이다. 랜덤 포레스트는 의사결정나무를 기본 모델로 사용하며, 수백 개의 의사결정나무 알고리즘을 결합한다. 다양성을 확보하기 위해 복원 추출 방식을 사용하는 배깅(Bagging)과 변수를 무작위로 추출하는 랜덤 서브스페이스(Random Subspace) 기법을 사용한다. 이를 통해 개별 의사결정나무의 성능이 다소 떨어지더라도, 여러 모델을 결합함으로써 전체적인 성능을 향상시킬 수 있다.



[그림 2-2] 랜덤 포레스트의 개념(김종성 외 5인, 2019)

2) SMOTE (Synthetic Minority Oversampling Technique)

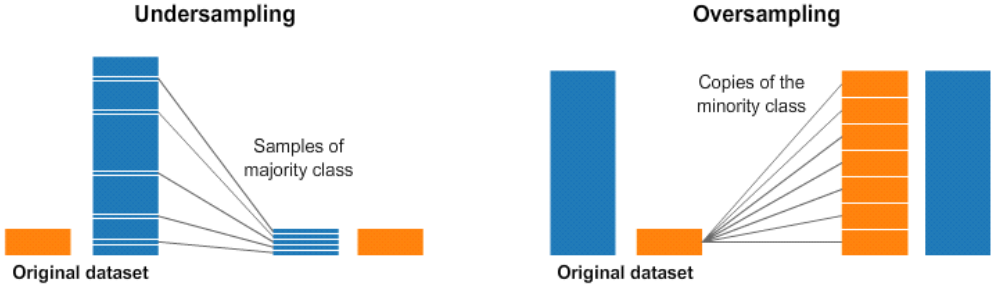
데이터의 각 클래스 개수가 크게 차이나면 머신러닝 모델의 예측 결과에 편향이 발생할 수 있는 문제가 있다. 이러한 불균형 문제를 해결하는 기법 중 하나로 SMOTE(Chawla et al., 2002)가 있다.

데이터 불균형을 해소하기 위해 다양한 방법이 사용될 수 있으며, 그중 가장 기본적인 방법은 [그림2-3]에 제시되어 있는 Undersampling과 Oversampling이다.

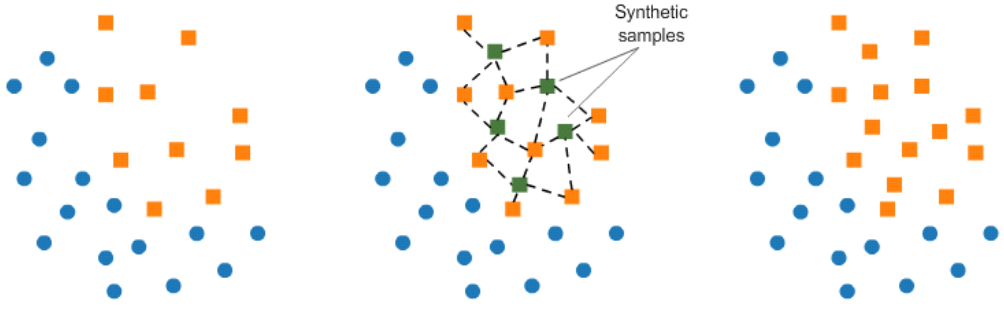
Undersampling은 다수 클래스의 임의 데이터를 제거하여 소수 클래스 수준으로 맞추는 방법이다. 이 방법은 충분한 훈련 데이터가 있을 때 연산 시간과 용량을 줄일 수 있지만, 중요한 데이터가 유실될 수 있는 위험이 있다. 또한, 임의로 선택된 데이터가 모집단을 대표하지 못할 가능성과 편향될 가능성도 있다. Oversampling은 소수 클래스를 다수 클래스 수준으로 복제하여 수를 늘리는 방법이다. 이 방법은 데이터를 잃지 않고 더 좋은 성능을 보이지만, 소수 클래스를 판별할 때 과적합(Overfitting)이 발생할 가능성이 있다.

SMOTE는 소수 클래스를 단순 복제할 때 발생할 수 있는 과적합 문제를 피하기 위해 사용된다. SMOTE는 소수 클래스 데이터를 K-최근접 이웃(KNN, K-Nearest Neighbors) 기법을 통해 선형 관계의 가중치를 두어 인공적으로 무작위 데이터를 생성하여 데이터 불균형을 해소하는 기법이다.

SMOTE의 동작 순서는 [그림 2-4]와 같다. 먼저 1)주어진 소수 클래스 내의 데이터를 확인한다. 2)각 소수 클래스 데이터 포인트에 대해 가장 가까운 K개의 이웃을 찾는다. 3)선택된 데이터 포인트와 이웃들 간의 선형 관계를 기반으로 새로운 데이터를 생성한다. 이는 두 데이터 포인트 사이에 임의의 점을 생성하는 방식으로 이루어지며, 생성된 점은 기존 데이터 포인트와 이웃 데이터 포인트 사이의 벡터를 따라 위치한다. 4)이러한 과정을 반복하여 소수 클래스의 데이터 수를 원하는 수준으로 증가시킨다.



[그림 2-3] Undersampling & Oversampling(Chawla et al., 2002)



[그림 2-4] Oversampling using SMOTE(Chawla et al., 2002)

제 3 절 제조산업에서의 AI활용 분석에 관한 기존 연구

다이캐스팅 산업의 품질 데이터 분석 및 시스템 적용과 관련된 다양한 선행 연구가 있다. 자동차 엔진 부품을 생산하는 주조 공장을 대상으로 주조 금형 온도 데이터를 모니터링하여 제품의 불량을 예측하기 위해 앙상블 기반 랜덤 포레스트(Random Forest) 알고리즘을 개발하여 현장에 적용한 사례가 있다(이준혁, 2019). 이 연구에서는 불량 예측 결과를 가시화하고, 이를 활용하여 공장 운영을 지원할 수 있는 대시보드를 개발했다. 불량과 주조 조건 간의 상관관계를 분석하여 품질 예측 알고리즘을 개발하고, 품질에 영향을 미치는 주요 파라미터를 식별하여 관련 주조 조건을 관리할 수 있도록 한 연구도 있다(박상우, 2019). 다른 연구로는, 다이캐스팅 장비의 주조 공정 변수 데이터와 센서 데이터를 활용해 C4.5 분류기와 AdaBoostC2 알고리즘을 이용하여 모델 학습을 수행(김준, 2020)하거나, XGBoost를 적용하여 다이캐스팅 불량 판정 모델을 개발한 사례가 있다(이주연, 2020). 다이캐스팅 공정 시뮬레이션을 통해 학습 데이터를 생성하고, 생성한 데이터를 이용하여 LSTM 기반 시계열 모델을 적용하여 모델 학습을 수행한 연구도 있다(이승로, 2021). 주조 공정의 실시간 수집되는 파라미터 데이터를 기반으로 머신러닝 기술을 이용하여 주조 시점에서 불량 발생을 SVM 모델과 AdaBoost 모델을 통해 실시간으로 예측할 수 있는 접근법을 제시한 사례도 있다(박철순, 2022). 또한, 현장의 데이터 취득 특성에 따라 비지도, 반지도, 그리고 소수의 불량 데이터를 활용한 지도 학습을 수행한 연구가 있다(이정수, 2022). 금형 내부의 압력 데이터를 이용하여 XGBoost, Random Forest 및 DNN 방법으로 불량을 예측한 사례가 있다(C.-H. Lin, 2022). 이 외에도, POP에서 수집된 주조 파라미터를 로지스틱 회귀, 앙상블 모델, 인공신경망을 사용하여 품질 예측 모델을 구축하고 검증한 사례가 있다(박상우, 2022).

기타 산업 에서는 PET병 제조공정에서 AI기반 머신비전을 통해 CNN과

이상치 감지 알고리즘을 통해 불량 검사 시스템을 개발하여, 육안 검사 보다 높은 불량 검출결과를 보인 사례가 있다(박태윤, 2023). 사출성형의 공정 최적화를 위해 Logistic Regression, Random Forest, SVM 알고리즘을 통해 머신러닝 모델을 개발하고, 성과지표로 정확도를 평가하여 최적 설정값을 제안한 사례가 있다(차형주, 2023). 사출 공정 데이터의 상관관계 분석 후 DNN, SVM, Random Forest 머신러닝 모델을 학습 및 성능 평가하여 불량률을 개선한 사례가 있다(이기성, 2022). 제조공정의 데이터를 Logistic Regression, Random Forest, SVM 알고리즘과 하이퍼 파라미터 최적화를 통해 성능 향상 방법을 제시하고 품질예측의 정확도를 높인 사례가 있다(김종훈, 2021). 또한 발포 공정의 데이터를 kNN 및 앙상블 알고리즘으로 학습 후 불량여부를 예측 한 사례가 있다(최낙훈, 2021).

제 3 장 연구방법 및 결과

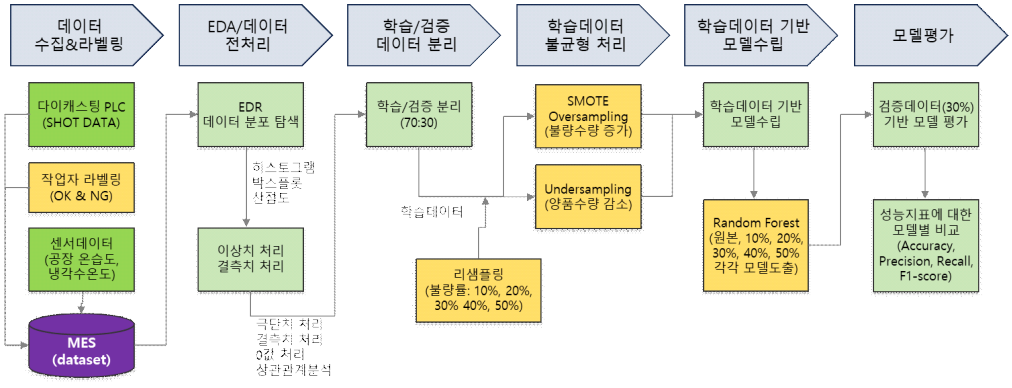
제 1 절 연구목표 및 연구모델

1) 연구 목표

고압 다이캐스팅 공정에서 발생하는 불량품을 예측하기 위해 주요 공정 변수를 분석하고, 랜덤 포레스트 모델을 기반으로 불량 예측의 정확성을 높이는 학습데이터의 리샘플링 방법을 연구

2) 연구 모델

연구 모델은 고압 다이캐스팅 공정의 데이터를 활용하여 랜덤 포레스트 알고리즘을 통한 불량 예측 모델의 개발과 성능 평가를 중심으로 구성된다. 이 과정은 다음과 같은 단계로 진행된다.



[그림 3-1] 연구모델

제 2 절 데이터 수집

다이캐스팅 공정에서 불량품을 예측하기 위해 주요 주조 파라미터를 기반으로 불량 예측 모델을 수립하였다. 이를 위해 POP 시스템에서 설비 주조 파라미터, 냉각수 온도, 공장 내 온도와 습도, 금형 정보를 수집하였으며, 숙련된 작업자의 주조 후 SHOT별 외관검사를 통한 양품/불량 분류결과를 활용하였다. 약 7개월간 동일한 모델과 사양의 2개의 설비에서 수집된 데이터 중 Count가 5000이상인 8개 금형에 대한 '105,748'개의 데이터를 수집하였으며. 설비 및 금형별로 수집된 DATA수는 아래와 같다.

[표 3-1] 금형 별 수집된 데이터 수

설비	금형	Count	양품수량	불량수량	불량률
A	A5	10,127	9,178	949	9.37%
	A8	22,979	22,229	750	3.26%
	A15	10,839	10,425	414	3.82%
	A19	5,693	4,872	821	14.42%
B	B1	31,611	30,698	913	2.89%
	B3	6,175	5,600	575	9.31%
	B20	10,656	10,228	428	4.02%
	B21	7,668	7,368	300	3.91%
A합계		49,638	46,704	2,934	5.91%
B합계		56,110	53,894	2,216	3.95%
A+B합계		105,748	100,598	5,150	4.87%

[표 3-2] 금형 별 수집된 불량정보

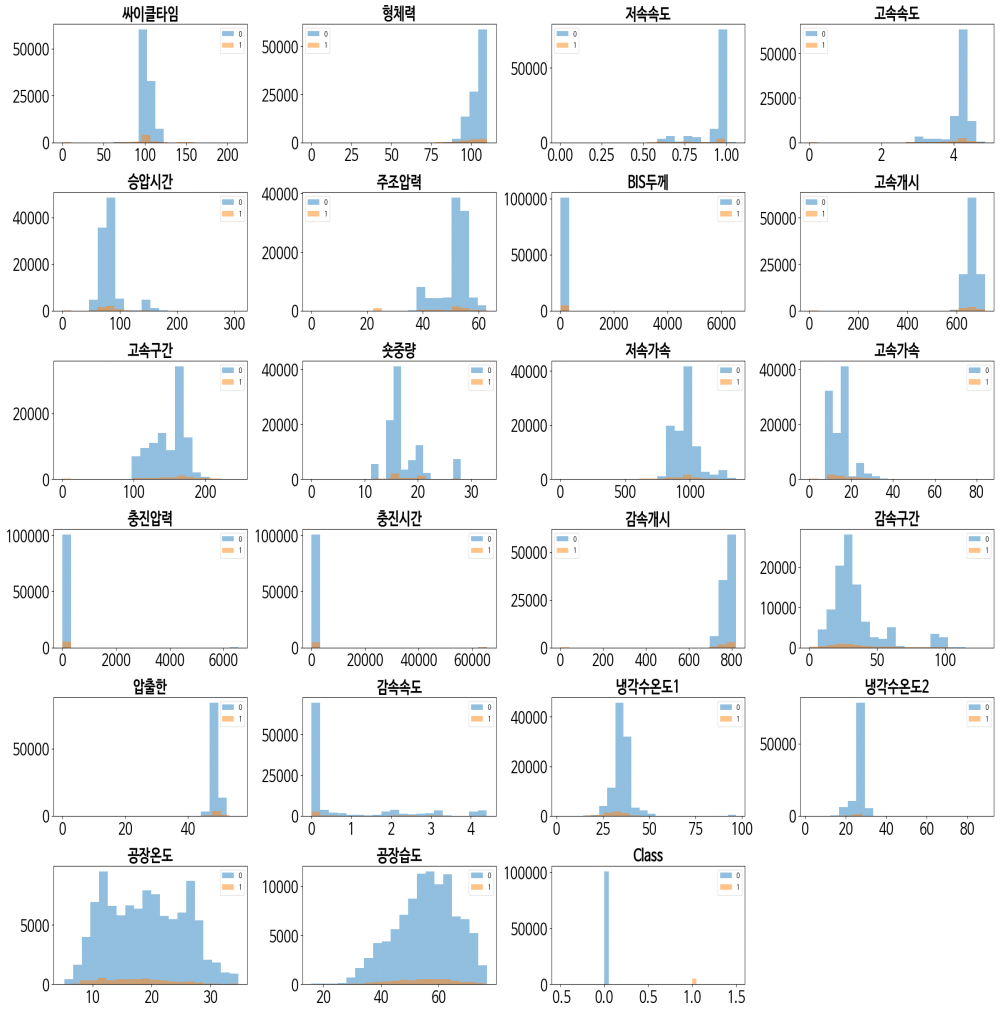
설비	금형	불량유형										
		미성형	미인출	소착	수축	오염	저압	절육	크랙	탕회	테스트	기타
A	A5	436	9	0	3	7	181	30	0	1	277	5
	A8	310	32	0	5	135	227	2	0	1	7	31
	A15	79	9	1	6	94	152	73	0	0	0	0
	A19	387	20	0	36	71	160	22	97	21	3	4
B	B1	183	16	6	2	33	271	393	0	0	0	9
	B3	450	3	0	15	12	79	14	0	0	0	2
	B20	125	14	2	10	2	238	30	0	0	0	7
	B21	158	3	0	8	7	120	4	0	0	0	0
A합계		1,212	70	1	50	307	720	127	97	23	287	40
B합계		916	36	8	35	54	708	441	0	0	0	18
A+B합계		2,128	106	9	85	361	1,428	568	97	23	287	58

제 3 절 데이터설명 및 통계적 분포

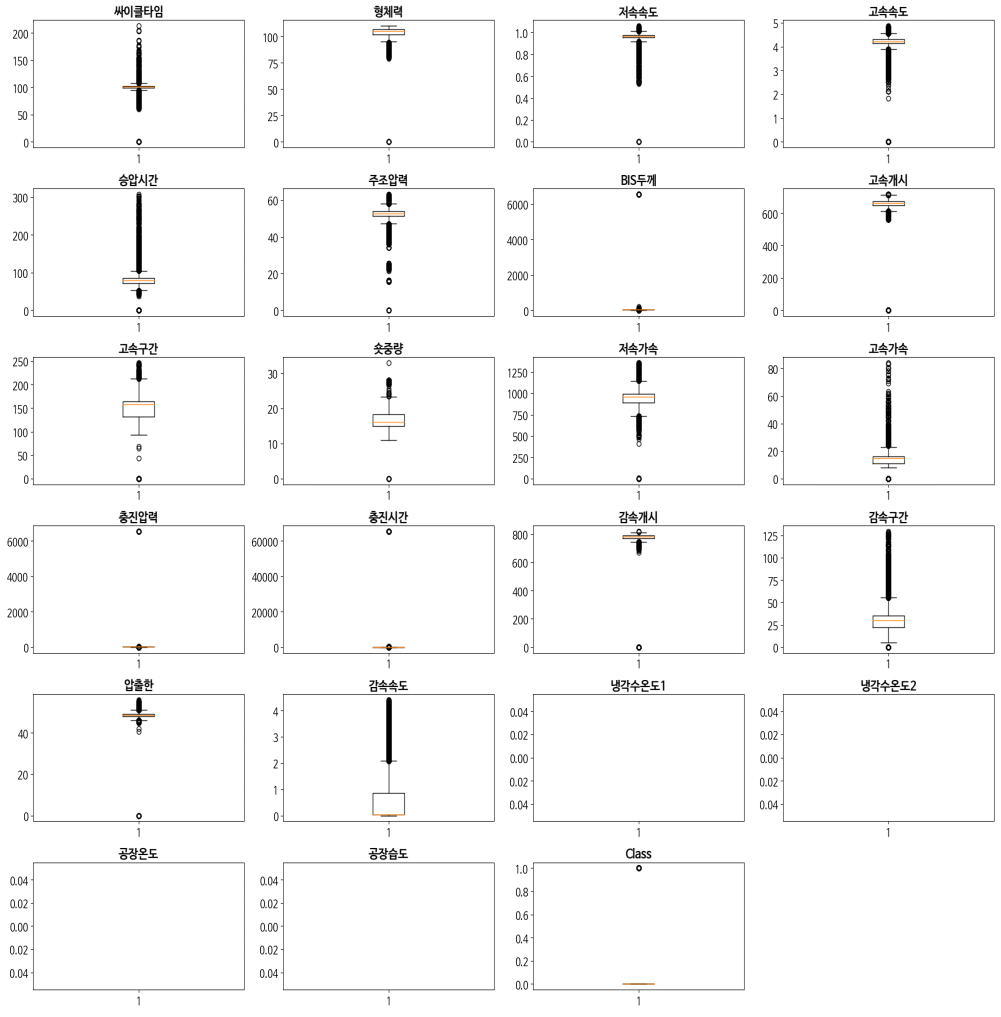
설비 A,B와 8개의 금형에서 수집된 데이터는 싸이클타임, 형체력, 저속속도 등 18개의 PLC데이터와, 냉각수온도, 공장온습도등 4개의 센서데이터, 1개의 Class(라벨)로 구성되어 있으며, 일부 결측치와, 이상치를 포함하고 있다. 변수정보 및 기초 통계량은 [표 3-3], [그림 3-2], [그림 3-2] 에서 확인 할 수 있다.

[표 3-3] 변수정보 및 기초 통계량

No.	변수	수집방법	단위	N	N*	평균	표준 편차	최소값	Q1	중위수	Q3	최대값
1	싸이클타임	PLC	s	105748	0	102.110	5.20848	0	99.7	101.9	103.1	213.1
2	형체력	PLC	%	105748	0	103.899	4.74493	0	102	105	107	110
3	저속속도	PLC	m/s	105748	0	0.926159	0.0999688	0	0.952	0.964	0.976	1.059
4	고속속도	PLC	m/s	105748	0	4.15515	0.363753	0	4.139	4.226	4.309	4.871
5	승압시간	PLC	ms	105748	0	82.2753	20.9949	0	72	79	85	307
6	주조압력	PLC	Mpa	105748	0	51.0640	5.82049	0	51.2	52.6	54	63.1
7	BIS두께	PLC	mm	105748	0	34.2271	173.762	0	27	29.5	32.4	6553.4
8	고속개시	PLC	mm	105748	0	661.471	36.7122	0	649.1	660.9	674.3	715.5
9	고속구간	PLC	mm	105748	0	148.855	24.0198	0	132.2	157.7	164.5	246.3
10	숏중량	PLC	kg	105748	0	17.3157	3.49636	0	15	16.1	18.4	33
11	저속가속	PLC	ms	105748	0	963.341	100.153	0	890	962	993	1357
12	고속가속	PLC	ms	105748	0	14.8382	5.03229	0	11	15	16	84
13	충진압력	PLC	Mpa	105748	0	16.3359	160.884	0	11.4	12.4	13.9	6553.5
14	충진시간	PLC	ms	105748	0	246.188	3602.57	0	41	47	50	65534
15	감속개시	PLC	mm	105748	0	776.375	38.7522	0	771.8	783.6	789.6	820.4
16	감속구간	PLC	mm	105748	0	33.9520	19.6438	0	22.3	30.2	35.5	129.6
17	압출한	PLC	-	105748	0	48.5661	1.54023	0	47.9	48.5	49.2	56
18	감속속도	PLC	m/s	105748	0	0.754504	1.27900	0	0.031	0.033	0.85	4.413
19	냉각수온도1	센서설치	도	105649	99	35.0727	5.82502	2	32.772	35	37.113	96.836
20	냉각수온도2	센서설치	도	105649	99	25.9907	2.89559	2	26	26.736	27	88.768
21	공장온도	센서설치	도	105669	79	18.9291	6.51558	4.891	13.183	18.822	24.402	34.699
22	공장습도	센서설치	%	105669	79	55.5848	10.5073	15.821	48.6405	56.311	63.209	76.785
23	Class	숙련된작업자	합/불	105748	0	0.0487007	0.215243	0	0	0	0	1



[그림 3-2] 변수에 대한 히스토그램



[그림 3-3] 변수에 대한 박스플롯

제 4 절 데이터 전처리

결측치 처리: 수집된 데이터에서 냉각수온도1,냉각수온도2, 공장온도, 공장습도의 결측치와, PLC로 수집되는 모든 데이터가 0으로 수집된 항목은 제거하였다.

이상치 탐지 및 처리: 승압시간, BIS두께, 저속가속, 저속변동, 고속변동, 충진압력, 충진시간에 있는 극단값의 이상치(65534, 6553.4, 6553.5)는 데이터 분포에 따라 평균치로 대체하거나, 극단값을 제한하였고, 일부는 삭제하였다.

PLC에서 수집되는 18개 변수 중, 고속속도가 0일때는 승압시간, 고속개시, 고속구간, 고속가속, 충진압력, 충진시간, 감속개시, 감속구간이 모두 0으로 표시되는 특징이 있었으며, 이에 해당하는 데이터는 188건이 있었고, 이중 97.9%(185건)가 불량이었다.

[표 3-4] 전처리 후 기초 통계량

No.	변수	N	N*	평균	표준 편차	최소값	Q1	중위수	Q3	최대값
1	싸이클타임	105263	0	102.146	5.02603	0	99.7	101.9	103.1	213.1
2	형체력	105263	0	103.917	4.56966	79	102	105	107	110
3	저속속도	105263	0	0.926811	0.0988458	0.533	0.952	0.964	0.976	1.059
4	고속속도	105263	0	4.15567	0.360137	0	4.139	4.226	4.309	4.871
5	승압시간	105263	0	82.2599	20.8744	0	72	79	85	307
6	주조압력	105263	0	51.0699	5.76988	15.7	51.2	52.6	54	63.1
7	BIS두께	105263	0	29.6703	5.78943	0.2	27.1	29.5	32.4	137.7
8	고속개시	105263	0	661.693	35.7869	0	649.4	660.9	674.3	715.5
9	고속구간	105263	0	148.755	23.8469	0	132.2	157.7	164.5	246.3
10	숏중량	105263	0	17.3248	3.49592	11	15	16.1	18.4	33
11	저속가속	105263	0	963.535	99.4596	8	890	962	993	1357
12	고속가속	105263	0	14.8185	5.00740	0	11	15	16	84
13	충진압력	105263	0	12.3814	2.02869	0	11.4	12.4	13.9	20.8
14	충진시간	105263	0	48.6132	28.6474	0	41	47	50	600
15	감속개시	105263	0	776.491	37.6152	0	771.8	783.6	789.6	820.4
16	감속구간	105263	0	33.9579	19.6338	0	22.3	30.2	35.5	129.6
17	압출한	105263	0	48.5710	1.42063	0	47.9	48.5	49.1	56
18	감속속도	105263	0	0.750160	1.27670	0	0.031	0.033	0.825	4.413
19	냉각수온도1	105263	0	35.0727	5.83006	2	32.771	35	37.1	96.836
20	냉각수온도2	105263	0	25.9907	2.89857	2	26	26.736	27	88.768
21	공장온도	105263	0	18.9368	6.52281	4.891	13.153	18.858	24.412	34.699
22	공장습도	105263	0	55.5971	10.5109	15.821	48.673	56.327	63.218	76.785
23	Class	105263	0	0.0460466	0.209587	0	0	0	0	1

[표 3-5] 전처리 후 금형 별 수집된 데이터 수

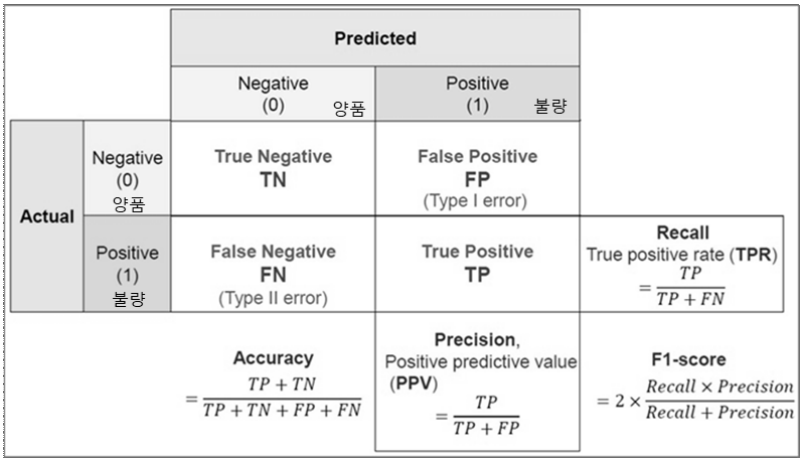
설비	금형	Count	양품수량	불량수량	불량률
A	A5	9,845	9,178	667	6.78%
	A8	22,851	22,112	739	3.23%
	A15	10,833	10,422	411	3.79%
	A19	5,687	4,871	816	14.35%
B	B1	31,561	30,652	909	2.88%
	B3	6,174	5,597	577	9.35%
	B20	10,647	10,218	429	4.03%
	B21	7,665	7,366	299	3.90%
A합계	A	49,216	46,583	2,633	5.35%
B합계	B	56,047	53,833	2,214	3.95%
A+B합계	AB	105,263	100,416	4,847	4.60%

[표 3-6] 전처리 후 금형 별 수집된 불량정보

설비	금형	불량유형										
		미성형	미인출	소착	수축	오염	저압	절육	크랙	탕회	테스트	기타
A	A5	433	8	0	3	7	180	30	0	1	0	5
	A8	309	31	0	5	133	227	2	0	1	0	31
	A15	79	8	0	6	94	150	73	0	0	0	1
	A19	386	20	0	36	71	159	22	97	21	0	4
B	B1	183	16	6	2	33	269	391	0	0	0	9
	B3	450	3	0	15	12	81	14	0	0	0	2
	B20	125	14	2	10	2	239	30	0	0	0	7
	B21	158	3	0	8	7	119	4	0	0	0	0
A합계		1,207	67	0	50	305	716	127	97	23	0	41
B합계		916	36	8	35	54	708	439	0	0	0	18
A+B합계		2,123	103	8	85	359	1424	566	97	23	0	59

제 5 절 성능평가지표의 정의 및 계산방법

이번 연구에서 랜덤 포레스트 모델의 성능을 평가하기 위해 사용되는 주요 성능 지표는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score)입니다. 이러한 지표들은 모델이 불량품을 얼마나 잘 예측하는지를 종합적으로 판단하는 데 중요합니다. 각 지표의 정의와 계산 방법은 다음과 같다.



[그림 3-4] 성능 평가지표의 정의 및 계산 방법

- 정확도 (Accuracy): 모든 예측 중 올바르게 예측한 비율
- 정밀도 (Precision): 모델이 불량으로 예측한 샘플 중 실제 불량인 비율
- 재현율 (Recall): 실제 불량 샘플 중 모델이 불량으로 예측한 비율
- F1-Score: 정밀도와 재현율의 조화 평균으로, 두 지표의 균형을 나타냄

제 6 절 학습데이터셋 구성 및 특징

1) 학습데이터의 불균형 처리

학습과 검증데이터를 7:3으로 나누었고, 학습데이터에는 SMOTE샘플링, 언더샘플링을 적용하였다. 오버샘플링 및 언더샘플링 적용 시 불량 비율은 10%, 20%, 30%, 40%, 50%를 각각 적용하였다.

2) 적용데이터 및 적용AI모델

A설비 4개 금형과, B설비 4개 금형, A설비 전체 1개, B설비 전체 1개, A+B 설비 전체 1개, 총 11개 금형 데이터 셋을 구성하였다. 금형 별로 학습데이터의 불량비율을 SMOTE샘플링(10%, 20%, 30%, 40%, 50%), 언더샘플링(10%, 20%, 30%, 40%, 50%)한 10개의 추가 학습용 데이터셋을 구성하였다. 불량률이 10%가 넘는 A19 금형의 경우 10%에 대한 샘플링 데이터는 제외하고 총 119개의 학습용 데이터셋을 도출하였다.

3) 검증 데이터셋

각 금형별로, 학습에 활용하지 않은 나머지 30%의 데이터를 검증 데이터로 활용하였다.

제 7 절 모델 생성 및 평가 결과

리샘플링을 하지 않은 원본 학습데이터와, 불량률을 10%, 20%, 30%, 40%, 50%로 SMOTE샘플링 및 언더샘플링을 적용한 학습데이터로 랜덤 포레스트 모델을 생성하였다. 랜덤 포레스트 모델 생성 시 n_estimators 는 99로 설정하였다. 생성된 랜덤 포레스트 모델을, 학습데이터와 검증데이터에 적용하고 평가하고, 금형에 대한 평가결과를 [표 3-7], [표 3-8], [표 3-9], [표 3-10]에 나타내었다.

[표 3-7] A5, A8, A15 금형에 대한 평가결과

No.	금형	학습데이터(70%)				학습데이터에 대한 평가결과				검증데이터(30%)			검증데이터에 대한 평가결과			
		학습데이터	양품	불량	불량률	Accuracy	Recall	Precision	F1_score	양품	불량	불량률	Accuracy	Recall	Precision	F1_score
1	A5	원본	6,425	466	6.76%	0.999855	0.997854	1	0.998926	2,753	201	6.80%	0.96107	0.452736	0.947917	0.612795
2		SMOTE샘플링 10%	6,425	713	9.99%	1	1	1	1				0.959716	0.462687	0.894231	0.609836
3		SMOTE샘플링 20%	6,425	1,606	20.00%	1	1	1	1				0.959377	0.492537	0.846154	0.622642
4		SMOTE샘플링 30%	6,425	2,753	30.00%	1	1	1	1				0.958362	0.517413	0.8	0.628399
5		SMOTE샘플링 40%	6,425	4,283	40.00%	1	1	1	1				0.95633	0.517413	0.764706	0.617211
6		SMOTE샘플링 50%	6,425	6,425	50.00%	1	1	1	1				0.954638	0.522388	0.734266	0.610465
7		언더샘플링 10%	4,194	466	10.00%	1	1	1	1				0.959039	0.467662	0.87037	0.608414
8		언더샘플링 20%	1,864	466	20.00%	1	1	1	1				0.952945	0.512438	0.715278	0.597101
9		언더샘플링 30%	1,087	466	30.01%	1	1	1	1				0.936696	0.58209	0.531818	0.555819
10		언더샘플링 40%	699	466	40.00%	1	1	1	1				0.90149	0.631841	0.369186	0.466055
11		언더샘플링 50%	466	466	50.00%	1	1	1	1				0.841571	0.691542	0.255046	0.372654
12	A8	원본	15,497	498	3.11%	1	1	1	1	6,615	241	3.52%	0.984102	0.651452	0.862637	0.742317
13		SMOTE샘플링 10%	15,497	1,721	10.00%	1	1	1	1				0.984102	0.701245	0.820388	0.756152
14		SMOTE샘플링 20%	15,497	3,874	20.00%	1	1	1	1				0.982351	0.717842	0.765487	0.740899
15		SMOTE샘플링 30%	15,497	6,641	30.00%	1	1	1	1				0.982205	0.73444	0.753191	0.743697
16		SMOTE샘플링 40%	15,497	10,331	40.00%	1	1	1	1				0.981622	0.738589	0.738589	0.738589
17		SMOTE샘플링 50%	15,497	15,497	50.00%	1	1	1	1				0.980455	0.73444	0.716599	0.72541
18		언더샘플링 10%	4,482	498	10.00%	1	1	1	1				0.979288	0.721992	0.698795	0.710204
19		언더샘플링 20%	1,992	498	20.00%	1	1	1	1				0.97112	0.751037	0.567398	0.646429
20		언더샘플링 30%	1,162	498	30.00%	1	1	1	1				0.957701	0.763485	0.441247	0.559271
21		언더샘플링 40%	747	498	40.00%	1	1	1	1				0.939323	0.79668	0.34347	0.48
22		언더샘플링 50%	498	498	50.00%	1	1	1	1				0.90738	0.850622	0.254975	0.392344
23	A15	원본	7,309	274	3.61%	0.999868	0.99635	1	0.998172	3,113	137	4.22%	0.975385	0.525547	0.827586	0.642857
24		SMOTE샘플링 10%	7,309	812	10.00%	0.999877	0.998768	1	0.999384				0.975692	0.591241	0.778846	0.672199
25		SMOTE샘플링 20%	7,309	1,827	20.00%	1	1	1	1				0.975385	0.583942	0.776699	0.666667
26		SMOTE샘플링 30%	7,309	3,132	30.00%	1	1	1	1				0.973846	0.583942	0.740741	0.653061
27		SMOTE샘플링 40%	7,309	4,872	40.00%	1	1	1	1				0.973231	0.576642	0.731481	0.644898
28		SMOTE샘플링 50%	7,309	7,309	50.00%	1	1	1	1				0.973231	0.576642	0.731481	0.644898
29		언더샘플링 10%	2,466	274	10.00%	1	1	1	1				0.975077	0.613139	0.75	0.674699
30		언더샘플링 20%	1,096	274	20.00%	1	1	1	1				0.965846	0.649635	0.585526	0.615917
31		언더샘플링 30%	639	274	30.01%	1	1	1	1				0.955077	0.678832	0.476923	0.560241
32		언더샘플링 40%	411	274	40.00%	1	1	1	1				0.938462	0.686131	0.374502	0.484536
33		언더샘플링 50%	274	274	50.00%	1	1	1	1				0.889846	0.715328	0.235012	0.353791

[표 3-8] A19, B1, B3 금형에 대한 평가결과

No.	금형	학습데이터(70%)				학습데이터에 대한 평가결과				검증데이터(30%)			검증데이터에 대한 평가결과			
		학습데이터	양품	불량	불량률	Accuracy	Recall	Precision	F1_score	양품	불량	불량률	Accuracy	Recall	Precision	F1_score
34	A19	원본	3,411	569	14.30%	1	1	1	1	1,460	247	14.47%	0.930873	0.623482	0.860335	0.723005
35		SMOTE샘플링 20%	3,411	852	19.99%	1	1	1	1				0.927944	0.635628	0.826316	0.718535
36		SMOTE샘플링 30%	3,411	1,461	29.99%	1	1	1	1				0.926772	0.647773	0.808081	0.719101
37		SMOTE샘플링 40%	3,411	2,274	40.00%	1	1	1	1				0.925015	0.684211	0.771689	0.725322
38		SMOTE샘플링 50%	3,411	3,411	50.00%	1	1	1	1				0.912712	0.684211	0.704167	0.694045
39		언더샘플링 20%	2,276	569	20.00%	1	1	1	1				0.9256	0.643725	0.80303	0.714607
40		언더샘플링 30%	1,327	569	30.01%	1	1	1	1				0.917399	0.696356	0.722689	0.709278
41		언더샘플링 40%	853	569	40.01%	1	1	1	1				0.894552	0.761134	0.608414	0.676259
42		언더샘플링 50%	569	569	50.00%	1	1	1	1				0.836555	0.817814	0.463303	0.591508
43	B1	원본	21,467	625	2.83%	1	1	1	1	9,185	284	3.00%	0.978245	0.323944	0.867925	0.471795
44		SMOTE샘플링 10%	21,467	2,385	10.00%	1	1	1	1				0.978667	0.380282	0.80597	0.516746
45		SMOTE샘플링 20%	21,467	5,366	20.00%	1	1	1	1				0.977822	0.390845	0.75	0.513889
46		SMOTE샘플링 30%	21,467	9,200	30.00%	1	1	1	1				0.977506	0.40493	0.72327	0.519187
47		SMOTE샘플링 40%	21,467	14,311	40.00%	1	1	1	1				0.976555	0.387324	0.696203	0.497738
48		SMOTE샘플링 50%	21,467	21,467	50.00%	1	1	1	1				0.976449	0.411972	0.676301	0.512035
49		언더샘플링 10%	5,625	625	10.00%	1	1	1	1				0.976238	0.390845	0.680982	0.496644
50		언더샘플링 20%	2,500	625	20.00%	1	1	1	1				0.967684	0.450704	0.460432	0.455516
51		언더샘플링 30%	1,458	625	30.00%	1	1	1	1				0.94445	0.464789	0.26087	0.334177
52		언더샘플링 40%	937	625	40.01%	1	1	1	1				0.898933	0.503521	0.149114	0.230088
53		언더샘플링 50%	625	625	50.00%	1	1	1	1				0.802936	0.573944	0.08543	0.148723
54	B3	원본	3,921	400	9.26%	1	1	1	1	1,676	177	9.55%	0.91959	0.254237	0.725806	0.376569
55		SMOTE샘플링 10%	3,921	435	9.99%	1	1	1	1				0.916892	0.259887	0.666667	0.373984
56		SMOTE샘플링 20%	3,921	980	20.00%	1	1	1	1				0.910416	0.293785	0.55914	0.385185
57		SMOTE샘플링 30%	3,921	1,680	29.99%	1	1	1	1				0.910416	0.355932	0.547826	0.431507
58		SMOTE샘플링 40%	3,921	2,614	40.00%	1	1	1	1				0.905019	0.384181	0.503704	0.435897
59		SMOTE샘플링 50%	3,921	3,921	50.00%	1	1	1	1				0.899622	0.389831	0.469388	0.425926
60		언더샘플링 10%	3,600	400	10.00%	0.99975	0.9975	1	0.998748				0.91959	0.265537	0.712121	0.386831
61		언더샘플링 20%	1,600	400	20.00%	1	1	1	1				0.907178	0.389831	0.518797	0.445161
62		언더샘플링 30%	933	400	30.01%	1	1	1	1				0.874258	0.446328	0.369159	0.404092
63		언더샘플링 40%	600	400	40.00%	1	1	1	1				0.832704	0.497175	0.28479	0.36214
64		언더샘플링 50%	400	400	50.00%	1	1	1	1				0.762008	0.610169	0.225	0.328767

[표 3-9] B20, B21, A 금형에 대한 평가결과

No.	금형	학습데이터(70%)				학습데이터에 대한 평가결과				검증데이터(30%)			검증데이터에 대한 평가결과			
		학습데이터	양품	불량	불량률	Accuracy	Recall	Precision	F1_score	양품	불량	불량률	Accuracy	Recall	Precision	F1_score
65		원본	7,163	289	3.88%	0.999866	0.99654	1	0.998267				0.973396	0.492857	0.831325	0.618834
66		SMOTE샘플링 10%	7,163	795	9.99%	1	1	1	1				0.973709	0.564286	0.77451	0.652893
67		SMOTE샘플링 20%	7,163	1,790	19.99%	1	1	1	1				0.973396	0.607143	0.73913	0.666667
68		SMOTE샘플링 30%	7,163	3,069	29.99%	1	1	1	1				0.973083	0.635714	0.717742	0.674242
69		SMOTE샘플링 40%	7,163	4,775	40.00%	1	1	1	1				0.97277	0.614286	0.722689	0.664093
70	B20	SMOTE샘플링 50%	7,163	7,163	50.00%	1	1	1	1	3,055	140	4.38%	0.968388	0.571429	0.661157	0.613027
71		언더샘플링 10%	2,601	289	10.00%	1	1	1	1				0.973396	0.621429	0.731092	0.671815
72		언더샘플링 20%	1,156	289	20.00%	1	1	1	1				0.968075	0.664286	0.628378	0.645833
73		언더샘플링 30%	674	289	30.01%	1	1	1	1				0.958059	0.678571	0.516304	0.58642
74		언더샘플링 40%	433	289	40.03%	1	1	1	1				0.94241	0.7	0.408333	0.515789
75		언더샘플링 50%	289	289	50.00%	1	1	1	1				0.906729	0.75	0.285326	0.413386
76		원본	5,163	202	3.77%	0.999814	0.99505	1	0.997519				0.973913	0.391753	0.974359	0.558824
77		SMOTE샘플링 10%	5,163	573	9.99%	1	1	1	1				0.973913	0.412371	0.930233	0.571429
78		SMOTE샘플링 20%	5,163	1,290	19.99%	1	1	1	1				0.974348	0.453608	0.88	0.598639
79		SMOTE샘플링 30%	5,163	2,212	29.99%	1	1	1	1				0.973478	0.443299	0.86	0.585034
80		SMOTE샘플링 40%	5,163	3,442	40.00%	1	1	1	1				0.975217	0.494845	0.857143	0.627451
81	B21	SMOTE샘플링 50%	5,163	5,163	50.00%	1	1	1	1	2,203	97	4.22%	0.973478	0.484536	0.810345	0.606452
82		언더샘플링 10%	1,818	202	10.00%	1	1	1	1				0.972609	0.412371	0.869565	0.559441
83		언더샘플링 20%	808	202	20.00%	1	1	1	1				0.96	0.505155	0.526882	0.515789
84		언더샘플링 30%	471	202	30.01%	1	1	1	1				0.931304	0.546392	0.317365	0.401515
85		언더샘플링 40%	303	202	40.00%	1	1	1	1				0.876957	0.639175	0.2	0.304668
86		언더샘플링 50%	202	202	50.00%	1	1	1	1				0.797391	0.659794	0.128773	0.215488
87		원본	32,623	1,828	5.31%	0.999913	0.998359	1	0.999179				0.972164	0.580124	0.864815	0.694424
88		SMOTE샘플링 10%	32,623	3,624	10.00%	0.999972	0.999724	1	0.999862				0.970674	0.601242	0.812081	0.690935
89		SMOTE샘플링 20%	32,623	8,155	20.00%	1	1	1	1				0.968642	0.62236	0.759091	0.683959
90		SMOTE샘플링 30%	32,623	13,981	30.00%	1	1	1	1				0.966813	0.637267	0.721519	0.676781
91		SMOTE샘플링 40%	32,623	21,748	40.00%	1	1	1	1				0.96512	0.64472	0.69385	0.668384
92	A	SMOTE샘플링 50%	32,623	32,623	50.00%	1	1	1	1	13,960	805	5.45%	0.964307	0.657143	0.678205	0.667508
93		언더샘플링 10%	16,452	1,828	10.00%	1	1	1	1				0.969861	0.608696	0.790323	0.687719
94		언더샘플링 20%	7,312	1,828	20.00%	0.999891	0.999453	1	0.999726				0.961327	0.659627	0.641304	0.650337
95		언더샘플링 30%	4,265	1,828	30.00%	1	1	1	1				0.93891	0.714286	0.461107	0.560429
96		언더샘플링 40%	2,742	1,828	40.00%	1	1	1	1				0.906197	0.763975	0.339779	0.470363
97		언더샘플링 50%	1,828	1,828	50.00%	1	1	1	1				0.858855	0.801242	0.25107	0.382336

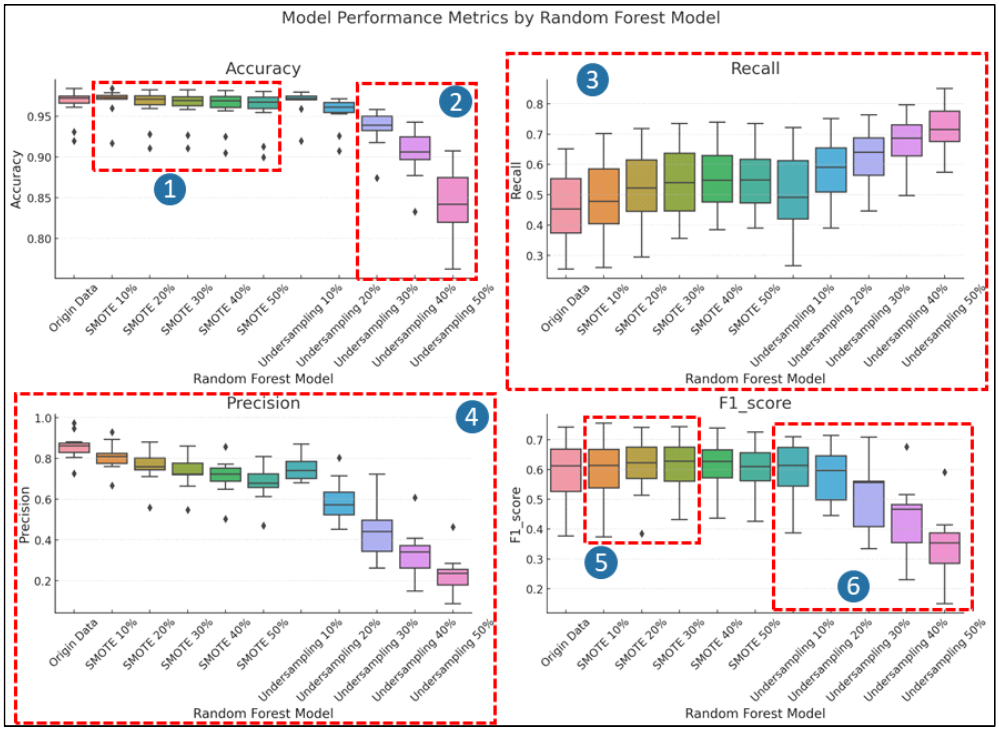
[표 3-10] B, AB 금형에 대한 평가결과

No.	금형	학습데이터(70%)				학습데이터에 대한 평가결과				검증데이터(30%)			검증데이터에 대한 평가결과			
		학습데이터	양품	불량	불량률	Accuracy	Recall	Precision	F1_score	양품	불량	불량률	Accuracy	Recall	Precision	F1_score
98	B	원본	37,684	1,548	3.95%	0.999847	0.996124	1	0.998058	16,149	666	3.96%	0.971038	0.354354	0.805461	0.492179
99		SMOTE샘플링 10%	37,684	4,187	10.00%	0.999952	0.999522	1	0.999761				0.971335	0.402402	0.761364	0.526523
100		SMOTE샘플링 20%	37,684	9,421	20.00%	0.999979	0.999894	1	0.999947				0.970681	0.436937	0.711491	0.541395
101		SMOTE샘플링 30%	37,684	16,150	30.00%	1	1	1	1				0.969194	0.448949	0.664444	0.535842
102		SMOTE샘플링 40%	37,684	25,122	40.00%	1	1	1	1				0.968659	0.456456	0.648188	0.535683
103		SMOTE샘플링 50%	37,684	37,684	50.00%	1	1	1	1				0.967113	0.460961	0.612774	0.526135
104		언더샘플링 10%	13,932	1,548	10.00%	1	1	1	1				0.970027	0.441441	0.690141	0.538462
105		언더샘플링 20%	6,192	1,548	20.00%	1	1	1	1				0.956111	0.512012	0.452255	0.480282
106		언더샘플링 30%	3,612	1,548	30.00%	0.999806	0.999354	1	0.999677				0.932977	0.591592	0.315452	0.411488
107		언더샘플링 40%	2,322	1,548	40.00%	1	1	1	1				0.90675	0.623123	0.239607	0.346122
108		언더샘플링 50%	1,548	1,548	50.00%	1	1	1	1				0.836396	0.692192	0.153309	0.251021
109	AB	원본	70,281	3,403	4.62%	0.999973	0.999412	1	0.999706	30,135	1444	4.57%	0.972165	0.451524	0.882273	0.597343
110		SMOTE샘플링 10%	70,281	7,809	10.00%	0.999987	0.999872	1	0.999936				0.972102	0.493075	0.826945	0.617787
111		SMOTE샘플링 20%	70,281	17,570	20.00%	0.999989	0.999943	1	0.999972				0.970297	0.522161	0.752495	0.616517
112		SMOTE샘플링 30%	70,281	30,120	30.00%	0.99999	0.999967	1	0.999983				0.96941	0.539474	0.721296	0.617274
113		SMOTE샘플링 40%	70,281	46,854	40.00%	0.999991	0.999979	1	0.999989				0.967732	0.547091	0.683983	0.607926
114		SMOTE샘플링 50%	70,281	70,281	50.00%	1	1	1	1				0.966117	0.548476	0.654545	0.596835
115		언더샘플링 10%	30,627	3,403	10.00%	0.999971	0.999706	1	0.999853				0.970898	0.515235	0.772586	0.618197
116		언더샘플링 20%	13,612	3,403	20.00%	0.999941	0.999706	1	0.999853				0.96105	0.590028	0.571812	0.580777
117		언더샘플링 30%	7,940	3,403	30.00%	0.999912	0.999706	1	0.999853				0.941385	0.639889	0.409756	0.499594
118		언더샘플링 40%	5,104	3,403	40.00%	0.999882	0.999706	1	0.999853				0.910478	0.690443	0.295232	0.413607
119		언더샘플링 50%	3,403	3,403	50.00%	1	1	1	1				0.853479	0.745845	0.201799	0.317652

제 8 절 평가 결과 분석

아래의 [그림 3-5]의 평가결과에 대한 박스플롯을 보면 ①SMOTE샘플링 10%~50% 까지의 모델은 높은 정확도를 유지하고 있으며, ②언더샘플링 모델은 학습DATA가 감소하면서 정확도가 낮아짐을 보인다. ③학습데이터의 불량률 증가 시 재현율은 증가하며, ④불량률 증가 시 정밀도는 감소함을 보인다, 특히 언더샘플링 모델은 정밀도가 크게 감소한다. ⑤SMOTE 10%~30% 모델에서 높은 F1-Score를 유지하고 있으며, ⑥언더샘플링모델은 전반적으로 F1-Score가 낮게 분포한다.

이러한 내용으로 보아 학습데이터 불량률의 불균형 해소를 위한 샘플링 기법은 모델성능에 크게 영향을 미치며 언더샘플링의 경우 SMOTE에 비해 정확도 및 정밀도, F1 Score가 크게 떨어질 수 있음을 보여준다.

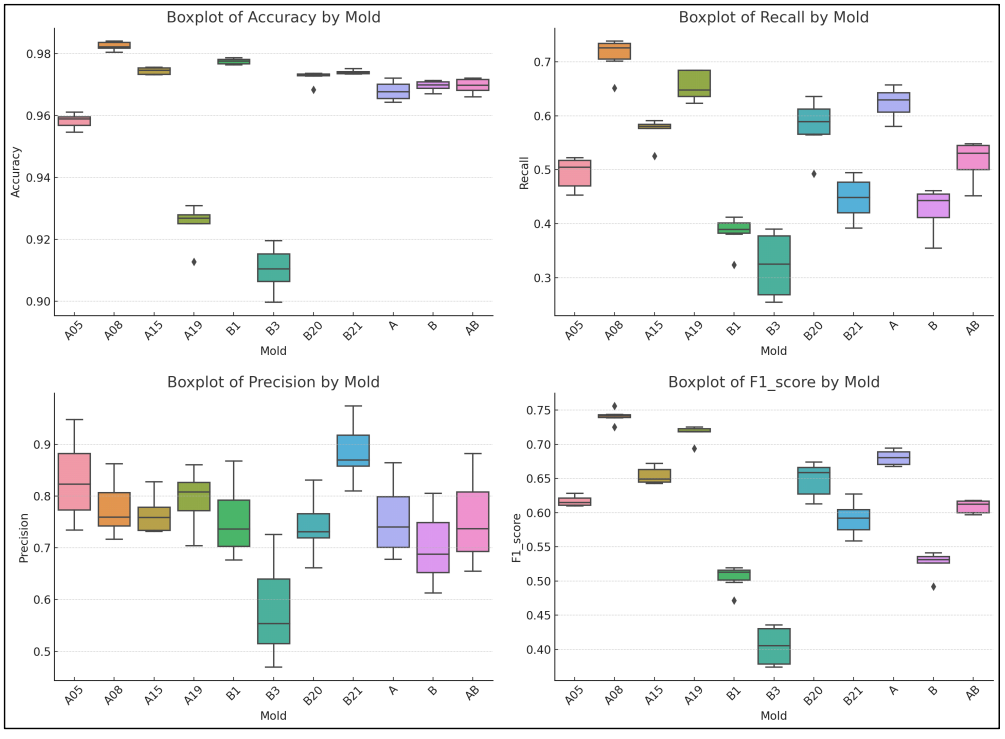


[그림 3-5] 랜덤 포레스트 모델의 평가결과에 대한 박스플롯

정확도 및 정밀도가 크게 떨어지는 언더샘플링 모델은 제외하고, SMOTE샘플링 모델에 대하여 [그림 3-6] 금형별 성과지표에 대한 박스플롯을 확인하였다.

정확도는 대부분의 금형이 비슷한 분포를 보이지만 A19와 B3금형은 상대적으로 낮은 정확도를 보여주고 있다. A08금형의 재현율이 가장 높게 나타난다. F1-score 및 재현율,정밀도는 금형마다 MIN/MAX 범위가 상이하며 이는 금형마다 데이터의 특징이 다르다는 것을 보여준다.

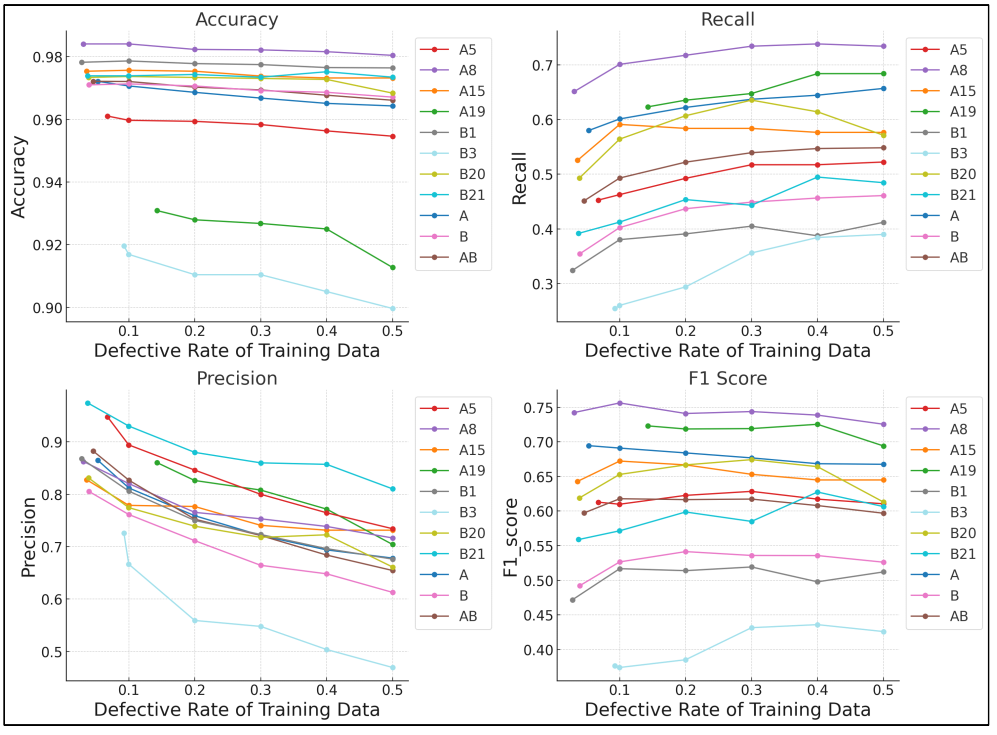
같은 설비에서 생산된 제품이라도 사용된 금형에 따라 랜덤 포레스트 모델의 검증 정확도가 다름을 알 수 있다. 따라서 모델 최적화 시 각 금형의 특성과 데이터 분포를 고려해야 한다.



[그림 3-6] 금형 별 성과지표에 대한 박스플롯

[그림 3-7]은 SMOTE샘플링 모델의 학습데이터의 불량률과 성과지표간의 관계를 보여주고 있다. 정확도는 불량률이 증가 시 약간 감소하는 경향을 보이며, 불량률이 증가할수록 재현율은 증가하고 정밀도는 감소한다. F1-score는 불량률이 10%가 될 때까지는 증가하는 경향을 보이나 이후에는 특별한 관계를 보이지 않고, 재현율 증가와 정밀도의 감소를 균형 있게 반영하고 있다.

생산 환경에서 불량 탐지는 매우 중요한 요소이므로, 모델의 재현율과 정밀도 사이의 균형을 맞추는 것이 중요하다. 이와 같은 결과는 SMOTE를 통한 불량비율의 균형이 모델의 성능을 최적화하고 실제 적용 가능성을 높이는 데 중요한 역할을 한다는 것을 보여준다. 따라서 모델 개발 및 운영 과정에서 이러한 성능지표와 불량률의 관계를 면밀히 검토하고 반영하는 것이 중요하다.



[그림 3-7] 학습데이터의 불량률과 성과지표의 관계

제 4 장 결론 및 시사점

본 연구는 학습데이터의 불량률과 다양한 샘플링 기법이 AI 모델의 성능에 미치는 영향을 분석하였다. 이를 통해 모델의 성능의 최적화를 위한 샘플링 방법의 가이드를 제시하고자 하였다. 다음은 본 연구의 주요 결론이다.

학습 데이터의 불량률이 높아질수록 모델의 재현율(Recall)은 향상되지만, 정밀도(Precision)와 정확도(Accuracy)는 감소하는 경향을 보였다. 이는 모델이 불량 샘플을 더 잘 예측 하지만 정상 데이터를 불량으로 잘못 예측하는 비율이 증가함을 의미한다.

이러한 결과는 SMOTE를 통한 불량 비율의 균형이 모델 성능을 최적화하고 실제 적용 가능성을 높이는 데 중요한 역할을 한다는 것을 보여준다. SMOTE샘플링 및 언더샘플링이 적용된 랜덤 포레스트 모델들을 비교한 결과, 평균적으로 SMOTE 30% 모델이 가장 높은 F1-score를 보였다. SMOTE 샘플링 기법은 재현율과 F1-score를 향상시키는데 효과적이었지만, 언더샘플링기법은 샘플링 비율이 높아질수록 성능이 급격히 떨어지는 경향을 보였다. 원본 데이터 기반 모델은 정밀도가 가장 높았고, 금형 별 모델 성능을 분석한 결과, 금형 별로 정확도, 재현율, F1-score 지표에서 큰 차이가 발생하였다. 이는 각 금형의 특성과 데이터 분포에 따라 모델 성능이 달라지므로, 금형 별로 최적화된 모델을 개발할 필요가 있음을 시사한다.

본 연구를 통해 다이캐스팅 공정의 불균형 데이터에 대하여, 학습 데이터의 불량률과 샘플링 기법이 AI 모델의 성능에 미치는 영향을 분석하였다. 모델 성능을 최적화하기 위해서는 데이터 전처리와 샘플링 기법을 적용한 모델 선

택이 필수적 이라는 것을 알 수 있었다. 각 금형의 특성에 맞춘 최적화된 모델을 개발함으로써, 제조 공정의 불량 예측 정확도를 높이고 품질 관리를 강화할 수 있을 것이다.

이 연구는 머신러닝 모델의 성능 최적화를 위한 샘플링 기법의 가이드를 제공하며, 제조업 분야에서 머신러닝 모델의 적용을 확대하는 데 기여할 것으로 기대된다. 앞으로의 연구에서는 품질에 영향을 미치는 다양한 제조 공정 변수를 추가하고, 충분한 데이터를 확보하여 모델 성능을 개선하는 것이 필요하다. 또한 모델을 실제 제조현장에 적용하여 성능을 검증하고, 보다 정교한 최적화 기법을 개발하는 것이 중요하다.

참 고 문 헌

1. 국내문헌

국가뿌리산업진흥센터. (2023).『2023 뿌리산업 백서』.

김문조. (2023). 주조 분야 인공지능 활용 사례.『한국주조공학회지』, 43(2), 77-83.

김재선, 박춘우, 박원석, 박영현, 조창현, 김동주. (2023). 공정 매개변수 및 열화상 이미지를 기반으로 한 다공성 결함 감지를 위한 고압 다이캐스팅 결함 예측 딥러닝 알고리즘에 관한 연구.『한국CDE학회 논문집』, 28(3), 222-231.

김정태, 이영철, 이정수. (2021). 다이캐스팅공장을 위한 스마트팩토리 추진실태 및 추진방안.『한국주조공학회지』, 41(2), 178-188.

김종성, 이준형, 김동현, 최창현, 이명진. (2019). 머신러닝 기반의 호우피해 발생 확률 예측 모형 개발.『한국방재학회논문집』, 19(6), 115-127.

김종훈, 오하영. (2021). 머신러닝을 활용한 제품 특성 예측모델의 성능향상 방법 연구.『한국정보통신학회논문지』, 25(6), 749-756.

김준, 강형석, 이주연. (2020). 다이캐스팅 공정의 품질고도화를 위한 지능화 분석 시스템 개발.『한국정밀공학회지』, 37(4), 247-254.

박상우. (2022). "기계학습 기반의 중소 다이캐스팅 공정 품질예측 방법론 개발: 설비 주조파라미터 데이터를 중심으로". 동국대학교 일반대학원. 박사학위논문.

박주남. (2022). "설명 가능한 인공지능을 활용한 다이캐스팅 공정 품질 예측". 성균관대학교 대학원. 석사학위논문.

박철순, 김홍섭. (2022). 주조공정 설비에 대한 실시간 모니터링을 통한 불량예측에 대한 연구.『한국산업경영시스템학회지』, 45(4), 157-166.

박태윤. (2023). "AI 솔루션을 활용한 스마트 공장 생산제품의 불량검출에 관한

연구". 숭실대학교 대학원. 석사학위논문.

박호준. (2017). "다이캐스팅용 마그네슘합금의 기계적 특성에 미치는 합금 원소의 영향". 서울대학교 대학원. 석사학위논문.

오동호. (2015). "다이캐스팅 공정의 혁신을 통한 품질개선에 관한 연구". 인하대학교 대학원. 석사학위논문.

윤정근. (2023). "HPDC 공정 Data 분석을 통한 불량 예측 AI 연구". 인하대학교 제조혁신전문대학원. 석사학위논문.

이기성, 이종찬. (2022). 사출 공정에서 불량 예측을 위한 데이터 분석 및 AI 적용 모델. 『한국산학기술학회논문지』, 23(12), 1001-1008.

이승로, 이승철, 한도석, 김낙수. (2021). 고압 다이캐스팅 공정에서 제품 결함을 사전 예측하기 위한 기계 학습 기반의 공정관리 방안 연구. 『한국주조공학회지』, 41(6), 521-526.

이정수, 이영철. (2021). 『비지도 딥러닝 기반 다이캐스팅 제품 불량 검출』, 광주: 대한기계학회 2021년 학술대회.

이정수, 최영심. (2022). 다단계 딥러닝 기반 다이캐스팅 공정 불량 검출. 『한국주조공학회지』, 42(6), 369-376.

이주연. (2020). 다이캐스팅 공정 지능화를 위한 데이터 수집, 처리, 분석 및 활용 기술 개발. 『한국생산제조학회지』, 29(6), 441-448.

이준혁. (2019). "품질 제어를 위한 빅데이터 기반 사이버물리제조시스템 프레임워크 설계 및 구현: 로트 생산 방식의 주조 공정을 대상으로". 성균관대학교 대학원. 석사학위논문.

전보민. (2023). "불균형 데이터의 예측 모델 성능 향상을 위한 오버샘플링 기법". 고려대학교 대학원. 석사학위논문.

차형주. (2023). "머신러닝을 활용한 사출성형공정의 품질예측". 한밭대학교 대학원. 석사학위논문.

천정권. (2022). 다이캐스팅 공정 및 품질관리. 『한국주조공학회지』, 42(6), 379-386.

최낙훈, 오종석, 안종록, 김기선. (2021). 머신러닝을 활용한 자동차 시트용 폴리우레탄 발포공정의 불량 예측 모델 개발. 『한국산학기술학회논문지』, 22(6), 36-42.

최정길. (2021). 주물산업의 스마트 팩토리 (1) - 스마트 팩토리 -. 『한국주조공학회지』, 41(6), 575-580.

최창옥, 김창규, 길상철. (2015). 다이캐스팅기술의 개발과 발전동향. 『한국주조공학회지』, 35(5), 128-140.

최창옥. (2016). 『알루미늄합금 다이캐스팅의 사출성형』, 한국과학기술정보연구원.

2. 국외문헌

Breiman L. (2001). Random forests.『Machine Learning』, 45(1), 5-32.

C.-H. Lin, Hu. Guo-Hsin, C.-W. Ho, Hu. Chia-Yen. (2022). Press Casting Quality Prediction and Analysis Based on Machine Learning. *electronics*, 11(14), 2204.

E.J. Vinarcik. (2003). High Integrity Die Casting Processes, New York, NY: John Wiley & Sons, Inc.

Mariana Belgiu, Lucian Dragut. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing,* 114, 24-31.

N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* 16, 321-357.

X. P. Niu, B. H. Hu, I. Pinwill and H. Li. (2000). Vacuum Assisted High Pressure Die Casting of Aluminium Alloys. *Journal of Materials Processing Technology,* 105(1-2), 119-127.

ABSTRACT

Study on Defect Prediction Performance in Die Casting Process with Imbalanced Datasets: Utilization of SMOTE and Random Forest

Lee, Sang-Min

Major in Smart Convergence Technical Consulting

Dept. of Smart Convergence Consulting

Graduate School of Knowledge Service Consulting

Hansung University

This study aims to provide a guide for sampling techniques to increase performance metrics using random forest and SMOTE(Synthetic Minority Over-sampling Technique) to predict defects in high pressure die casting process. The die casting process requires real-time data analysis and prediction, as even small changes in process variables can cause defects. In this study, 105,748 data points collected over a period of 7 months were analyzed for various process variables. In the data preprocessing process, missing values and outliers were removed, and SMOTE and undersampling techniques were applied to solve the problem

of data imbalance. As a result of training and evaluation using random forest models, the model with SMOTE technique showed high prediction performance, especially the model with SMOTE 10% to 30% sampling ratio maintained high F1-score. This study presented a data-driven approach to improve the accuracy of defect prediction in the die casting process and contribute to quality control in the manufacturing process. In future research, it is necessary to add more variables and sufficient data to further improve the model performance and increase the applicability of the model in real industrial sites.

【Key words】 Machine learning, Random forest, Die-casting, SMOTE, Defect prediction,