

기계학습과 SHAP 기반 콘크리트 블록 강도와 흡수율 예측 : 생산공정 PLC 데이터 활용 사례

손현승*, 조광문
국립목포대학교 컴퓨터학부 교수

Machine Learning and SHAP-Based Prediction of Concrete Block Strength and Water Absorption: A Case Study using Production Process PLC Data

Hyun Seung Son*, Kwangmoon Cho
Professor, School of Computer Science and Engineering, Mokpo National University

요약 콘크리트 블록 제조 공정에서 강도 및 흡수율 시험은 양생 기간과 파괴 시험이 수반되어 시간과 비용 부담이 크고, 생산 전수에 대한 실시간 품질 피드백을 제공하기 어렵다. 이러한 한계를 보완하기 위해 본 논문은 배합기, 생산 관련 설비, 틸팅기, 양생실에서 발생하는 공정 로그를 시간 단위로 통합하고, 이를 실측 시험 결과와 연계한 통합 데이터셋을 구축한다. 이후 다중 출력 회귀 프레임워크를 적용하여 GBDT, RF, ET, XGB 등 후보 모델들을 비교한 결과, 설비 A에서는 ET가 평균 R^2 0.8846(평균 MAE 0.1661), 설비 B에서는 GBDT가 평균 R^2 0.9558(평균 MAE 0.0729)로 높은 예측 성능을 보였다. 또한 SHAP 분석을 통해 설비 A에서는 진동시간, 몰탈 공급시간, 온도 변동과 같은 운전 조건 변수가, 설비 B에서는 생산 수량, 호퍼 계량값, 믹싱 횟수와 같은 투입량 및 생산 상태 변수가 상대적으로 중요한 영향 요인임을 확인하였다. 본 논문은 PLC 공정 데이터와 품질 시험 데이터를 통합하고 예측 모델과 설명 가능한 분석을 결합함으로써, 품질을 사전에 예측하고 공정 관리 우선순위를 설정하는 데 활용 가능한 근거를 제시했다는 점에서 의의가 있다.

주제어 : 기계학습, SHAP, 콘크리트 블록, 압축강도 예측, PLC 데이터

Abstract In concrete block manufacturing, strength and water absorption tests require curing periods and destructive testing, which incur substantial time and cost and make it difficult to provide real-time quality feedback for all produced units. To address this limitation, this paper integrates process logs collected from the mixer, production-related equipment, the tilting machine, and the curing chamber at an hourly resolution, and constructs an integrated dataset by linking them with measured test results. Using a multi-output regression framework, we compared candidate models including GBDT, RF, ET, and XGB. The results show that, for Facility A, ET achieved strong performance with an average R^2 of 0.8846 (average MAE of 0.1661), while for Facility B, GBDT achieved high performance with an average R^2 of 0.9558 (average MAE of 0.0729). Moreover, SHAP analysis confirmed that operating-condition variables such as vibration time, mortar feeding time, and temperature fluctuations were relatively more influential for Facility A, whereas input-quantity and production-state variables such as production count, hopper weighing values, and mixing count were relatively more influential for Facility B. By integrating PLC process data with quality test data and combining predictive modeling with explainable analysis, this paper provides practical evidence to support proactive quality prediction and the prioritization of process-control factors.

Key Words : Machine Learning, SHAP, Concrete Block, Compressive Strength Prediction, PLC data

본 과제(결과물)는 2025년도 교육부 및 전라남도의 재원으로 전라남도RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-14-001)

*교신저자 : 손현승(hson@mnu.ac.kr),

접수일 2025년 12월 19일 수정일 2026년 02월 06일 심사완료일 2026년 02월 06일

1. 서론

콘크리트 블록은 도로 경계석 보도 포장 용벽 등 다양한 인프라에 광범위하게 사용되며, 제품의 압축강도와 흡수율은 구조적 성능과 내구성을 좌우하는 핵심 품질 지표이다. 현재 현장에서 이루어지는 콘크리트 블록 강도와 흡수율 평가는 주로 규격에 따른 공시체 제작과 양생 이후의 파괴 시험 및 흡수율 시험에 의존한다. 그러나 이러한 실험은 7일 또는 28일 이상의 양생 기간과 반복적인 파괴 시험을 필요로 하며, 시험체 제작 인력과 장비 운용 비용까지 고려하면 시간과 비용 부담이 매우 크다[1].

이러한 시험 절차는 일정 기간의 양생과 시험 장비 운영 인력 투입 시험 편차 관리 등이 복합적으로 요구되므로 시간과 비용이 많이 들 뿐 아니라, 생산된 전수에 대해 실시간으로 품질을 피드백하기 어렵다는 한계를 갖는다. 실제 현장에서는 생산 로트마다 충분한 시험을 수행하기 어려워, 일부 샘플 시험 결과에 의존하거나 경험적 배합 조정에 의존하는 경우가 많아 품질 편차가 발생할 위험이 존재한다[1].

최근에는 이러한 한계를 보완하기 위해 재료 배합 데이터와 양생 기간을 활용하여 콘크리트 압축강도를 예측하는 기계학습 또는 딥러닝 기반 연구가 국내외에서 활발하게 수행되고 있다. 국내에서는 시멘트, 혼화재, 골재, 물, 감수제, 재령 등 배합 설계 변수를 입력으로 회귀 분석 모델이나 양상을 모델을 적용하여 압축강도를 예측하고, 통계적 회귀모형과 기존 인공신경망 기법을 비교하는 연구가 있다[2-3]. 또한 딥러닝을 적용하여 배합 설계와 강도 간의 비선형 관계를 학습하고, 목표 강도에 대응하는 배합을 역으로 산정하는 시도도 이루어지고 있다[4].

국제적으로도 공개 콘크리트 데이터셋이나 현장 실험 데이터를 기반으로 다양한 기계학습 알고리즘을 비교 및 평가하는 연구가 늘어나고 있다. 예를 들어 XGBoost, Random Forest, 서포트 벡터 회귀, 신경망 등을 이용해 일반 콘크리트 및 고성능 콘크리트의 압축강도를 예측하고, 평가 지표(R^2 , RMSE, MAE 등)를 통해 최적 모델을 선정하는 연구[5-8], 설명 가능한 부스팅 모델을 도입해 혼합 비율 인자를 정량적으로 분석하는 연구[6] 등이 있다. 이러한 연구들은 실험적 강도 시험의 부담을 줄이고, 설계 단계에서 강도 수준을 예측함으로써 설계 효율성과 품질 관리의 신뢰도를 높이는 데 기여하고 있다.

한편 최근에는 단순 예측 성능을 넘어, AutoML과 설명 가능한 인공지능(explainable Artificial Intelligence, XAI)을 결합하여 콘크리트 강도 예측 모델의 해석 가능

성을 높이려는 시도도 증가하고 있다. Optuna[10]와 같은 하이퍼파라미터 최적화 도구를 이용하여 다양한 앙상블 알고리즘을 자동 탐색하고, SHAP(Shapley Additive Explanations)[11]를 활용해 시멘트, 물-결합재비, 재령 등 입력 변수의 기여도와 상호작용을 정량적으로 분석하는 연구가 대표적이다. 또한 재료나 구조 성능 예측 모델에 PDP(Partial Dependence Plot), SHAP 등 XAI 기법을 적용해 보강 변수나 설계 변수의 영향도를 도출하는 연구도 진행되고 있어, 토목 및 건축 재료 분야에서 설명 가능한 기계학습의 활용 가능성을 보여준다[6, 8-9].

그러나 기존 연구의 대부분은 배합 설계 변수와 양생 기간과 같은 실험실 수준의 재료 정보에 집중하고 있으며, 실제 블록 생산 공정에서 수집되는 공정 제어 데이터까지 통합적으로 고려한 사례는 찾아보기 어렵다. 선행 연구들이 사용하는 데이터셋은 대개 배합표와 실험실 압축강도 결과로 구성되어 있으며, 실제 블록 제조 공정에서 PLC(Programmable Logic Controller)가 계측하는 1차 및 2차 진동시간, 몰탈 공급시간, 양생실 온도 및 시간과 같은 공정 변수는 입력 요인으로 포함되지 않는다. 즉, 배합 설계가 동일하더라도 진동 조건, 공급 패턴, 양생 이력의 차이에 따라 강도와 흡수율이 달라질 수 있음에도, 생산설비 및 공정 데이터를 결합한 기계학습과 SHAP 기반 분석은 국내외 문헌에서 거의 보고되지 않는다.

따라서 본 논문에서는 기존 배합 중심 강도 예측 연구의 한계를 보완하고, 실제 콘크리트 블록 공장의 PLC 공정 데이터를 포함한 다차원 생산 데이터 기반으로 강도와 흡수율을 동시에 예측할 수 있는 기계학습 및 SHAP 분석을 통해 재료 배합과 공정 제어 변수 각각이 품질 특성에 미치는 영향을 정량적으로 규명하는 것을 목표로 한다. 이러한 접근은 실험 비용 절감과 함께, 공정 조건을 포함한 통합 품질 지표 기반의 블록 생산 관리, 그리고 공정 최적화를 통한 목표 강도 및 흡수율 달성 전략 수립에 기여할 수 있을 것으로 기대한다.

2. 관련연구

2.1 콘크리트 강도 예측을 위한 기계학습 연구

권민호와 강민곤은 공개 데이터와 국내 현장 레미콘 데이터를 통합하여 다양한 회귀 기반 기계학습 모델로 콘크리트 압축강도를 예측하는 연구를 수행하였다. CatBoost, LightGBM, Random Forest, XGBoost 등 최신 회귀 알고리즘을 교차검증과 그리드 서치를 통해

최적화하고, 기존 인공신경망 모델보다 낮은 오차와 높은 정확도를 확보했으며, 특히 XGBoost 회귀 모델이 가장 우수한 성능을 보인다고 보고하였다[2]. 이 연구는 혼화재, 골재, 재령 등의 배합 정보를 기반으로 현장 적용이 가능한 예측 모델을 구축했다는 점에서 의미가 있다.

이상민 등은 공개 콘크리트 배합 데이터와 실험으로 측정된 공시체 데이터를 통합해 여러 기계학습 모형의 압축강도 예측 성능을 체계적으로 비교하였다. 입력 변수로 시멘트, 고로슬래그 미분말, 플라이 애시, 물, 고성능 감수제, 굵은 골재, 잔골재, 재령을 사용하고, 선형회귀, 결정트리, 서포트 벡터 회귀, 앙상블, 신경망, 가우시안 프로세스 회귀 모델을 동일 데이터셋에서 비교한 결과, 신경망과 가우시안 프로세스 회귀 계열이 전반적으로 가장 우수한 예측 성능을 보이는 것으로 나타났다[12]. 이 논문은 동일 데이터에 대해 여러 기계학습 모형을 비교함으로써 모델 선택의 기준을 제시했다는 점에서 의미가 있다.

최근에는 데이터 구성과 전처리 전략 자체에 초점을 둔 데이터 중심 머신러닝 연구도 등장하고 있다. 김경철 등은 건설 폐기물 순환자원을 활용한 콘크리트의 압축강도 예측에서 데이터 품질과 특성 공학의 중요성을 강조하면서, 다양한 머신러닝 알고리즘을 적용해 재료 종류와 배합비 변화가 강도에 미치는 영향을 정량화하였다[13]. 이와 같은 연구들은 기존의 단순 회귀식 대신 데이터 기반 회귀모형을 통해 압축강도 예측 정확도를 크게 향상시키고 있다.

국외에서도 콘크리트 압축강도 예측에 기계학습을 적용한 연구가 증가하고 있다. Fu 등은 다양한 트리 기반 앙상블 알고리즘과 신경망을 활용하여 콘크리트 압축강도를 예측하고, 모델 구조와 특성 중요도 분석을 결합한 설명 가능 기계학습 프레임워크를 제시하였다[7]. Zhang 등은 고성능 콘크리트를 대상으로 랜덤 포레스트와 그라디언트 부스팅 계열 모델을 활용하여 압축강도 예측 성능을 향상하고, 모델 내부 동작을 해석 가능한 형태로 제시하였다[14].

요약하면, 국내 선행 연구들은 레미콘 콘크리트나 일반 구조용 콘크리트를 대상으로 배합 정보를 중심으로 한 압축강도 예측에 집중해 왔으며, 최신 기계학습 알고리즘을 활용해 예측 성능을 향상하는 데 주력하고 있다. 그러나 품질 규격상 중요한 또 다른 물성인 흡수율을 동시에 예측 대상으로 설정한 연구는 상대적으로 부족하다.

이들 연구는 공통으로 물시멘트비, 시멘트와 혼화재

비율, 골재 비율, 재령 등 전통적인 배합 설계 변수를 핵심 입력으로 사용하면서도, 모델 선택과 하이퍼파라미터 최적화를 통해 고정밀 예측을 구현한 점에서 국내 연구와 유사하다. 다만 대부분의 연구가 압축강도 단일 지표에만 초점을 두고 있으며, 제품 유형별 특성 예를 들어 블록류 제품의 흡수율이나 기건비중과 같이 품질 규격에서 요구하는 다차원 물성을 동시에 예측하는 시도는 상대적으로 드물다.

2.2 SHAP를 이용한 콘크리트 물성 예측 연구

최근 재료공학 분야에서는 단순한 예측 정확도를 넘어서, 각 입력 변수가 결과에 미치는 영향을 해석하기 위한 설명 가능 인공지능 기법이 활발히 도입되고 있다. 특히 게임이론 기반 기법인 SHAP는 개별 특성이 예측값에 기여하는 정도를 정량화할 수 있어 콘크리트 물성 연구에서도 널리 활용되고 있다.

Wu 등은 지속 가능 콘크리트의 압축강도를 예측하기 위해 하이브리드 머신러닝 모델을 구축한 뒤, SHAP를 이용해 시멘트, 혼화재, 물 시멘트비, 치환재 비율 등이 강도에 미치는 기여도를 정량적으로 분석하였다[15]. 이를 통해 특정 혼화재의 치환률이 모델 출력에 미치는 영향이 재령과 상호작용을 한다는 점 등 전통적 회귀분석으로 파악하기 어려운 비선형 관계를 시각화하였다.

Fu 등은 설명 가능 기계학습 모델을 이용해 콘크리트 압축강도를 예측하고, SHAP를 활용해 다양한 트리 기반 모델에서 공통으로 중요한 변수와 모델별로 차별적인 변수 영향 패턴을 비교하였다[7]. 이 연구는 동일 데이터에 대해 여러 알고리즘을 적용한 후, SHAP 분석을 통해 각 알고리즘이 입력 변수를 해석하는 방식의 차이를 정량적으로 보여준다는 점에서 의미가 있다.

또 다른 연구에서는 고성능 콘크리트나 탄소 포집 양생 콘크리트를 대상으로 XGBoost, 랜덤 포레스트, LightGBM 등과 SHAP를 결합해 고성능 콘크리트 강도에 대한 물 시멘트비, 실리카 폼, 고로슬래그 등의 영향도를 시각화하고, 최적 배합 설계 방향을 제시하고 있다[15].

이러한 선행 연구들은 SHAP를 통해 각 재료 변수의 개별적 영향뿐 아니라 변수 간 상호작용 효과를 분석함으로써, 단순한 블랙박스 예측 모델을 넘어 설계 의사결정을 지원하는 해석 도구로서의 가능성을 보여준다. 그러나 대부분의 연구가 압축강도 단일 지표에만 주목하고 있으며, 흡수율과 같은 내구성 관련 지표에 대해 SHAP 분석을 수행한 사례는 제한적이다.

3. 데이터 수집 및 통합 방법

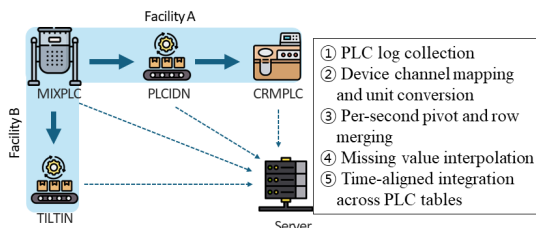
3.1 PLC 기반 데이터 수집 환경

본 논문에서 사용한 생산공정 데이터는 콘크리트 블록 공장의 각 설비에 연결된 PLC에서 수집한 로그를 기반으로 한다. 공장은 배합실(MIXPLC), 양생실(CRMPLC), 생산관련(PLCIDN), 틸팅기(TILTIN) 등 네 가지 주요 설비별로 별도의 PLC 로그 테이블을 운영하고 있으며, 각 테이블은 <Table 1>과 같은 공통 구조를 가진다.

<Table 1> Column definitions for PLC log data

Column	Description	Data type / format	Example
idtime	Timestamp of the PLC record with millisecond	YYYY-MM-DD HH:MM:SS.sss	2025-12-11 14:23:05.123
deviceid	Integer identifier of the PLC input/output channel	unsigned int	1
rcvdata	Raw value received from the given deviceid at idtime	unsigned int	1070

로그 파일은 일 단위로 CSV 형식으로 축적되며, 본 논문에서는 2025년 생산 기간 중 강도와 흡수를 시험체가 확보된 날짜의 로그를 중심으로 데이터를 추출한다. 데이터처리 단계에서는 이들 CSV 파일을 입력으로 받아, 장치 채널 매핑과 전처리 및 피벗 과정을 수행한다. [Fig. 1]은 PLC 로그로부터 분석용 시계열 데이터를 생성하는 전체 흐름을 개념적으로 나타낸 것이다.



[Fig. 1] Preprocessing flow of concrete block process data based on PLC logs

3.2 장치 채널 매핑 및 단위 정규화

원시 로그의 `deviceid`는 숫자 코드로만 제공되므로, 해석 가능한 변수 이름으로의 매핑이 필요하다. 이 단계에서는 PLCIDN.csv, MIXPLC.csv, CRMPLC.csv, TILTIN.csv에 따라 서로 다른 매핑 디셔너리를 선택하여 `deviceid`를 의미 있는 한글 변수명으로 변환한다.

예를 들어 성형기 본체 로그(PLCIDN)의 경우 일부 채널은 <Table 2>와 같이 매핑한다.

<Table 2> Example of PLC Channel-to-Variable Mapping

Deviceid	Description	Units
1	Product name	
2	Operating time	1 second
3	Production quantity	1 piece
4	Production cycle time	0.1 second
5	Stacking count	1 piece
6	Feeding time	0.1 second
7	1st vibration time	0.1 second
8	2nd vibration time	0.1 second

배합실(MIXPLC)는 “호퍼1 계량 설정값”, “설비A 시멘트계량값”, “설비B 물계량값”과 같은 계량 관련 채널이, 틸팅기(TILTIN)는 “생산량”, “틸팅기가동시간”, “적재수량” 등 생산량 및 틸팅 동작 관련 채널이 다수 포함된다. 이러한 매핑을 통해 최종 피벗 결과에서는 각 열 이름이 실제 공정 변수 의미를 직관적으로 반영하도록 구성한다.

일부 로그에서 온도 센서는 16비트 부호 없는 정수(UINT) 형식으로 저장되며, 실제 온도는 2의 보수 표현을 사용한다. 이를 복원하기 위해 (Equation 1)과 같은 변환 수식을 사용한다. 2^{bits} 는 값의 최댓값이고 2^{bits-1} 은 양수/음수를 나누는 기준값이다. 예를 들어 16비트 값일 때 최댓값은 65536이고 양수/음수를 나누는 기준값은 32768이다.

$$x_{signed} = \begin{cases} x & (x < 2^{bits-1}) \\ x - 2^{bits} & (x \geq 2^{bits-1}) \end{cases} \text{ (Equation 1)}$$

PLC 로그에서 여러 시간 관련 채널은 0.1초 단위로 저장된다. 예를 들어 “생산사이클시간”, “공급시간”, “1차진동시간”, “2차진동시간”, “몰탈공급시간”, “임펠라시간”, “탈형1차상승시간”, “진동기 운전시간”, “압축상승시간”, “탈형하강시간” 등이 이에 해당한다. 이들 열은 수치형으로 변환한 뒤 10으로 나누어 초 단위로 통일하였다. 마찬가지로 시멘트와 물 계량값 등 일부 계량 채널은 0.1 kg 단위로 기록된다. 이들 데이터들은 값을 10으로 나누어 kg 단위로 정규화하였다.

3.3 초 단위 피벗과 사이클 기준 레코드 생성

원시 PLC 로그는 `(idtime, deviceid, rcvdata)` 형

태의 긴 형식으로 되어 있어, 하나의 시점에 여러 채널 값이 여러 행에 나뉘어 존재한다. 기계학습에 활용하기 위해서는 시간축을 기준으로 각 시점별 모든 채널 변수가 한 행에 모인 넓은 형식으로 변환할 필요가 있다. 이를 위해 처리 함수는 <Table 3>과 같은 절차로 동작한다.

이 알고리즘은 먼저 idtime 값을 날짜 및 시간 형식으로 파싱한 뒤, 소수점 이하를 제거하여 초 단위 시각으로 정규화한다. 이어서 사용자가 지정한 연도(예: 2025년)에 해당하는 레코드만을 남겨 분석 대상 기간을 한정한다. 그런 다음 동일한 초 단위 시각과 deviceid 조합에 대해 가장 마지막에 기록된 rcvdata를 대표값으로 선택하고, 이를 이용해 deviceid를 열 이름으로 하는 피벗 연산을 수행함으로써 시각별, 채널별 값이 한 행에 정리된 wide_named 데이터를 생성한다. 이때 각 deviceid에는 사전에 정의된 한글 채널명을 매핑하여 열 이름이 실제 공정 변수를 의미하도록 하고, 동시에 채널명을 적용하지 않은 원본 형태의 wide_id 데이터도 함께 저장하여, 필요시 원시 deviceid 기준 분석이나 매핑 검증에 활용할 수 있도록 한다.

<Table 3> Algorithm for per-second pivoting and cycle-based record generation

```
PivotPLCLogsBySecond(L, Y, NameMap):
1: // Timestamp parsing and normalization
2: for each row r in L do
3:   t_raw ← ParseDateTime(r.idtime)
4:   // drop fractional seconds
5:   t_sec ← FloorToSecond(t_raw)
6:   r.time_sec ← t_sec
7: end for
8: // Filter by target year
9: F ← { r ∈ L | Year(r.time_sec) = Y }
10: // Select last rcvdata per (time_sec, deviceid)
11: G ← GroupBy(F, keys = (time_sec, deviceid))
12: R ← empty table
13: for each group g in G do
14:   // latest original timestamp
15:   r_last ← argmax_{r ∈ g} r.idtime
16:   AppendRow(R, r_last)
17: end for
18: // Pivot with raw deviceid as columns
19: WideId ← Pivot( table = R,
20:   index = time_sec, columns = deviceid,
21:   values = rcvdata)
22: // Create copy and apply channel names
23: WideNamed ← CopyTable(WideId)
24: for each column c in WideNamed.columns do
25:   if c ∈ NameMap then
26:     RenameColumn(WideNamed, c, NameMap[c])
27:   end if
28: end for
29: return WideNamed, WideId
```

같은 초 단위로 기록된 데이터라도 실제 공정에서는 한 사이클 동안 여러 건의 로그가 밀집하여 발생하는 경

우가 있다. 이를 하나의 레코드로 통합하기 위해서는 시간 열(time_col) 기준으로 정렬된 데이터에서 인접 행의 시간 간격이 `window_seconds`를 초과할 때마다 새로운 클러스터를 시작하고, 클러스터 내부의 여러 행을 한 행으로 합치는 방식을 적용한다.

3.4 순차 행 병합과 결측치 보간

초 단위 피벗 후에도 일부 채널 값은 특정 시점에만 기록되거나, 센서 이상으로 인해 결측치가 발생할 수 있다. 이를 완화하기 위해 순차 행 병합 알고리즘과 결측치 채우기 전략 두 단계의 통합 절차를 추가로 적용한다.

3.4.1 순차 행 병합 알고리즘

순차 행 병합 알고리즘은 동일 테이블 내에서 인접한 행들을 순차적으로 비교하여, 첫 행의 빈 칸을 둘째 행의 값으로 채우는 병합을 수행한다. 알고리즘의 핵심 아이디어는 다음과 같다. 먼저 idtime을 기준으로 정렬된 데이터에서 현재 행과 바로 다음 행을 순차적으로 비교하는 방식으로 동작한다. 이때 idtime을 제외한 어떤 열에서든 두 행 모두에 값이 존재하는 경우, 해당 열은 이미 정보가 채워져 있다고 판단하여 두 행을 더 이상 병합하지 않고 현재 행을 그대로 유지한다.

반대로 현재 행의 특정 열이 결측적이고 다음 행의 동일 열에만 값이 존재할 때는 그 값을 현재 행으로 복사하여 정보를 보완한다. 이러한 절차를 전체 행에 대해 반복 적용함으로써, 가능한 한 적은 수의 행으로 정보를 압축하면서도 서로 다른 값이 충돌하여 잘못된 정보가 섞이는 상황은 피하도록 한다. 이를 의사코드로 정리하면 <Table 4>와 같다.

<Table 4> Sequential row merging algorithm

```
Algorithm MergeSequentialRows(df):
1: df ← df sorted by idtime
2: current ← first row of df
3: merged_rows ← empty list
4: for each next row in df (from 2nd to last) do
5:   if ∃ column col ≠ "idtime" such that
6:     current[col] ≠ NULL and row[col] ≠ NULL then
7:     append current to merged_rows
8:     current ← row
9:   else
10:    for each column col ≠ "idtime" do
11:      if current[col] = NULL and row[col] ≠ NULL then
12:        current[col] ← row[col]
13:      end if
14:    end for
15:  end if
16: end for
17: append current to merged_rows
18: return DataFrame(merged_rows)
```

이 알고리즘은 시간축을 따라 이동하면서 “전후 행 중 하나에만 값이 있는 컬럼”을 자연스럽게 통합하여, 센서 기록이 시간적으로 분산된 경우에도 한 행에 정보가 모이도록 만들어 준다.

3.4.2 결측치 채우기 전략

병합된 데이터 내에 남아 있는 결측치는 변수 유형과 테이블별 특성을 고려하여 보간하도록 한다. 먼저 수치형 열의 경우 선형 보간을 사용하여 이전 시점과 이후 시점의 값을 이용해 결측 구간을 자연스럽게 채운다. 다만 특정 테이블(예: MIXPLC)에 포함된 설정값, 계량 설정값과 같이 실제 공정에서 불연속적이거나 계단형으로 변하는 변수에 대해서는 이러한 연속 보간이 부적절할 수 있으므로, 직전 값을 그대로 이어받는 항상 앞 방향 채우기 전략을 별도로 적용한다.

문자열 열은 라벨, 품명과 같은 범주형 정보가 시간에 따라 갑자기 변하지 않는다는 가정을 반영하여 항상 앞 방향 채우기를 적용한다. 또한 보간이 완료된 후에는 수치 결과를 사람이 읽기 쉬운 형태로 저장하기 위해, 모든 수치값을 소수 둘째 자리까지 반올림하되, 정수인 경우는 불필요한 소수점 표기를 제거한다.

3.5 PLC 간 시간 기준 통합

앞 절까지의 전처리와 병합 및 보간은 각 PLC 테이블 별로 독립적으로 수행된다. 최종적으로는 네개의 설비 로그(PLCIDN, MIXPLC, CRMP, TILTIN)에 대해 생성된 파일을 `idtime` 기준으로 외부 조인하여, 공정 전 구간의 상태를 통합적으로 나타내는 하나의 학습용 테이블을 구성하였다.

각 설비가 서로 다른 시간 간격으로 데이터를 기록하더라도, 초 단위로 통일된 `idtime` 키와 순차 병합 및 보간 과정을 통해 동일한 행에서 서로 다른 설비의 상태가 결합된다. 이와 같이 구성된 최종 데이터셋은 콘크리트 블록 강도와 흡수율을 동시에 예측하는 기계학습 모델의 입력으로 사용되며, 이후 4장에서 모델 학습과 성능 평가, SHAP 기반 변수 영향 분석에 활용된다.

4. 다중 출력 회귀 모델 설계

이 장에서는 3장에서 구축한 통합 데이터셋을 이용해 콘크리트 블록의 강도와 흡수율을 동시에 예측하는 다중 출력 회귀 모델의 구조와 학습 절차를 설명한다. 제안 모

델은 배합 정보, 그리고 PLC에서 수집된 공정 집계 변수들을 입력으로 받아, 한 생산 배치에 대응하는 강도 및 흡수율 시험값 여섯 개를 벡터 형태로 예측하도록 구성하였다.

4.1 입력 변수와 목표 변수

본 논문에서 사용한 수치형 입력 변수는 생산 지표, 혼합 설비 계량 설정값 통계, 설비A와 설비B 계량 및 믹싱 집계, 생산량 집계 등 네 가지 그룹으로 구성한다. 먼저 배합 및 생산 관련 기본 변수로는 배합 비율을 나타내는 연속형 지표 혼합비율(mixrat)과, 한 배치에서 실제로 생산되거나 적재된 블록 개수(stccnt)를 포함한다. mixrat은 시멘트, 골재, 물 등의 상대적 비율을 압축한 지표로서 배합 특성을 요약하며, stccnt는 동일 조건에서 생산된 양을 반영하여 생산 규모와 품질 간 관계를 파악하는 데 활용된다.

혼합 설비(MIXPLC) 측면에서는 호퍼1~3 계량 설정값 평균(mean)과 표준편차(std) 값과 시멘트 계량 설정값 평균과 표준편차, 물 계량 설정값 평균과 표준편차를 정의하여, 일정 시간 구간 동안의 계량 설정값에 대한 평균과 표준편차를 계산한다. 이들 통계량은 목표 계량 값이 얼마나 안정적으로 유지되었는지, 계량 설정값에 어느 정도 변동성이 존재하는지를 정량적으로 보여 주며, 배합 과정의 일관성을 평가하는 데 사용된다.

설비 A와 설비 B에 대해서는 실제 계량 및 믹싱 관련 집계 변수들을 추가한다. 구체적으로 믹싱횟수 최대값(max)을 통해 각 설비에서 수행된 최대 믹싱 횟수를 산출하고, 시멘트계량값, 물계량값, 호퍼1~3 계량값을 통해 각 설비에서 관측된 최대 계량값을 추출한다. 이러한 최대값 지표는 실제 투입된 재료량과 공정 강도를 반영함으로써, 재료 투입 조건과 최종 물성 간의 관계를 분석하는 데 기여한다.

마지막으로 생산량 관련 집계 변수로 설비A와 B의 생산량을 정의하여, 주어진 시간 구간 내에서 설비A와 설비B가 기록한 최대 생산량을 계산한다. 이는 혼합 및 성형 공정의 부하 수준과 생산 속도를 간접적으로 나타내는 지표로, 생산량이 증가하는 상황에서 강도 및 흡수율 예측 성능이 어떻게 변화하는지, 그리고 과부하 상태가 품질 저하와 관련되는지를 분석하는 데 활용할 수 있다.

이와 같이 제품 특성, 배합 정보, 공정 계량값 집계를 한 번에 입력으로 사용함으로써, 단순 배합표뿐 아니라 실제 공장 공정의 변동성이 예측 모델에 반영되

도록 한다.

목표 변수는 한 생산 배치에 대해 수행된 강도 및 흡수율 시험값 여섯 개로 구성하였다. 먼저 강도(strength) 1, 강도 2, 강도 3는 동일 배치에서 채취한 시편에 대해 측정된 압축강도 시험값으로, 시편 간 반복 측정값 또는 재령 조건별 강도 결과를 나타낸다. 또한 흡수(water absorption) 1, 흡수 2, 흡수 3은 동일 배치에 대해 수행된 흡수율 시험값으로, 시편별 또는 시험 조건별 흡수율 측정 결과를 의미한다. 본 연구에서는 이 여섯 개의 실수값을 하나의 벡터로 묶어 (Equation 2)와 같이 다중 출력 회귀 문제로 정의하였다.

$$y = (s_1, s_2, s_3, a_1, a_2, a_3) \quad (\text{Equation 2})$$

4.2 다중 출력 회귀 프레임워크

강도와 흡수율은 동일한 배합과 공정 조건의 영향을 동시에 받는 물성이므로 서로 높은 상관성을 가지는 경우가 많다. 따라서 두 특성을 각각 독립적인 단일 출력 모델로 예측하기보다는, 하나의 모델이 여섯 개의 출력(강도 3개, 흡수율 3개)을 동시에 예측하도록 다중 출력 구조로 설계하는 것이 더 합리적이다. 이러한 접근은 동일한 입력 특성으로부터 강도와 흡수율 사이의 상관 구조를 부분적으로 공유할 수 있게 하며, 결과적으로 두 물성의 예측값이 서로 모순되지 않도록 일관성을 유지하는데에도 유리하다. 또한 이후 SHAP 기반 해석 단계에서 동일한 입력 변수가 강도와 흡수율에 각각 어떤 방향과 크기로 기여하는지를 동일한 틀에서 비교할 수 있어, 공정 변수의 영향 패턴을 보다 체계적으로 분석할 수 있다는 장점이 있다.

이를 위해 본 논문에서는 사이킷런의 Multi Output Regressor를 기반으로 한 다중 출력 회귀 프레임워크를 채택한다. 이 방법은 하나의 입력 벡터 x 에 대해 출력 차원 수만큼 동일한 구조의 회귀 모델을 독립적으로 학습시키는 방식으로 동작한다. 즉, 여섯 개의 목표 변수를 예측하기 위해 $i=1, \dots, 6$ 에 해당하는 개별 회귀기 $f_i(x)$ 를 각각 학습하고, 최종적으로는 (Equation 3) 형태의 예측 벡터를 생성한다.

$$\hat{y} = (f_1(x), f_2(x), \dots, f_6(x)) \quad (\text{Equation 3})$$

출력별로는 서로 다른 회귀기가 학습되지만, 입력 데이터에 대한 전처리 과정과 특징 공간, 그리고 학습 평가를 위한 데이터 분할은 모든 출력에서 동일하게 공유하도록 구성한다.

4.3 전처리 및 학습 파이프라인

예측 모델은 전처리 모듈과 회귀 모델을 하나의 파이프라인으로 결합하여 학습된다. 전처리 단계는 수치형 변수를 처리하는 구조를 갖는다.

수치형 입력 변수는 모든 연속형 특성을 대상으로 하였으며, 결측값은 중앙값 기반 대체(median)를 적용하여 처리한다. 이는 센서 오류나 데이터 누락으로 인해 일부 값이 결측으로 발생할 수 있다는 점을 고려하여, 평균보다 이상치에 덜 민감한 중앙값을 사용함으로써 보다 견고하게 결측치를 보정하기 위한 것이다. 또한 본 연구에서 사용한 기본 회귀 모델이 트리 기반 알고리즘으로 구성되어 있어 입력 변수의 스케일 변화에 상대적으로 불변이라는 특성을 반영하여, 별도의 정규화나 표준화와 같은 추가 스케일링은 수행하지 않았다.

이러한 전처리 과정을 구현하기 위해 Column Transformer를 사용하였으며, 이를 통해 범주형 변수와 수치형 변수를 각각 다른 전처리 파이프라인으로 전달한 뒤 변환 결과를 하나의 특성 행렬로 결합한다. 전처리된 특성과 회귀 모델은 사이킷런의 Pipeline으로 통합하여 관리하였고, 파이프라인 전체를 학습과 예측의 기본 단위로 사용함으로써 전처리와 모델 학습이 일관된 절차로 수행되도록 한다. 또한 하이퍼파라미터 탐색 역시 개별 모델이 아니라 파이프라인 단위로 수행하여, 전처리 과정까지 포함한 최적 설정을 탐색할 수 있도록 구성하였다. 마지막으로 데이터는 모델 학습용과 성능 평가용으로 8:2 비율로 분할하여 실험을 진행한다.

4.4 후보 회귀 모델과 경량 하이퍼파라미터 탐색

본 논문에서는 콘크리트 블록의 강도와 흡수율을 예측하기 위해 트리 기반 앙상블 회귀 모델을 후보군으로 설정하고 성능을 비교한다. 후보 모델은 그래디언트 부스팅(Gradient Boosting Decision Tree, GBDT), 랜덤 포레스트(Random Forest, RF), 엑스트라 트리(Extra Trees, ET), 그리고 엑시지부스트(XGBoost, XGB)로 구성한다.

GBDT는 잔차를 순차적으로 보정하는 부스팅 방식의 모델로, 비교적 적은 수의 강한 트리들을 결합하여 비선형 관계를 효과적으로 근사한다. 본 논문에서는 초기 설정으로 약 300회의 트리(추정기)와 최대 깊이 5 내외의 트리 구조를 사용하고, 학습률은 0.05 수준으로 설정하였다. RF는 데이터와 특성을 무작위로 부분 샘플링하여 다수의 결정 트리를 학습한 뒤 그 예측을 평균하는 배깅

기반 모델로, 초기 설정으로 400개의 트리를 사용하고 최대 깊이는 제한하지 않았으며 병렬 연산을 활성화하였다. ET는 RF와 유사하지만, 분할 후보를 보다 무작위로 생성하여 분산을 줄이는 특성을 가지며, 초기 설정으로 400개의 트리와 병렬 연산을 사용한다. XGB는 부스팅 기반 모델로 성능이 우수하다는 보고가 많아 후보로 포함하였으며, 초기 설정은 300회의 부스팅 단계, 학습률 0.05, 최대 깊이 5, 그리고 샘플링 비율(데이터 및 특성) 0.9 수준으로 설정한다.

각 모델에 대해서는 <Table 5>와 같이 완전 탐색 대신 경량 하이퍼파라미터 탐색을 수행한다. 탐색은 전처리 단계와 다중 출력 회귀 모델을 포함한 전체 파이프라인에 대해 랜덤 탐색(Randomized SearchCV)을 적용하는 방식으로 구성하며, 반복 횟수는 최대 10회, 교차검증은 3겹(cv=3)으로 설정한다. 또한 계산 효율을 위해 병렬 처리는 가능한 모든 코어를 사용하도록(n_jobs=-1) 설정한다. 탐색을 통해 가장 높은 교차검증 점수를 기록한 파이프라인의 회귀기와 하이퍼파라미터를 선택하고, 이를 전체 학습 세트에 대해 재학습하여 최종 모델로 확정한다. 특정 모델에 대해 탐색 공간이 정의되지 않았거나 라이브러리 제약이 있는 경우에는 기본 하이퍼파라미터를 사용한다.

<Table 5> Hyperparameter search space for randomized search

Model	Search Parameters
GBDT	n_estimators = {200, 300, 500} learning_rate = {0.05, 0.1} max_depth = {3, 4, 5}
RF/ET	n_estimators = {200, 400, 600} max_depth = {None, 10, 20} max_features = {None, "sqrt", "log2", 0.7}
XGB	n_estimators = {200, 400, 600} learning_rate = {0.03, 0.05, 0.1} max_depth = {3, 4, 5} subsample = {0.8, 0.9, 1.0} colsample_bytree = {0.8, 0.9, 1.0}

탐색 결과 가장 높은 교차검증 점수를 기록한 파이프라인의 회귀기와 하이퍼파라미터를 추출하여, 이후 전체 학습 세트에 대해 재학습을 수행한다. 만약 특정 모델에 대해 별도의 탐색 공간이 정의되지 않은 경우에는 기본 하이퍼파라미터를 사용한다.

4.5 성능 평가 지표와 모델 선택 기준

각 후보 모델의 성능 평가는 테스트 세트에서 목표 변수별로 결정계수(R²)와 평균 절대 오차(MAE)를 계산하는

방식으로 수행하였다. 목표 변수 (j)에 대한 R² 값은 (Equation 4)와 같이 정의된다.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (\text{Equation 4})$$

여기서 y_{ij}는 실제 강도 또는 흡수율 값, \hat{y}_{ij} 는 예측 값, \bar{y}_j 는 목표 변수 (j)의 평균값이다. R²가 1에 가까울수록 해당 목표 변수에 대한 예측 성능이 우수함을 의미한다. MAE는 예측 오차의 절대값 평균으로 (Equation 5)와 같다. 단위가 원래의 측정 단위와 동일하다는 점에서 해석이 직관적이다.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (\text{Equation 5})$$

이와 같이 본 연구의 예측 모델은 통합된 제품 배합 공정 데이터를 입력으로, 다중 출력 회귀와 경량 AutoML 방식의 모델 비교 및 하이퍼파라미터 탐색을 통해 강도와 흡수율을 동시에 예측하도록 설계한다. 다음 장에서는 선정된 최종 모델을 대상으로 SHAP 기반 설명 가능성 분석을 수행하여, 재료 및 공정 변수가 강도와 흡수율에 미치는 영향 구조를 정량적으로 규명한다.

5. 실험 결과 및 SHAP 해석

5.1 실험 설정

3장에서 설명한 데이터 수집과 전처리 과정을 거쳐, 배합 정보 및 PLC 공정 집계 변수를 통합한 데이터셋을 구축하였다. 분석 대상 데이터의 기간과 건수는 3장의 표에 제시한 바와 같으며, 각 레코드는 하나의 생산 배치 또는 시험 편차가 존재하지 않도록 통합한 생산 구간에 대응한다. 목표 변수는 압축강도(s1, s2, s3)와 흡수율(a1, a2, a3) 여섯 개로 정의하였다.

예측 모델 학습을 위해 전체 데이터셋을 <Table 6>과 같이 학습 세트와 테스트 세트로 8 대 2 비율로 분할한다. 분할 시에는 무작위 시드를 고정하여 실험을 재현할 수 있게 하였으며, 학습 세트는 모델 학습과 간단한 하이퍼파라미터 탐색에, 테스트 세트는 최종 일반화 성능 평가에만 사용하였다. 여기서 설비 A는 상대적으로 크기가 작은 콘크리트 블록(소형 제품)을 생산하는 설비이며, 설비 B는 크기가 큰 콘크리트 블록(대형 제품)을 생산하는 설비이다.

본 논문에서 PLC 공정 데이터는 2022년부터 2025년 까지 장기간 축적되어 있으나, 압축강도 및 흡수율과 같은 품질 특성은 전 생산 로트에 대해 상시 측정되지 않고 별도의 품질 시험을 통해 획득된다. 특히 본 논문에서 활용 가능한 압축강도 및 흡수율 품질시험 데이터가 약 6 개월치에 한정되어 있어, PLC 로그와 로트 단위로 매칭 가능한 라벨 데이터 수가 제한되었고 그 결과 학습 데이터셋 규모가 상대적으로 작다는 한계가 있다.

데이터가 제한된 조건에서 단일 분할 평가는 성능 변동이 커질 수 있어, 본 논문에서는 KFold 기반 반복 교차검증을 사용한다. 사용한 기본 설정은 k=5분할이며, 시드는 실험 재현을 위해 고정한다.

<Table 6> Train/test split of the dataset

Facility	Train	Test	Total
A	80	20	100
B	98	25	123
Total	178	45	223

5.2 모델별 성능 비교 및 예측 결과

후보 모델별 성능 평가는 테스트 세트에서 목표변수마다 결정계수(R²)와 평균 절대 오차(MAE)를 계산하는 방

식으로, 5-fold 교차검증하여 결과를 얻고, 성능을 평균으로 결과를 출력하였다. <Table 7>과 <Table 8>은 설비 A와 B에 대한 네 가지 후보 모델에 대해 여섯 개 목표 변수별 R²와 MAE, 그리고 여섯 변수에 대한 평균 R²를 정리한 것이다.

설비 A와 설비 B에서 수집된 PLC 기반 공정 데이터를 이용하여, 후보 회귀 모델(GBDT, RF, ET, XGB)의 예측 성능을 비교하였다. 평가지표로는 결정계수(R²)와 평균절대오차(MAE)를 사용하였다. R²는 1에 가까울수록 설명력이 높고, MAE는 0에 가까울수록 실제값과의 오차가 작음을 의미한다. 각 설비에 대해 강도 관련 지표(s1, s2, s3)와 흡수율 관련 지표(a1, a2, a3)별 성능을 산출한 뒤 평균(avg) 값으로 종합 비교하였다.

설비 A의 평균 성능(avg) 기준으로, ET가 가장 높은 R²(0.8846)을 기록하여 전반적인 설명력이 가장 우수한 모델로 확인되었다. 반면 GBDT는 평균 MAE가 가장 낮아(0.1543) 절대 오차 관점에서 가장 안정적인 예측을 보였다. RF는 평균 R²가 0.8497으로 준수했으나 평균 MAE가 0.2849로 상대적으로 커, 오차 최소화 관점에서는 불리하였다. XGB는 평균 R²(0.8197)이 다른 모델 대비 낮고, 평균 MAE(0.1866)도 ET/GBDT 대비 개선 여지가 나타났다.

<Table 7> Comparison of prediction model for concrete block strength and water absorption in Facility A

Model	metric	s1	s2	s3	a1	a2	a3	avg
GBDT	R ²	0.9153	0.873	0.9041	0.9273	0.9884	0.6385	0.8744
	MAE	0.2777	0.3432	0.2795	0.0079	0.0034	0.0139	0.1543
RF	R ²	0.8785	0.8522	0.8632	0.9062	0.9356	0.6622	0.8497
	MAE	0.4974	0.6079	0.5579	0.0162	0.0119	0.0178	0.2849
ET	R ²	0.9144	0.8944	0.8972	0.9697	0.9902	0.6414	0.8846
	MAE	0.3041	0.3431	0.329	0.0051	0.0035	0.0117	0.1661
XGB	R ²	0.8992	0.8358	0.7664	0.9095	0.9314	0.5756	0.8197
	MAE	0.3001	0.3669	0.42	0.0098	0.0088	0.0141	0.1866

<Table 8> Comparison of prediction model for concrete block strength and water absorption in Facility B

Model	metric	s1	s2	s3	a1	a2	a3	avg
GBDT	R ²	0.9768	0.9719	0.9828	0.9818	0.9185	0.9029	0.9558
	MAE	0.1292	0.1502	0.1063	0.0144	0.0134	0.0237	0.0729
RF	R ²	0.8627	0.8741	0.8981	0.7653	0.7513	0.6767	0.8047
	MAE	0.4784	0.4739	0.4605	0.0523	0.0577	0.0716	0.2657
ET	R ²	0.9513	0.9491	0.9449	0.9739	0.9286	0.8518	0.9333
	MAE	0.2151	0.2306	0.2301	0.0151	0.0216	0.0302	0.1238
XGB	R ²	0.958	0.9522	0.9585	0.9447	0.9097	0.8609	0.9307
	MAE	0.1877	0.196	0.1805	0.0201	0.0236	0.0311	0.1065

세부 항목별로는 강도(s1~s3)에서 모든 모델이 대체로 높은 R2(대략 0.81~0.88 수준)를 보여 강도 예측은 비교적 안정적으로 수행되는 경향이 있었다. 그러나 흡수율(a1~a3) 중 a3 항목에서 R2가 전반적으로 크게 하락하는 패턴이 관찰되었다(GBDT 0.6385, RF 0.6622, ET 0.6414, XGB 0.5756). 이는 설비 A에서 a3에 해당하는 목표변수가 다른 항목 대비 예측 난이도가 높음을 이야기한다.

설비 B에서는 평균 기준으로 GBDT가 R2=0.9558, MAE=0.0729로 가장 우수하여 사실상 전 항목에서 매우 높은 정확도를 달성하였다. 다음으로 ET(R2= 0.9333, MAE=0.1238)와 XGB(R2=0.9307, MAE=0.1065)가 높은 성능을 보였다. 반면 RF는 평균 R2=0.8047, MAE=0.2657으로 가장 낮은 성능을 보여, 설비 B 데이터에서는 트리 기반 배경(Random Forest) 단독 적용 시 한계가 확인되었다.

특히 설비 B의 GBDT, ET, XGB는 R2가 0.93 이상에 근접하는 수준으로, 강도 및 흡수율 예측 모두에서 안정적인 결과를 보였다. 이는 동일한 모델 계열이라도 설비별 데이터 분포, 공정 변동성, 로그 품질 차이에 따라 성능이 크게 달라질 수 있음을 보여준다. 각 모델별 성능 비교에 사용한 최적의 하이퍼파라미터 값은 <Table 9>와 같다.

<Table 9> Best hyper-parameter settings for each candidate model in Facilities A and B

Model	Facility	Best hyper-parameter
GBDT	A/B	n_estimators = 500 learning_rate = 0.05 max_depth = 3
RF	A/B	n_estimators = 400 max_depth = 20 max_features = "log2"
ET	A	n_estimators = 400 max_depth = 20 max_features = "log2"
	B	n_estimators = 600 max_depth = 10 max_features = 0.7
XGB	A	n_estimators = 400 learning_rate = 0.1 max_depth = 4 subsample = 1.0 colsample_bytree = 0.8
	B	n_estimators = 200 learning_rate = 0.03 max_depth = 5 subsample = 0.8 colsample_bytree = 0.8

이 결과는 설비 A와 설비 B의 PLC 기반 공정 데이터를 이용해 후보 모델(GBDT, RF, ET, XGB)에 대해 하이퍼파라미터 탐색을 수행한 결과로 각 모델에서 최적으로 선정된 하이퍼파라미터 조합을 정리한 것이다. 결과적으로, 모델별 설비 A와 B에서 최적 하이퍼파라미터 조합이 서로 달랐으며, 설비별 데이터 특성에 따라 튜닝 결과가 달라지는 경향을 보였다. 반면 GBT와 RF는 설비 A/B 모두에서 동일한 최적 설정이 도출되어 비교적 일관된 하이퍼파라미터 구성을 나타냈다.

5.3 SHAP 기반 전역 변수 중요도 분석

최종 모델이 트리 기반 앙상블 모델이므로, 설명 가능 인공지능 기법인 SHAP(Shapley Additive Explanations)를 이용하여 입력 변수들이 예측 결과에 미치는 영향을 해석하였다. 다중 출력 회귀 구조에서는 목표 변수마다 별도의 추정기가 존재하므로, 강도와 흡수율에 대해 각각 대표 목표 변수를 선택하여 SHAP 분석을 수행하였다.

본 논문에서는 다중 출력(강도 3종, 흡수율 3종) 모델의 SHAP 결과를 한 장의 그림으로 요약하기 위해, 각 출력별 SHAP 값을 그대로 합치는 방식을 적용하였다. 구체적으로는 Multi Output Regressor의 각 목표변수에 대해 SHAP 값을 계산한 뒤, 출력별 SHAP 값 행렬을 모아 절대값을 취하고, 동일한 샘플-특성 위치에서 출력들에 대한 평균 절대 SHAP 값을 계산하여 단일 집계 SHAP 행렬을 생성하였다. 이후, 이 집계 SHAP 값을 이용해 요약 플롯을 1장으로 시각화함으로써, 특정 목표변수에 국한되지 않고 전체 강도와 흡수율 예측에 공통적으로 크게 기여하는 변수를 직관적으로 비교할 수 있도록 하였다. 이 방식은 집계 과정에서 부호 정보는 상쇄될 수 있어 방향성보다는 영향의 크기 중심으로 해석한다.

SHAP 요약 플롯은 변수별 평균 |SHAP| 기준의 중요도와 샘플별 기여도 분포를 동시에 나타낸다. y축은 중요도가 큰 순서대로 정렬된 입력 변수이며, x축의 SHAP 값이 클수록 해당 변수가 모델 예측에 미치는 영향의 크기가 크다. 점의 색상은 변수값의 크기를 의미하며(빨간색: High, 파란색: Low), 같은 변수라도 운영 조건에 따라 기여도 크기가 달라질 수 있음을 분포 형태로 확인할 수 있다.

[Fig. 2]에서는 운전잔여분 계열 변수(unit 1~5)가 상위 다수 포함되어 예측에 가장 큰 영향을 미치는 것으로 나타났다.



[Fig. 2] SHAP summary plot for Facility A



[Fig. 3] SHAP summary plot for Facility B

다음으로 2차 진동시간(평균, 표준편차)와 몰탈 공급 시간(평균, 표준편차)이 상위 중요 변수로 확인되어, 설비 A에서는 공정 시간 및 진동 조건의 변동이 목표 물성 예측에 크게 작용함을 시사한다. 또한 온도(센서 4) 표준

편차가 상위권에 포함되어, 공정 중 온도 변동성이 제품 강도와 흡수율 예측에 유의미한 설명력을 제공하는 것으로 해석된다. 반면, 호퍼 계량 설정값 및 실제 계량값(호퍼, 시멘트)은 중상위 또는 중위권에 분포하여, 설비 A에서는 투입량 자체보다 운전 시간 및 공정 조건(진동, 공급, 온도)이 상대적으로 더 큰 예측 단서로 활용되는 경향을 보였다.

[Fig. 3]에서는 적재된 블록 개수(stccnt)가 다른 변수 대비 매우 큰 SHAP 분포 범위를 보여 모델 예측을 지배하는 핵심 변수로 나타났다. 그 다음으로 호퍼 실제 계량값, 1차 진동시간(평균), 믹싱 횟수가 상위권을 구성하여, 설비 B에서는 원재료 투입량과 혼합 및 진동 조건이 예측 성능에 직접적으로 기여하는 패턴이 뚜렷하다. 또한 시멘트와 물 계량 설정값의 표준편차가 중요 변수로 포함되어, 설비 B에서는 설정값의 안정성이 제품 물성 예측에 의미 있는 신호로 작용함을 보여준다. 전반적으로 설비 B는 상위 변수에서 SHAP 값의 분산이 크고 영향 범위가 넓어, 특정 공정 변수의 조건 변화가 예측값을 크게 좌우하는 구조적 특성이 관찰된다.

설비 A의 SHAP 결과에서는 운전잔여분, 진동시간, 몰탈 공급시간, 온도 변동 등 공정 시간 및 운전 조건 변수가 예측에 가장 크게 기여하였다. 반면 설비 B에서는 stccnt와 호퍼 실제 계량값, 믹싱 횟수, 진동시간 등 생산 상태 및 투입량 변수의 영향이 두드러졌다. 즉, 설비별로 제품 강도 및 흡수율 예측에 핵심적으로 작용하는 공정 인자가 상이하므로, 해석 및 공정 최적화 또한 설비 특성에 맞춘 접근이 필요하다.

6. 결론 및 향후 연구

본 논문은 콘크리트 블록 제조 공정에서 수집되는 배합기, 생산설비, 탈링기, 양생기의 PLC 데이터를 통합하고, 이를 실측 압축강도 및 흡수율 시험 결과와 연계하여 다중 물성 예측이 가능한 학습 데이터셋을 구축하였다. 또한 다중 출력 회귀 모델(GBDT, RF, ET, XGB)을 비교함으로써 설비별 데이터 특성에 적합한 예측 모델을 도출하고, 최종 모델에 대해 SHAP 분석을 적용하여 공정 변수가 물성 예측에 기여하는 구조를 정량적으로 제시하였다.

성능 측면에서 설비 A에서는 ET가 평균 R^2 0.8846 (평균 MAE 0.1661)로 상대적으로 우수한 예측력을 보였고, 설비 B에서는 GBDT가 평균 R^2 0.9558(평균

MAE 0.0729)로 강도 및 흡수율 전반에서 매우 안정적인 성능을 나타냈다. 해석 결과, 설비 A는 진동시간, 공급시간, 온도 변동 등 공정 시간 및 운전 조건 변수의 영향이 두드러진 반면, 설비 B는 생산 상태와 투입량 관련 변수의 영향이 상대적으로 크게 나타나 설비별로 핵심 관리 인자가 상이함을 확인하였다. 이는 현장에서 품질 편차의 원인을 데이터 기반으로 진단하고, 설비 특성에 맞춰 우선 관리해야 할 공정 조건을 도출함으로써 시험 비용과 리드타임을 줄일 수 있는 가능성을 보여준다.

다만 본 논문은 단일 공장의 특정 기간 데이터에 한정되어 있으며, 일부 공정 변수는 센서 정밀도와 누락값 처리 방식의 영향을 받을 수 있다는 한계가 있다. 향후 연구에서는 다양한 공장과 제품군으로 데이터 범위를 확장해 모델 일반화를 검증하고, 강도 및 흡수율뿐 아니라 비용과 탄소 배출량까지 함께 고려하는 다목적 최적화 프레임워크로 확장할 필요가 있다.

REFERENCES

- [1] M.Shaaban, K.Shaban, A.F.Deifalla and M.Alam, "Machine learning approaches for forecasting compressive strength of high-strength concrete," *Scientific Reports*, Vol.15, Article 25567, 2025.
- [2] M.H.Kwon and M.Kang, "Predicting Compressive Strength of Concrete Using Regression Model in Machine Learning," *Journal of the Korean Society for Advanced Composite Structures*, Vol.13, No.4, pp.24-35, 2022.
- [3] J.Jeon, H.S.Kim and C.Kim, "Study on Prediction of Compressive Strength of Concrete based on Aggregate Shape Features and Artificial Neural Network," *Journal of The Korea Institute for Structural Maintenance and Inspection*, Vol.25, No.5, pp.135-140, 2021.
- [4] J.Choi, *Deep Learning-based Prediction of Concrete Strength and Mix Proportion Estimation*, M.S. thesis, Hanyang University, 2023.
- [5] S.Elhishi, A.M.Elashry and S.El-Metwally, "Unboxing machine learning models for concrete strength prediction using XAI," *Scientific Reports*, Vol.13, Article 19892, 2023.
- [6] S.O.Abioye, Y.O.Babatunde, O.A.Abikoye, A.N.Shaibu and B.J.Bankole, "Optimized machine learning algorithms with SHAP analysis for predicting compressive strength in high-performance concrete," *AI in Civil Engineering*, Vol.4, No.16, 2025.
- [7] H.Fu, X.Zhou, P.Xu and D.Sun, "Prediction of Compressive Strength of Concrete Using Explainable Machine Learning Models," *Materials*, Vol.18, No.21, Article 5009, 2025.
- [8] M.S.Khan, T.Peng, M.A.Khan, A.Khan, M.Ahmad, K.Aziz, M.M.S.Sabri and N.S.Abd El-Gawaad, "Explainable AutoML models for predicting the strength of high-performance concrete using Optuna, SHAP and ensemble learning," *Frontiers in Materials*, Vol.12, Article 1542655, 2025.
- [9] Y.Kim and J.Shin, "Data-Driven Analysis of Retrofit Effects on RC Columns Under Blast Loading: A PDP-Based xAI Approach," *Journal of the Computational Structural Engineering Institute of Korea*, Vol.38, No.5, pp.317-324, 2025.
- [10] T.Akiba, S.Sano, T.Yanase, T.Ohta and M.Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.2623-2631, 2019.
- [11] S.M.Lundberg and S.I.Lee, "A unified approach to interpreting model predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp.4768-4777, 2017.
- [12] S.M.Lee, H.S.Sung and H.K.Kang, "Comparison of Performance for Predicting Compressive Strength of Concrete Using Machine Learning," *Journal of the Korea Concrete Institute*, Vol.34, No.5, pp.505-513, 2022.
- [13] K.C.Kim, B.J.Ji, Y.K.Hwang, Y.S.Yoon and K.T.Koh, "Predicting Concrete Compressive Strength using Data-centric Machine Learning Approach," *Journal of the Korean Recycled Construction Resources Institute*, Vol.12, No.4, pp.449-458, 2024.
- [14] Y.Zhang, W.Ren, Y.Chen, Y.Mi, J.Lei and L.Sun, "Predicting the compressive strength of high-performance concrete using an interpretable machine learning model," *Scientific Reports*, Vol.14, No.1, Article 28346, 2024.
- [15] Y.Wu and Y.Zhou, "Hybrid machine learning model and Shapley additive explanations for compressive strength of sustainable concrete," *Construction and Building Materials*, Vol.330, Article 127298, 2022.

손 현 승(Hyun Seung Son) [정회원]



- 2007년 2월 : 홍익대학교 컴퓨터 정보통신공학과 (공학사)
- 2009년 2월 : 홍익대학교 전자전산공학과 (공학석사)
- 2015년 8월 : 홍익대학교 전자전산공학과 (공학박사)
- 2021년 3월 ~ 현재 : 국립목포대학교 컴퓨터학부 조교수

<관심분야>

딥러닝 및 기계학습, 스마트양식, 메타모델, 모델변환, 테스트, SW재사용

조 광 문(Kwangmoon Cho) [종신회원]



- 1995년 8월 : 고려대학교 전산과 학과(이학박사)
- 1995년 9월 ~ 2005년 2월: 삼성 전자 통신연구소 선임연구원
- 2000년 3월 ~ 2005년 2월 : 백석대학교 정보통신학부 교수
- 2005년 3월 ~ 현재 : 국립목포대학교 컴퓨터학부 교수

<관심분야>

사물인터넷, 통신 소프트웨어, 인공지능 교육, 웹 서비스