**Prediction of concrete compressive strength using a Deepforest‑based model**

* 콘크리트 압축 강도 시험

1. 비파괴 검사(NDT, Non-Destructive Testing)

* 콘크리트의 물리적 properties를 기반으로 압축강도를 정량화
* 손상을 주지 않고 물성 간의 관계 추론

1. 반발경도법(rebound method)

* Robound instrument를 이용하여 콘크리트 경도 측정, 탄산화 깊이를 기준으로 압축강도 추정
* 간단하고 유연하지만 정확도가 다소 떨어짐

1. 초음파 속도법(ultrasonic pulse velocity)

* 콘크리트 내에서의 초음파 속도 변화를 관측하여 강도를 추정
* 재료 구성, 수분 함량, 양생 기간 등의 변수가 속도에 영향을 미침
* 압축강도와의 관계를 정확하게 파악하기 어려운 경우가 多

1. 방사선 조사법(ray method)

* 감마선을 이용한 기법
* 방사선 안전 문제로 실제 적용은 드뭄

1. 미세손상 검사(Micro-Damage Testing)

* 체계적인 시험을 통해 콘크리트 구조물의 강도와 결함을 평가
* 높은 탐지 정확도

1. Core Drilling

* 콘크리트 부재에서 코어 샘플을 채취한 뒤, 이를 파괴 시험을 통해 콘크리트 강도를 산정
* 매우 정밀한 실험 결과를 제공하지만, 상대적으로 복잡한 작업 공정과 높은 비용이라는 단점 有

1. Pull-out

* 구조 부재 내에 미리 리벳 부품을 삽입한 후, 시편을 뽑아낼 때 콘크리트가 견디는 파괴력을 측정
* 이 데이터를 기반으로 압축강도를 추정
* 코어 드릴링보다 작업이 간단하고 비용이 낮으며 적용이 쉬움
* 다양한 실험 변수에 영향을 받아 정확도가 낮을 수 있음

1. Nail shooting

* 해머 축을 이용해 시험용 못을 콘크리트에 박은 다음, 다이얼 게이지를 통해 박힌 깊이를 측정
* 이 깊이를 통해 압축강도를 추정
* Factor

1. blast furnace slag (BFS, 고로 슬래그)

* 철광석을 고온에서 녹여 제철할 때 나오는 부산물로, 시멘트 혼화재로 사용

1. fly ash (비산재)

* 석탄을 연소할 때 생기는 미세한 가루로 재활용하여 콘크리트에 혼입함
* 시멘트를 일부 대체하여 사용되고, 콘크리트 혼화재로 분류됨

1. water

* 물이 많아지면 시공은 쉬워지나 강도와 내구성은 약해짐

1. Super plasticizer (SP, 고성능 감수제)

* 콘크리트 유동성 개선 첨가제로 W/C 비율을 낮추면서도 작업성을 유지하거나 향상시킴
* 일반 감수제는 물 제거율이 5~10%, 고성능은 15~30%

1. coarse aggregate (굵은 골재)

* 자갈, 쇄석 등의 큰 입자 골재로 콘크리트에서 보이는 덩어리 자갈
* 강도에 큰 영향을 주는 물리적 뼈대 역할을 함
* 굵은 골재의 강도, 형태, 입도 분포에 따라 콘크리트의 특성이 달라짐

1. fine aggregate (잔골재)

* 모래에 해당하는 굵은 골재보다 작은 입자로, 시멘트 페이스트와 굵은 골재 사이를 메워주는 재료임
* 표면 마감, 점성, 수밀성에 중요한 역할을 함

1. Age (양생 기간, 경과 시간)

* 콘크리트를 타설한 후 경과된 일수 또는 시간을 말함
* 타설 후 3일, 7일, 28일, 90일이 일반적인 시험일임
* 양생이 진행됨에 따라 강도는 점점 증가함

1. Cement

* 다양한 요인들과 콘크리트 압축강도 간의 상관관계
* Factor 간 범위가 달라서 편향될 가능성이 있으므로 정규화 필요

1. 텍스트, 스크린샷, 소프트웨어, 컴퓨터 아이콘이(가) 표시된 사진

   AI 생성 콘텐츠는 정확하지 않을 수 있습니다.blast furnace slag (고로 슬래그)

* 슬래그가 50~150 정도 혼입되면 강도가 높아지는 경향이 일부 보임
* 전체적으로 큰 상관관계는 뚜렷 X

1. fly ash

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.100이상부터 강도 하락 경향을 보임
* 전반적으로 고강도와 큰 관련 X

1. Water

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* 음의 상관관계를 보임

1. super plasticizer

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.값이 일정 부분에 집중되어 있음
* 강도와의 관계는 불규칙하나 소량 첨가시 작업성 & 강도 향상

1. Coarse Aggregate

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.전반적으로 명확한 상관관계는 X
* 너무 많거나 적으면 강도에 부정적일 수 있음

1. Fine Aggregate

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.전반적으로 명확한 상관관계는 X
* 너무 많거나 적으면 강도에 부정적일 수 있음

1. Age

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.1 ~ 30일 사이 강도가 빠르게 증가하고 이후에 완만해짐
* 양의 상관관계를 가짐

1. Cement

* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.시멘트 양이 많아지면 강도가 높아지는 경향을 보임
* 강한 양의 상관관계는 아니지만 경향성 은 존재함
* Random Forest (RF, 랜덤포레스트)
* 여러 개의 CART 결정 트리(decision tree)를 결합하여 예측을 수행
* 비선형적이고 고차원의 복잡한 문제를 다루는 데 강점을 가짐
* 일반적으로 다수의 결정 트리를 생성하기 위해 무작위 벡터를 여러 번 생성하여 각 트리를 구성
* 트리들의 예측을 종합하여 최종 결과를 도출
* RF와 Extra trees의 차이점

|  |  |  |
| --- | --- | --- |
|  | Random Forest | Extra Trees |
| Training Data set의 선택 방식 | Bootstrap sampling을 통해 data set 구성 | 원래의 training data set을 그대로 사용 |
| 분할 지점 결정 | 전통적인 decision Tree처럼 최적의 특성 분할 지점(Optimal factor partition point)을 결정함 | 특성 값을 임의로 선택하여 분할 지점을 결정함  -> RF보다 DT 구조가 더 큼 |
| 장단점 | - | - RF보다 분산을 줄일 수 있음  - 대신 편향이 더 커짐  - 일반화 능력이 RF보다 우수한 경우도 있음 |

\*bootstrap Sampling

- 원본 데이터 set에서 복원 추출 방식으로 같은 크기의 새로운 데이터 set을 여러 번 만드는 통계적 기법

-> 각 tree를 훈련시킬 때, 다른 bootstrap sample을 써서 다양성 부여

-> Data가 부족한 상황에서도 복수의 모델 훈련을 가능하게 해줌

\*Generalization ability (일반화 능력)

- 머신러닝 모델이 훈련 데이터에만 과도하게 맞추지 않고, 새로운 데이터에 대해서도 잘 예측할 수 있는 능력

* Deep forest
* RF와 심층 ANN의 장점을 결합하여 cascaded forest을 통해 전통적인 트리 기반 모델을 심층적으로 통합
* 다중 스케일 스캐닝을 통해 비선형 매핑 능력을 향상시킴
* 딥러닝에 비해 하이퍼파라미터가 적고 학습이 쉬움
* cascade forest)라고도 불리며, RF와 ExtraTrees의 개별적인 장점을 종합적으로 통합하기 위해 고안
* Deepforest의 각 계층(layer)은 두 개의 RF 기반 모델과 두 개의 ExtraTrees 기반 모델로 구성됨
* 각 layer에서의 콘크리트 압축강도 예측값은 이 네 가지 모델의 평균값으로 계산됨
* 모델 학습 중에는, 계층별(layer-by-layer) 연산을 수행하며, 이전 계층에서의 연산 결과를 다음 계층으로 전달
* 동시에 각 계층의 예측 오차가 이전 계층과 얼마나 차이가 나는지를 평가

1. 현재 계층의 예측 오차가 이전 계층보다 작아진다면

* 다음 계층의 계산을 계속 수행

1. 현재 계층의 예측 오차가 더 커진 경우

* 다음 계층으로의 계산을 중단
* 이후, 전체 계층 중에서 가장 낮은 예측 오차를 보인 계층을 출력 계층으로 선택하여 최종 예측값을 생성

텍스트, 도표, 평면도, 기술 도면이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. 최적 계층 수

* 해당 계층이 출력 계층으로 선택되었을 때 모델의 예측 오차(평균 제곱 오차, MSE)가 최소가 되는 계층 수
* Performance evaluation
* Grid Search 기법을 활용하여 각 모델의 최적 하이퍼파라미터를 탐색함

: 여러 하이퍼파라미터의 후보값을 grid 형태로 조합해서 그 중 best를 찾는 방식

* 다양한 모델의 성능 평가를 위해 Fivefold Cross-Validation을 수행함

: Training Data set을 5개로 분할한 후, 4개는 training용으로, 1개는 test 용으로 사용

* 이 과정을 5회 반복한 후, 결과의 평균을 구해 전체적인 성능 평가

1. Linear (선형 회귀)
2. KNN (K-Nearest Neighbors, K-최근접 이웃)

* 가장 가까운 K개의 이웃 데이터를 찾아서 그들의 값을 회귀하는 방식

1. DT (Decision Tress, 결정 트리)
2. SVR (Support Vector Regression , 서포트 벡터 회귀)

* SVM (Support Vector Machine)이라는 분류 알고리즘의 회귀 버전
* 너무 멀지 않은 선 안에 최대한 많은 데이터가 들어오게끔 하는 회귀선(또는 곡선)을 찾는 것

1. LASSO (Least absolute shrinkage and selection operator)

* 불필요한 변수를 자동으로 제거해주는 선형 회귀 알고리즘
* 변수의 계수들을 강제로 0에 가깝게 줄이면서(Shrinkage) 필요없는 변수의 계수는 아예 0으로 만듬 = 변수 선택 + 과적합 방지를 동시에 해주는 알고리즘임

1. MLP (Multilayer Perceptron)

* 인공신경망(ANN)의 가장 기본적인 형태로, 여러 층을 통해 입력값에서 복잡한 패턴을 학습

1. Extra Trees

* RF와 유사하지만 더 randomness가 강함
* 학습 속도가 빠르고 과적합 방지 효과가 좋음

1. RF

* 다수의 결정트리를 생성하고 예측값을 평균/다수결로 예측하는 앙상블 모델
* Bootstrap (중복추출) + 랜덤한 변수 선택을 이용해 다양한 트리 구성
* 비선형 문제도 잘 해결함

1. AdaBoost (Adaptive Boosting)

* 약한 학습기(예: 결정트리)를 순차적으로 연결하여 성능을 향상
* 이전 모델이 틀린 데이터에 가중치를 두어 다음 모델이 더 집중하게 함
* 단순한 모델을 여러 번 결합해도 높은 성능을 보이지만, 노이즈에 약하고 과적합 가능성이 있음

1. GBM (GradientBoosting)

* AdaBoost와 유사하지만, 오차에 대한 기울기(gradient)를 이용해 잔차(residual)를 줄이는 방식 -> 이 방향으로 다음 트리를 계속 학습
* 정확도가 매우 높고 조정 가능한 파라미터가 많다
* 학습 속도는 느리고 과적합 방지 기능이 필요함

1. Bagging (Bootstrap Aggregating)

* 여러 학습기(보통 결정트리)를 무작위로 샘플링된 데이터(=bootstrap sampling)로 학습시키고, 결과를 평균/다수결로 종합
* 과적합이 감소하고 모델 안정성이 향상됨
* 단순한 기법이지만 예측 성능을 높임

1. Deepforest

* Random Forest + Extra Trees를 계단식 층(layer) 구조로 연결한 모델
* 각 층마다 RF 2개, ExtraTrees 2개 사용하고 평균 예측값으로 다음 층에 전달
* 딥러닝보다 학습은 쉽고 하이퍼파라미터가 적음
* 과적합이 적고 복잡한 관계 학습이 가능함
* standard indicators (표준 지표) -> 4가지

1. 결정 계수(R²)

* 텍스트, 폰트, 도표, 라인이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.범위 ( -∞, 1)
* 값이 1에 가까울수록 예측기의 성능이 우수함을 의미함

1. 평균 제곱 오차(MSE)
2. 평균 절대 오차(MAE)
3. 제곱근 평균 제곱 오차(RMSE)

* 얘네 셋은 0에 가까울수록 예측 오차가 작다는 것을 의미함
* 결과 - 다양한 특성 인자에 따른 모델 성능
* 다양한 인자 조합을 사용하여 DT 기반 모델을 훈련 data set에 적용함
* 모델을 서로 다른 인자 조합으로 평가하기 위해 fivefold cross-validation을 수행하였으며, 비교를 위한 표준 지표들도 도입함

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* age만을 사용하여 훈련시켰을 때 R^2 = 0.39 로 가장 높음
* age가 가장 중요한 인자임
* 물과 시멘트의 R^2 또한 0.22, 0.24로 우수한 편임
* 수화 반응을 통해 시멘트가 경화되어 일정 수준 이상의 강도가 확보됨
* 사용되는 물과 시멘트의 비율, 종류 둥을 제어하여 강도와 내구성 향상 가능
* 물 인자는 상관계수가 -0.29로 가장 낮았지만, 음의 상관관계를 보여 오히려 뛰어난 예측 성능을 보였다.
* BFS + SP + age + cement + water 조합이 가장 높은 결정 계수를 달성함
* 결과 - 훈련 데이터셋에 대한 모델 성능
* 12개의 회귀 모델에 앞서 선택된 최적의 인자 조합을 사용함
* 공정한 성능 비교를 위해 동일한 훈련 데이터셋을 동일한 fold로 나누고 5겹 교차 검증을 수행

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* 각 모델의 최적의 하이퍼파라미터

텍스트, 스크린샷, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* Deepforest가 가장 우수한 성능을 보임
* RF 가 두번째로 우수한 성능을 보임
* 인공 신경망 기반 모델인 MLP는 가장 낮은 성능을 보임
* 해당 문제에 신경망 모델을 적절하지 않음을 알 수 있음
* 결과 - 독립 테스트 데이터셋에서의 모델 성능 비교
* 전체 학습 데이터셋을 사용하여 12개의 회귀 모델을 각기 최적의 하이퍼파라미터로 학습시킨 후, 독립 테스트 데이터셋에서 각 모델의 성능을 비교

텍스트, 스크린샷, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* Deepforest가 가장 높은 결정계수 0.91을 기록함
* ExtraTrees, RF, Bagging, Deepforest와 같은 트리 기반 모델들이 모두 결정계수가 0.90이상으로 나오는 우수한 성능을 보임
* LASSO 기반 모델은 가장 낮은 결정 계수 0.61을 보였다
* 트리 기반 모델이 압축 강도 예측에 더 적합함

∵ 트리 기반 모델이 예측에 가장 관련 있는 특성들을 자동으로 선택할 수 있기 때문임

+ 노이즈와 이상치에 대해 어느정도 안전함

* 트리 구성 과정에서 데이터 필터링 및 차원 축소를 진행하기 때문임

텍스트, 스크린샷, 그래프, 다채로움이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

* 약 5.8%의 샘플만이 10을 초과하는 절대 차이를 보였으며 이는 모델의 성능이 우수함을 보여줌
* 텍스트, 라인, 도표, 그래프이(가) 표시된 사진

  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.최첨단(State-of-the-art) 예측기와의 비교
* Deepforest 기반 회귀기의 성능을 더욱 탐색하기 위해, 최첨단 예측기들과 비교를 수행
* 모델은 각각 GEP 기반 모델, RF 기반 모델, SVR 기반 모델, ELM 기반 모델에 기반하여 구축됨
* 모든 모델은 동일한 학습 데이터셋으로 학습되었으며, 동일한 독립 테스트 데이터셋에서 평가됨
* Deepforest 기반 모델은 상관계수가 0.95 이상으로 가장 높았으며, 중심 RMS 오차 및 표준편차 또한 5개 모델 중 가장 낮았다
* 결론
* 모델 훈련을 위해, 고로 슬래그, 고성능 감수제, 양생 일수, 시멘트, 물의 다섯 가지 요소를 12개의 회귀 모델에 적용하였고, 하이퍼파라미터 최적화를 위해 grid 서치를 사용
* 검증 실험을 통해 Deepforest가 최적 회귀기로 선정
* 독립 테스트 데이터셋에서 R² = 0.91의 성능을 달성
* Deep forest 모델이 콘크리트 강도 예측에 효과적임
* 실험 결과에 따르면, 콘크리트 강도에 가장 큰 영향을 미치는 요인은 양생 일수이며, 그 다음은 시멘트임
* 고로 슬래그, 고성능 감수제, 양생 일수, 시멘트, 물 요소는 콘크리트 강도 예측에 있어 가장 우수한 지표로 확인됨
* 굵은 골재 및 잔골재를 추가했을 경우 R² 값이 향상되지 않았음
* 두 요소가 예측 정확도를 향상시키는 데 유의미한 영향을 미치지 않는다
* Deepforest 기반 회귀기는 훈련 세트와 독립 테스트 세트에서 모두 우수한 결과를 보였고, 명시적인 하이퍼파라미터 조정이 필요하지 않아 모델 최적화의 어려움을 줄임
* 이 모델(deep forest)의 장점

1. 보다 정확한 콘크리트 강도 예측 모델 구축을 통해 예측 정확도 향상
2. 보다 정밀한 콘크리트 배합 비율 산정을 통해 콘크리트의 성능 및 비용 최적화
3. 생산 과정에서 주요 파라미터를 보다 정확하게 제어함으로써 생산 효율성과 품질을 향상

* 머신러닝 기반 강도 예측 모델을 실제 환경에 적용할 경우 고려사항

1. 머신러닝 알고리즘의 기반이 되는 고품질 데이터 확보는 필수적
2. 실험 데이터의 정확성과 완전성을 보장해야 함
3. 서로 다른 머신러닝 알고리즘은 서로 다른 데이터셋과 문제에 적합할 수 있음

* 실제 상황에 따라 적절한 알고리즘을 선택하는 것이 필요

1. 머신러닝 모델은 높은 예측 정확도를 달성하기 위해 광범위한 훈련 및 검증이 필요하며, 충분한 시간과 자원의 투입이 요구됨
2. 강도 예측 모델을 실제 환경에 적용하려면 데이터 수집, 모델 배포, 유지보수 등 다양한 실제 적용상의 도전 과제를 함께 고려해야 함