- 연구에 사용한 머신러닝 기법 정리

선형 회귀 (Linear Regression):

-가장 간단하고 기본적인 회귀 알고리즘으로 종속 변수와 한 개 이상의 독립 변수간의 관계를 모델링하는데 사용되는 기법이다. 이때 입력 변수와 출력 변수의 선형관계를 파악하며 적은 계산으로 빠르게 학습이 가능합니다

*처음에 우리는 선형적인 관계가 있을 것으로 생각하여 linear Regression을 통해서 접근했다는 것 얘기하면 좋을 듯 -> 거기서 발전시켜 오버피팅이 일어나지 않게 하기위해 릿지회귀로 진행을 했다고 추가하면 될 듯

릿지 회귀(Ridge Regression)

-Ridge Regression은 Linear Regression의 한 변형으로 Overfitting을 방지하기 위한 정규화 기법중 하나입니다. 과적합이란 머신러닝 모델이 학습데이터에 너무 맞춰져 새로운 데이터에 대한 일반화 성능이 낮아지는 현상을 말하는데, 선형회귀 모델에 규제를 추가하여 이러한 성능저하가 일어나는 것을 저지해 주는 역할을 합니다.

*선형 회귀, 릿지 회귀로 진행하였을 때 그래프를 보여주면서 nSub, nARC, thARC 과 투과율과의 관계에서 선형적인 관계를 찾을 수 없어서 비선형 관계를 모델링 할 수 있는 밑에 나오는 방법(의사결정 트리 회귀, 랜덤 포레스트 회귀)들을 사용하게 되었다고 하면 될 듯

의사결정 트리 회귀 (Decision Tree Regression):

-비선형 관계를 모델링할 수 있는 알고리즘 중 하나입니다. 결정트리를 사용하여 연속적인 종속변수를 예측하는 기법으로서 데이터를 기반으로 의사결정 트리를 구축하여 예측합니다. 이 Dexision Tree Regression에서 중요한 하이퍼파라미터는 트리의 깊이, 분할 조건등이 있다. 이 회귀 방법에서도 과적합(Overfitting)이 일어나기 쉬워 하이퍼 파라미터르 조절하여 모델의 성능 및 과적합을 관리할 수 있다.

*이 기법을 썼을 때 하이퍼 파라미터의 값을 조절하지 않았더니 바로 오버피팅이 발생하였음

* 의사결정 트리 회귀(단일 트리 구조 -> 결정 과정을 이해하기 쉬움, 시각화 가능하여 설명력이 좋음, 적은 규모의 데이터 셋, 간단한 모델을 원할 때 유리)

* 랜덤포레스트 회귀 (여러 개의 트리를 앙상블하여 최종예측을 만들기 때문에 해석이 더 어려움, 다수의 트리가 결합된 구조로 시각화하기가 복잡함, 대규모 데이터셋에서 높은 성능, 작은 데이터셋에서는 과적합의 위험이 있음)

-대규모의 데이터셋이나 예측 성능이 우선일 때는 랜덤 포레스트가 더 적합할 수 있다!

랜덤 포레스트 회귀 (Random Forest Regression):

랜덤 포레스트 회귀(Random Forest Regression)는 결정 트리를 기반으로 하는 앙상블(Ensemble) 학습 방법 중 하나로, 여러 개의 결정 트리를 결합하여 높은 예측 성능을 제공하는 회귀 분석 기법입니다. 이는 분산을 줄이고 일반화 성능을 향상시키는 효과를 가져올 수 있다. 이 기법은 계산 비용이 높기 때문에 대용량 데이터셋에 적용할 때 속도가 느릴 수 있다.

신경망 기반 회귀 (Neural Network Regression):

Neural Network을 사용하여 연속적인 종속 변수를 예측하는 회귀 분석 기법입니다. Neural Network는 인공지능 모델 중 하나로, 생물학적 뉴런의 작동 원리를 모방하여 구성된 계층적인 모델로 입력 데이터를 학습하고 예측하기 위해 가중치와 활성화 함수를 사용합니다.

복잡한 비선형 관계를 대응할 수 있지만, 많은 데이터와 계산 리소스가 필요할 수 있습니다. 마찬가지로 과적합의 위험이 있어 적절한 규제와 조절이 필요합니다.

최적화

-머신러닝에서 최적화는 모델의 성능을 향상시키기 위해 모델의 파라미터를 조정하거나 학습 과정을 최적으로 만드는 프로세스를 의미합니다. 모델의 성능을 최대화하거나 손실을 최소화하는 것이 최적화의 주요 목표이고 이를 위해서 목적 함수(Objective Function) 또는 손실 함수(Loss Function)를 정의하고 이 함수를 최소화하거나 최대화하는 파라미터 값을 찾습니다

*우리는 모델에서 필요로 하는 파라미터에 적절한 값을 넣기 위하여 목적함수를 정의하여 이 함수에서 T_mean이 최댓값이 나올 수 있는 파라미터를 찾아 모델에 적용하였다

최적화의 종류

1. **목적 함수(Objective Function) 또는 손실 함수(Loss Function):**
   - 최적화의 목표를 정의하는 함수로, 모델의 성능을 측정하는 데 사용됩니다.
   - 분류 문제에서는 정확도를, 회귀 문제에서는 평균 제곱 오차(Mean Squared Error, MSE) 등을 목적 함수로 사용합니다.
2. **파라미터 조정:**
   - 모델의 성능을 향상시키기 위해 모델의 파라미터 값을 조정합니다.
   - 파라미터는 모델의 가중치(weight) 및 편향(bias) 등을 포함하며, 이 값을 최적으로 조정함으로써 모델의 예측 성능을 향상시킬 수 있습니다.
3. **경사 하강법(Gradient Descent):**
   - 최적화의 핵심 알고리즘 중 하나로, 기울기(경사)를 활용하여 목적 함수의 최솟값을 찾아가는 과정입니다.
   - 학습률(learning rate)과 함께 사용되며, 학습률은 얼마나 큰 보폭으로 이동할지 결정합니다.
4. **미분과 기울기(Gradient):**
   - 목적 함수를 최소화하기 위해 기울기를 계산하고, 이를 통해 최적의 파라미터 값을 찾습니다.
   - 경사 하강법에서는 목적 함수를 파라미터에 대해 편미분하여 기울기를 계산합니다.
5. **하이퍼파라미터 튜닝:**
   - 최적화 과정에서 학습률 및 정규화 강도와 같은 하이퍼파라미터를 조절하여 모델의 성능을 최적화합니다.
   - 하이퍼파라미터 튜닝은 반복적인 실험과 검증을 통해 이루어집니다.

스코어

머신러닝에서 스코어 함수(Score Function)는 주어진 입력 데이터에 대해 모델의 예측 성능을 측정하는 함수를 의미합니다. 스코어 함수는 모델의 특성과 문제의 특성에 따라 선택되며, 모델을 훈련하고 평가할 때 중요한 역할을 합니다. 선택된 스코어 함수를 최적화하여 모델의 성능을 향상시키는 것이 우리가 만든 모델의 타당성을 검증하는 데에 있어 중요한 역할을 한다고 볼 수 있습니다.