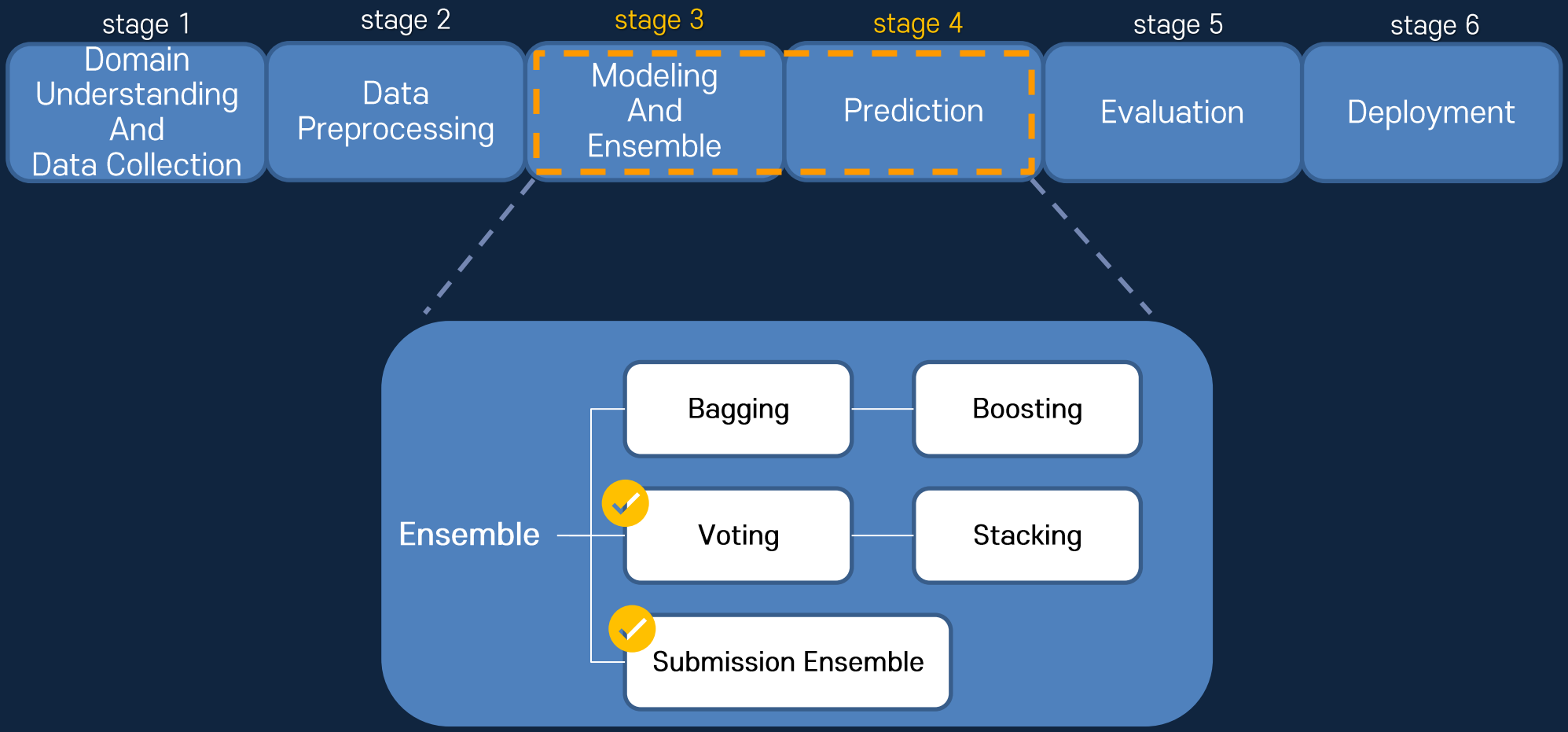
ML SESSION

#8 Ensemble

DATA&ANALYTICS

INDEX

0 ML FLOW


stage 1
Domain Understanding And Data Collection
stage 2
Data Preprocessing
stage 3
Modeling And Ensemble
stage 4
Prediction
stage 5
Evaluation
stage 6
Deployment
Ensemble
Bagging
Boosting
Voting
Stacking
Submission Ensemble

1 Ensemble

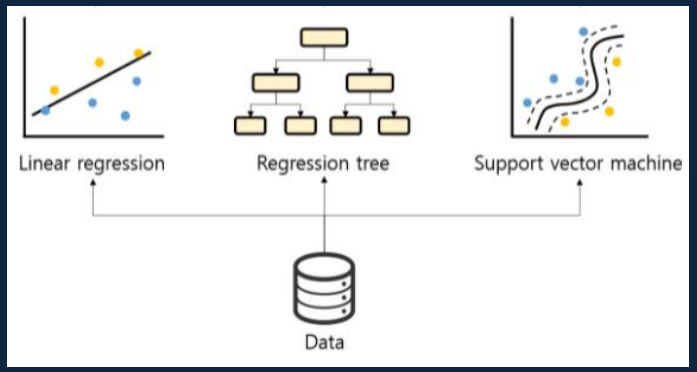
Ensemble이란?

하나가 아닌 여러 모델을 써서 예측력을 높이는 것

- ▷ 단일 모델보다 성능이 좋음
- ▷ 단일 모델의 약점을 보완, 단일 알고리즘의 약점을 보완

하나가 아닌 여러 모델을 써서 과적합을 막을 수 있음

- ▷ 데이터를 다양한 관점으로 바라보고 조합함
- ▷ 더 나은 일반화를 가능케함



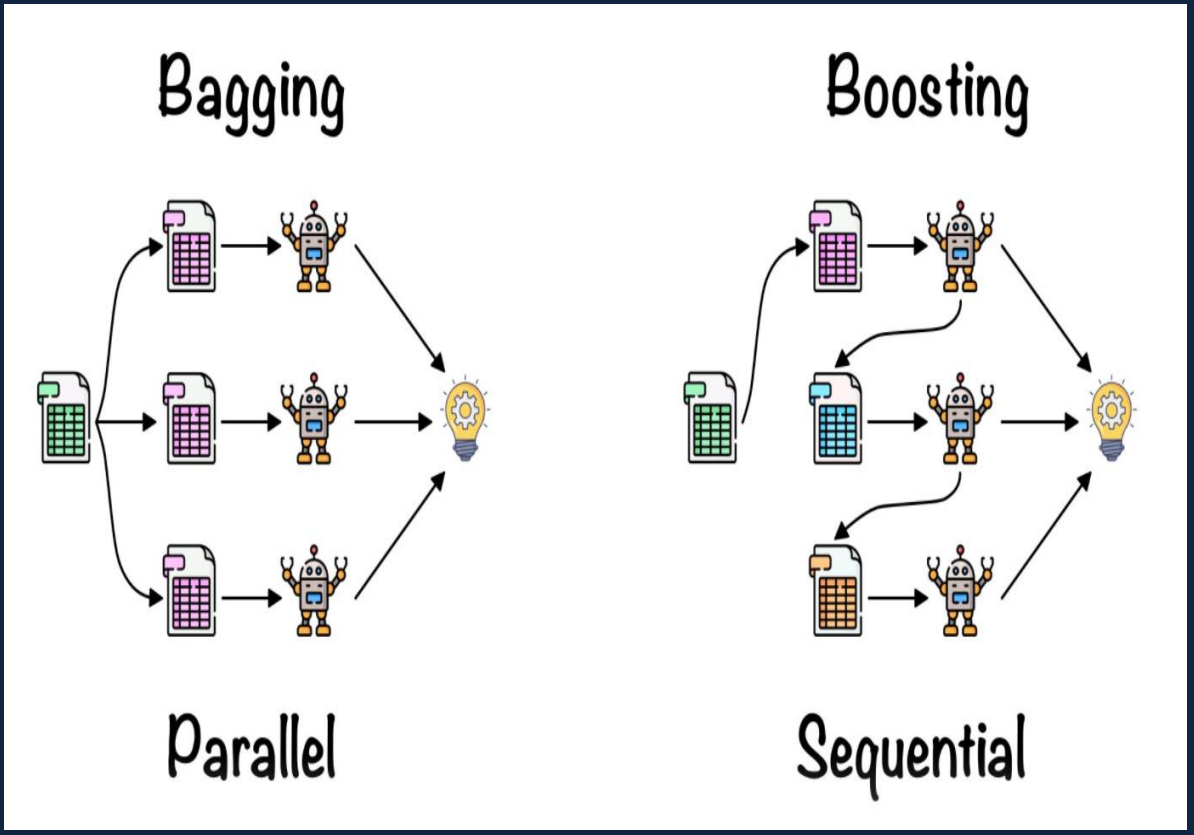
여러 모델?

같은 여러 모델을 사용하는 경우 ▷ Bagging, Boosting

다른 여러 모델을 사용하는 경우 ▷ Voting, Stacking

2 Ensemble

같은 여러 모델을 사용하는 경우



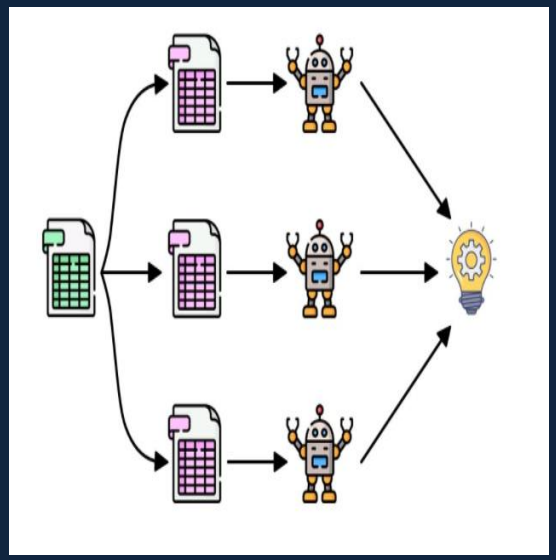
앙상블을 구성하는 모델들이 모두 같다

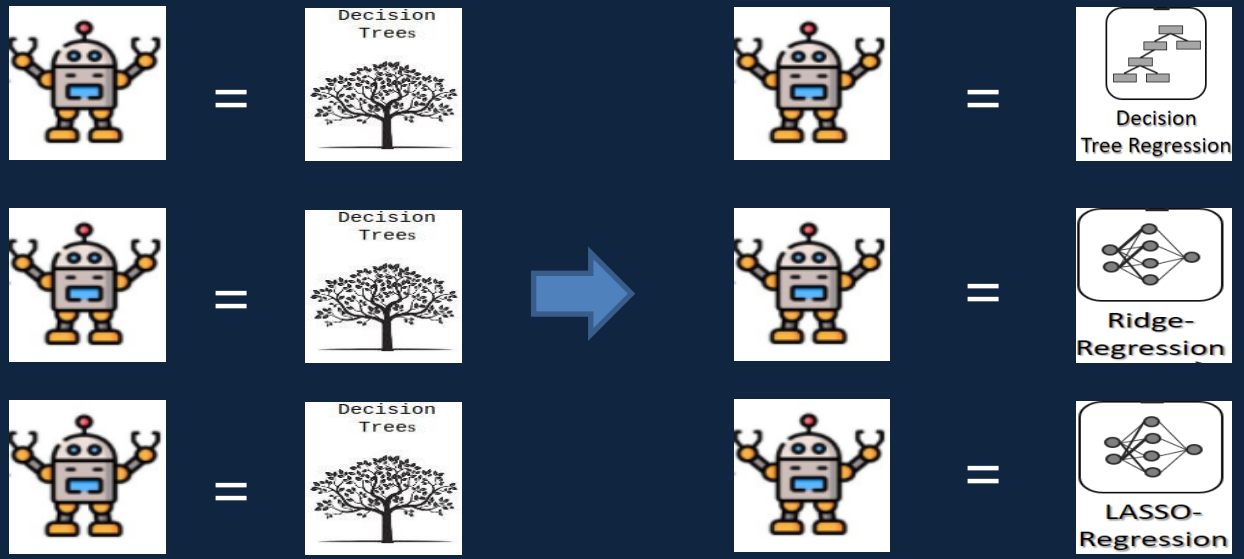
트리계열 모델들이 앙상블 되어
최종 모델을 구성하고 있다.

Bagging : RandomForest, ExtraTrees
Boosting : AdaBoost, GBM, XGB, LGBM

2 Ensemble

다른 여러 모델들을 사용하는 경우



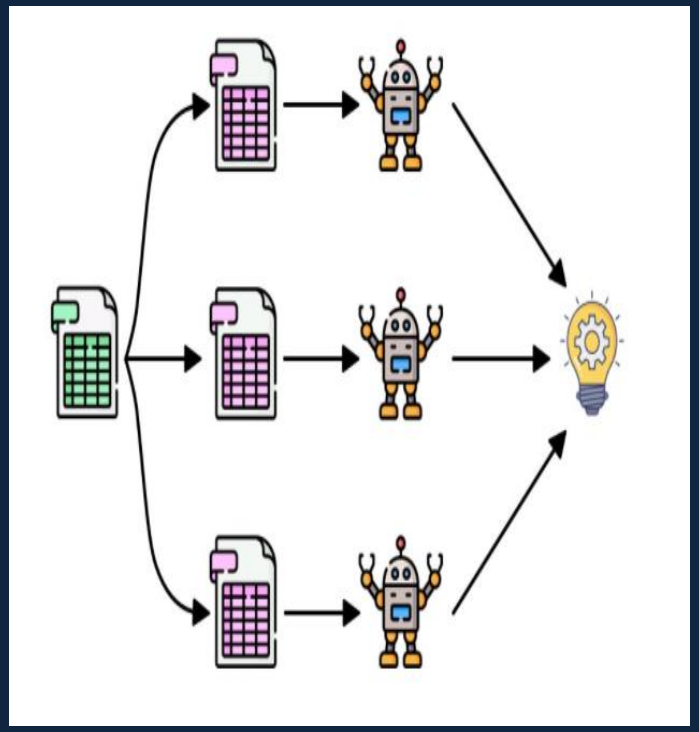


앙상블을 구성하는 모델들이 모두 다르다

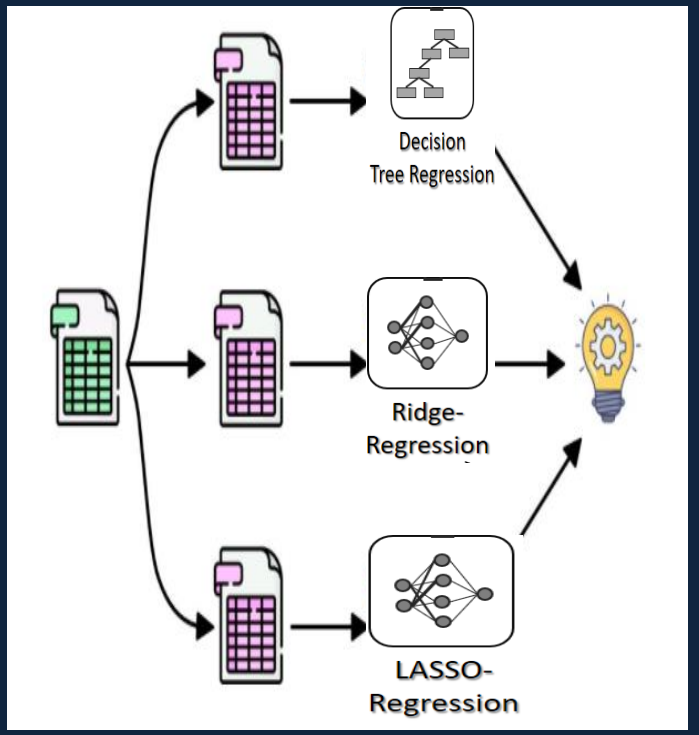
3 Voting

Bagging vs Voting

Bagging



Voting



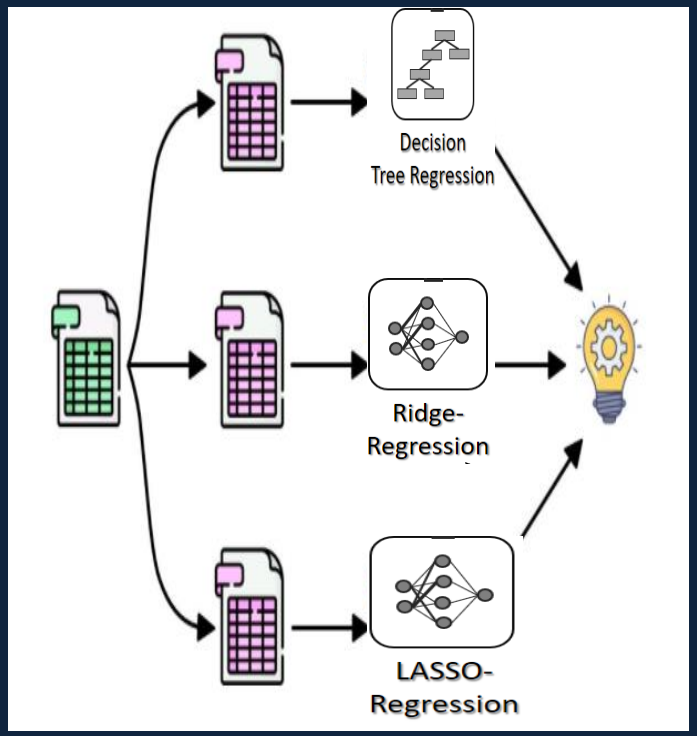
공통점 : Aggregation
- 분류: 최빈값
- 회귀: 평균

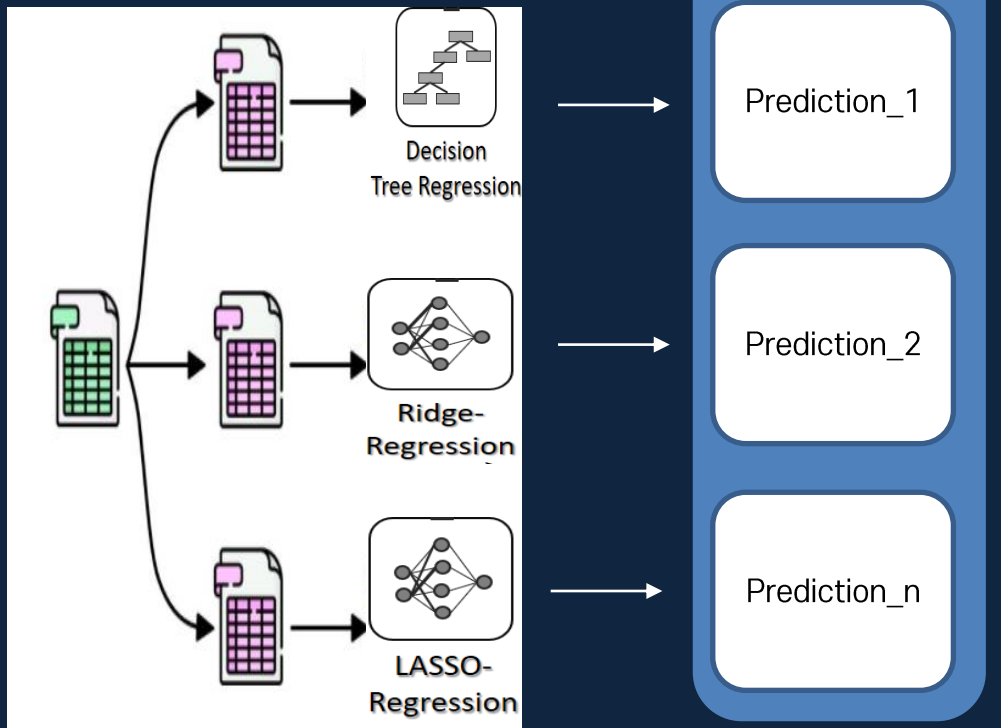
차이점 : 모델 종류
- Bagging : 같은 모델
- Voting: 다른 모델

4 Stacking

Stacking의 아이디어

Voting

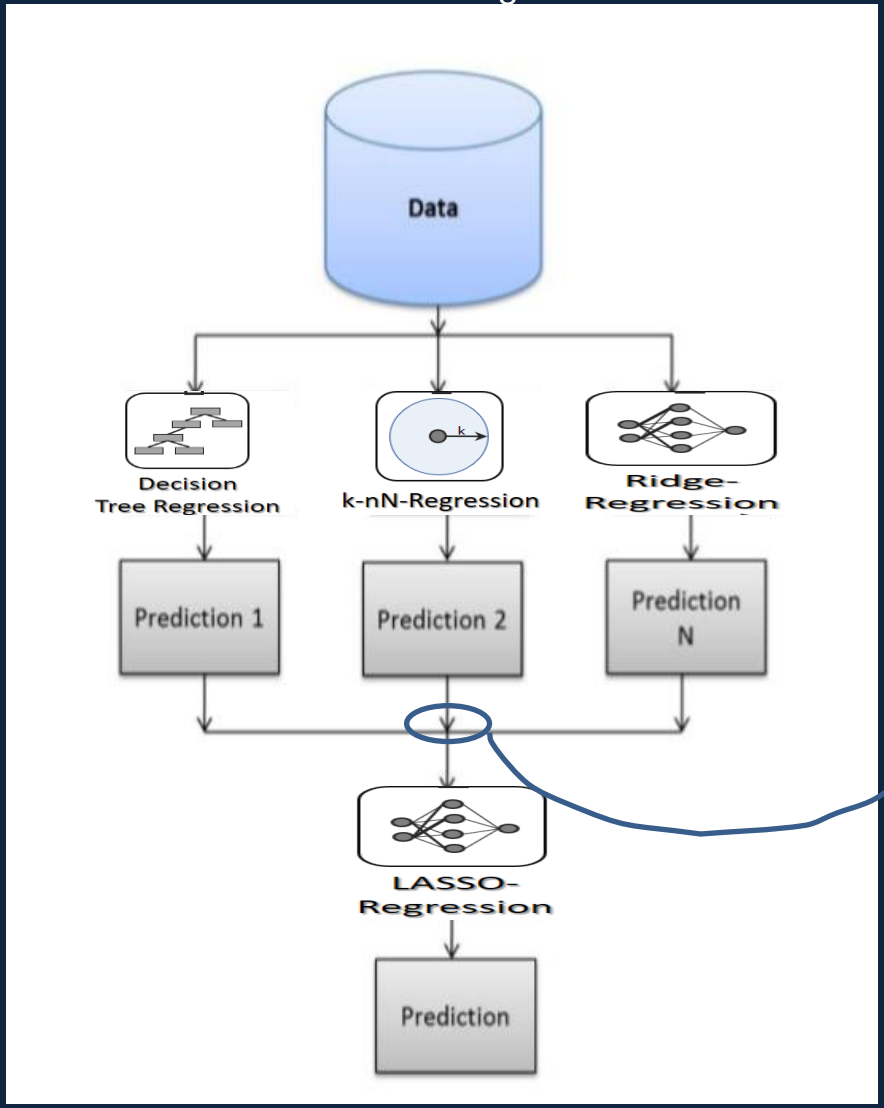




prediction들을 이용하여
앙상블 할 수 없을까?

4 Stacking

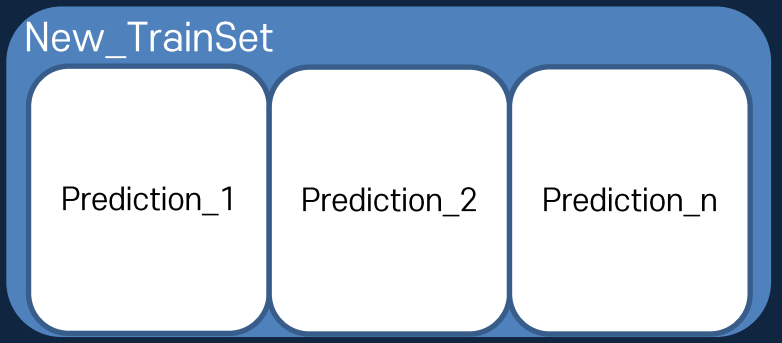
Stacking



① 앙상블 시킬 다양한 모델들을 가져옴

② 모델들을 학습시켜 Predictions를 찍음

③ Predictions를 모아서 새로운 TrainSet으로 만듦

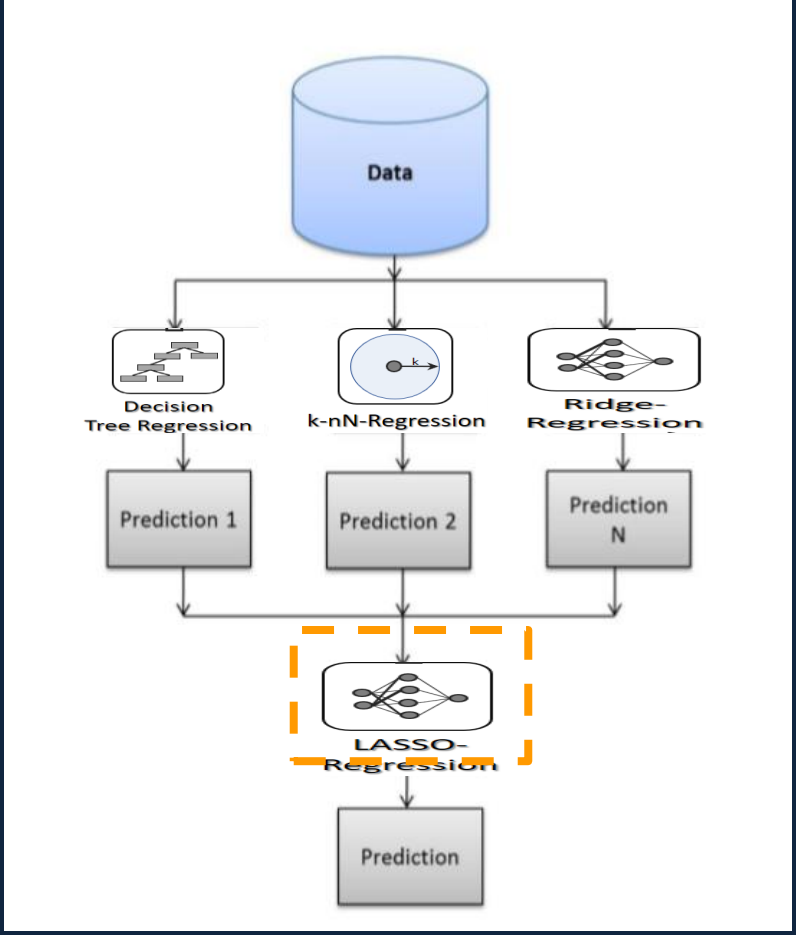


④ 마지막으로 설정한 모델로
새로운 TrainSet를 학습 및 예측

※ 마지막 모델을 일반적으로 Meta Model이라 부름

4 Stacking

Meta 모델의 활용



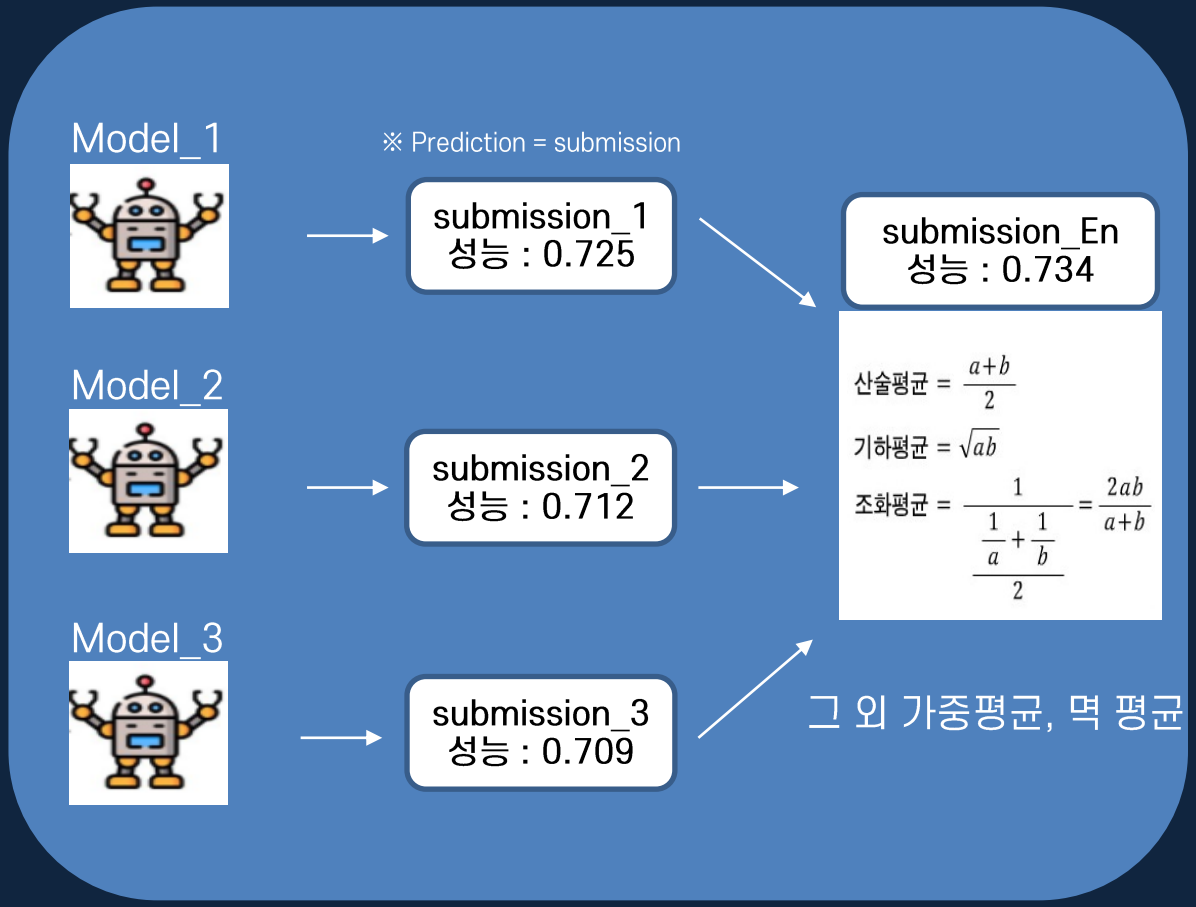
① Meta Model 1 : 성능 좋은 단일 모델

② Meta Model 2 : Voting

③ Meta Model 3 : Stacking

2 Layers Ensemble

5 Submission Ensemble



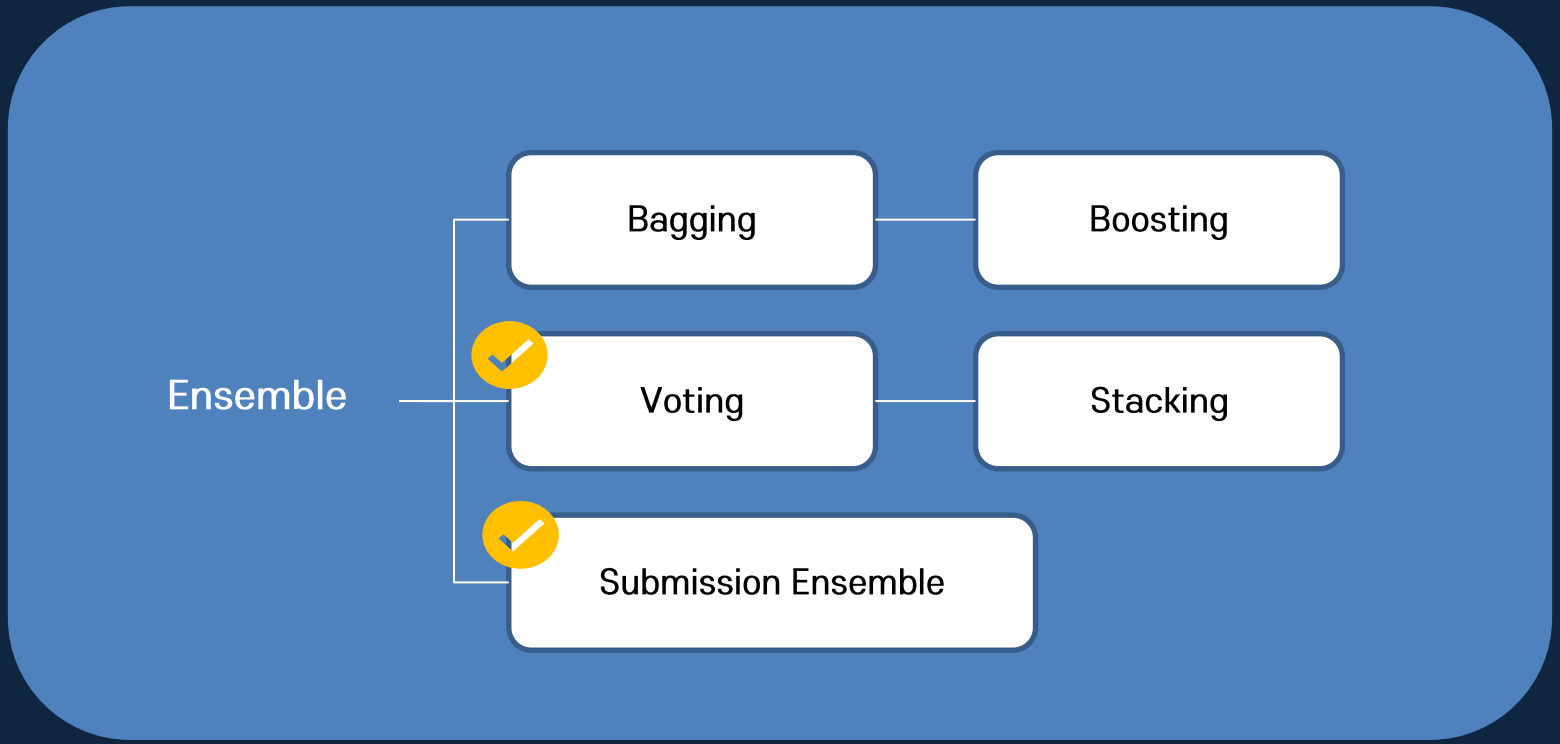
※ Submission Ensemble이 잘 작동하려면

- 모델들이 학습한 데이터가 조금씩 달라야 함
- 데이터를 학습하는 모델들이 알고리즘적으로 서로 다름
- 학습할 데이터도 조금씩 다르고 모델들도 알고리즘적으로 서로 다름
- ML_submission, DL_submission을 모두 활용하기도 함

※ 성능이 어느 정도 비슷한 Prediction끼리 해야함

5 Submission Ensemble

정리



THANK YOU