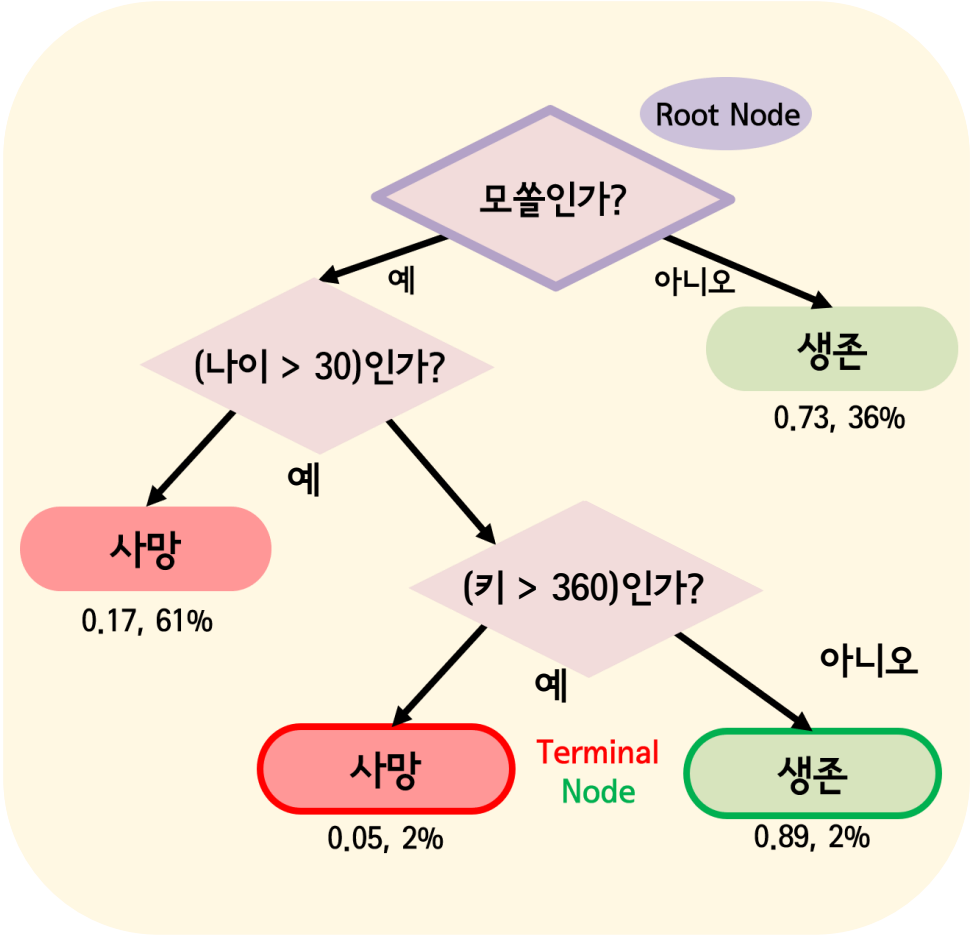
PRESENTATION

4팀

김동영
강수경
김재희
유경민
최윤혜

INDEX

CART - 분류모델



각 node는 서로 배타적이다

Terminal node 들의 합
=
Root node의 데이터 수

중복되어 분할 X

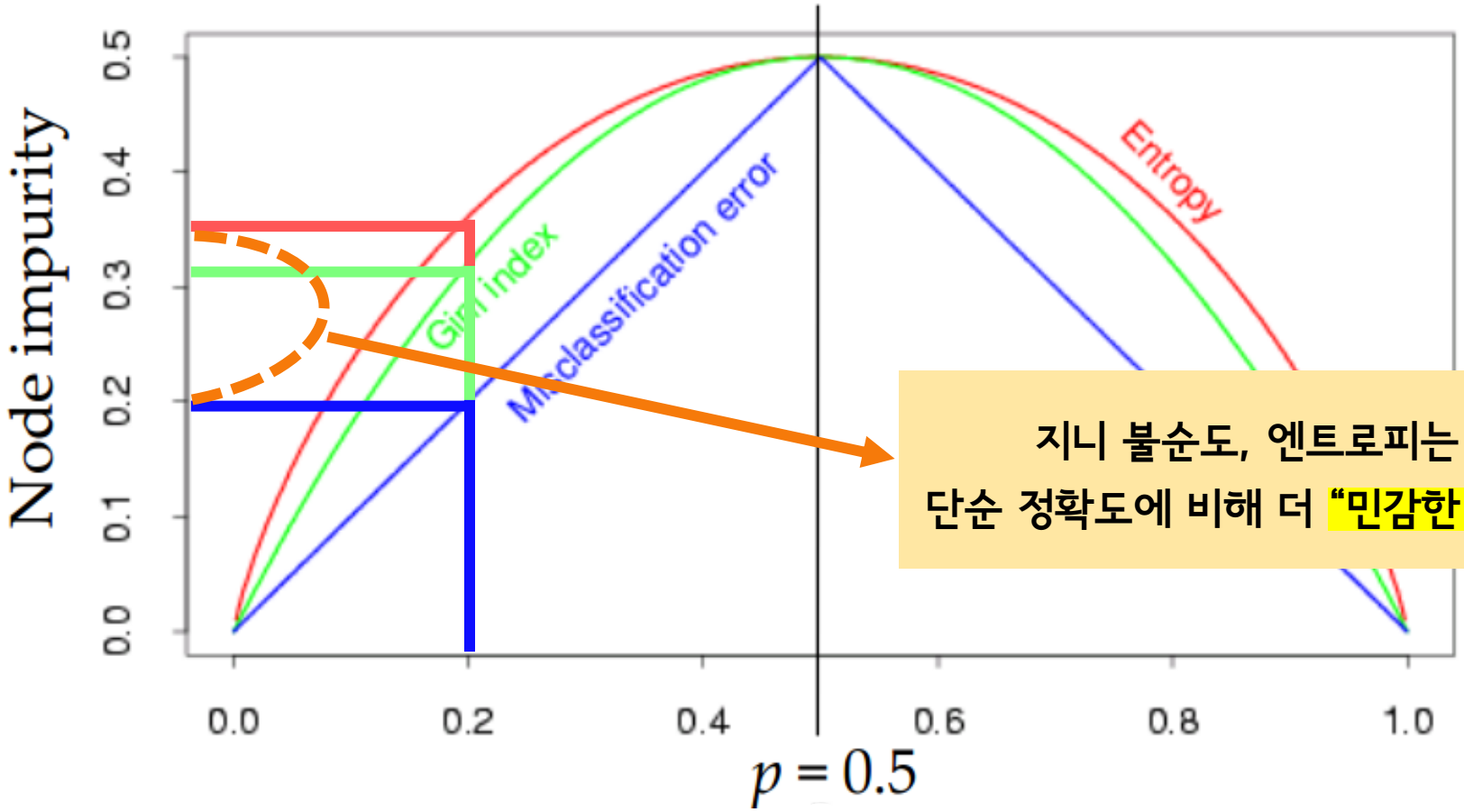
"모든 데이터는 분할에 사용"

불순도 / 불확실성

이 모델은 각 노드의 순도가 증가하는 방향
(불순도/불확실성이 감소하는 방향) 으로 학습을 진행한다.

단순 정확성 (accuracy) 보다는
"지니 불순도"사용!

불순도 / 불확실성

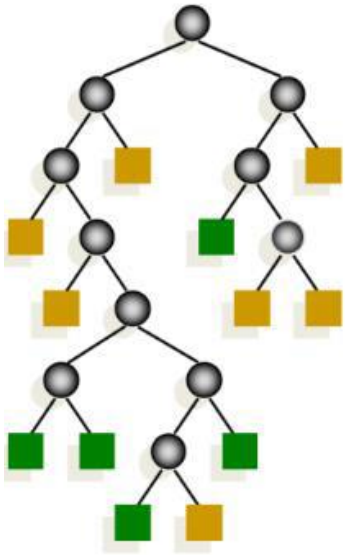


지니 불순도, 엔트로피는
단순 정확도에 비해 더 **"민감한 지표"**

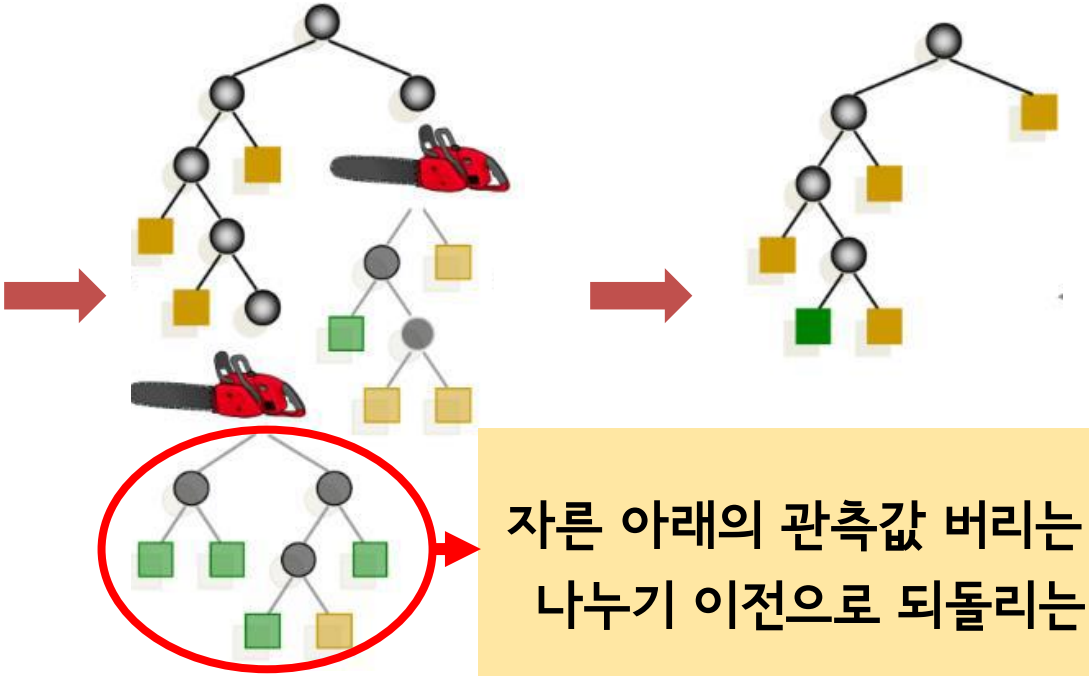
Entropy 부록참조

가지치기 (pruning)

트리를 최대로 나누기!

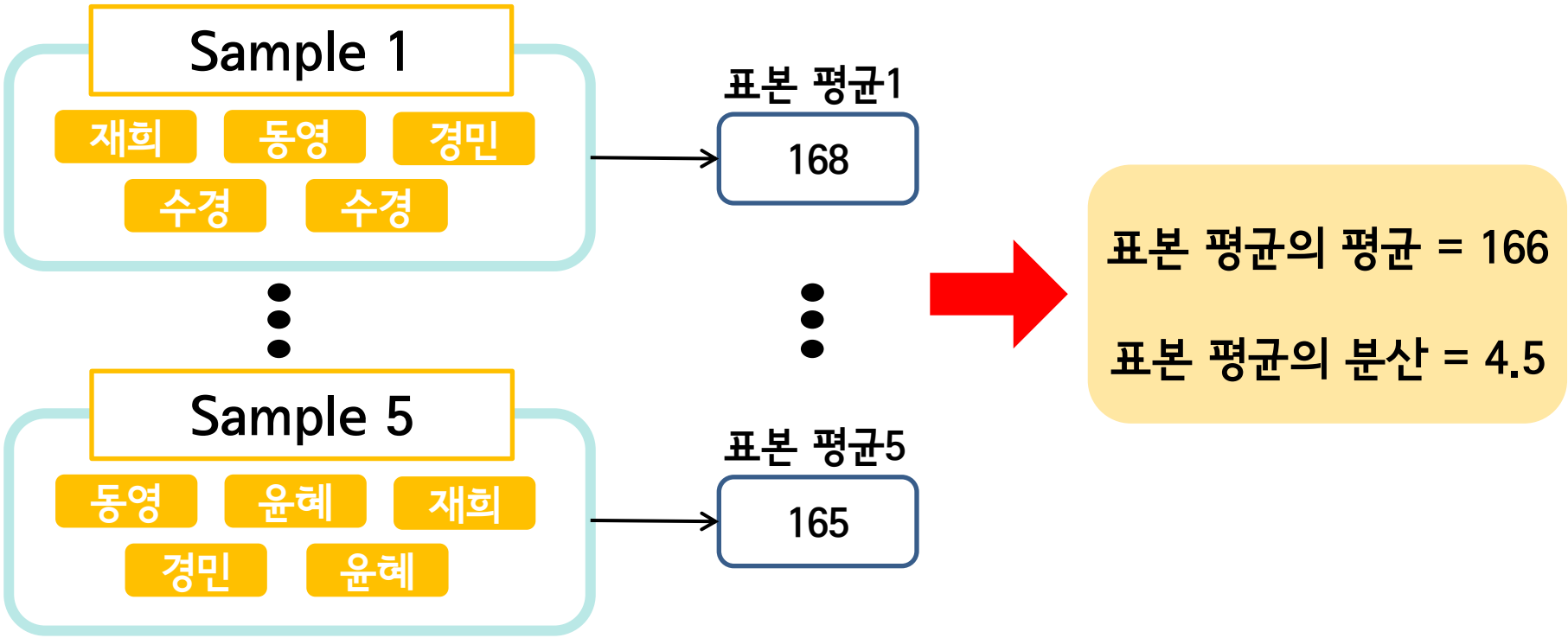


순도가 높아지는 방향으로 트리 자르기!



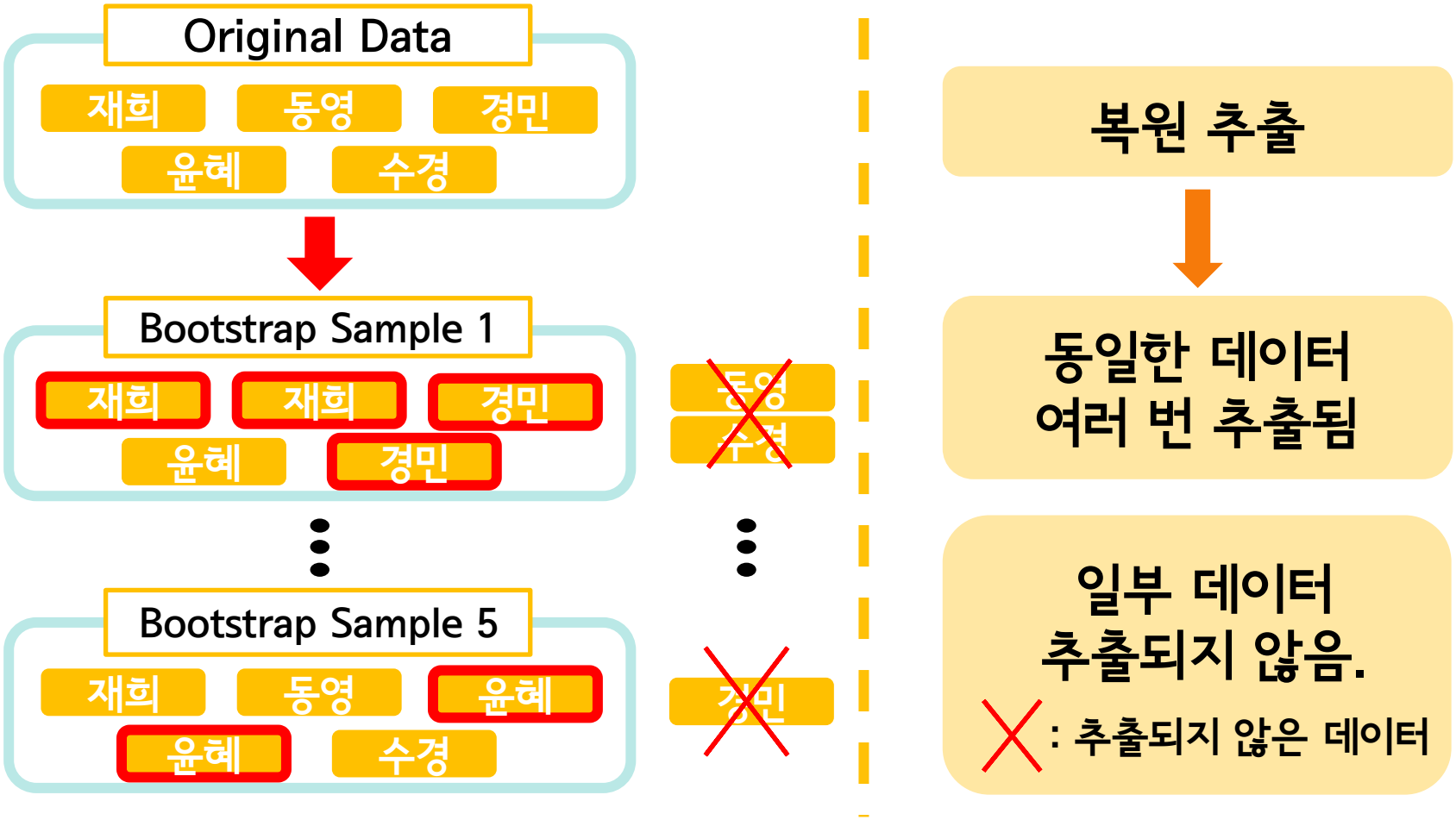
자른 아래의 관측값 버리는 것 X
나누기 이전으로 되돌리는 것!

Resampling



여러 개의 샘플을 추출하면 추리 통계량을 알아낼 수 있다!

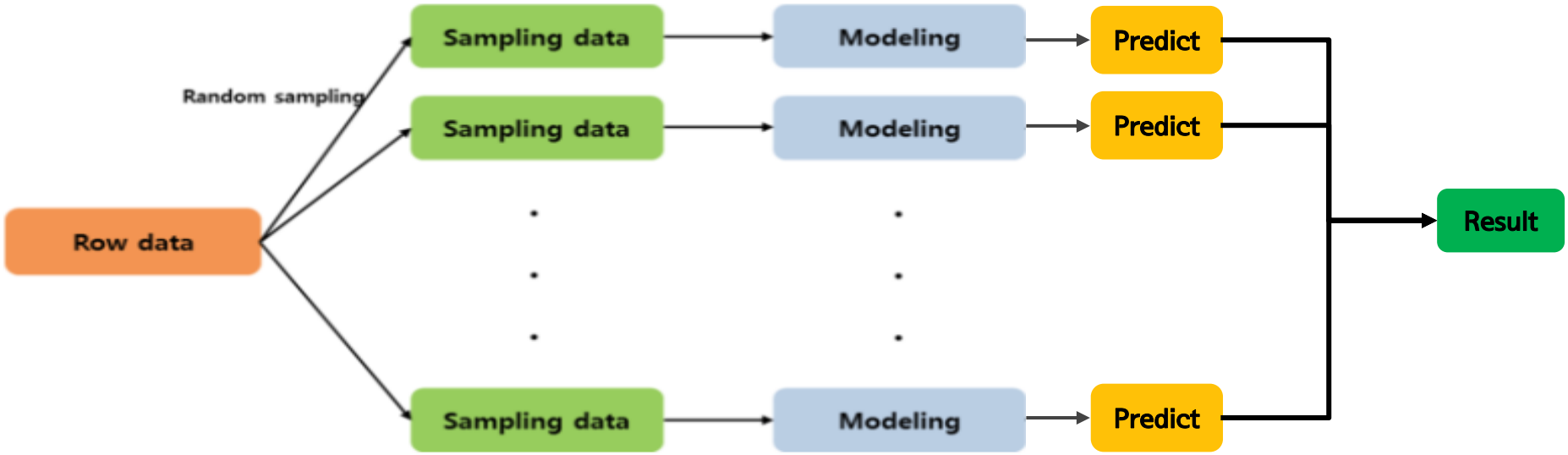
Bootstrap


Original Data
재희
동영
경민
윤혜
수경
Bootstrap Sample 1
재희
재희
경민
윤혜
경민
동영
수경
Bootstrap Sample 5
재희
동영
윤혜
윤혜
수경
경민
복원 추출
동일한 데이터
여러 번 추출됨
일부 데이터
추출되지 않음.
: 추출되지 않은 데이터

Bagging

Bagging

Bagging = Boostrap + Aggregating



Bootstrap으로 얻은 표본에 대해 모델링한 결과를 합침

Bagging

Bagging 예측값 산출 방법

회귀 문제

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

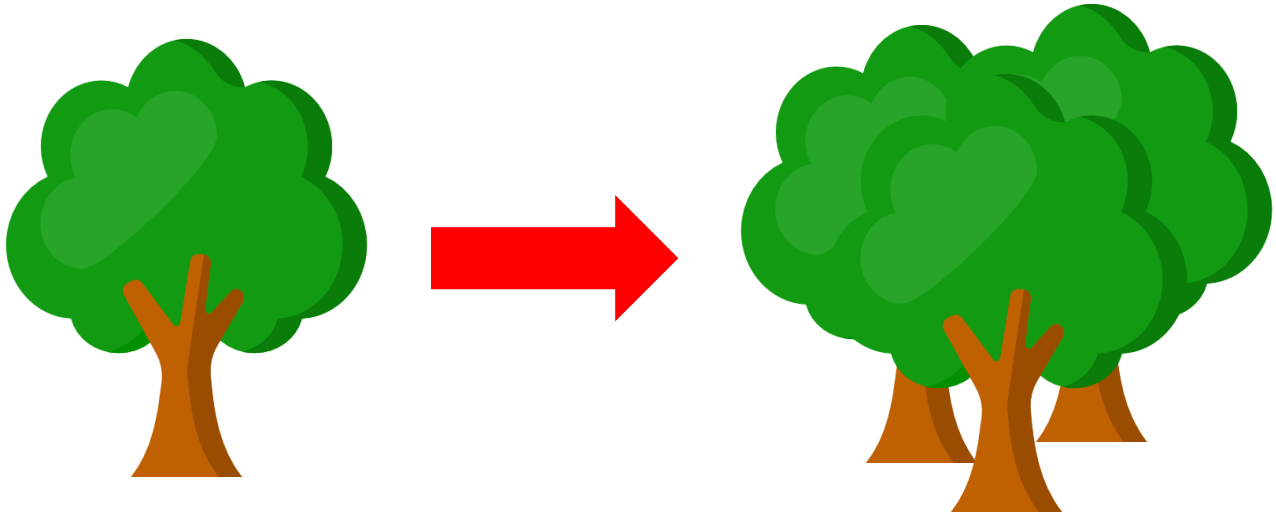
B개 트리의 예측값의 **평균**

분류 문제

$$\arg\max_{k} \sum_{b=1}^{B} I(\hat{f}^{*b}(x) = k)$$

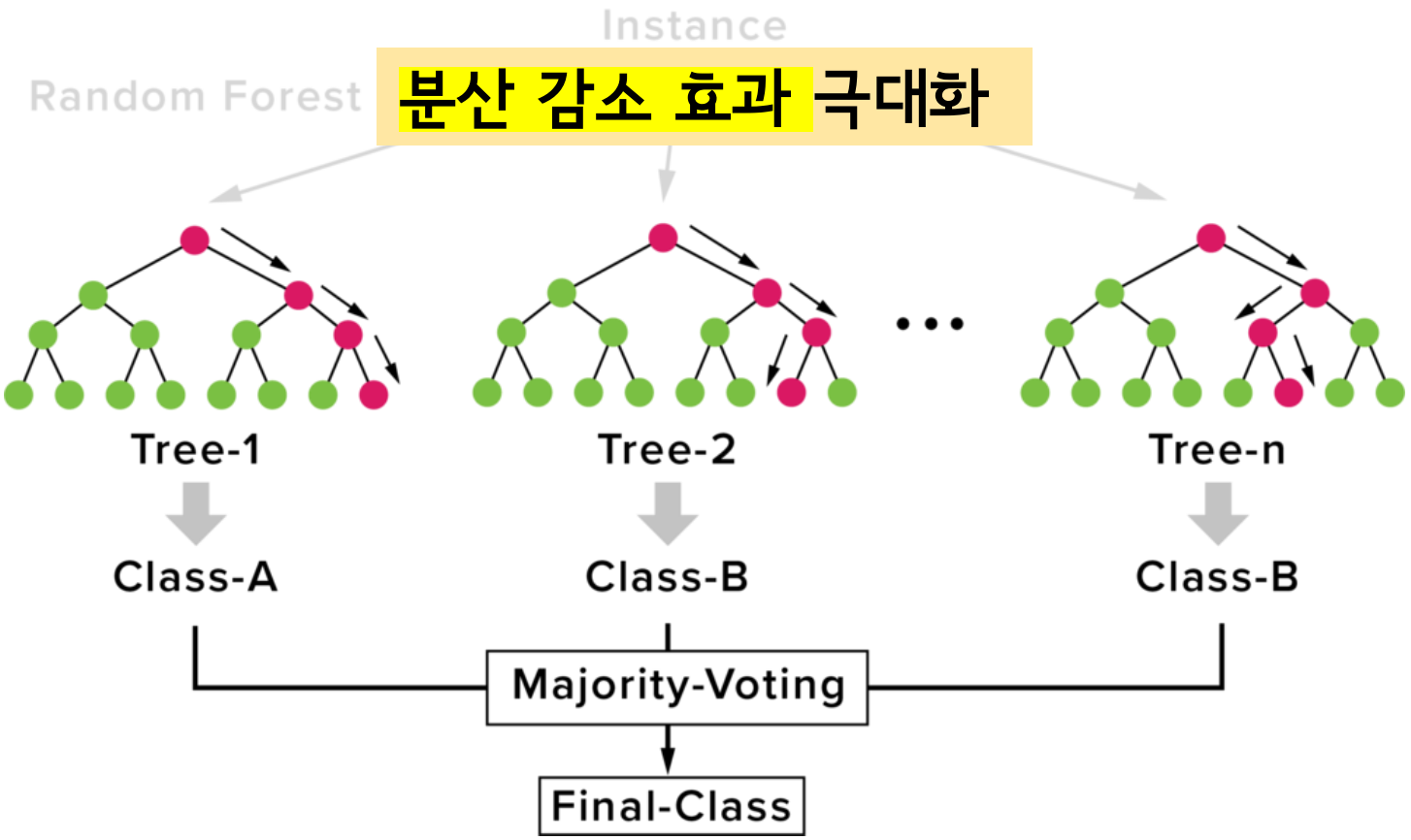
B개 트리의 예측값 중
다수결로 가장 많이 나온 값

Bagging

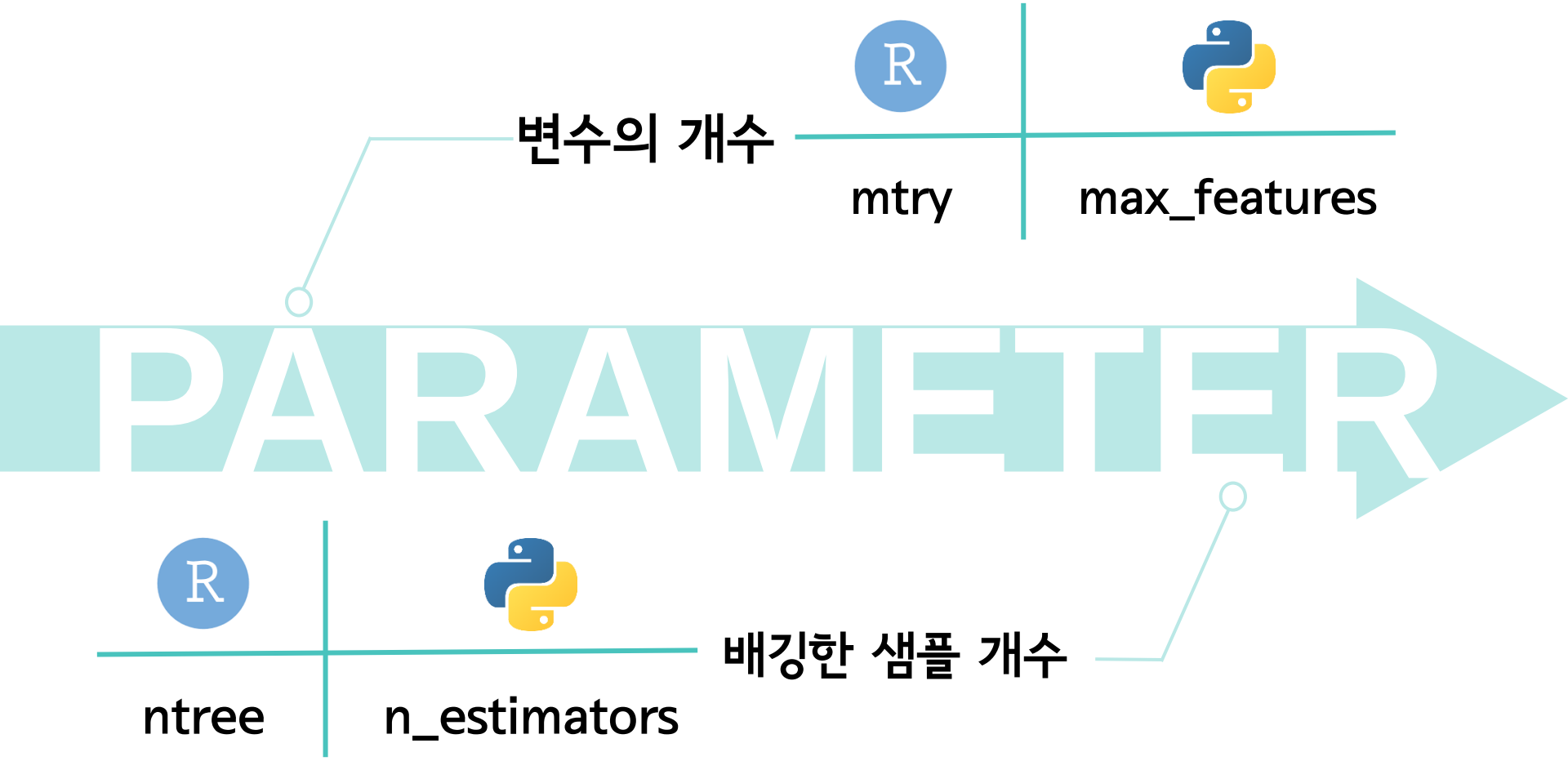


Bagging 결과 예측력 UP, 해석력 DOWN 됨

Random Forest



Random Forest Parameter



평가지표

회귀 모델에서 사용되는 대표적 평가지표

MSE
Mean Square Error

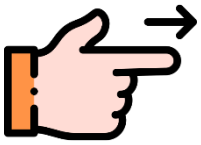
MAE
Mean Absolute Error

평가지표

분류 모델에서 사용되는 대표적 평가지표

Accuracy

f1 score



범주팀의 클린업을 기대해주세요~!