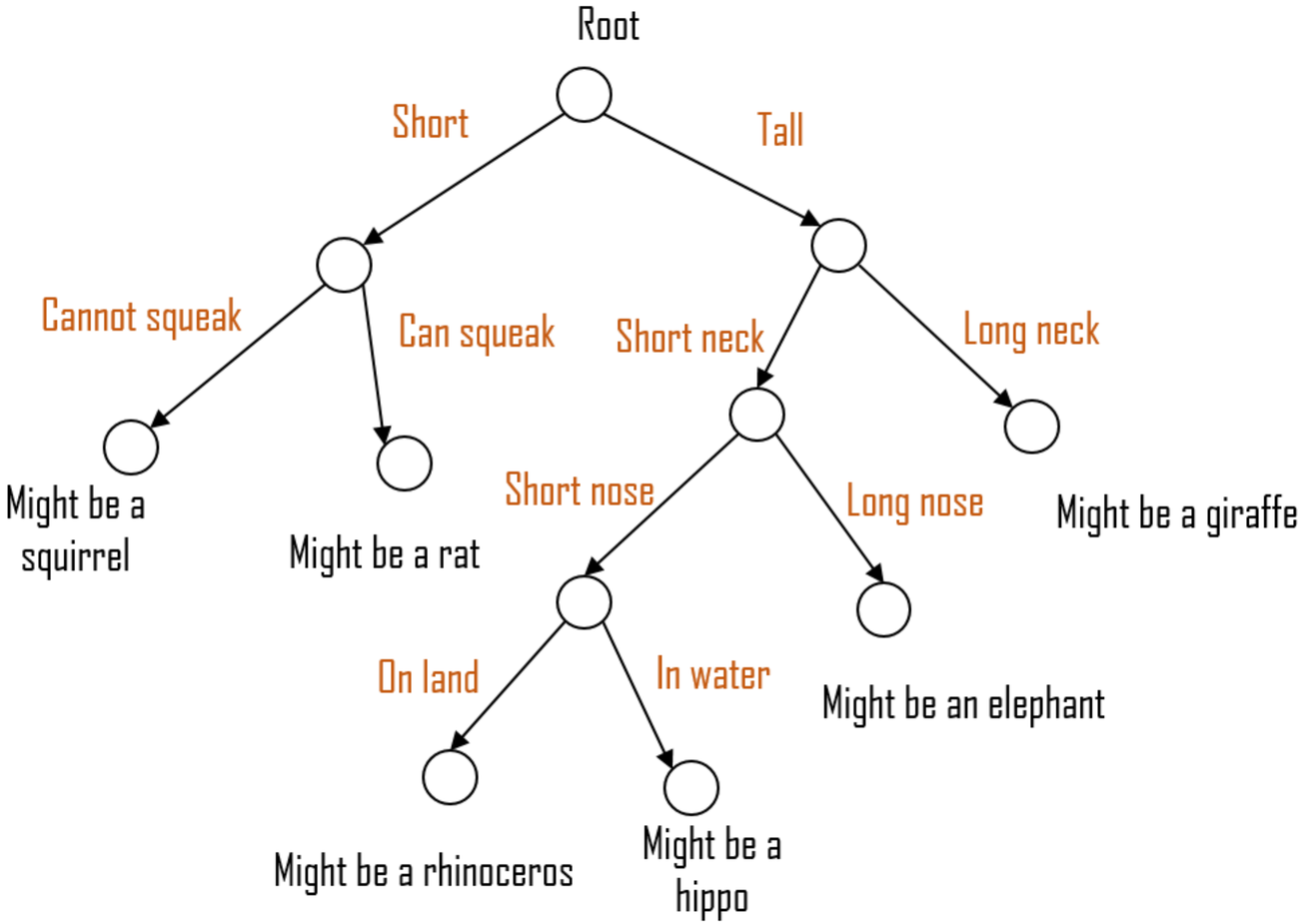
# Random Forest

데이터 분석 모델링반(ML1)

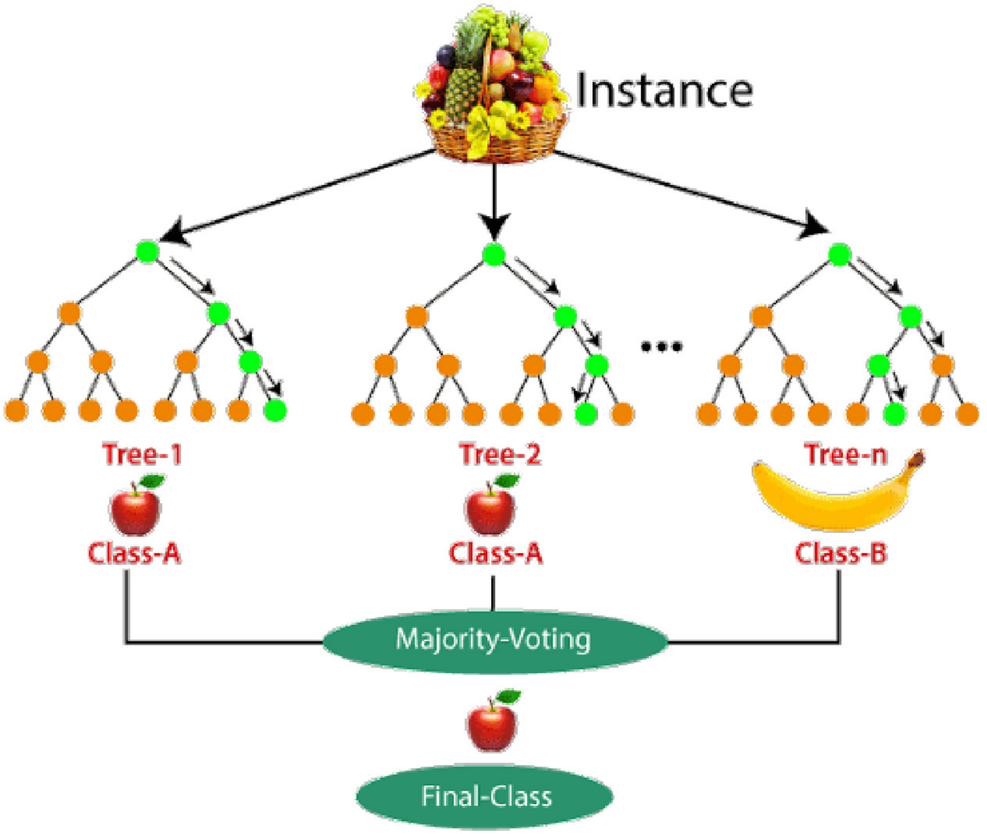
# 기존 Decision Tree




기존 DT
단일모델 : 하나의 트리구조, 규칙기반

기존 DT의 변동성이 큰 문제들? -> 보완하기 위해

앙상블, 배깅, 부트스트랩



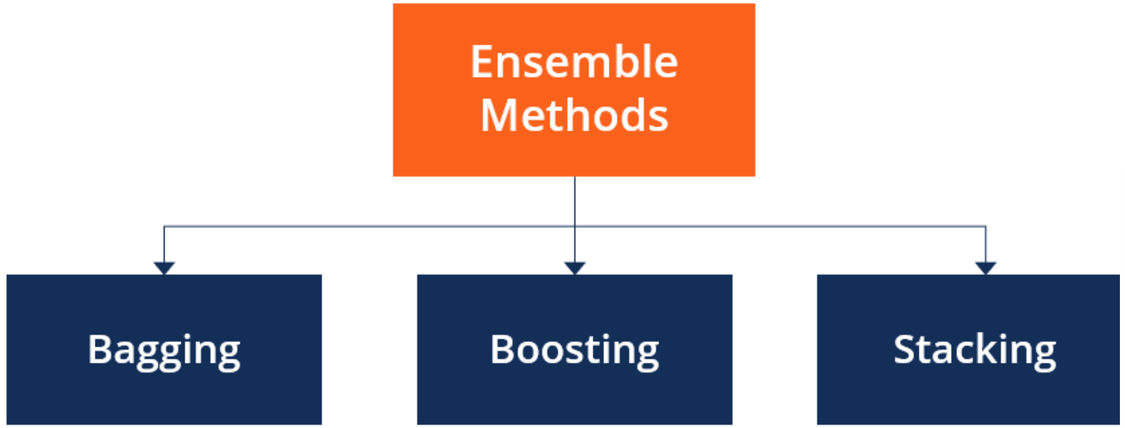

여러 개의 결정트리 결합하여 예측 성능 향상
분류(Classification), 회귀(Regression) 모두 사용 가능

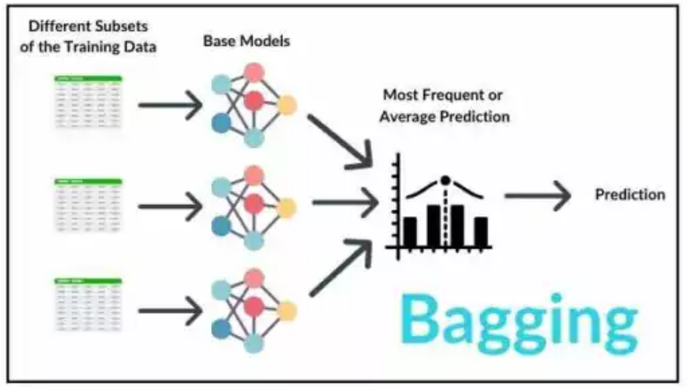
# 앙상블 학습 (Ensemble Learning)

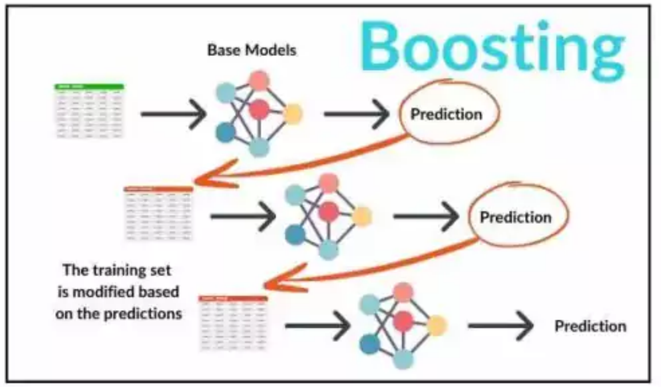
여러 개 모델을 사용할 수 있고
여러 개 데이터셋을 나눠서 사용할 수도 있다.

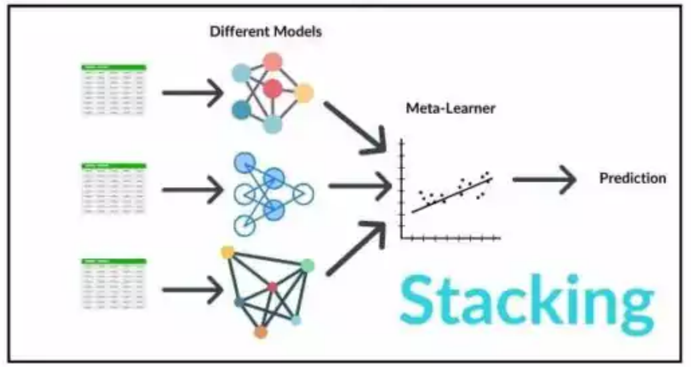
여러 개의 모델, 데이터를 통해서 결합해서 만들어지는 과정을 앙상블

배깅은 동일한 알고리즘 사용하지만-> 다양한 데이터의 샘플을 통해 학습 시킨 후 예측을 평균화, 다수결 투표를 통해 최종 예측



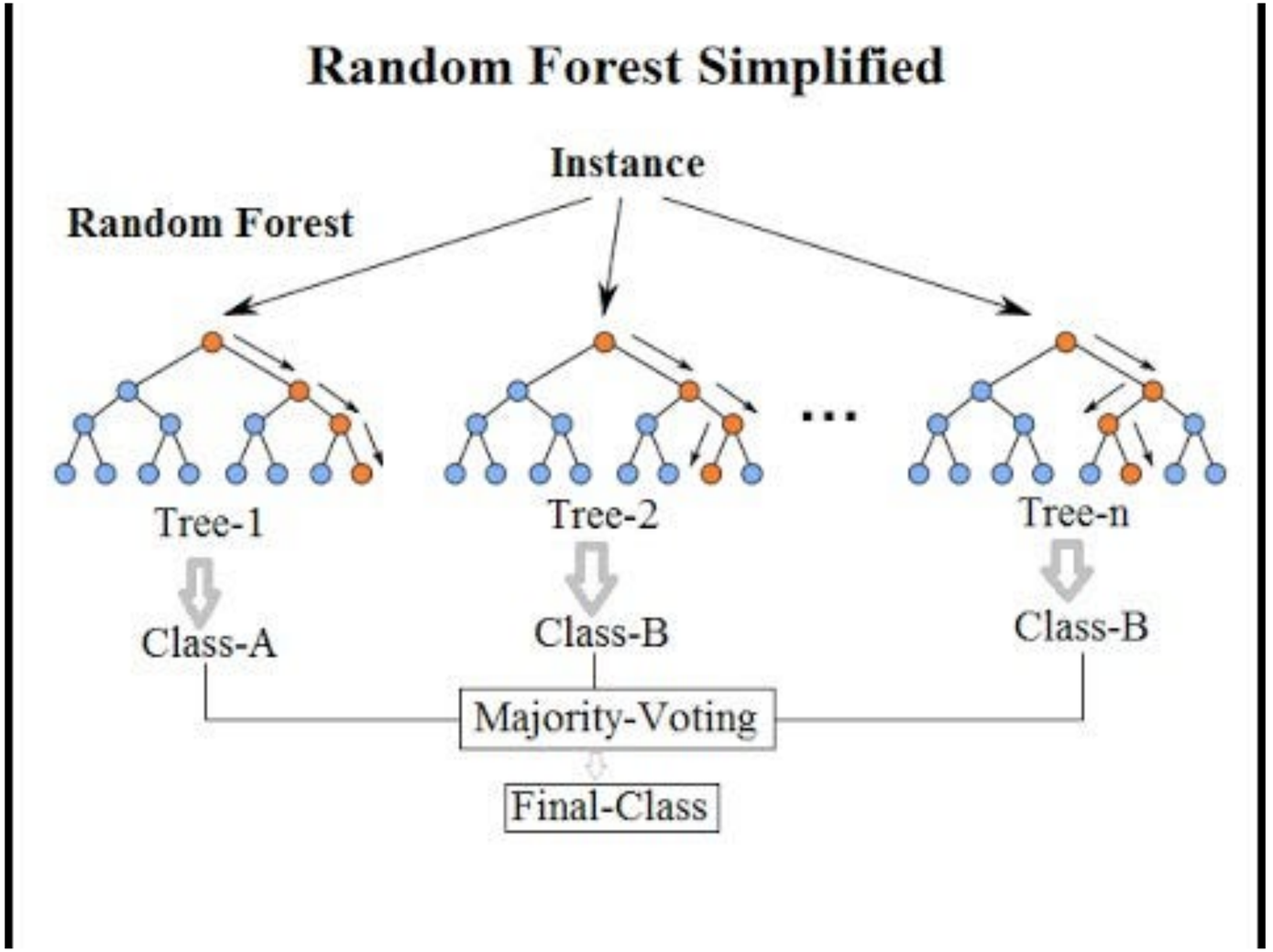



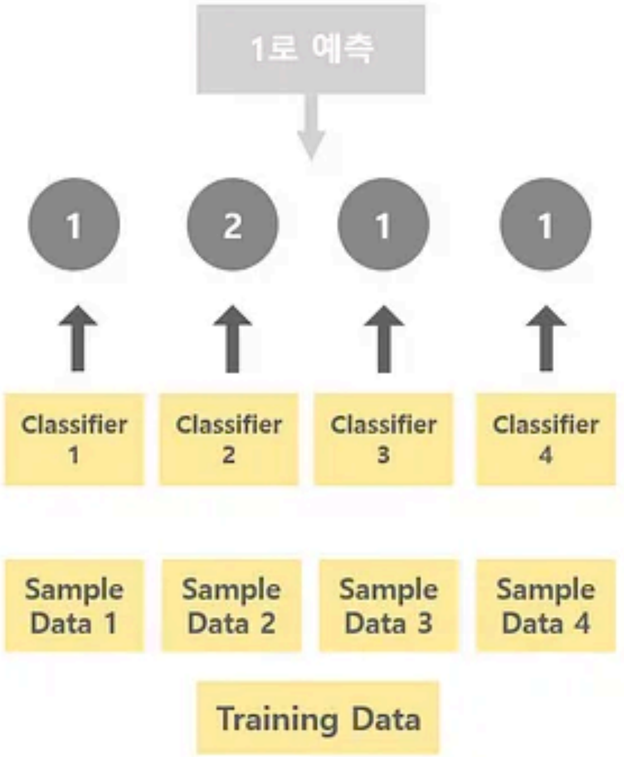




앙상블 학습은 여러 개의 모델을 결합하여 하나의 예측 모델을 만드는 방법

배깅 (Bagging, Bootstrap Aggregating)

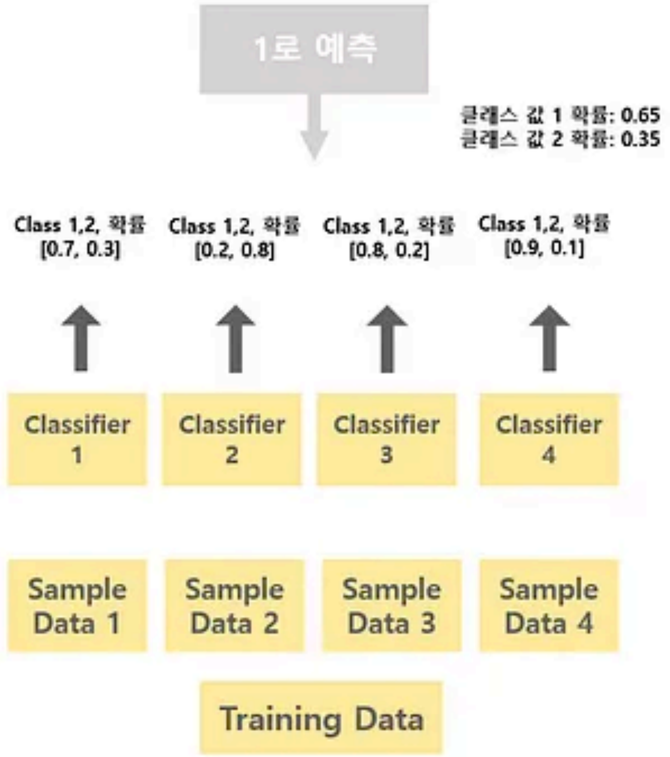


다수결로 인해서 최종 예측은 레이블 값 1로 결정



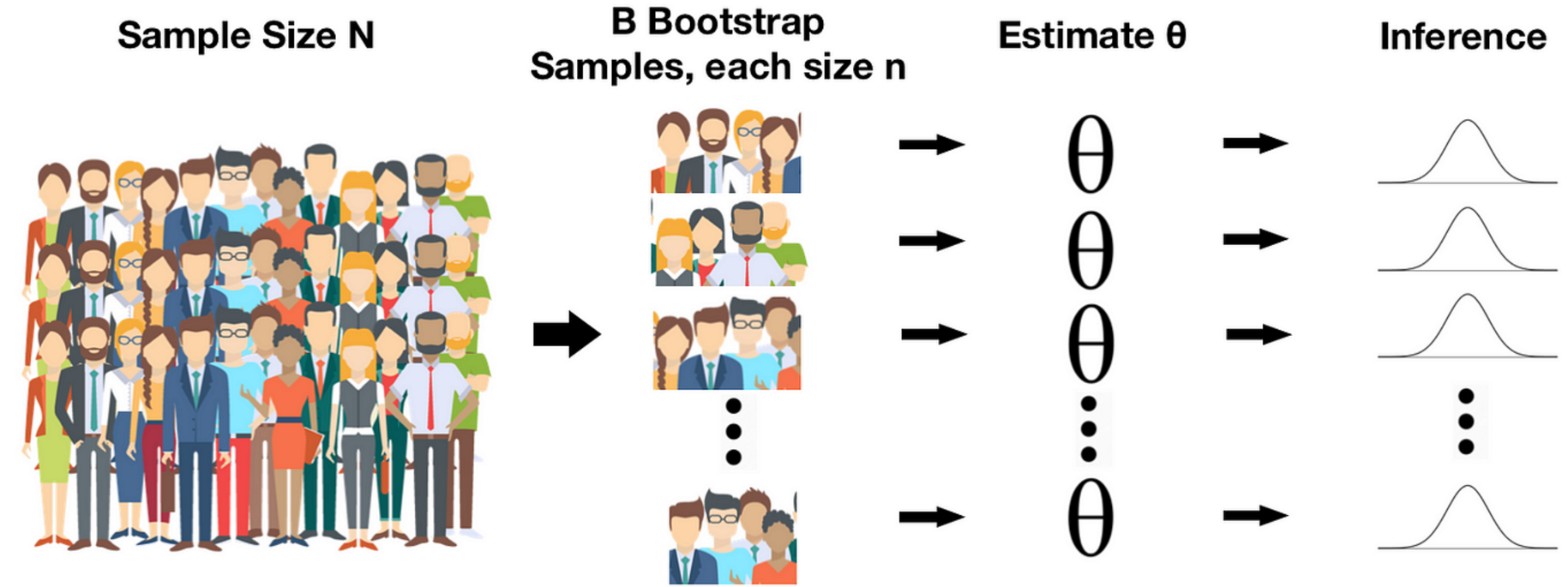

(하드 보팅)

Classifier 들의 class 확률을 평균을 취해서 결정




(소프트 보팅)

부트스트랩 샘플링(Bootstrap Sampling)



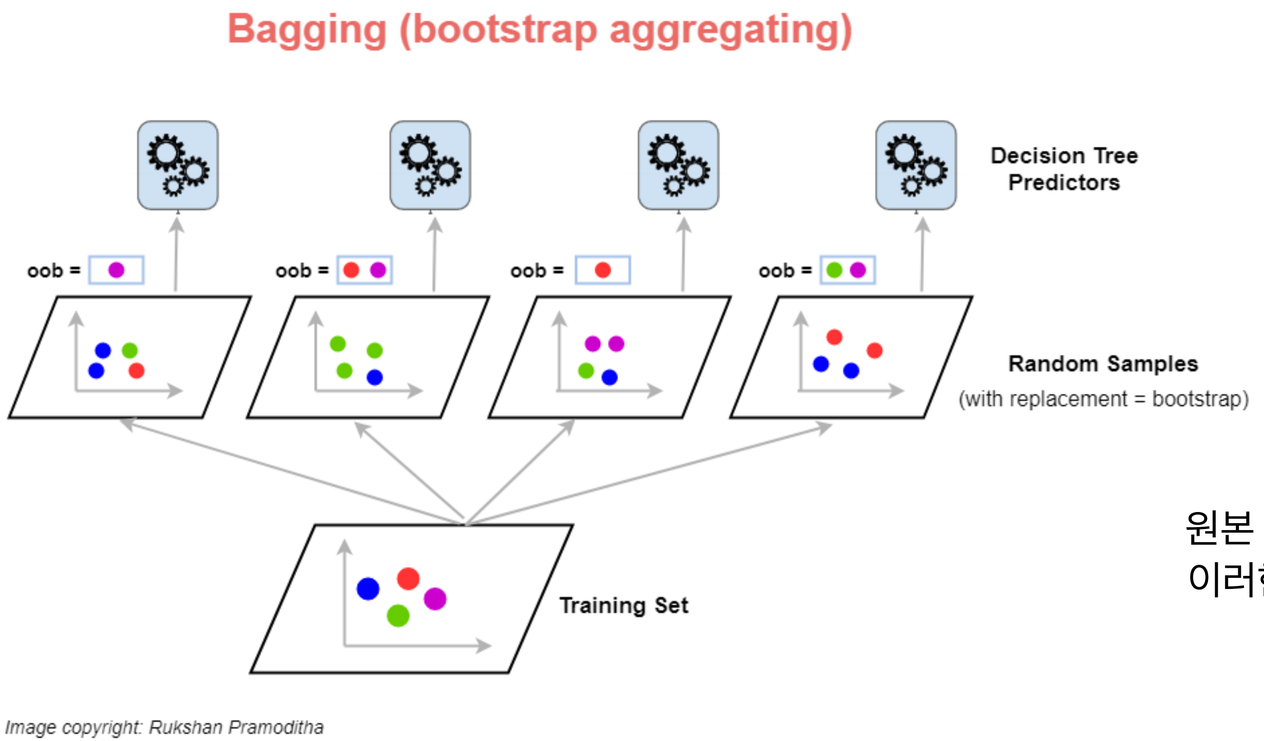

신뢰성을 높이기 위해 자주 사용하는 기법, 원본 데이터 셋에서 무작위로 데이터 선택해 새로운 샘플을 여러 번 생성하는 방법
이러한 과정에서 중복을 허용한다는 점이 특징

Why 부트스트랩?

부트스트랩의 샘플링 역할
신뢰구간 및 표준 오차 추정 : 무작위 샘플링을 통해 각 샘플에 대해 관심 있는 통계량 계산, 이를 기반으로 신뢰구간 표준오차 추정

데이터 분포 추정 : 부트스트랩 샘플링은 데이터의 분포를 직접적으로 가정하지 않기 때문에, 비정형적인 경우도 가능

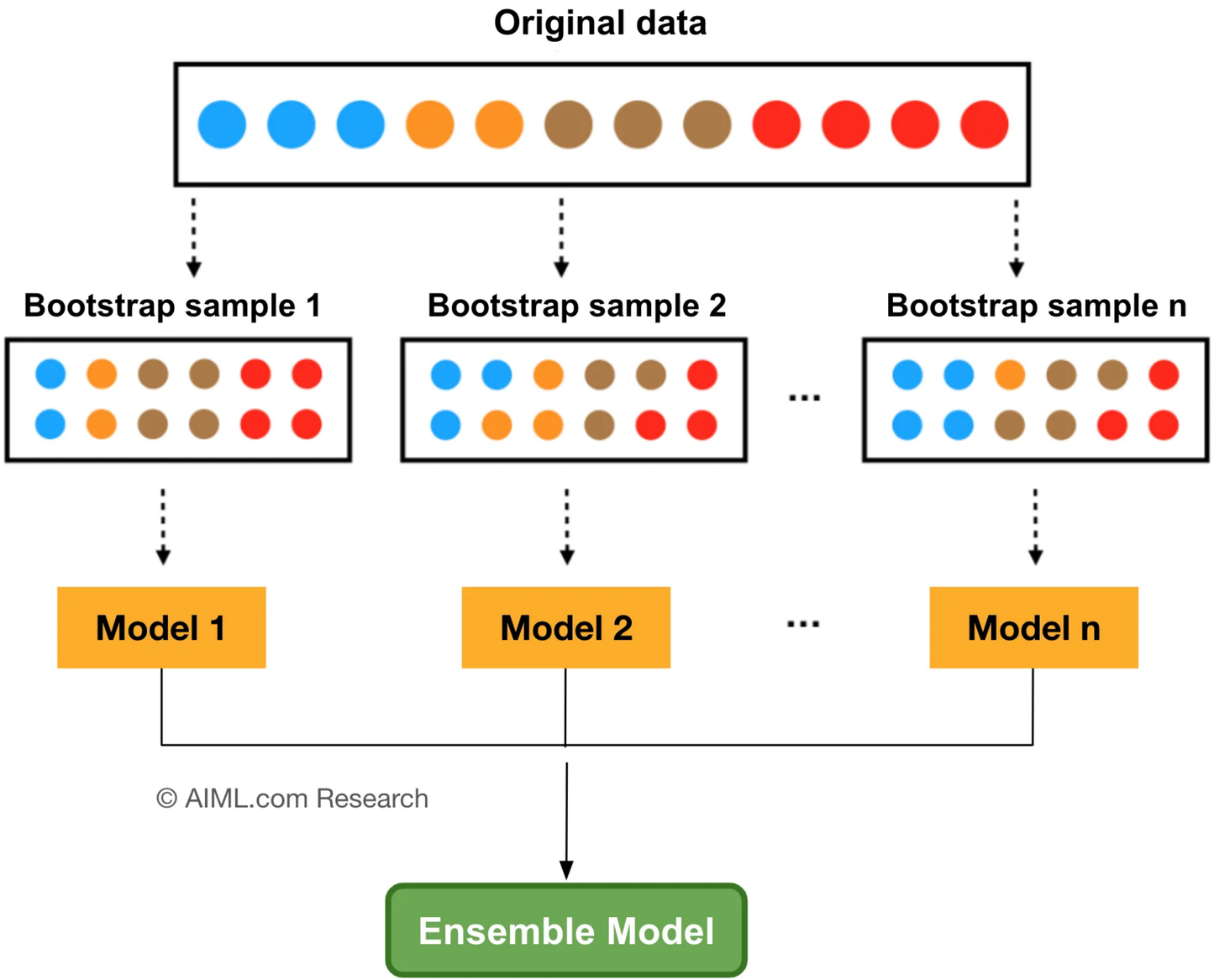
표본의 크기가 작을 때 활용 : 표본 크기가 작아도 부트스트랩 샘플링을 통해 여러 번의 샘플링을 수행할 수 있다.




OOB? ( Out - of Bag ) 샘플 평가

원본 데이터셋의 일부 데이터 포인트는 선택되지 않고 샘플에 포함되지 않음
이러한 데이터 포인트를 OOB샘플,, 모델 학습하지 않았으므로 평가에 사용

# Bagging




**Bootstrapping**

(Sampling with replacement)

**Model Training**

(Different models such as Decision Trees, SVM, Logistic can be used)

**Aggregation**

(Classification: Majority vote is taken
Regression: Average of outputs)

**Bagging**

# Bagging vs Boosting