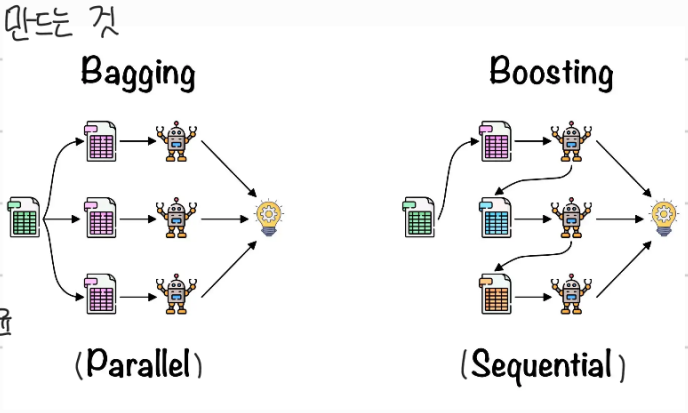
Ensemble ⇒ 각 모델의 장점을 합쳐서 예측하면 좋지 않을까?

- 목표: weak learner를 잘 조합하여 strong learner를 만드는 것
- 목적: 분산↓ & 편차↓ ┌ Bagging (병렬적)
  └ Boosting (직렬적)
- 왜 사용하는가? ① $E_{Ensemble} \leq E_{Avg}$
  앙상블 모델의 에러 평균 / 여러 개별 모델 사용 후 에러 평균
  ② Law of Large Numbers



- 분산과 편차에 따른 모델 복잡도

		예) Training set	Test set	
variation이 크다 = overfitting	"모델 복잡도↓"	99%	80%	~> data 개수↑
bias가 크다 = underfitting	"모델 복잡도↑"	80%	79%	~> 네트워크 복잡도↑

⇒ bias와 variation이 모두 낮아야 좋은 모델

* 앙상블 분산, 앙상블은 여러 개의 모델을 결합하여 하나의 예측을 만들어내는데 이때 각 개별 모델의 예측들이 얼마나 다른지를 나타내는 개념

* 앙상블 편차, 앙상블 모델의 예측과 실제 값 사이의 차이

- 배깅과 부스팅모델 각각의 개념과 차이점

Bagging : bootstrap 학습 후 결과물을 aggregation 하는 방법
- Majority voting (hard) "다수결로"
- Weighted voting (soft) "Classifier들의 class 확률을 평균 취해서"
- Stacking

⇒ low bias, high variance에 잘 맞음 » Random Forest

Boosting : 이전 분류기의 학습 결과를 토대로 다음 분류기의 학습 데이터의 샘플 가중치를 조정해 학습을 진행하는 방법

⇒ low variance, high bias에 잘 맞음 » AdaBoost, GBM, XGBoost

< 차이점 >

- (배깅) 균일한 확률분포에 의해 훈련집합을 생성함 🔴
  (부스팅) 분류하기 어려운 훈련 집합 생성함 🔵
- (배깅) 과대적합에 강함 🔴
  (부스팅) 오답에 더 집중할 수 있기 때문에 높은 정확도를 가지지만 과대적합의 가능성이 있음 🔵
- (배깅) 특정영역에서 정확도 낮음 🔴
  (부스팅) 이상치, 결측치에 취약함 🔵
- 부스팅이 배깅보다 일반적으로 시간이 오래 걸림