② 배깅 (Bagging)

- 앞서 모형 결합에서 사용하는 독립적인 모형의 수가 많을 수록 "앙상블"이 좋아질 가능성이 높다는 것이었음.
  ↳ 하지만, 각각 과소모형을 사용하는데 한계가 있으므로 보통은 "배깅" 방법을 사용하여 같은 종류의 모형을 쓰지만, 서로 다른 결과를 출력하는 다수의 모형을 만든다. ⇒ what is bagging?

- "배깅" (bagging)은 i) 중복된 모형과 ii) 중복된 오차를 사용하는 피셔 부트스트래핑(Bootstrapping)과 유사하게 트레이닝 Data를 랜덤하게 선택해서 다수결 모형을 추정.

- Trainy Data를 선택하는 방법에 따라 아래와 같이 부른다.
  i) 같은 Sample Data를 중복하여 사용하지 X ⇒ Pasting
  ii) 같은 Sample Data를 중복하여 사용하면 ⇒ Bagging
  iii) 데이터가 아니라 "다차원 독립변수"들에 "일부 차원"을 선택하면 ⇒ Random Subspaces
  iv) Data Sample과 독립변수 차원 모두 일부만 random하게 사용하면 ⇒ Random Patches
  ↳ 앙상블 학습시에는 트레이닝 Data 뿐만 아니라 다른 Data를 사용할 수 있다.
  ↳ ⇒ OOB (out-of-Bag)

③ 랜덤 포레스트 (Random Forest)

↳ 의사결정 모형(Decision Tree)을 개별 모형으로 사용하는 모형 결합 방법을 말한다.
↳ 배깅과 마찬가지로 Data 샘플의 일부만 선택하여 사용한다.
↳ 하지만, 노드분리시 모든 독립 변수들을 비교하여 최선의 독립 변수를 선택하는 것이 아니라, 독립변수의 차원을 랜덤하게 감소시켜 그 중에서 독립변수를 선택한다.

⇒ 이렇게 함으로써, 개별모형들 사이의 상관관계가 줄어들기 때문에, 모형 성능의 변동이 감소하는 효과가 있다.

↳ 이러한 방법을 극단적으로 이용한 것이 "Extremely Randomized Trees" 모형
↳ 각 노드에서 랜덤하게 독립변수를 선택한다.

↳ 랜덤 포레스트의 "장점"은 각 독립변수의 중요도 (feature importance)를 계산할 수 있다.
⇒ 포레스트 안에서 사용된 모든 노드에 대해 어떤 독립변수를 사용했고 그 노드에서 얻은 정보량 (Information Gain)을 구할 수 있으므로 각각의 독립변수들이 얻어낸 Information Gain의 평균을 비교하면, 어떤 독립변수가 중요한지 비교 가능.