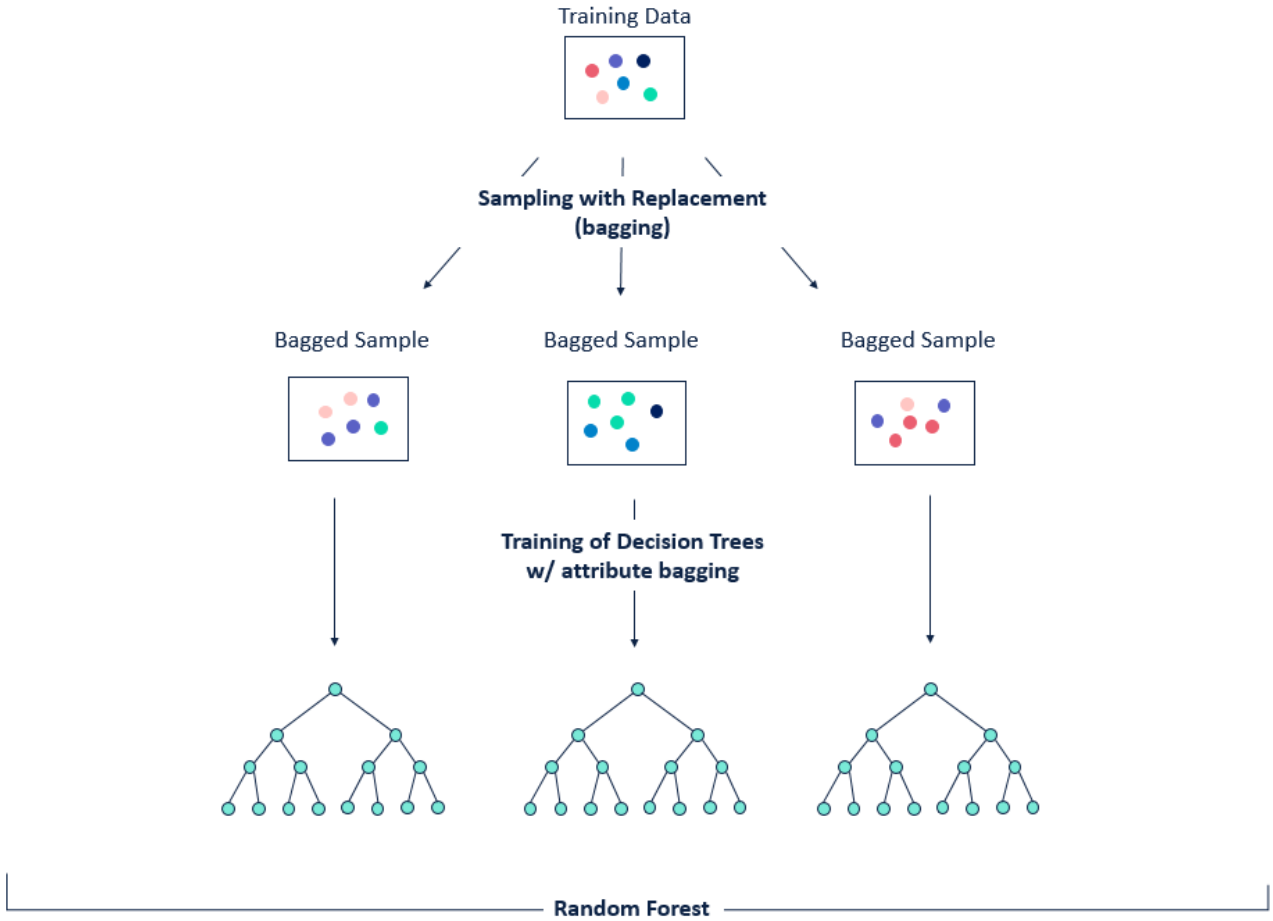
# RandomForest(랜덤 포레스트)

## 1.설명

의사 결정 나무(Decision Tree)가 모여 랜덤 포레스트를 구성한다. 이를 통해 의사결정 나무의 하나의 훈련 데이터에 오버 피팅되는 경향의 문제를 여러개의 의사결정 나무를 통해 해결하는 것이다.




출처 : (https://injo.tistory.com/30)

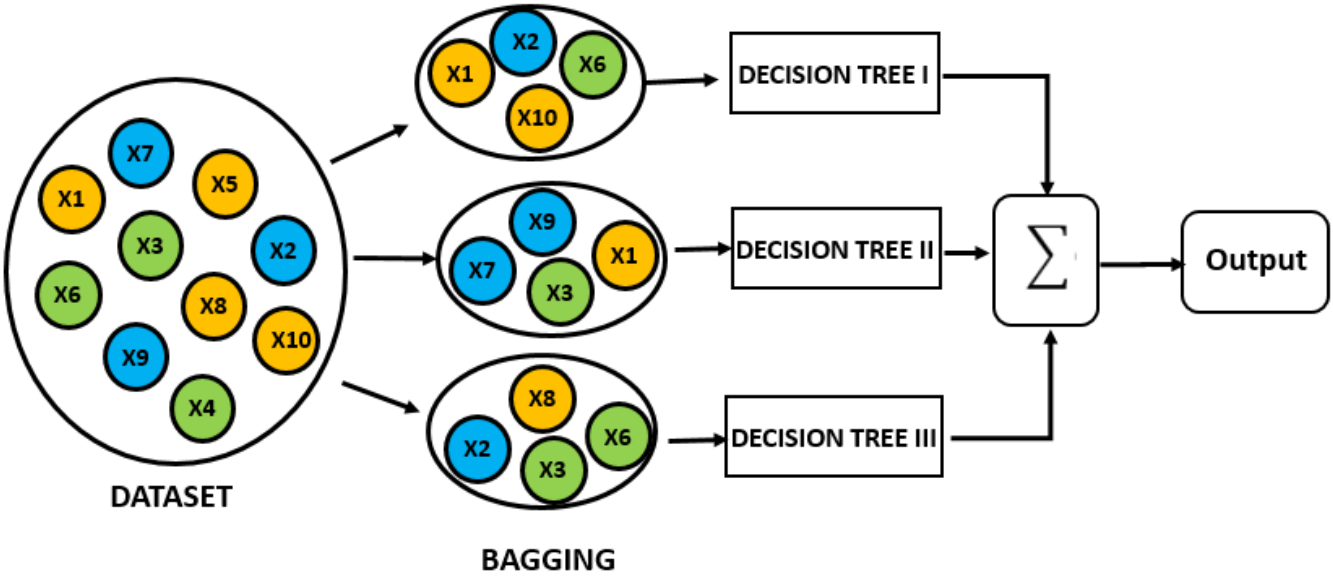
여러 개의 의사결정 나무를 형성하고 새로운 데이터 포인트를 각 트리에 통과시키며, 각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류 결과로 선택한다. 랜덤 포레스트가 생성한 일부 트리는 overfitting될 수 있지만, 많은 수의 트리를 생성함으로써 overfitting이 예측하는데 있어 큰 영향을 미치지 못하도록 예방한다.

---

## 2.특징

### 2-1.Bagging

Bagging은 트리를 만들 때 training set의 부분집합을 활용하여 형성하는 것을 말한다. 예를 들어 training set에 1000개의 데이터가 있다면, 각 트리를 생성할 때, 100개의 데이터만 임의로 선택하여 트리를 만드는 데 활용할 수 있다. 즉 **모든 트리는 각기 다른 데이터를 바탕으로 형성되지만, 모두 training set의 부분 집합이다.**




출처 : (https://eunsukimme.github.io/ml/2019/11/26/Random-Forest/)

데이터를 임의로 선택할 때 중요한 것 중복을 허용한다는 것이다.(with replacement) 위의 그림을 보면 X2가 1,3 의사결정 나무에 중복되어 선택되었다. 중복을 허용함으로써 training set에서 100개만 뽑기보다 1000개씩 매번 뽑아도 unique한 데이터 셋을 형성할 수 있으므로 n개의 training set이 있다면, 임의로 n개의 데이터를 중복 허용하여 선택함으로써 각 트리를 형성한다(부트스트래핑-bootstraping).


< 부트스트래핑 샘플링 방식 >

출처 : (https://kimdingko-world.tistory.com/180)

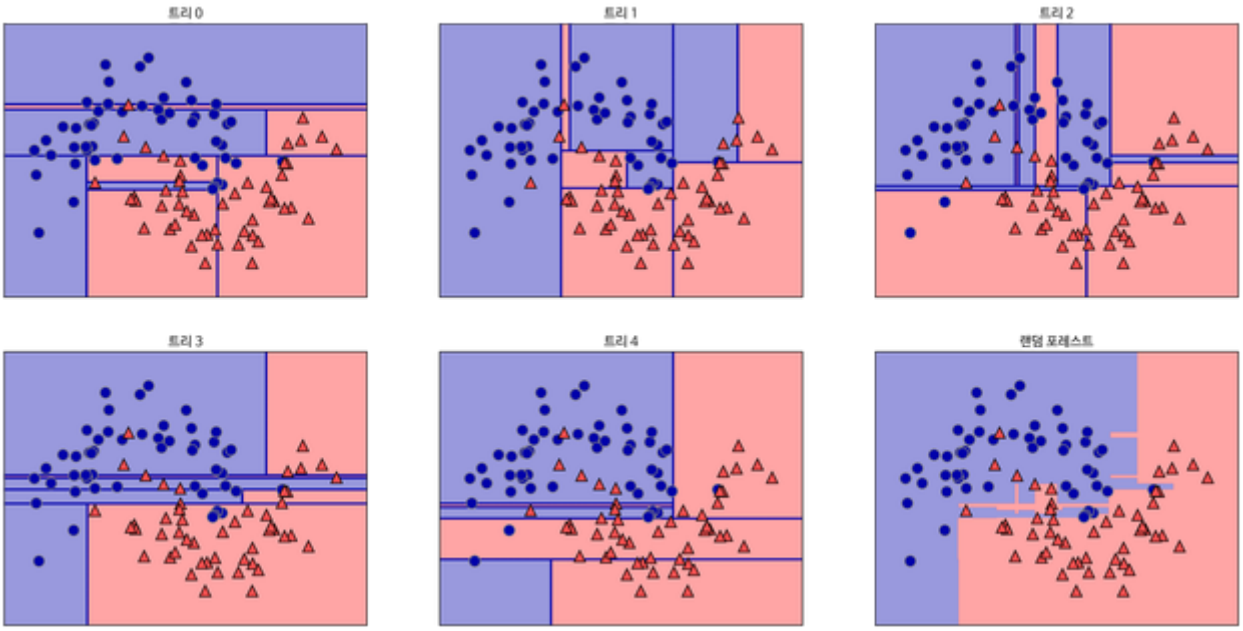
## 2-2.Bagging Feature

트리를 형성할 때 데이터 셋에만 변화를 주는 것이 아니라, feature 선택하는 데 있어서도 변화를 줄수 있다. Feature의 부분집합을 활용할 수 있다.

> 예시) 자동차 등급 분류
> Feature : 가격, 유지 보수 비용, 문의 갯수, 탑승 인원 수, 트렁크 사이즈, 안전 등급
> bagging feature : (안전 등급, 문의 갯수, 가격), (트렁크 사이즈, 문의 갯수, 유지보수 비용), .....

이렇게 다양한 bagging feature를 트리가 만들어 질 때까지 반복한다. 일반적으로 M개의 feature가 존재할 때, 임의로 선택하는 feature의 수는 sqrt(M) 개이다.

> 예시) 25개의 features -> 5개의 feature 선택




출처 : (https://bkshin.tistory.com/entry/머신러닝-5-랜덤-포레스트Random-Forest와-앙상블Ensemble?category=1057680)

## 2-3 Parameters

- n_estimators : 랜덤 포레스트 안의 결정 트리 갯수.
  n_estimators가 클수록 좋다. 결정 트리가 많을 수록 더 깔끔한 Decision Boundary가 나오지만 메모리와 훈련시간이 증가한다.
- max_features : 무작위로 선택할 features 갯수.
  max_features = total features면 모든 features 사용해서 결정트리 만든다. bootstrap parameter가 false면 비복원 추출을 하여 전체 feature를 이용하여 트리를 만든다. 반면 bootstrap = true(default)면 전체 feature가 복원추출하여 트리를 만든다. max_feature값이 크면 랜덤 포레스트들이 매우 비슷해지고 가장 두드러진 특성에 맞게 예측한다. max_feature값이 작으면 랜덤 포레스트의 트리들의 모양이 서로 매우 달라지고 오버피팅이 줄어든다.(대개 true 사용)

## 2-4. 장단점

- 장점 : 의사 결정 트리의 쉽고 직관적인 장점을 그대로 가지고 있고, 앙상블 알고리즘 중 비교적 빠른 수행 속도를 가지고 있다. 다양한 분야에서 좋은 성능과 정확도가 나온다. 대용량 데이터 처리에 효과적이고 많은 입력 변수를 다룰 수 있다.간편하고 빠르다.

- 단점 : 하이퍼 파라미터가 많아 튜닝 시간이 많이 소요된다. 속도와 메모리의 비용이 상대적(linear에 비해)크다. (트리가 더 많을 경우 정확도는 높아지나 시간과 리소스 소모가 크다.),트리 깊이와 갯수 설정을

잘못하면 과적합 발생

## 적합한 사례

- 대용량의 임의의 매개 변수를 가진 데이터에도 적합. 빠른 연산, 높은 정확도
- 엑스박스 키넥트에서 30만장의 사진에서 2000개의 픽셀을 임의 추출하여 3개의 깊이를 가진 트리 20여개를 구성하여 이를 토대로 입력되는 사진을 5밀리초안에 배경을 제외하고 신체 트랙킹 가능
- CT에서 해부학 구조 검출,위치 파악
- MRI에서 악성 신경교종 검출