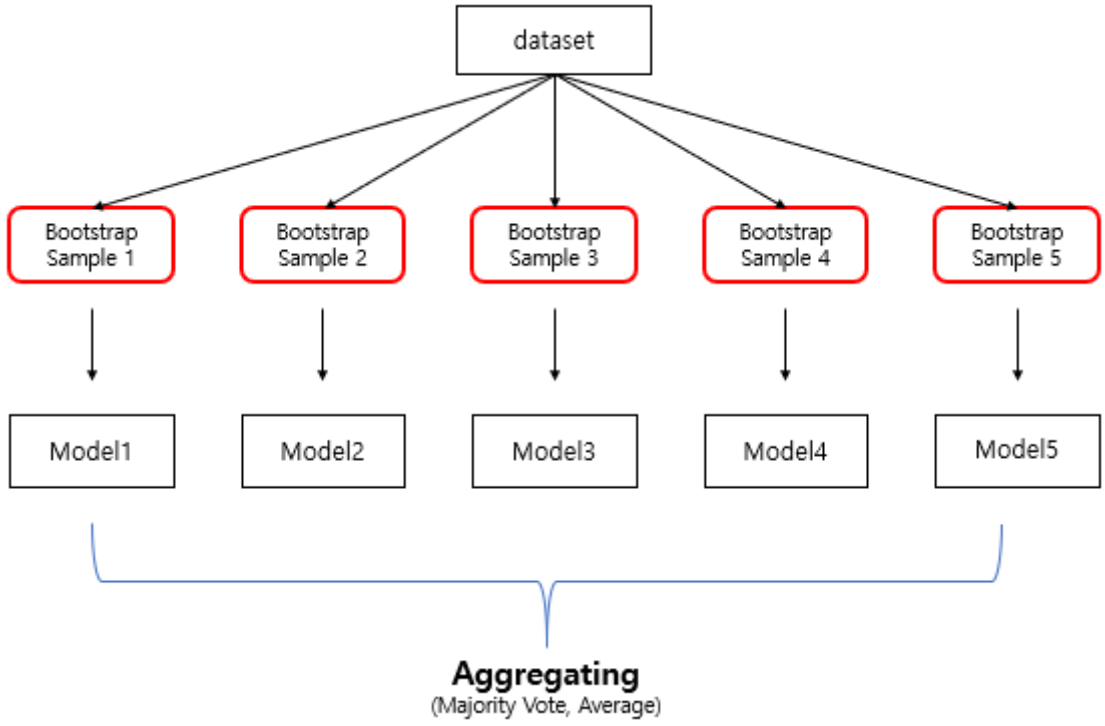
# Bagging vs voting sampling

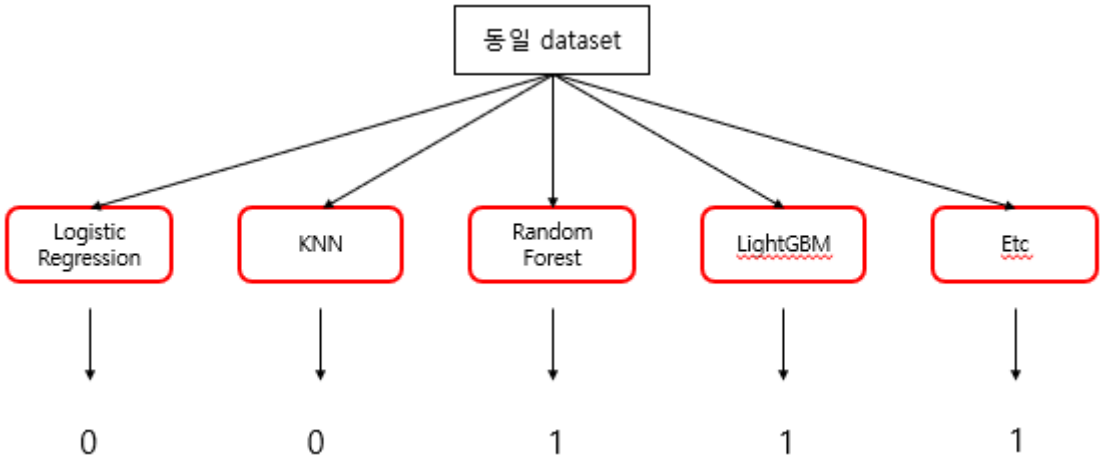
## Bagging

- bagging = bootsratp aggregating 의 줄임말로써, 주어진 데이터에 대해서 여러개의 붓스트랩(boostrap) 자료를 생성하고 각 붓스트랩 자료를 모델링 한 후 결합하여 최종의 예측 모형을 산출하는 방법!
- 붓스트랩 자료 : 단순 복원 임의 추출(random sampling)을 통해 원자료(raw data)로부터 크기가 동일한 여러 개의 표본 자료를 말한다



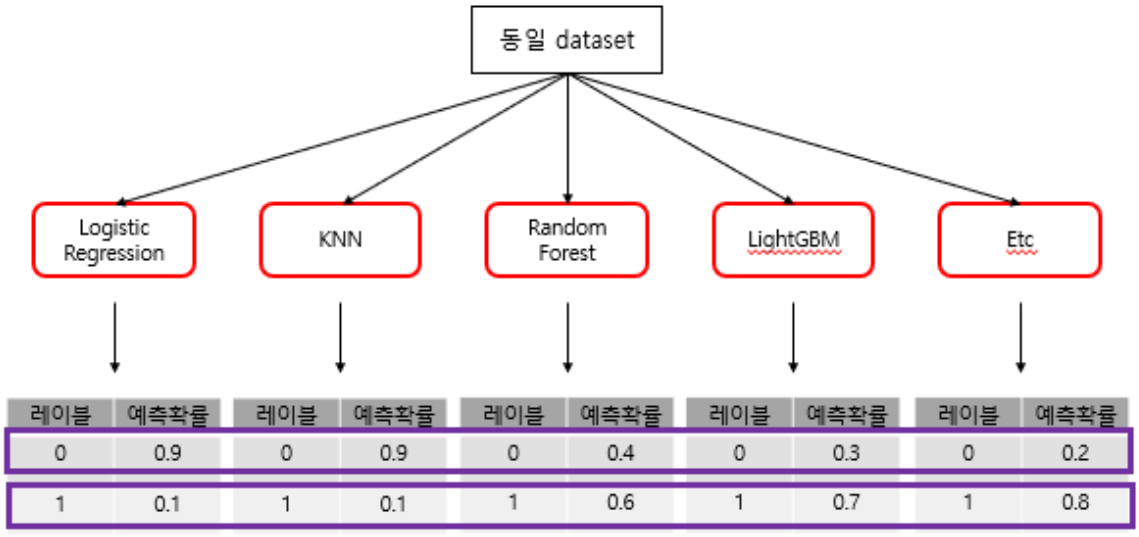

## Voting

- 여러 개의 분류기를 통해 예측하는것을 의미한다.




하드 보팅 : 위와 같이 예측을 진행하되 다수결의투표에 따라 분류하게된다.

Soft voting역시 그림으로 표현하자면 아래와 같다.



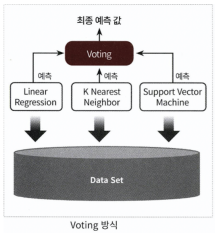

| 레이블 | 예측확률 | 레이블 | 예측확률 | 레이블 | 예측확률 | 레이블 | 예측확률 | 레이블 | 예측확률 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.9 | 0 | 0.9 | 0 | 0.4 | 0 | 0.3 | 0 | 0.2 |
| 1 | 0.1 | 1 | 0.1 | 1 | 0.6 | 1 | 0.7 | 1 | 0.8 |

레이블 0 예측확률 평균: 0.54
레이블 1 예측확률 평균: 0.46

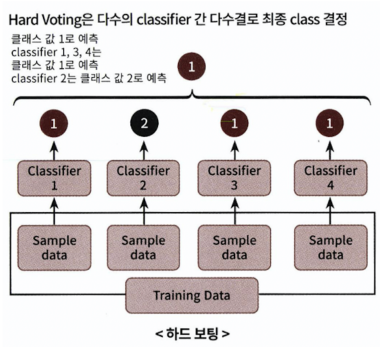
아래의 그림에서 하드 보팅이 마치 sampling되는 것처럼 오해를 불러일으킬수도 있는데 실제 저 박스의 의미는 전체 데이터가 들어간다고 이해를 하면 될것 같다.

## Q2

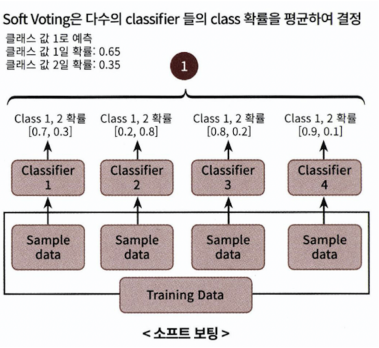
- 보팅도 배깅과 마찬가지로 샘플링 데이터를 다르게 가져가는가?



Voting 방식

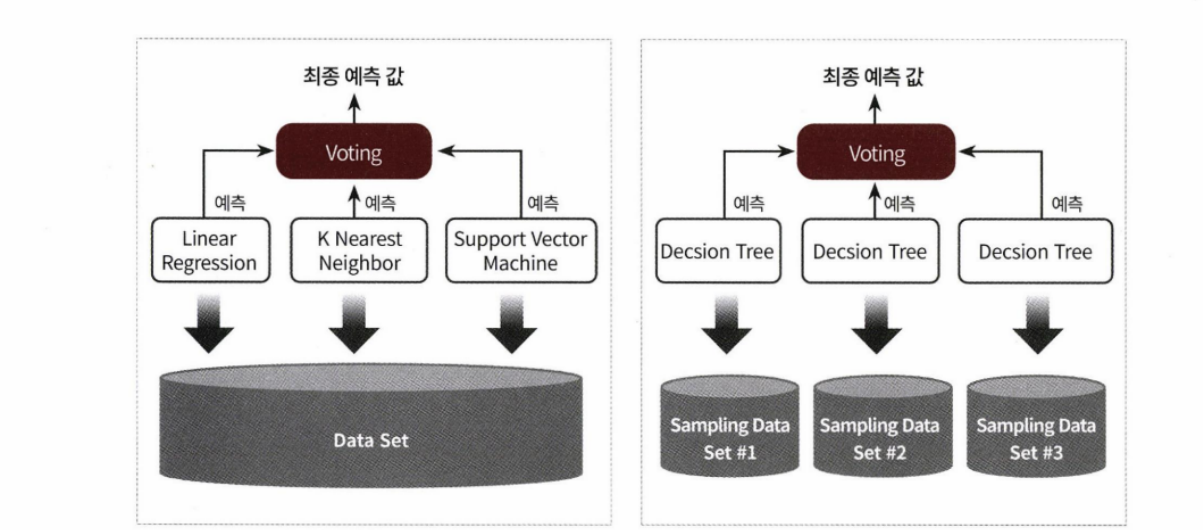


<하드 보팅>



<소프트 보팅>

만약 아래와 같이 하드보팅, 소프트 보팅이 전체 데이터에 대해 샘플링된 데이터라면 아래 책의 bagging 표현과 같이 #n의 방식으로 표현되었을 것이다.




ex) 실제로 sklearn의 votingclassifier의 파라미터에는 전체 데이터가 반복적으로 들어감을 확인할 수 있다. 또한 샘플링 하는 파라미터가 전혀 없다.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html