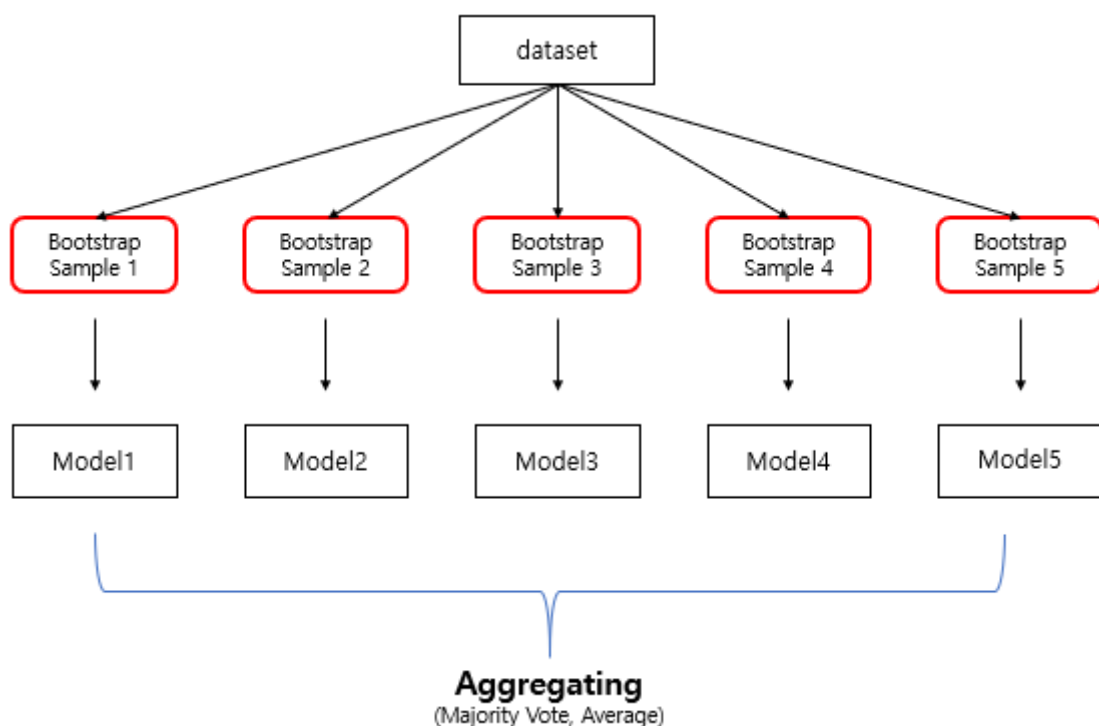


Bagging vs voting sampling

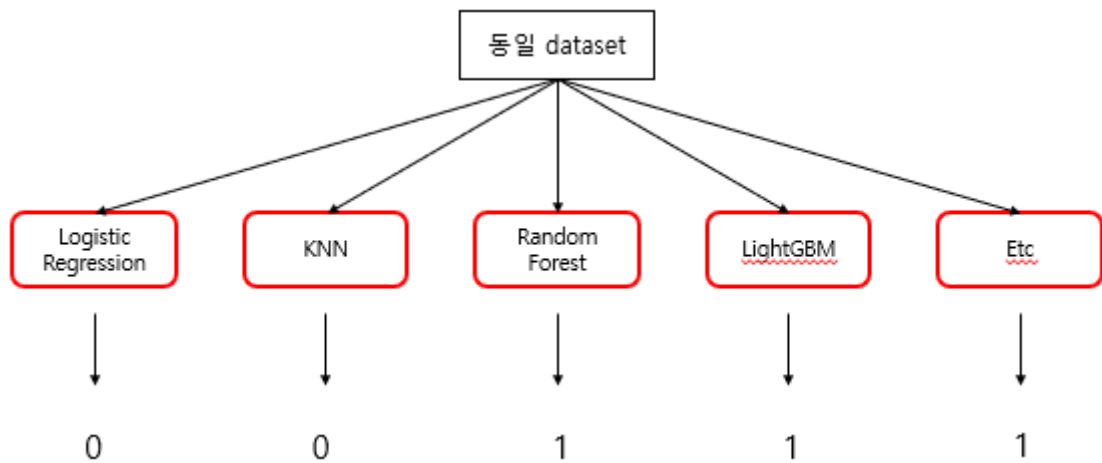
Bagging

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



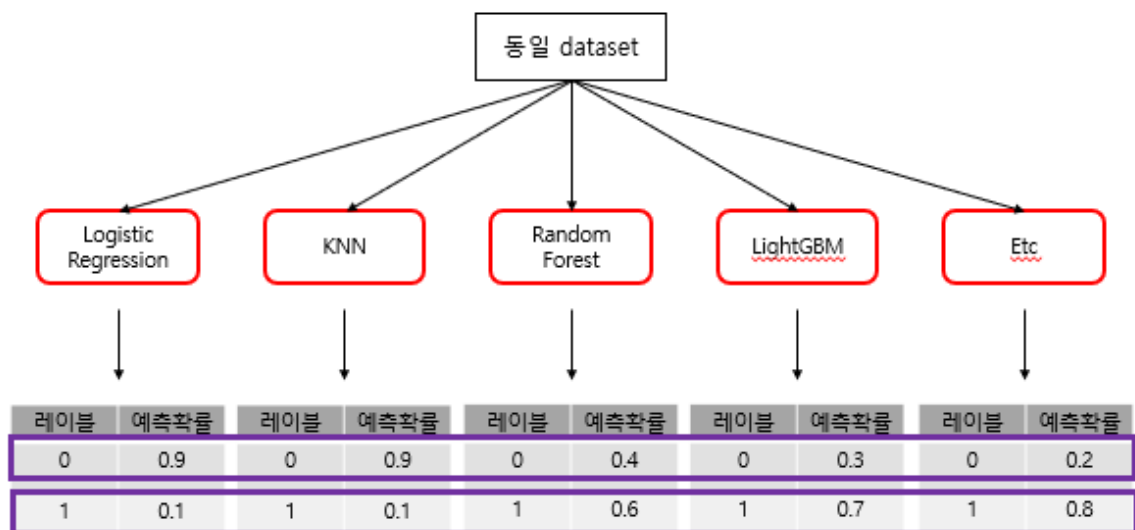
Voting

- 여러 개의 분류기를 통해 예측하는것을 의미한다.



하드 보팅 : 위와 같이 예측을 진행하되 다수결의투표에 따라 분류하게된다.

Soft voting역시 그림으로 표현하자면 아래와 같다.

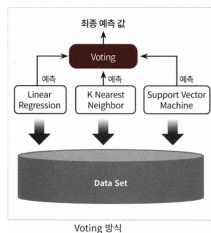


레이블 0 예측확률 평균: 0.54
레이블 1 예측확률 평균: 0.46

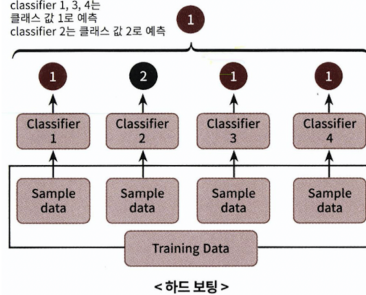
아래의 그림에서 하드 보팅이 마치 sampling되는 것처럼 오해를 불러일으킬수도 있는데 실제 저 박스의 의미는 전체 데이터가 들어간다고 이해를 하면 될것 같다.

Q2

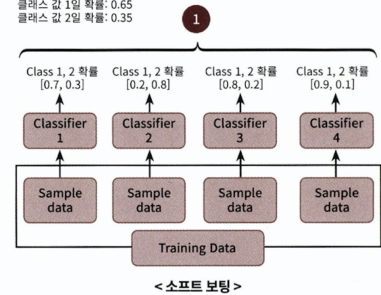
- 보팅도 배깅과 마찬가지로 샘플링 데이터를 다르게 가져가는가?



Hard Voting은 다수의 classifier 간 다수결로 최종 class 결정
클래스 값 1로 예측
classifier 1, 3, 4는
클래스 값 1로 예측
classifier 2는 클래스 값 2로 예측



Soft Voting은 다수의 classifier 들의 class 확률을 평균하여 결정
클래스 값 1로 예측
클래스 값 1일 확률: 0.65
클래스 값 2일 확률: 0.35



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

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html)

[learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html)