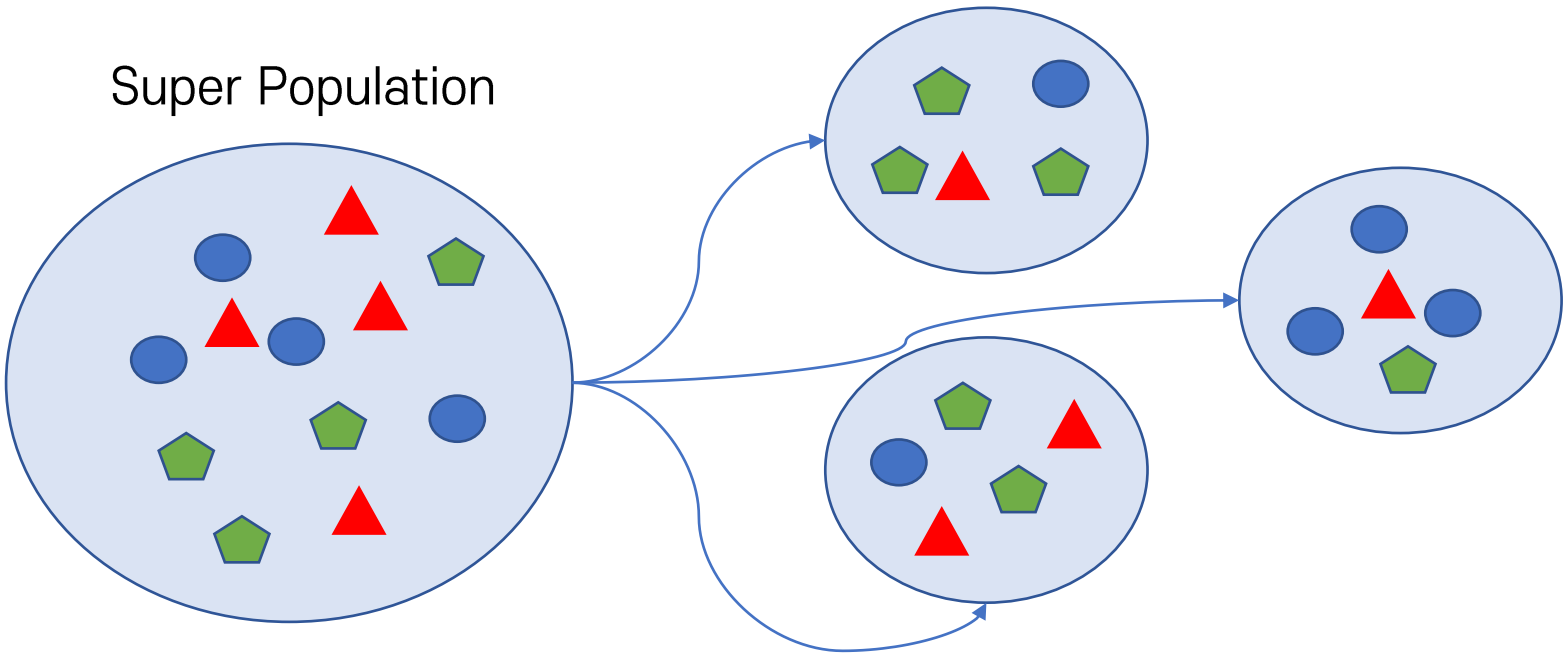
배깅, 부스팅

배깅 개념

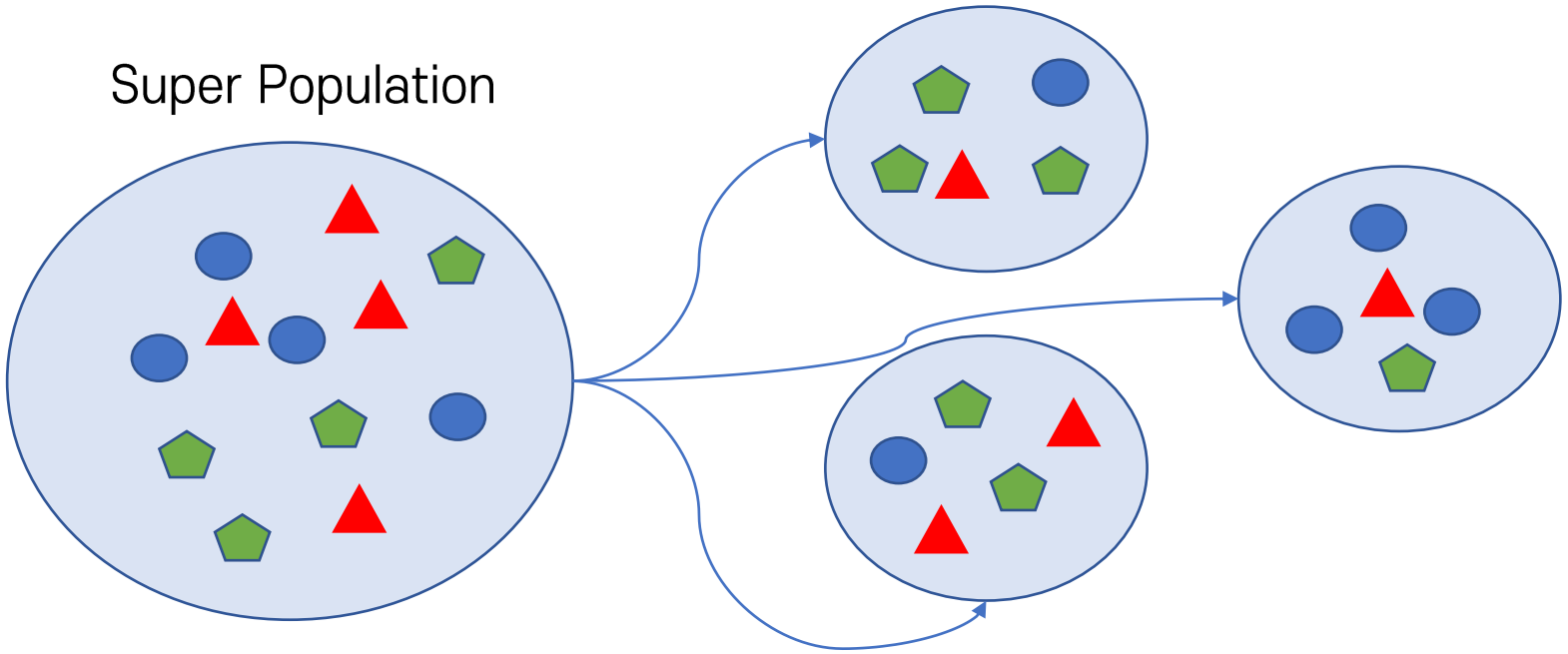
붓스트랩 (Bootstrap)

•랜덤 샘플링의 일종으로 가설 검증을 하거나 통계 계산을 하기 전에 **단순임의복원추출법**(중복허용)을 적용하여 여러 개의 동일한 크기의 표본 자료를 획득하는 방법

•주어진 데이터를 원래의 모집단을 대표하는 독립 표본으로 가정하고, 그 자료로부터 중복을 허용한 무작위 재추출을 하여 복수의 자료를 획득하고 각각에서 통계량을 계산



배깅 개념



Out of bag sample

: bootstrapping을 수행하면 평균적으로 각 예측기에 훈련 데이터의 63% 정도만 샘플링 되는데,

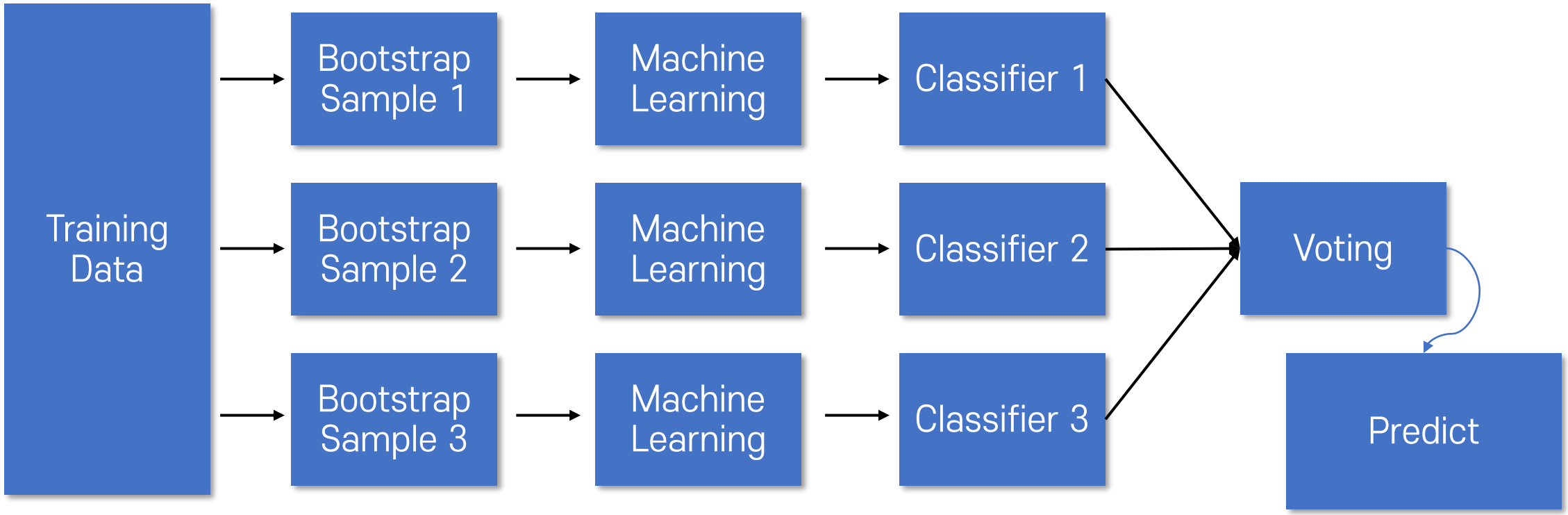
이 때 선택되지 않은 나머지 37%를 Out of bag 샘플이라고 함 (예측기마다 남겨진 37%는 모두 다름)

oob_score = True를 하면 oob 데이터를 모델의 평가를 위해 사용함 (검증 세트나 교차검증 필요 없음)

배깅 개념

•주어진 자료를 모집단으로 생각하여 주어진 자료에서 **여러 개의 붓스트랩 자료를 생성**하고 각 붓스트랩 자료에 예측 모형을 만든 후 결합하여 최종 예측모형을 만드는 방법

•분산을 줄이고 정확도를 개선하여 모델의 안정성을 크게 높여 과적합(overfitting)을 피하도록 함



배깅 개념

- 최적 의사결정 나무 구축에서 가장 어려운 가지치기를 진행하지 않고 약한 학습자인 나무를 최대로 성장시킨 후 보팅함
- 훈련자료의 모집단의 분포를 몰라 평균예측모형을 구할 수 없다는 문제를 해결하기 위해 훈련 자료를 모집단으로 생각하고 평균예측모형을 구하여 분산을 줄이고 예측력을 향상시킴

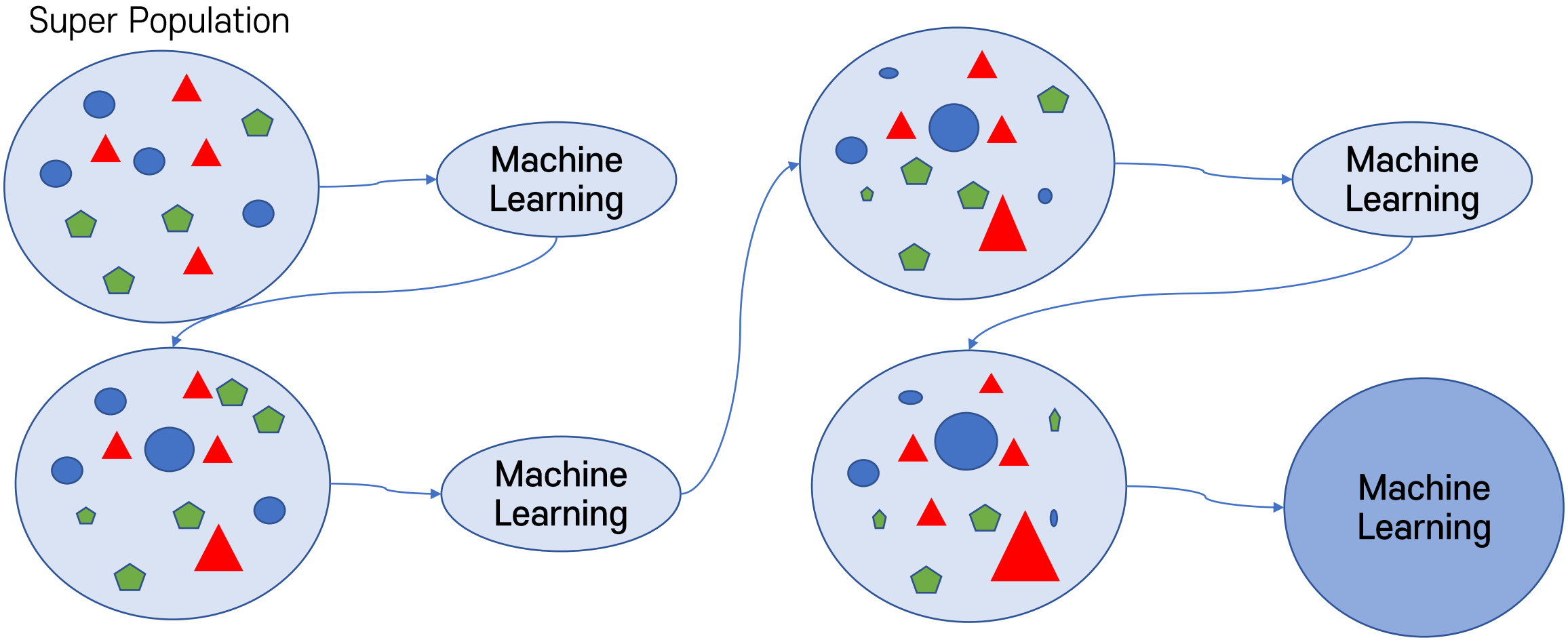
- **부스팅과의 차이점**
  - 주어진 자료보다 분산이 적은 앙상블 모델을 얻는 데 중점을 둠
  - 각 붓스트랩에 대해 붓스트래핑 및 모델링 과정이 병렬적으로 수행됨

배깅 개념

```
sklearn.ensemble.BaggingClassifier(
        base_estimator=None,
        n_estimators=10,
        max_samples=1.0, max_features=1.0,
        bootstrap=True,
        bootstrap_features=False,
        oob_score=False)
```

부스팅 개념

- 예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법
- 붓스트랩을 병렬로 수행(각 모델을 독립적으로 구축)하는 배깅과 달리 순차방식으로 학습을 진행함



부스팅 개념

- 배깅에 비해 모델의 장점을 최적화하고
  train 데이터에 대해 오류가 적은 결합모델을 생성할 수 있다는 장점이 있음
- train 데이터에 과적합할 위험이 있음

분류의 경우

sklearn.ensemble.AdaBoostClassifier(base_estimator = None, n_estimators = 50, learning_rate=1.0)

회귀의 경우

sklearn.ensemble.AdaBoostRegressor(base_estimator = None, n_estimators = 50, learning_rate=1.0)

base_estimator : 부스팅에서 수행할 분류기 (None이면 DecisionTreeClassifier/DecisionTreeRegressor를 수행)

n_estimators : 부스팅이 종료되는 최대 분류기의 수 (int, default=50)

부스팅 개념

- 배깅에 비해 모델의 장점을 최적화하고
  train 데이터에 대해 오류가 적은 결합모델을 생성할 수 있다는 장점이 있음
- train 데이터에 과적합할 위험이 있음

분류의 경우

sklearn.ensemble.AdaBoostClassifier(base_estimator = None, n_estimators = 50, learning_rate=1.0)

회귀의 경우

sklearn.ensemble.AdaBoostRegressor(base_estimator = None, n_estimators = 50, learning_rate=1.0)

base_estimator : 부스팅에서 수행할 분류기 (None이면 DecisionTreeClassifier/DecisionTreeRegressor를 수행)

n_estimators : 부스팅이 종료되는 최대 분류기의 수 (int, default=50)