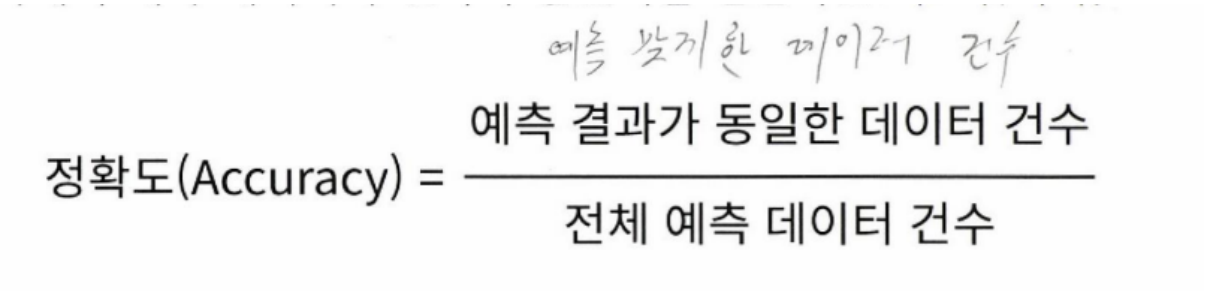
3장 평가

- 머신 러닝 모델은 분류냐 회귀냐에 따라 여러 종류로 나뉠수 있다.
- 회귀 : 실제값과 예측값의 오차 평균값에 기반한 여러 지표들이 있음 - 5장
- 분류 : 특히 이진 분류(0/1)의 상황에서는 정확도 보다는 다른 성능 평가 지표가 더 중요시되는 경우가 많다.
  - 정확도(Accuracy)
  - 오차행렬(Confusion Matrix)
  - 정밀도(Precision)
  - 재현율(Recall)
  - F1 스코어
  - ROC AUC

01. 정확도(Accuracy)



- 사이킷런의 BaseEstimator

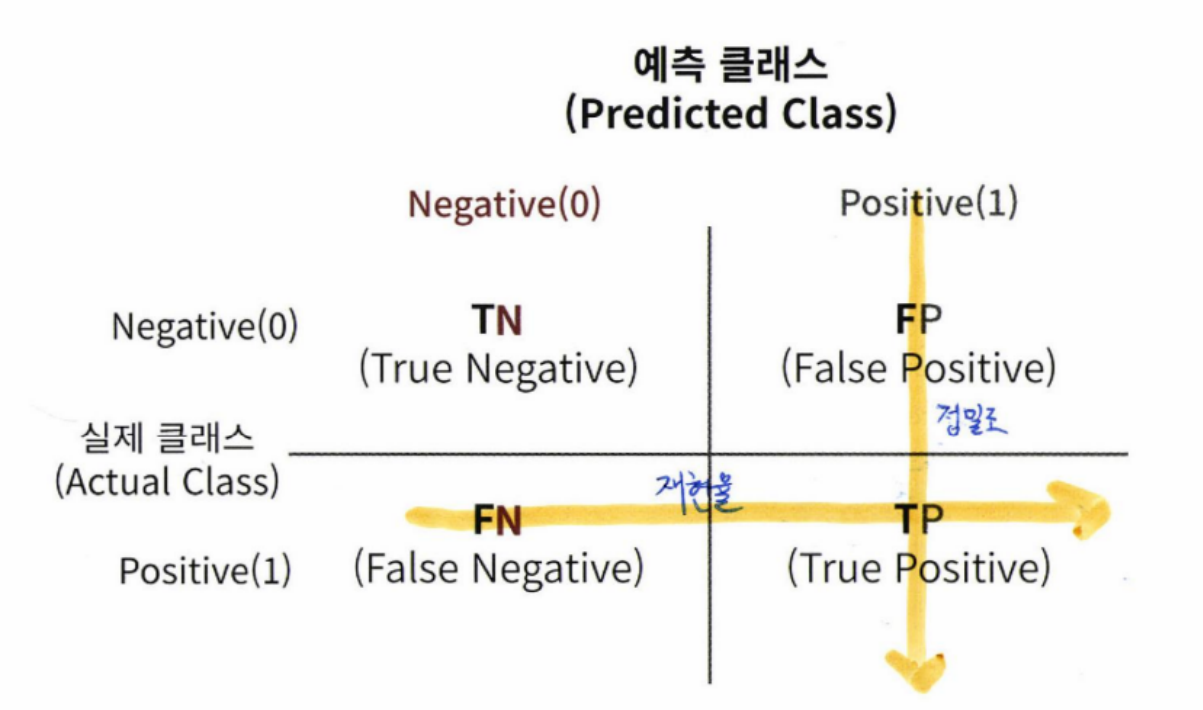
  사이킷런은 BaseEstimator를 상속받으면 Customized 형태의 Estimator를 개발자가 생성할 수 있습니다.

  → 그 동안 sklearn은 정해진 API를 따른다고만 생각했는데 지정하는게 가능합니다.

- 정확도는 불균형한(imbalanced) 레이블 값 분포에서 ML 모델의 성능을 판단할 경우, 적합한 평가 지표가 아닙니다.

02. 오차 행렬

- 학습된 분류 모델이 예측을 수행하면서 얼마나 헷갈리고, 있는지도 함께 보여주는 지표 (어떤 유형의 예측 오류가 발생하는지)



정확도 = 예측 결과와 실제 값이 동일한 건수/ 전체 데이터 수 = (TN + TP) / (TN + FP + FN + TP)

정밀도 = TP / (FP + TP)

재현율 = TP / (FN + TP)

03. 정밀도와 재현율

정밀도 = TP / (FP + TP)

재현율 = TP / (FN + TP)

- 정밀도/재현율 트레이드 오프
- 정밀도와 재현율의 맹점

04. F1 스코어

05. ROC 곡선과 AUC

06. 파마 인디언 당뇨병 예측

07. 정리