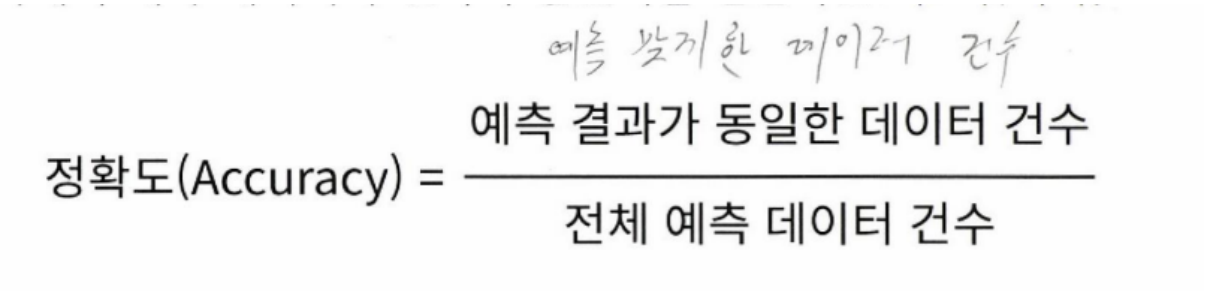
3장 평가

- 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

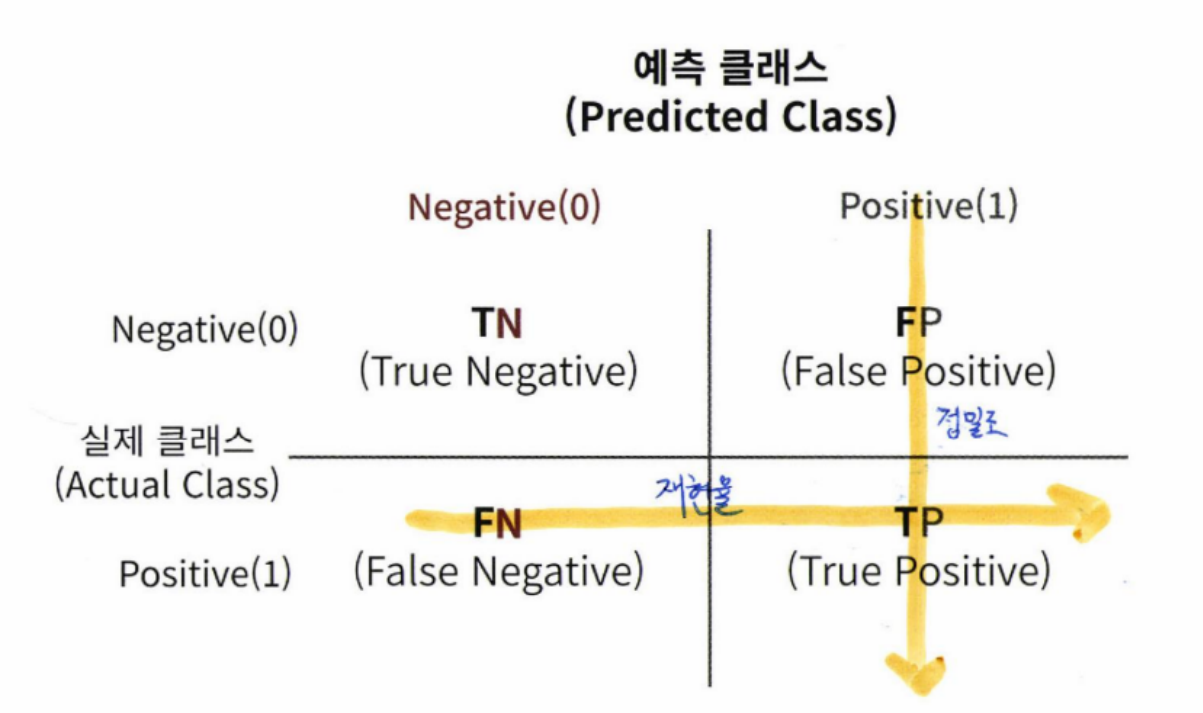
  사이킷런은 BaseEstimator를 상속받으면 Customized 형태의 Estimator를 개발자가 생성할 수 있습니다.

  → 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

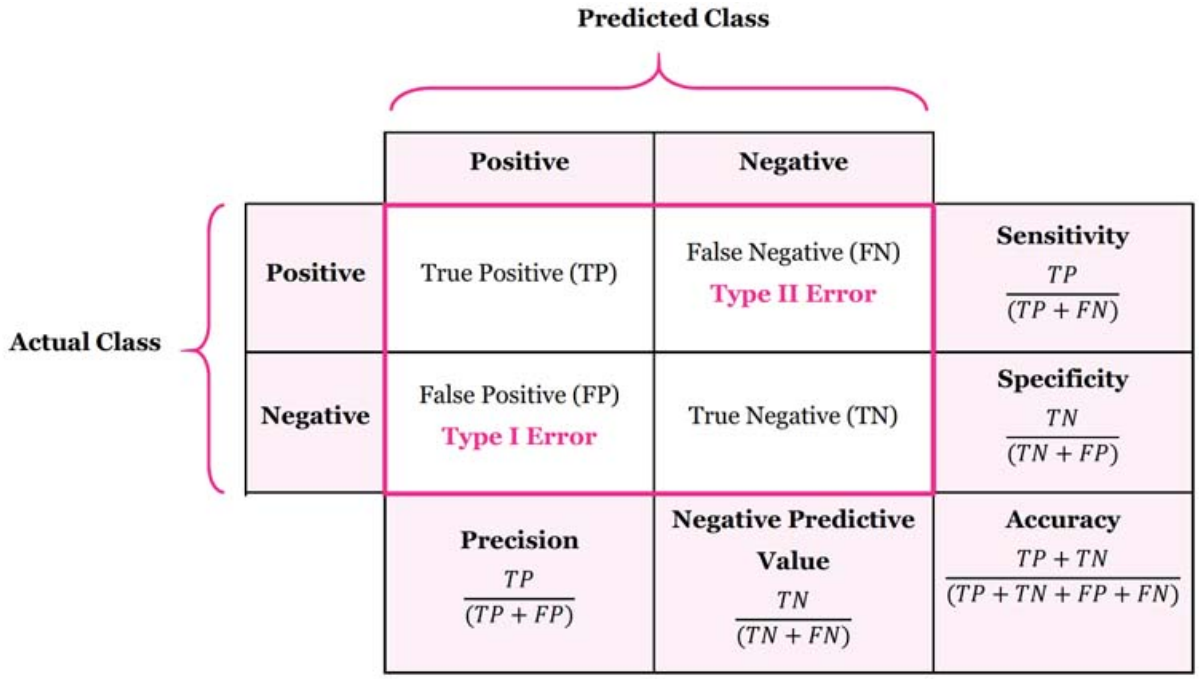
- 학습된 분류 모델이 예측을 수행하면서 얼마나 헷갈리고, 있는지도 함께 보여주는 지표 (어떤 유형의 예측 오류가 발생하는지)



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

정밀도(Precision) = TP / (FP + TP)

재현율(Sensitivity) = TP / (FN + TP)



03. 정밀도와 재현율

정밀도 = TP / (FP + TP)

재현율 = TP / (FN + TP)

- 정밀도/재현율 트레이드 오프
  - 정밀도와 재현율은 상호 보완적인 평가 지표이기 때문에 어느 한쪽을 높이면 다른 하나의 수치는 떨어지기 쉽다.
  - 예시 : 임계값(threshold)을 설정하여 특정 값 분류 실습을 한다.

    → 임계값에 따라 재현율과 정밀도의 차이가 생긴다.

    사이킷런은은 precision_recall_curve() API를 제공한다.

입력 파라미터	y_true: 실제 클래스값 배열 (배열 크기= [데이터 건수]) probas_pred: Positive 칼럼의 예측 확률 배열 (배열 크기= [데이터 건수])
반환 값	**정밀도**: 임곗값별 정밀도 값을 배열로 반환 **재현율**: 임곗값별 재현율 값을 배열로 반환

- 정밀도와 재현율의 맹점
  - 정밀도가 100%가 되는 방법

확실한 기준이 되는 경우만 Positive로 예측하고 나머지는 모두 Negative로 예측한다.

- 재현율이 100%가 되는 방법

  모든 환자를 Positive로 예측하면 됩니다.

04. F1 스코어

- F1 스코어 : 정밀도와 재현율을 결합한 지표
- 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 높은 값을 가진다.

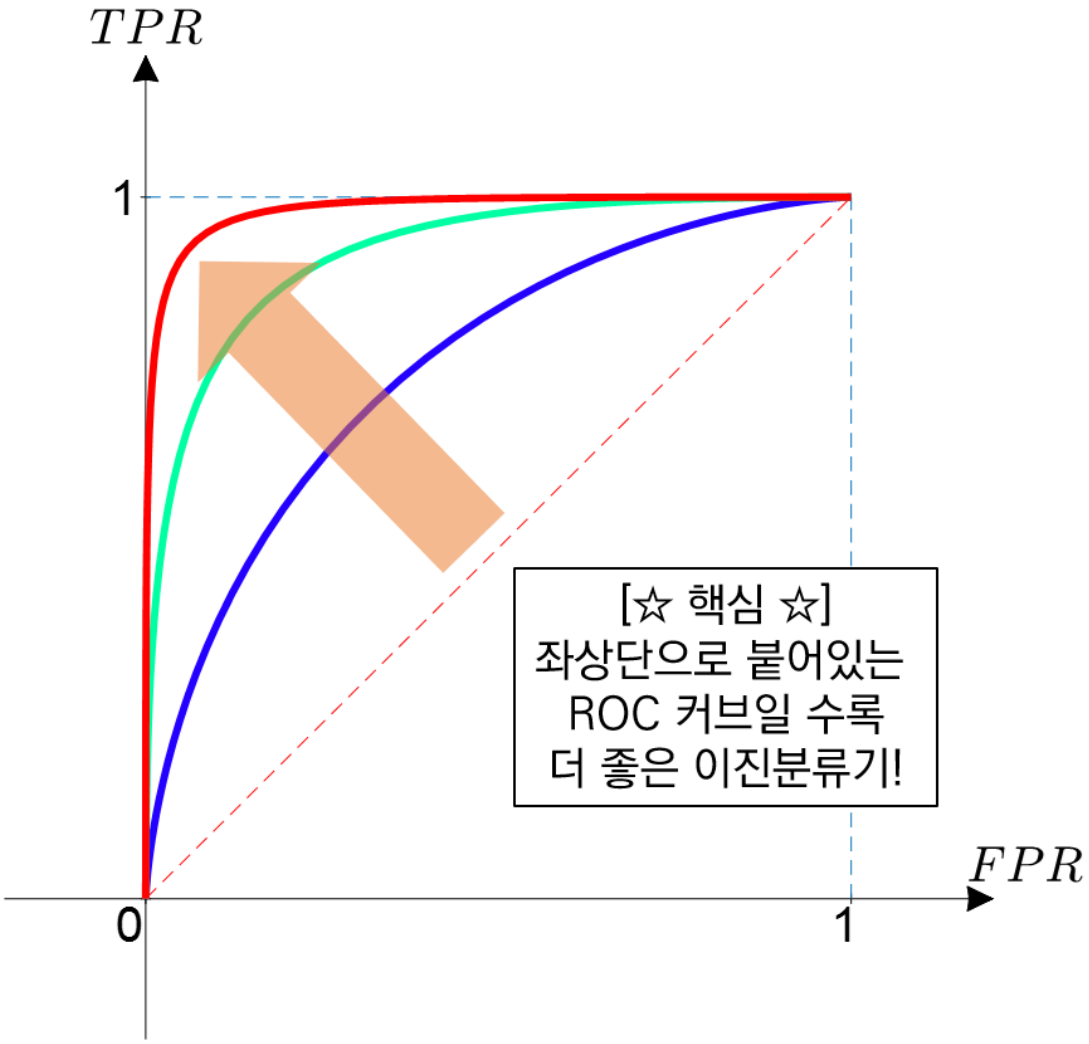
$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 * \frac{precision * recall}{precision + recall}$$

05. ROC 곡선과 AUC

- 이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표
- ROC(Receiver Operation Characteristic Curve) 곡선 : FPR(False Positive Rate)이 변할 때, TPR(True Positive Rate)이 어떻게 변하는지를 나태는 곡선

  FPR을 X축으로, TPR을 Y축으로 잡으면 FPR의 변화에 따른 TPR의 변화가 곡선 형태로 나타난다.

  - 민감도(TPR)는 실제값 Positive(양성)가 정확히 예측돼야 하는 수준을 나타낸다(질병이 있는 사람은 질병이 있는 것으로 양성 판정)
  - 특이성(TNR)은 실제값 Negative(음성)가 정확히 예측돼야 하는 수준을 나태난다.(질병이 없는 건강한 사람은 질병이 없는 것으로 음성 판정)
  - FPR은 FP / (FP + TN)이므로 1 - TNR 또는 1- 특이성으로 표현된다.



- AUC(Area Under Curve)

  ROC 곡선 자체는 FPR과 TPR의 변화 값을 보는 데 이용하며 분류의 성능 지표로 사용되는 것은 ROC 곡선 면적에 기반한 AUC 값으로 결정한다.

  일반적으로 1에 가까울수록 좋은 수치이다.

06. 파마 인디언 당뇨병 예측

- 데이터 셋 : 북아메리카 피마 지역 원주민의 Type-2 당뇨병 결과 데이터
- → 고립된 유전적 특성 때문에 식습관과 유전을 근거로한 당뇨 원인 파악에 좋은 데이터로 알려져 있다.

- Pregnancies: 임신 횟수
- Glucose: 포도당 부하 검사 수치
- BloodPressure: 혈압(mm Hg)
- SkinThickness: 팔 삼두근 뒤쪽의 피하지방 측정값(mm)
- Insulin: 혈청 인슐린(mu U/ml)
- BMI: 체질량지수(체중(kg)/(키(m))^2)
- DiabetesPedigreeFunction: 당뇨 내력 가중치 값
- Age: 나이
- Outcome: 클래스 결정 값(0또는 1)

이 정도의 데이터값을 인지하고 코드를 통한 분석을 해보겠습니다.