

# **Machine Learning Model Evaluation**

# 분류(Classification) 성능 평가 지표

---

## 분류 성능 평가 지표

- 정확도(Accuracy)
- 오차행렬(Confusion Matrix)
- 정밀도(Precision)
- 재현율(Recall)
- F1 스코어
- ROC AUC

# 정확도(Accuracy)

---

## 정확도

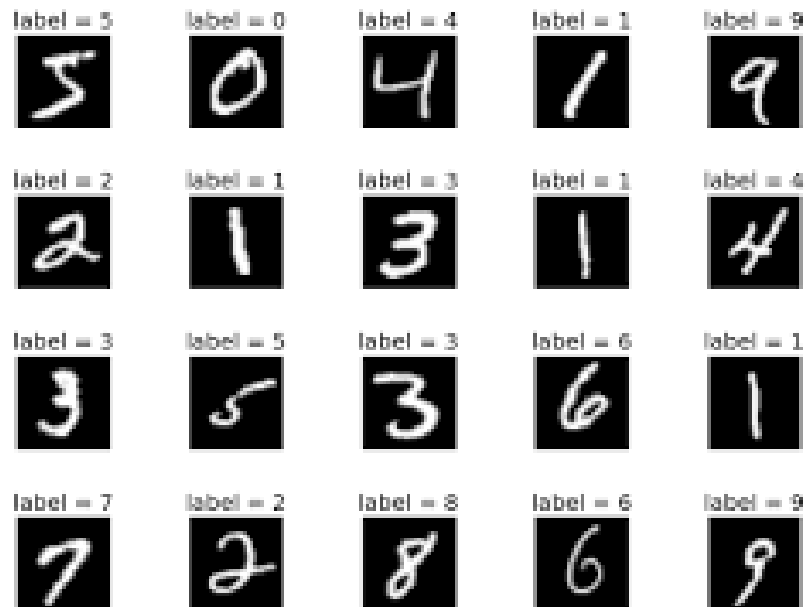
- 정확도는 직관적으로 모델 예측 성능을 나타내는 평가 지표
- 이진 분류의 경우 데이터의 구성에 따라 ML 모델의 성능을 왜곡할 수 있기 때문에 정확도 수치 하나만 가지고 성능을 평가하지 않음
- 특히 정확도는 불균형한(imbalanced) 레이블 값 분포에서 ML 모델의 성능을 판단할 경우, 적합한 평가 지표가 아님
- $$\text{정확도(Accuracy)} = \frac{\text{예측 결과가 동일한 데이터 건수}}{\text{전체 예측 데이터 건수}}$$

# 정확도의 문제점

## 정확도

- 타이타닉 생존자 예측에서 여성은 모두 생존으로 판별
- If Sex == '여성' 생존
- MNIST 데이터셋을 multi classification에서 binary classification으로

변경



# 오차 행렬(Confusion Matrix)

---

## 오차 행렬

- 오차 행렬은 이진 분류의 예측 오류가 얼마인지가 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표
- Positive(양성), Negative(음성) vs. TRUE(맞게 분류), FALSE(틀리게 분류)
- 4분면의 왼쪽과 오른쪽은 예측된 class를 기준으로 Negative와 Positive로 분류
- 위와 아래는 실제 class를 기준으로 Negative와 Positive로 분류

# 오차 행렬(Confusion Matrix)

## 오차 행렬

- 예측 class와 실제 class에 따라 TN, FP, FN, TP 형태로 오차 행렬의 4분면

예측 클래스(Predicted Class)

Negative(0)

Positive (1)

Negative(0)

TN

FP

(True Negative)

(False Positive)

실제 클래스  
(Actual Class)

FN

TP

(False Negative)

(True Positive)

Positive(1)

# 오차 행렬(Confusion Matrix)

## 오차 행렬

1. TN은 예측값을 Negative 값인 0으로 예측했고 실제값 또한 Negative 값인 0일 때
2. FP은 예측값을 Positive 값인 1으로 예측했는데 실제값은 Negative 값인 0일 때
3. FN은 예측값을 Negative 값인 0으로 예측했는데 실제값은 Positive 값인 1일 때
4. TP은 예측값을 Positive 값인 1으로 예측했고 실제값 또한 Positive 값인 1일 때

예측 클래스(Predicted Class)			
		Negative(0)	Positive (1)
실제 클래스 (Actual Class)	Negative(0)	검출되지 않아야 할 것이 검출되지 않음 Negative Negative (True Negative) TN	틀린결정 Negative Positive (False Positive) FP
	Positive(1)	검출 되어야 할 것이 검출되지 않음 Positive Negative (False Negative) FN	옳은결정 Positive Positive (True Positive) TP

# 오차 행렬(Confusion Matrix)

## 오차 행렬을 통한 정확도 지표 문제점 인지

		예측 클래스	
		Negative	Positive
실제 클래스	Negative	<b>TN</b> 예측 : Negative (7 이 아닌 Digit) 405 개	<b>FP</b> 예측 : Positive ( Digit 7 ) 0
	Positive	<b>FN</b> 예측 : Negative (7 이 아닌 Digit) 45 개 실제 : Positive ( Digit 7 )	<b>TP</b> 예측 : Positive ( Digit 7 ) 0 실제 : Positive ( Digit 7 )

TP는 0임.  
Positive로 예측이  
한건도 성공하지  
않음

FP가 0이므로  
Positive로 예측자  
체를 수행하지 않  
음을 알 수 있음

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



# 오차 행렬(Confusion Matrix)

## 오차 행렬을 통한 정확도 지표 문제점 인지

		예측 클래스	
		Negative	Positive
실제 클래스	Negative	<b>TN</b> 예측 : Negative (7 이 아닌 Digit) 405 개	<b>FP</b> 예측 : Positive ( Digit 7 ) 0
	Positive	<b>FN</b> 예측 : Negative (7 이 아닌 Digit) 45 개 실제 : Positive ( Digit 7 )	<b>TP</b> 예측 : Positive ( Digit 7 ) 0 실제 : Positive ( Digit 7 )

TP는 0임.  
Positive로 예측이  
한건도 성공하지  
않음

FP가 0이므로  
Positive로 예측자  
체를 수행하지 않  
음을 알 수 있음

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

# 정밀도(Precision)과 재현율(Recall)

## 정밀도와 재현율

- 정밀도는 예측을 Positive로 한 대상 중에 예측을 실제 값이 Positive로 일치한 데이터의 비율을 뜻함
- 재현율은 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻함
- 정밀도 =  $TP / (FP + TP)$ 
  - `precision_score()`
- 재현율 =  $TP / (FN + TP)$ 
  - `recall_score()`

		예측 클래스	
		Negative	Positive
실제 클래스	Negative	<b>TN</b> 예측 : Negative (7 이 아닌 Digit) 405 개	<b>FP</b> 예측 : Positive ( Digit 7 ) 0
	Positive	<b>FN</b> 예측 : Negative (7 이 아닌 Digit) 45 개	<b>TP</b> 예측 : Positive ( Digit 7 ) 0
		실제 : Negative (7 이 아닌 Digit)	실제 : Positive ( Digit 7 )

# 정밀도(Precision)과 재현율(Recall)

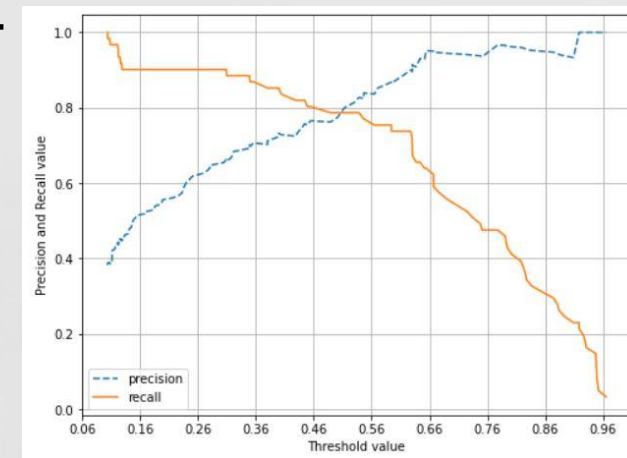
## 업무에 따른 재현율과 정밀도의 상대적 중요도

- 재현율이 상대적으로 더 중요한 지표인 경우는 실제 Positive 양성인 데이터 예측을 Negative로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 :  
암진단, 금융사기 판별
- 정밀도가 상대적으로 더 중요한 지표인 경우는 실제 Negative 음성인 데이터 예측을 Positive 양성으로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 : 스팸 메일
- 불균형한 레이블 클래스를 가지는 이진 분류 모델에서는 많은 데이터 중에서 중점적으로 찾아야 하는 매우 적은 수의 결과값에 Positive를 설정해 1값을 부여하고, 그렇지 않은 경우는 Negative로 0 값을 일반적으로 부여

# 정밀도(Precision)과 재현율(Recall)

## 정밀도/재현율 트레이드오프(Trade off)

- 분류하려는 업무의 특성상 정밀도 또는 재현율이 특별히 강조되어야 할 경우 분류의 결정 임계값(Threshold)을 조정해 정밀도 또는 재현율의 수치를 높일 수 있음
- 정밀도와 재현율은 상호 보완적인 평가 지표이기 때문에 어느 한쪽을 강제로 높이면 다른 하나의 수치는 떨어지기 쉬움
- 정밀도/재현율의 트레이드오프(Trade-off)



# 정밀도(Precision)과 재현율(Recall)

## 정밀도/재현율 맹점

- 정밀도를 100%로 만드는 법
  - 확실한 기준이 되는 경우만 Positive로 예측 하고 나머지는 모두 Negative로 예측
  - 정밀도  $TP / (TP + FP)$  전체 환자 1000명 중 확실한 Positive 징후만 가진 환자는 단 1명이라고 하면 이 한 명만 Positive로 예측하고 나머지는 모두 Negative로 예측하더라도 FP는 0, TP는 1이되므로 정밀도는  $1/(1+0)$ 으로 100%가 됨

# 정밀도(Precision)과 재현율(Recall)

## 정밀도/재현율 맹점

- 재현율을 100%로 만드는 법
  - 모든 환자를 Positive로 예측
  - 재현율  $TP / (TP + FN)$  전체 환자 1000명을 다 Positive로 예측
  - 이중 실제 양성인 사람이 30명 정도라도 TN이 수치에 포함되지 않고 (FN은 아예 0이므로  $30/(30+0)$ 이므로 100%가 됨

# F1 Score

---

## F1 Score

- F1 스코어(Score)는 정밀도와 재현율을 결합한 지표
- F1 스코어는 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가짐
- F1 스코어의 공식

$$F1 = \frac{2}{\left(\frac{1}{recall} + \frac{1}{precision}\right)} = 2 * (precision * recall) / (precision + recall)$$

# F1 Score

---

## F1 Score

- F1 스코어(Score)는 정밀도와 재현율을 결합한 지표
- A 예측 모델의 경우 정밀도 0.9 재현율이 0.1로 극단적인 차이가 나고,
- B 예측 모델은 정밀도가 0.5, 재현율이 0.5로 정밀도와 재현율이 큰 차이가 없다면
- A 예측 모델의 F1스코어는 0.18이고, B예측 모델의 F1 스코어는 0.5로 B 모델이 A모델에 비해 매우 우수한 F1 스코어를 가짐
- 사이킷런은 F1 Score를 위해 `F1_score()` 함수를 제공



# ROC 곡선과 AUC

---

## ROC 곡선과 AUC

- ROC 곡선(Receiver Operation Characteristic Curve)과 이에 기반한 AUC 스코어는 이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표
- 일반적으로 의학 분야에서 많이 사용되지만, 머신러닝의 이진 분류 모델의 예측 성능을 판단하는 중요한 평가 지표

# ROC 곡선과 AUC

---

## ROC 곡선과 AUC

- ROC 곡선은 FPR(False Positive Rate)이 변할 때 TPR(True Positive Rate)이 어떻게 변하는지를 나타내는 곡선
- FPR을 X축으로, TPR을 Y축으로 잡으면 FPR이 변화에 따른 TPR의 변화가 곡선 형태로 나타냄
- 분류의 성능 지표로 사용되는 것은 ROC 곡선 면적에 기반한 AUC값으로 결정
- AUC(Area Under Curve) 값이 ROC 곡선 밑의 면적을 구한 것이므로 일반적으로 1에 가까울수록 좋은 수치

# ROC 곡선

---

## FPR의 변화에 따른 TPR의 변화 곡선

- TPR은 Ture Positive Rate의 약자이며, 이는 재현율을 나타냄
- TPR은  $TP / (FN + TP)$ , 즉 재현율은 민감도로도 불림
- FPR은 실제 Negative(음성)을 잘못 예측한 비율을 나타냄
- 즉 실제로 Negative인데 Positive 또는 Negative로 예측한 것 중 Positive로 잘못 예측 비율
- FPR은  $FP / (FP + TN)$

# 사이킷런 ROC 곡선 및 AUC 스코어

---

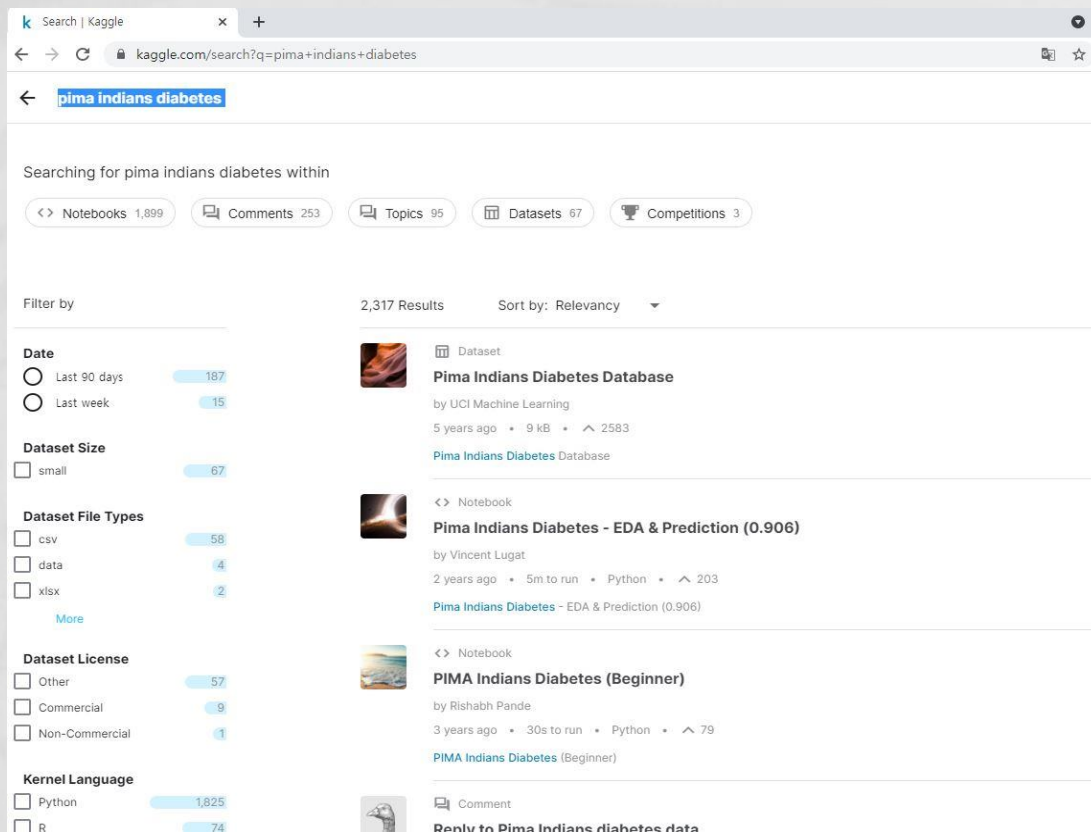
## 사이킷런 ROC 곡선 및 AUC 스코어

- 사이킷런은 임계값에 따른 ROC 곡선 데이터를 `roc_curve()`로
- AUC 스코어를 `roc_auc_score()` 함수로 제공

# 피마 인디언 당뇨병 예측

## 피마 인디언 당뇨병 예측

- 피마 인디언 당뇨병(Pima Indian Diabetes) 데이터 세트를 이용해 당뇨병 여부를 판단하는 머신러닝 예측 모델을 수립하고 평가지표를 적용



# 피마 인디언 당뇨병 예측

## 피마 인디언 데이터 세트

### Machine Learning Repository

UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets.php

UCI Machine Learning Repository  
Center for Machine Learning and Intelligent Systems








About Citation Policy Donate a Data Set

Repository Web

View ALL Datasets

Check out the [beta version](#) of the new UCI Machine Learning Repository we are currently testing! [Contact us](#) if you have any issues, questions, or concerns. [Click here to try out the new site.](#)

Browse Through: 588 Data Sets

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes
Classification (442) Regression (137) Clustering (117) Other (56)	 <a href="#">Abalone</a>	Multivariate	Classification	Categorical, Integer, Real	4177	8
Attribute Type Categorical (38) Numerical (396) Mixed (55)	 <a href="#">Adult</a>	Multivariate	Classification	Categorical, Integer	48842	14
Data Type Multivariate (456) Univariate (27) Sequential (57) Time-Series (121) Text (66) Domain-Theory (23) Other (21)	 <a href="#">Annealing</a>	Multivariate	Classification	Categorical, Integer, Real	798	38
Area Life Sciences (138) Physical Sciences (57) CS / Engineering (215) Social Sciences (38) Business (44) Game (11) Other (80)	 <a href="#">Anonymous Microsoft Web Data</a>		Recommender-Systems	Categorical	37711	294
# Attributes Less than 10 (151) 10 to 100 (266)	 <a href="#">Arrhythmia</a>	Multivariate	Classification	Categorical, Integer, Real	452	279
	 <a href="#">Artificial Characters</a>	Multivariate	Classification	Categorical, Integer, Real	6000	7
	 <a href="#">Audiology (Original)</a>	Multivariate	Classification	Categorical	226	

# 평가

---

## 평가

- 이진분류에서 정밀도, 재현율, F1 스코어, AUC 스코어가 주로 성능평가 지표로 활용
- 오차 행렬은 실제 클래스 값과 예측 클래스 값의 True, False에 따라 TN, FP, FN, TP로 매핑되는 4분면 행렬을 제공
- 정밀도와 재현율은 Positive 데이터 세트의 예측 성능에 좀 더 초점을 맞춘 지표이며, 분류 결정 임계값을 조정해 정밀도 또는 재현율은 수치를 높이거나 낮출 수 있음

# 평가

---

## 평가

- F1스코어는 정밀도와 재현율이 어느 한쪽으로 치우치지 않을 때 좋은 값을 가짐
- AUC스코어는 ROC 곡선 밑의 면적을 구한 것으로 1에 가까울 수록 좋은 수치



# 정리

---

## 정리

- 분류(Classification) 성능 평가 지표
- 정확도(Accuracy)
- 오차행렬(Confusion Matrix)
- 정밀도(Precision)
- 재현율(Recall)
- F1 스코어
- ROC AUC