



## 2장. 딥러닝 살펴보기

# Wrap Up

---

1. 인공지능 ← 머신러닝 ← 딥러닝
2. 텐서플로우의 케라스는 매우 단순하며 강력합니다.
3. 신경망 모델 사용을 위해서 GPU 보유 여부는 매우 중요합니다.
4. GPU를 보유하고 있지 않더라도 우리에게 **구글 코랩과 캐글 노트북**이 존재합니다.
5. 무료 GPU는 사용 시간이 제한되어 있으니, 주의해서 사용해야 합니다.
6. 캐글의 최대 강점인 **캐글 노트북은 전문가들의 노하우를 볼 수 있는 최고의 기술서**입니다.
7. 공식 홈페이지를 보는 습관은 실력 향상을 위한 좋은 방법입니다.

# 2장의 내용?

---

- 스포츠 경기에서 이기려면?
  - 선수 능력 파악, 적재적소 배치와 경기 흐름을 파악
  - 우리 팀 선수뿐만 아니라 적 팀의 선수 능력을 정확하게 파악해야 함
- 동일하게, 신경망 개념도 중요하지만 그 외의 다양한 요소를 고민할 수 있어야 함
- 이번 장에에서는 다음 내용을 다뤄봅니다.
  - 머신러닝 프로세스
  - 다양한 용어 살펴보기
    - 데이터 준비하기
    - 학습하기
    - 평가하기
  - 데이터셋 및 커뮤니티 살펴보기

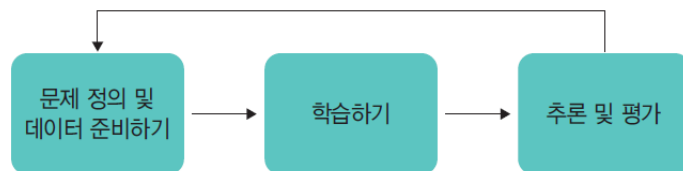
용어를 자세히 아는 것도 중요하지만,  
많은 용어를 아는 것도 굉장히 중요합니다.

# 머신러닝 프로세스 간략히 살펴보기

- ML(Machine Learning) Process

- 문제 정의 및 데이터 준비하기 → 학습하기 → 추론 및 평가 → ...


데이터 준비로  
학습하기로  
etc.



[그림 2-1] 머신러닝 프로세스(간략)

- 잘 구축된 프로세스가 결과적으로 성능이 우수한 모델을 만듦
  - 어떻게? **‘잘’** 해야함.
  - 직접 문제를 정의해보고, 해결해보는 과정을 **경험**하자!
- 이 책에서 다루는 대부분의 프로세스는 위 그림의 과정을 따름

# 문제 정의 및 데이터 준비하기

- 문제를 어떻게 정의하느냐에 따라 우리가 가야할 길이 달라짐
    - 올바른 방향으로 나아가기 위해 탐색적 데이터 분석(EDA)를 시작
    - 가장 중요한 것은 **명확한 문제 정의**
      - 무엇이 문제인가?
      - 난 문제를 어떻게 풀고 싶은가?
      - 사용자는 누구인가?
    - 숫자? 문자? 이미지? ...
    - 이진 분류? 다중 분류? 회귀? 생성?
  - 모델을 선택하기 전에, 데이터를 아~주 자세히 들여다보자!
    - 변형해야 할까, 어떤 전처리 방법을 선택할까
    - 데이터가 속한 도메인에는 어떤 처리 방법이 존재할까
-  **캐글을 활용하자!**
- 데이터에 관한 고민은 끊임없이 반복될 것



[그림 2-2] 내가 데이터를 만들 때

# 학습하기

- 모델 선택에는 어떤 질문이 필요할까요
  - 선택한 모델이 주로 어떤 데이터에 적용되었나요? 사례가 있을까요?
  - 모델이 얼마나 깊어야 하나요? 어떤 환경에서 사용될까요?
  - 옵티마이저는요? 손실 함수는요?
  - 실험 환경에 적합한 모델인가요?

- 깊은 고민도 좋지만, **SOTA 모델**을 활용해보자  
(또는 **미리 학습된 모델**)

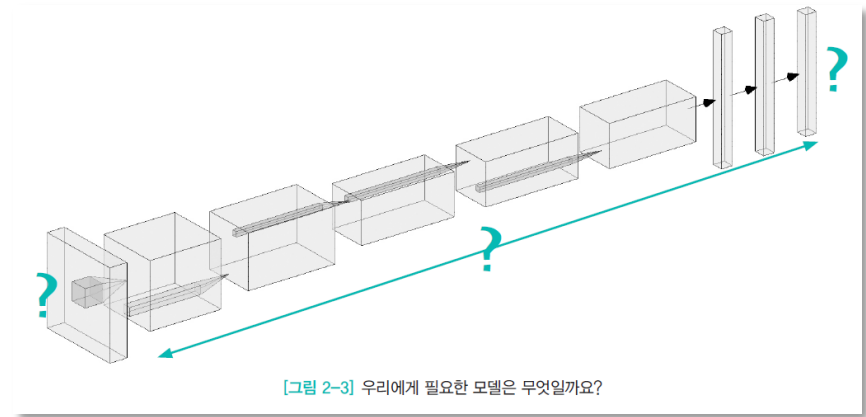
- 이미지? VGG, ResNet, Inception etc.
- 자연어 처리? ELMO, BERT, GPT etc.

- 모델을 선택했다면, **하이퍼파라미터를 조정**하는 단계는 필수!

- 배치 크기, 학습률, 층 파라미터 등

- 모델 선택은 **내부 요소와 외부 요소를 복합적으로 고려**해야 함

- 하이퍼파라미터와 모델 크기(실험 환경에 적합한지)



# 추론 및 평가

---

- 추론(Inference)
  - 학습된 모델로부터 정답이 없는 데이터에 대해 정답을 만드는 행위
- 추론을 통해 얻은 정답을 어떻게 평가? 햄버거 가게를 생각해보자.

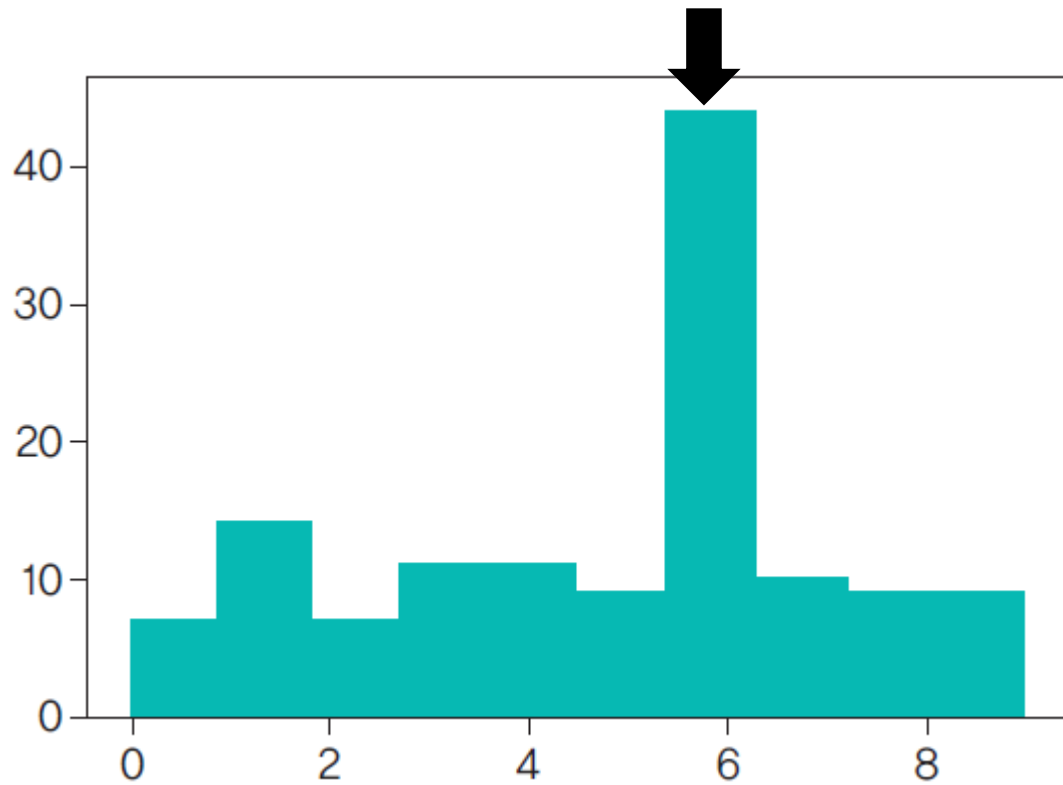
- 역할: 매니저(나), 직원 A, B
- 오늘 햄버거 만드는 실력을 보고 A, B 중 한 명을 승진시키겠다.
- A 직원: 10개를 다 만들었지만, 5개만 완벽
- B 직원: 다 만들지 못했지만, 6개가 완벽

Q. 누구 선택?

# 용어 살펴보기

---

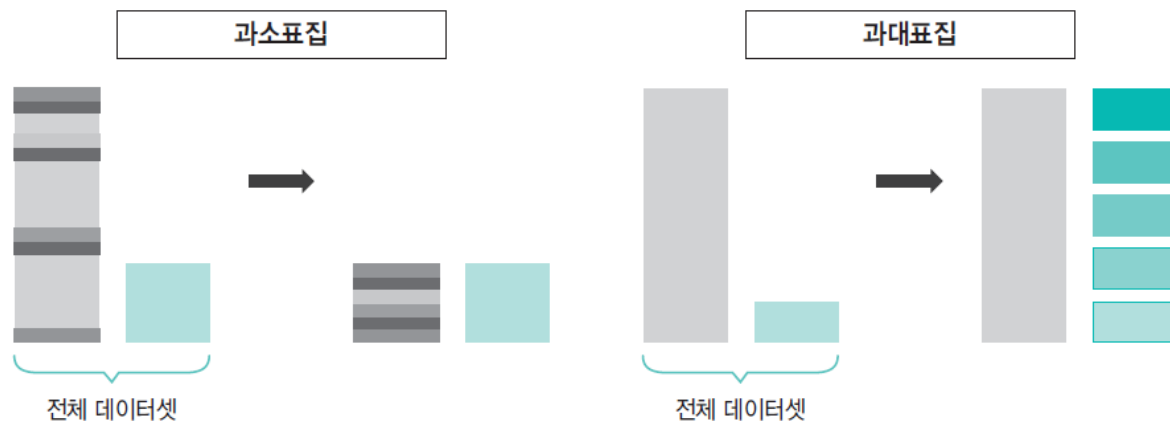
- 어떤 용어가 떠오르나요?





# 용어: 데이터 준비하기

- 클래스 불균형(Class Imbalance)
  - 클래스가 불균형하게 분포되어 있음
  - 은행 거래 사기, 희귀 질병, 기계 불량음 등의 사례
  - 이상 탐지(Anomaly Detection)
- 과소표집(UnderSampling)과 과대표집(OverSampling)
  - 과소표집 : 다른 클래스에 비해 상대적으로 많이 나타나 있는 클래스의 개수를 줄이는 것
  - 과대표집 : 개수가 적은 클래스를 복제하는 것



# 용어: 데이터 준비하기

- 회귀(Regression)
  - 여러 개의 특징을 통해 연속적인 숫자로 이루어진 정답을 예측
  - 영화 관객 수, 축구 선수 연봉, 주식 가격 등
  - 0과 1을 예측하는 로지스틱 회귀(Logistic Regression)

빵(g)	고기(g)	치즈(g)	가격(W)
300	500	100	5500
200	1000	50	10500
500	500	500	8000
150	300	150	4000
200	200	200	?

[그림 2-6] 회귀의 예: 햄버거 가격 예측

- 분류(Classification)

- 햄버거를 선택한다면?

햄버거	음료
1	0

- 불고기버거를 선택한다면?

불고기버거	치킨버거	치즈버거
1	0	0

- 불고기버거 세트를 선택했다면?

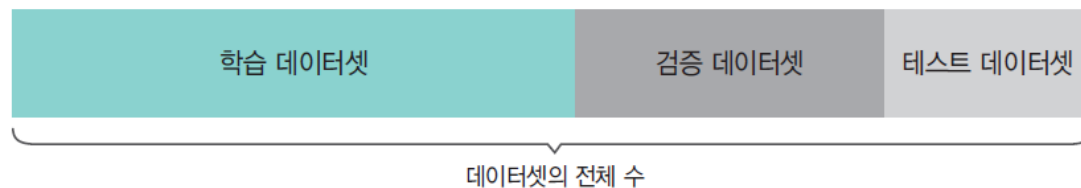
불고기버거	치킨버거	치즈버거	콜라	환타	사이다
1	0	0	1	0	0

\*선택: 1 미선택: 0

[그림 2-7] 이진 분류, 다중 분류, 다중 레이블 분류의 예

# 용어: 데이터 준비하기

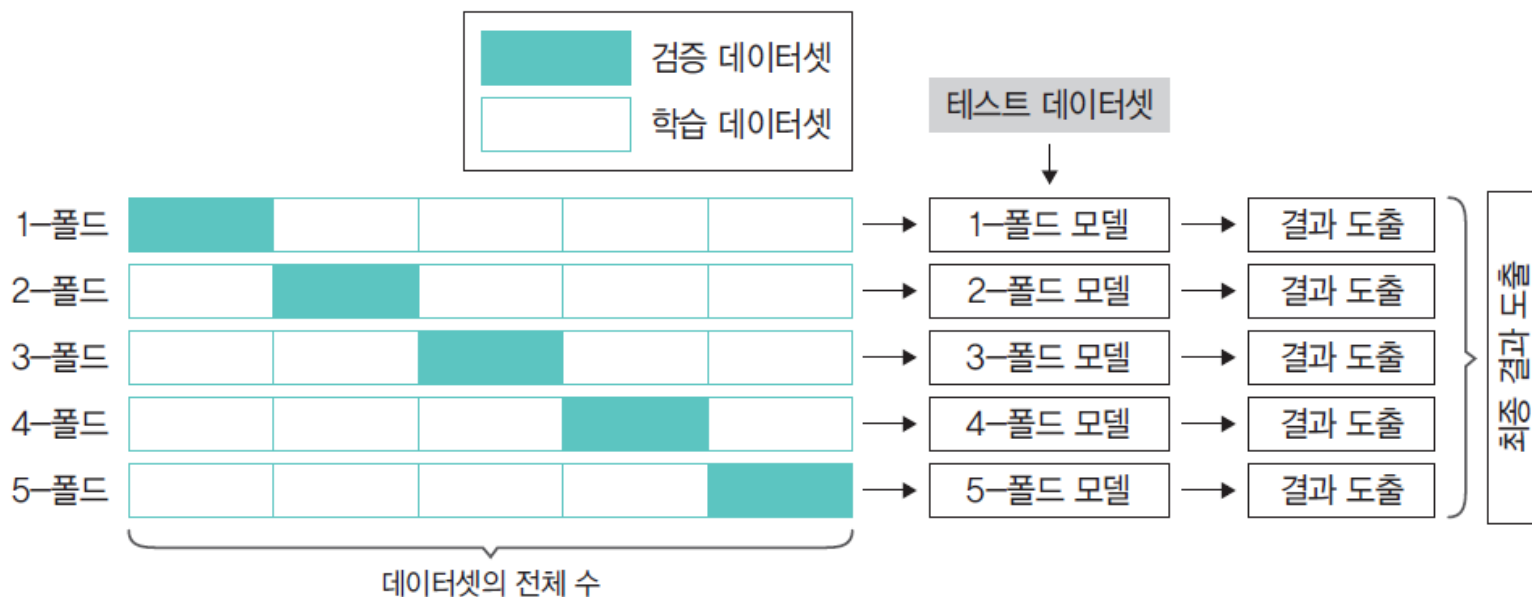
- 원핫 인코딩(One-Hot Encoding)
  - 하나의 클래스만 1이고 나머지 클래스는 전부 0인 인코딩(Encoding)
  - 불고기버거: [1, 0, 0]
  - 치킨버거: [0, 1, 0]
  - 치즈버거: [0, 0, 1]
- 교차 검증(Cross-Validation)
  - 모델의 타당성을 검증
    - ✓ 학습 데이터: 모델 학습에 사용
    - ✓ 검증 데이터: 모델의 검증을 위해 사용, 주로 학습 도중에 사용
    - ✓ 테스트 데이터: 모델의 최종 성능 평가에 사용
  - **테스트 데이터는 최종 평가 이전에는 절대로 사용하면 안된다**



[그림 2-8] 홀드아웃 교차 검증, 데이터 분리

# 용어: 데이터 준비하기

- K-Fold Cross-Validation
  - 모델은 많은 데이터를 보여줄수록 성능이 올라감
  - K-Fold 방법을 사용해서 최대한 많은 데이터를 볼 수 있도록 우리가 도와주자
  - K는 주로 3 ~ 10을 사용



[그림 2-11] K-폴드 과정 ②

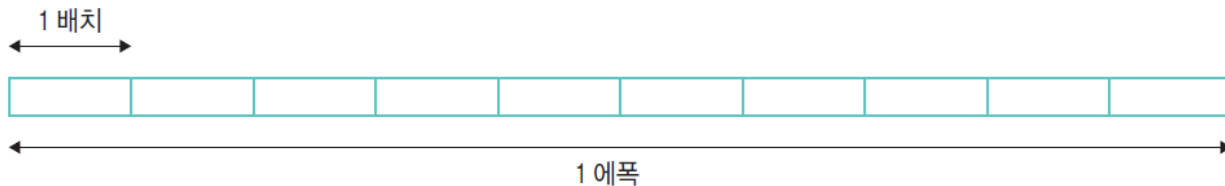
# 용어: 학습하기

---

- 하이퍼파라미터(Hyperparameter)
  - 경험에 의해 결정되는 요소
  - 학습률, 배치 크기, 에폭 등
  - 적합한 값을 찾기 위해 반복적인 실험과 많은 시간 투자가 필수
  - 9장의 케라스 튜너를 참고!
- 배치(Batch) & 배치 크기(Batch Size)
  - 데이터를 한 개만 사용하기엔 정확한 정답을 찾는 데 방해가 될 수 있고, 전부를 사용하기엔 시간이 너무 오래 걸리기 때문에 배치를 사용
  - 1,000개 데이터에서 배치가 10이면, 각 배치는 100개 데이터를 보유
  - 배치 크기는 기존 사례를 참고하거나 주로 2 제곱수를 사용(16, 32, 64, ...)

# 용어: 학습하기

- 에폭(Epoch) & 스텝(Step)
  - 에폭: 전체 데이터를 사용하여 학습하는 횟수
    - 전체 데이터를 10회 반복? 10 에폭
  - 스텝: 모델이 가진 가중치를 1회 업데이트하는 것



[그림 2-12] 배치와 에폭

햄버거 100개가 있고, 이를 전체 데이터라고 하겠습니다.

또한, 햄버거는 100개의 단위로 다시 제공받을 수 있으며, **100개를 전부 먹으면 1 에폭**이라고 정의하겠습니다.

햄버거 1,000개를 먹기 위해서는 **10회 반복(10 에폭)**해야 합니다. 이 과정에서 한번에 100개를 먹는 것은 너무 많은 것 같습니다. 20개씩 나눠서 먹겠습니다. 여기서 **배치는 5(100/20)**이며, **배치 크기는 20**이라고 할 수 있습니다.

# 용어: 학습하기

---

- 지도 학습(Supervised Learning)
    - 학습 데이터에 정답이 포함된 것
    - 모델에게 햄버거 사진을 보여주면서 햄버거라고 알려줌
  - 비지도 학습(UnSupervised Learning)
    - 학습 데이터에 정답이 포함되어 있지 않은 것
    - 모델에게 햄버거를 종류별로 여러 개 주고 같은 종류끼리 묶어보라고 하는 것
  - 햄버거 사진을 주고, 모델에게 다시 햄버거 사진을 그려보라고 하는 것
    - 생성 모델(Generative Model)
  - 에이전트가 주어진 환경에 대해 어떠한 행동을 결정하고, 이를 통해 얻는 보상으로 학습하는 것
    - 강화 학습(Reinforcement Learning)
-

# 용어: 학습하기

---

- 과대적합(Overfitting)

- 학습 데이터에서는 좋은 성능을 보이지만, 새로운 데이터에 대해서는 좋은 성능을 보이지 못하는 결과
- 학습 데이터를 단순히 외웠다고 표현할 수 있으며, 모델이 문제를 일반화(Generalization)하지 못했음

- 학습 데이터를 다양하게, 많이 수집합니다.
- 정규화(Regularization)를 사용합니다 → 규칙을 단순하게
- 트리플 치즈버거와 같은 이상치는 제거합니다.(데이터가 많다면 제거하는 방법은 좋지 않습니다)

- 과소적합(Underfitting)

- 모델이 학습 데이터를 충분히 학습하지 않아 모든 측면에서 좋지 않은 성능을 보여주는 결과
- 모델은 아직 성능이 개선될 여지가 남아있는 상태

- 학습 데이터를 다양하게, 많이 수집합니다.
  - 더 복잡한 모델을 사용합니다.
  - 모델을 충분히 학습시킵니다.
-



# 용어: 평가하기

- 혼동행렬(Confusion Matrix)
  - 모델의 성능 평가에 사용

		예측된 정답	
		True	False
실제 정답	True	TP	FN
	False	FP	TN

[그림 2-13] 혼동행렬

- 유통기한이 지난 햄버거를 유통기한이 지난 햄버거로 분류: **TP(True Positive)**
- 정상 햄버거를 정상 햄버거로 분류한 경우: **TN(True Negative)**
- 유통기한이 지난 햄버거를 정상 햄버거로 잘못 분류한 경우: **FN(False Negative)**
- 정상 햄버거를 유통기한이 지난 햄버거로 잘못 분류한 경우: **FP(False Positive)**

# 용어: 평가하기

- 정확도(Accuracy)
  - 전체 데이터 중에서 실제 데이터의 정답과 모델이 예측한 정답이 같은 비율

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- 데이터가 불균형할 때 사용하는 경우, 잘못된 지표로써 사용될 수 있음

		예측된 정답	
		유통기한이 지난 햄버거	정상 햄버거
실제 정답	유통기한이 지난 햄버거	90	0
	정상 햄버거	10	0

[그림 2-14] 정말로 정확도가 90%일까요?

# 용어: 평가하기

- 정밀도(Precision) & 재현율(Recall)
  - 정밀도 : True라고 예측한 정답 중에서 실제로 True인 비율
  - 재현율 : 실제 데이터가 True인 것 중에서 모델이 True라고 예측한 비율

		예측된 정답	
		True	False
실제 정답	True	TP	FN
	False	FP	TN

$$\text{정밀도} = \frac{TP}{TP + FP}$$

		예측된 정답	
		True	False
실제 정답	True	TP	FN
	False	FP	TN

$$\text{재현율} = \frac{TP}{TP + FN}$$

[그림 2-15] 정밀도와 재현율

# 용어: 평가하기

- 정밀도(Precision) & 재현율(Recall)
  - 재고 관리 직원과 고객 응대 직원, 어떤 기계를 제공해야 할까요?

A		예측된 정답	
		유통기한이 지난 햄버거	정상 햄버거
실제 정답	유통기한이 지난 햄버거	30	10
	정상 햄버거	30	30

$$acc = \frac{60(30 + 30)}{100(30 + 10 + 30 + 30)} = 60\%$$
$$precision = \frac{30}{60(30 + 30)} = 50\%$$
$$recall = \frac{30}{40(30 + 10)} = 75\%$$

B		예측된 정답	
		유통기한이 지난 햄버거	정상 햄버거
실제 정답	유통기한이 지난 햄버거	30	30
	정상 햄버거	10	30

$$acc = \frac{60(30 + 30)}{100(30 + 10 + 30 + 30)} = 60\%$$
$$precision = \frac{30}{40(30 + 10)} = 75\%$$
$$recall = \frac{30}{60(30 + 30)} = 50\%$$

[그림 2-16] 두 햄버거 기계의 성능

- 재고 관리 직원에게는 정상 햄버거를 유통기한이 지난 햄버거라고 판별하여 버리지 않도록 정밀도를 고려한 기계를 제공
- 고객 응대 직원에게는 유통기한이 지난 햄버거를 정상 햄버거라고 판별하여 고객에게 주지 않도록 재현율을 고려한 기계를 제공
- 정밀도와 재현율은 Trade-off 관계!

# 용어: 평가하기

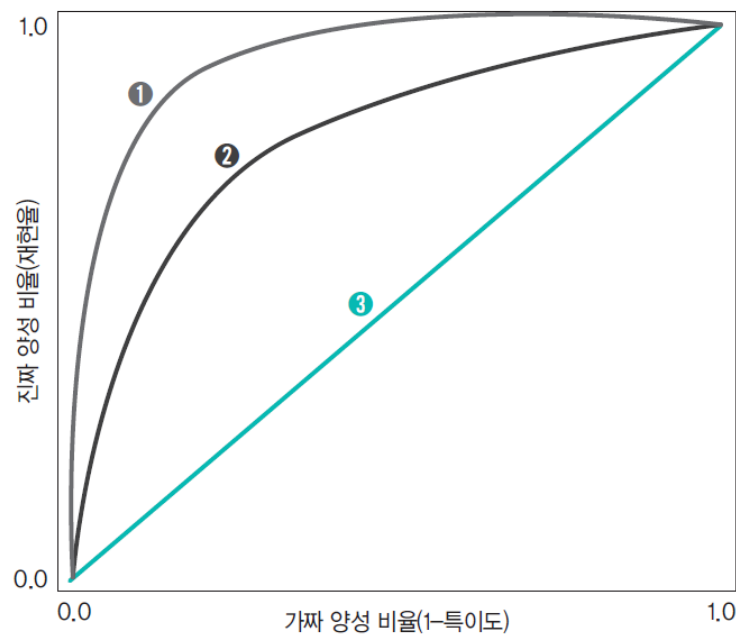
- F1-Score

- 정밀도와 재현율의 중요성이 같다고 가정하고, 두 지표의 조화평균으로 새로운 지표를 제공

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

- ROC 곡선

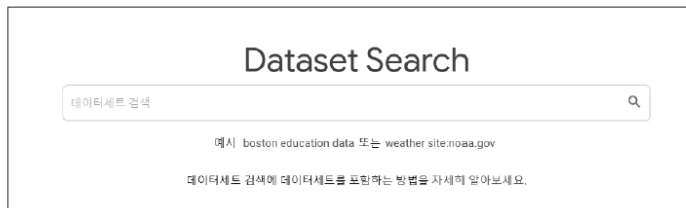
- 코드로 실습해보자.
- 코드 결과에서 어떤 모델을 선택하는 것이 좋을지 토론해보세요.
- 책 내용이 정답은 아님!



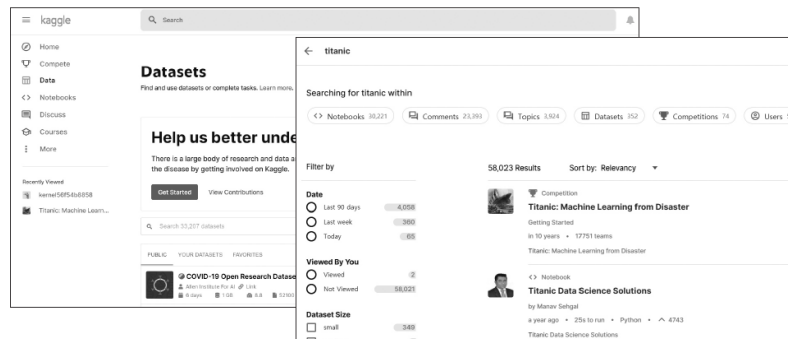
[그림 2-17] ROC 곡선

# 데이터셋 살펴보기

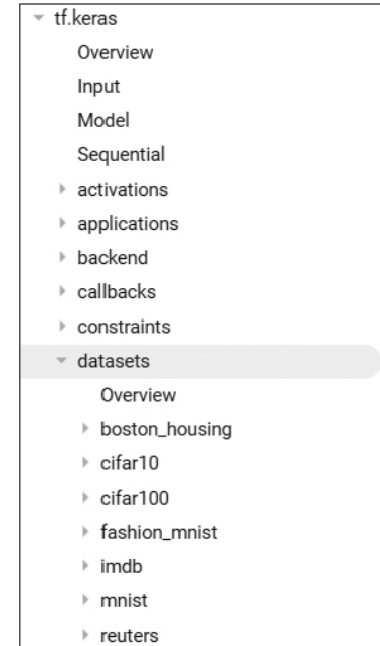
- 프로젝트를 수행하기 위해 데이터를 수집하는 것은 정말 어려운 일
  - 에이~ 그냥 있는거 다운받고, 크롤링 코드짜서 수집해보면 되잖아요.
  - ??? : 일단 해봐.
- 기존 사례에서 사용된 데이터셋을 먼저 적용해보는 것도 프로젝트를 성공으로 이끄는 지름길
  - 구글 데이터셋 검색
  - 캐글, AI hub, 공공 데이터 포털 등



[그림 2-19] 구글 데이터셋 검색 메인 화면



[그림 2-20] 캐글 데이터셋 활용



[그림 2-21] tf.keras.datasets

# 커뮤니티 살펴보기

- 공유와 소통을 할 수 없다면, 이제 살아남기 힘들 것
  - 커뮤니티를 활용하여 공유와 소통을 실천하고, 적극적으로 질문하자
  - 커뮤니티뿐만 아니라 주변 사람과도 공유와 소통을!



**캐글 코리아**  
**Kaggle Korea**  
Non-Profit Facebook Group Community

함께 공부해서, 함께 나눕시다  
Study Together, Share Together



# 요약 정리

---

1. 머신러닝 프로세스는 간략하게 **[문제 정의 및 데이터 준비하기 → 학습하기 → 추론 및 평가]**로 나눌 수 있습니다.
2. **[문제 정의 및 데이터 준비하기]**는 명확한 문제 정의와 데이터 전처리가 매우 중요합니다.
3. **[학습하기]**는 본격적으로 모델을 선택하고, 학습시키는 단계입니다. 하이퍼파라미터 실험 환경 등을 고려하여 학습시간을 효율적으로 활용할 수 있도록 해야 합니다.
4. **[추론 및 평가]**는 올바른 지표를 통해 모델의 성능을 신뢰할 수 있어야 합니다. 주어진 상황에 맞는 지표를 선택하는 것은 매우 어렵고 중요합니다.
5. 세 가지 절로 나누어 여러 가지 용어를 알아보았습니다.
  - 데이터 준비하기: 클래스 불균형, 과소표집과 과대표집, 회귀와 분류, 원핫 인코딩, 교차 검증
  - 학습하기: 하이퍼파라미터, 배치와 배치크기, 에폭과 스텝, 지도 학습, 비지도 학습, 과대적합과 과소 적합
  - 평가하기: 혼동행렬, 정확도, 정밀도와 재현율, F1-Score, ROC 곡선



# 요약 정리

---

6. **구글 데이터셋 검색과 캐글**은 데이터셋을 탐색하고 수집할 최적의 장소입니다. 그 외에도 공공 데이터 포털, AI Hub가 있습니다.
7. 문제는 **공유와 소통**을 통해 더 빠르게 해결될 수 있습니다. 국내에 이를 위한 다양한 커뮤니티가 존재한다는 점을 잊지마세요.