

# preLab 4차

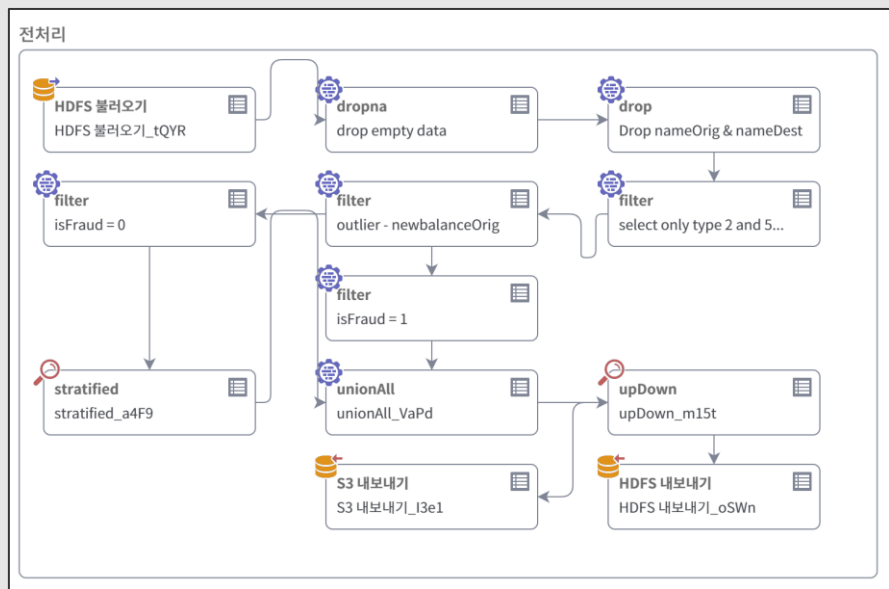


# 순서

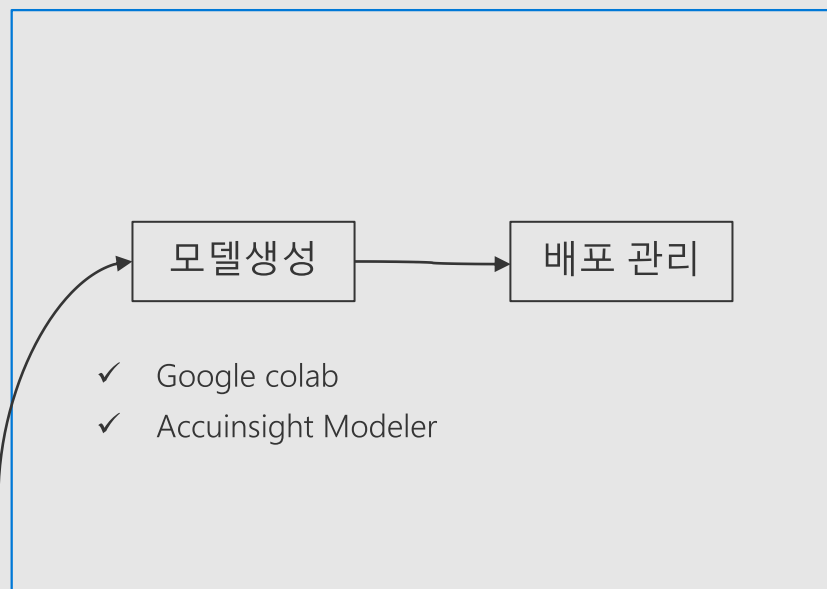
- ✓ 1. 개요
- ✓ 2. Supervised Learning
  - Class Imbalance
  - Modeling Using Accuinsight Modeler
  - 성능 평가

# 개요

## Pipeline



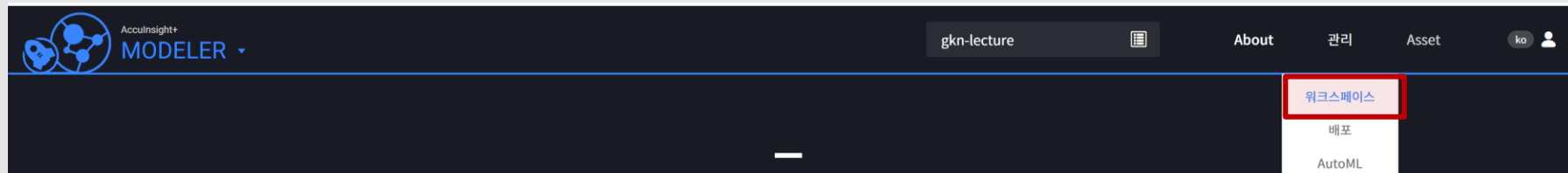
## Modeler



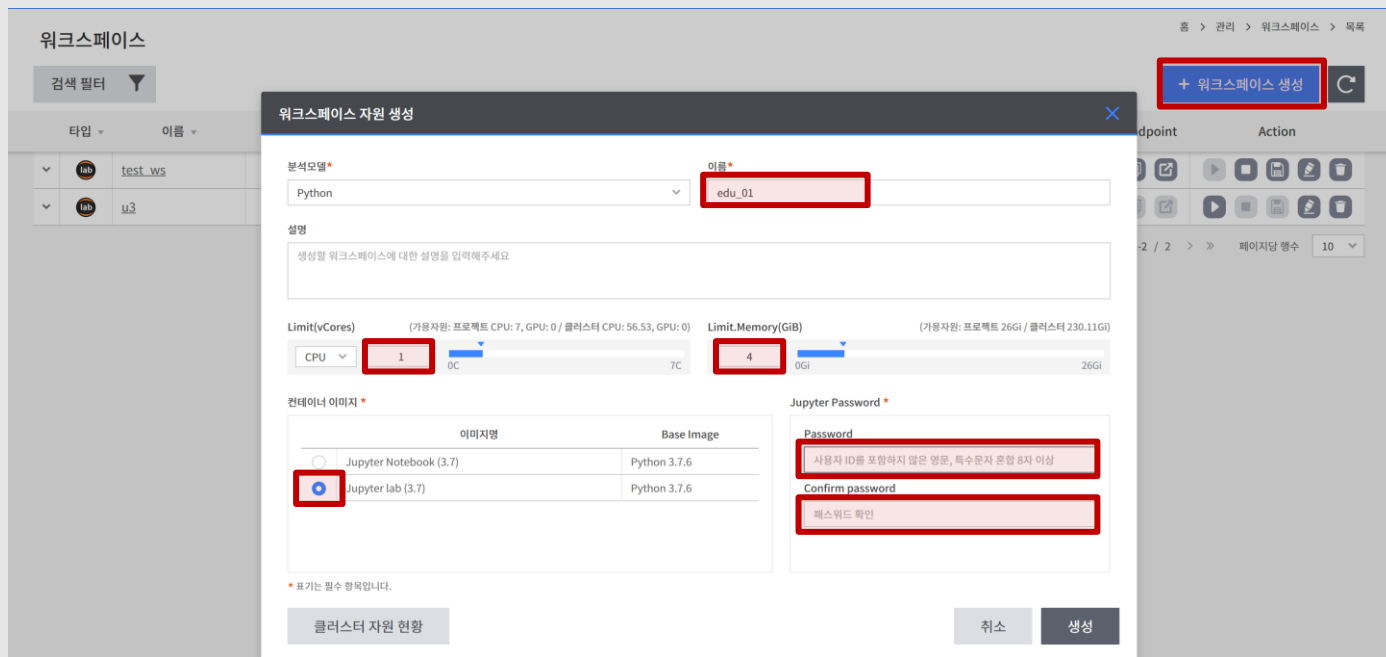
csv file

# 개요 - 준비사항

✓ Modeler 로그인



✓ 워크스페이스 생성



# 개요

## ✓ Anomaly Detection이란,

- Normal(정상) sample과 Abnormal(비정상, 이상치, 특이치) sample을 구별해 내는 문제
- 제조 불량탐지, 금융 사기탐지, 의료 영상, Social Network 등 다양한 분야에서 응용 됨

## ✓ 해결하기 위한 방안

- Supervised Learning
- Semi-supervised Learning
- Unsupervised Learning

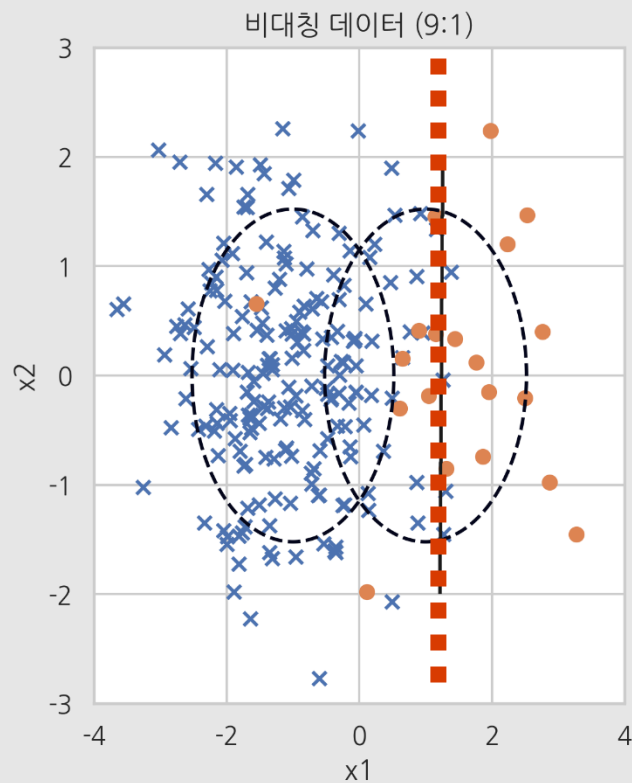
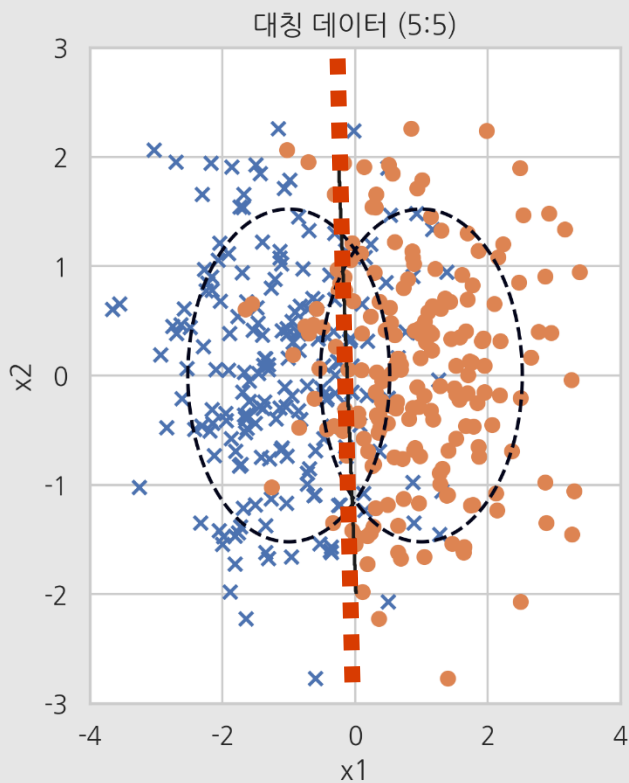
# Supervised Learning

# 다룰 내용

- ✓ Class Imbalance
- ✓ Modeling Using Accuinsight Modeler
- ✓ 성능 평가

# Class Imbalances

✓ 왜 문제인가?

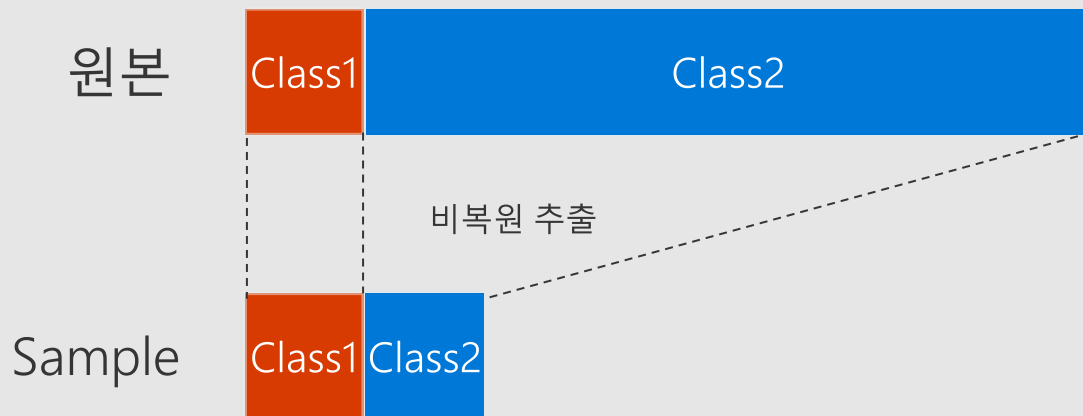
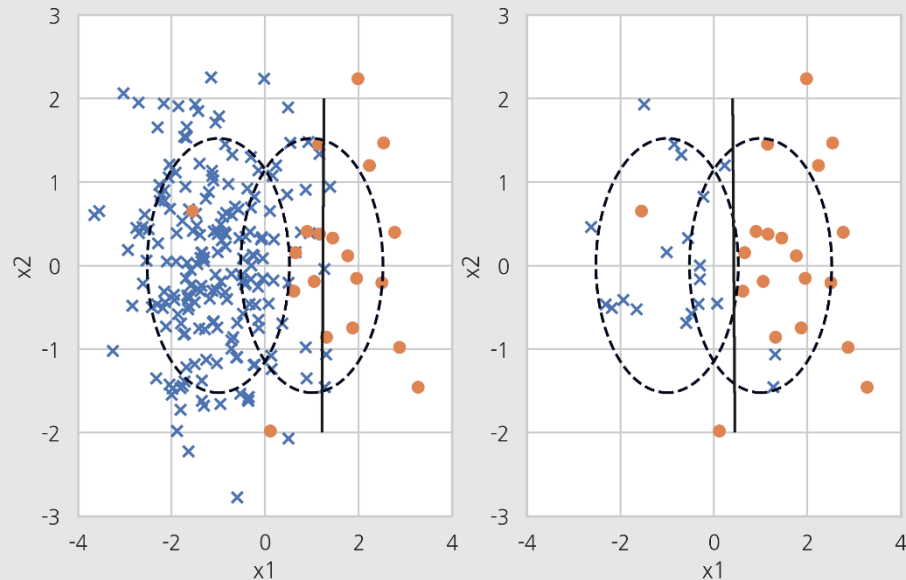




# Class Imbalances

## ✓ Down Sampling

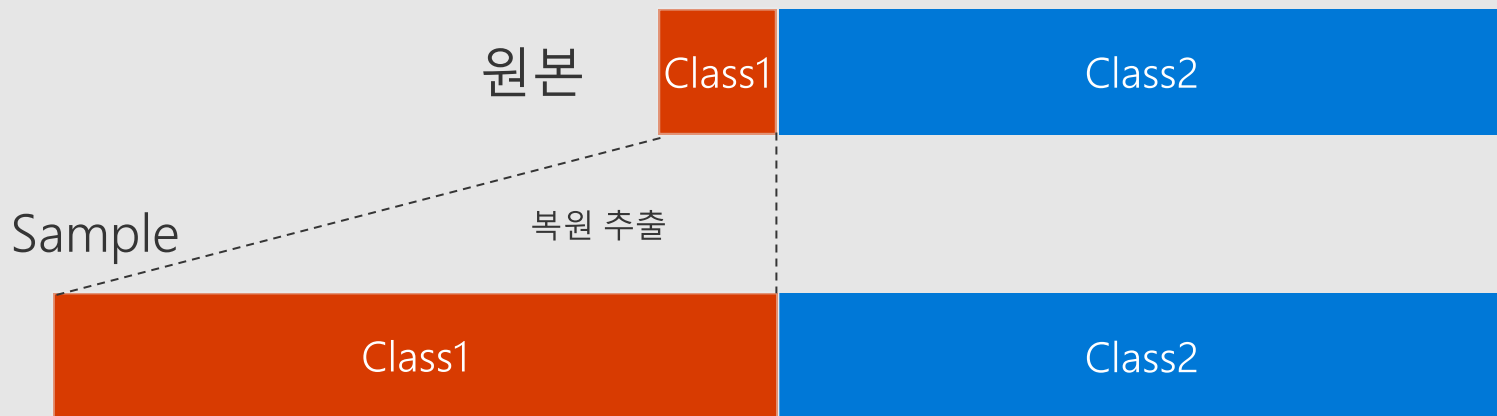
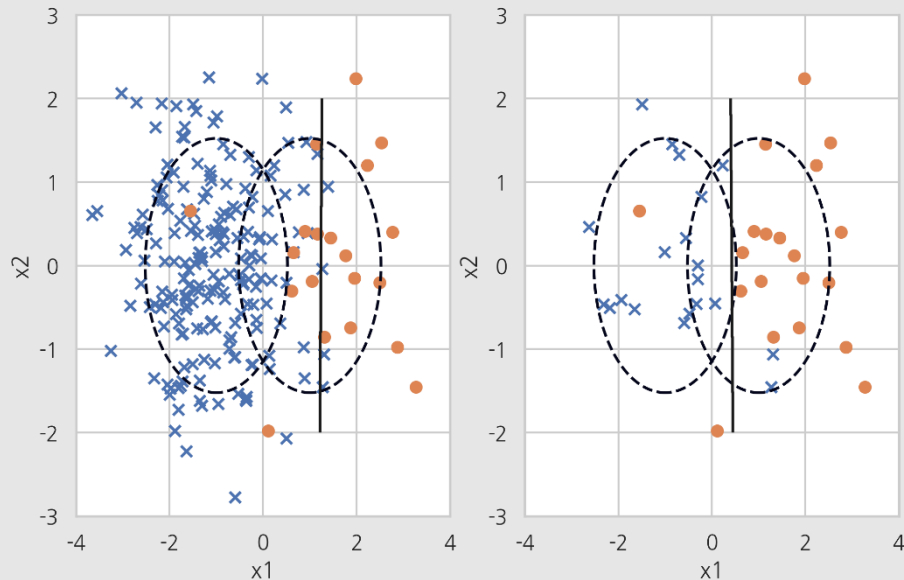
- 다수 Class의 데이터를  
소수 Class 수 만큼 random sampling



# Class Imbalances

## ✓ Up Sampling

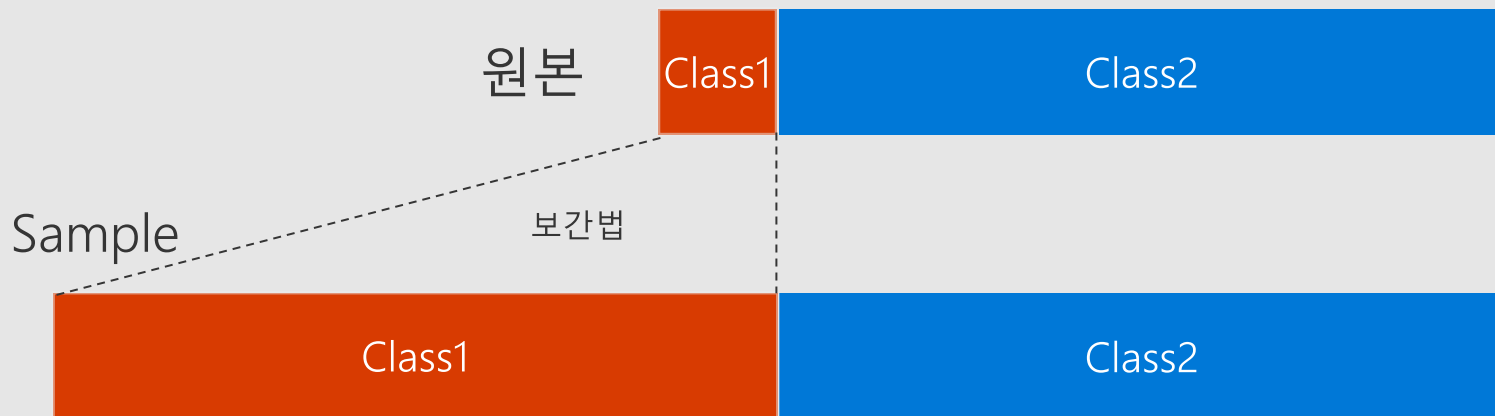
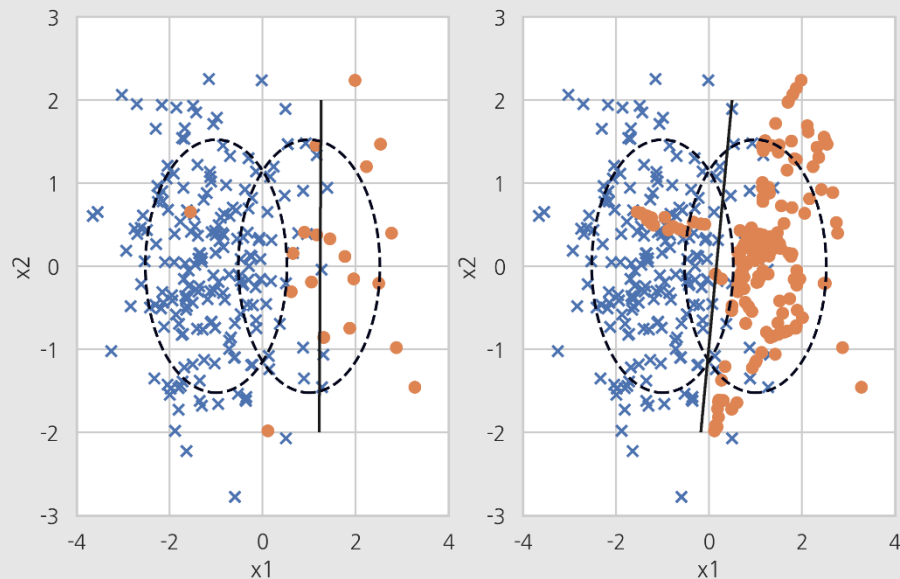
- 소수 Class의 데이터를  
다수 Class 수 만큼  
random sampling(복원 추출)



# Class Imbalances

## ✓ SMOTE(Synthetic Minority Oversampling TEchnique)

- 기존 소수 샘플을  
보간법(Interpolation)으로 새로운  
데이터를 만들어 냄



# 실습 : Class Imbalance & 모델링 연습

# Modeling Using Accuinsight Modeler

## ✓ 코드 구조

필요한 라이브러리 import

데이터 불러오기

데이터 준비

모델링 로깅 설정

모델링

# Modeling Using Accuinsight Modeler

✓ 어떻게 데이터를 가져올 것인가?

The screenshot displays the Accuinsight Modeler Pipeline interface. The top navigation bar includes a 'PIPELINE' dropdown menu (marked with a red box and a blue circle with the number 1), a 'Select Project' button, and links for '클러스터', '워크플로우', 'ML 관리', '브라우저', and '설정'. A dropdown menu for 'HDFS' (marked with a red box and a blue circle with the number 2) is open, showing options: HDFS, ICOS, Hive, and S3. Below the navigation bar, the main area is titled 'Pipeline' and features a '다양한 데이터를 예측 분석할 수 있는 배치/인터랙티브 방식의 Workflow Designer' link. The 'HDFS 브라우저' section shows a search bar with 'gkn\_cluster2' and a search button. A file list is displayed, including folders like 'apps', 'tmp', 'user', 'dpcore', 'hadoop', 'LabTest', 'kaggle', 'one-hot-multicolumn', 'pre\_processed\_data', and files like 'SUCCESS' and 'part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv' (marked with a red box and a blue circle with the number 3). A blue circle with the number 4 is positioned near a 'HDFS 경로 복사' button.

# Modeling Using Accuinsight Modeler

✓ 어떻게 데이터를 가져올 것인가?

Name	Last Modified
data_from_hdfs	16 minutes ago
filestorage	6 hours ago
runs	2 hours ago
sample	6 hours ago
sample-notebooks	41 minutes ago
1.Class Imbalance.ipynb	2 hours ago
2.샘플링과 모델링_연습.ipynb	2 hours ago
boston-house.ipynb	5 minutes ago
hdfs_FD_info.json	6 hours ago

## 2. 전처리된 데이터 다운로드

```
[2]: ## hdfs://172.31.10.128:8020/tmp/accu-edu01/dataset/boston.csv  
## hdfs://10.31.200.106:8020/user/hadoop/LabTest/pre_processed_data/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv
```

## 1 hdfs 경로 붙여넣기

```
[9]: hdfs_host = '10.31.200.106'  
hdfs_file_path = '/user/hadoop/LabTest/pre_processed_data/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv'
```

2

```
[7]: import json  
import os  
from collections import OrderedDict  
  
a = OrderedDict()  
a['host'] = hdfs_host  
a['port'] = '8020'  
a['filePath'] = hdfs_file_path  
a['target'] = 'isFraud'  
  
json_file_name = 'hdfs_FD_info.json'  
storage_info_json_path = os.path.join(os.getcwd(), json_file_name)  
storage_info_json_path  
  
with open(storage_info_json_path, 'w', encoding='utf-8') as save_file:  
    json.dump(a, save_file, indent='\t')
```

3

```
from Accuinsight.Lifecycle.tensorflow import accuinsight  
  
accu = accuinsight()  
  
accu.get_file('/home/work/hdfs_FD_info.json')  
  
Downloading file... part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv  
/home/work/data_from_hdfs/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000_20210426.csv
```

# Modeling Using Accuinsight Modeler

✓ 어떻게 데이터를 가져올 것인가?

Name	Last Modified
data_from_hdfs	21 minutes ago
filestorage	6 hours ago
runs	2 hours ago
sample	6 hours ago
sample-notebo...	an hour ago
1.Class Imbalan...	2 hours ago
2.샘플링과 모델...	2 hours ago
boston-house.i...	10 minutes ago
hdfs_FD_info.json	6 hours ago

2

```
import json
import os
from collections import OrderedDict

a = OrderedDict()
a['host'] = hdfs_host
a['port'] = '8020'
a['filePath'] = hdfs_file_path
a['target'] = 'isFraud'

json_file_name = 'hdfs_FD_info.json'
storage_info_json_path = os.path.join(os.getcwd(), json_file_name)
storage_info_json_path

with open(storage_info_json_path, 'w', encoding='utf-8') as save_file:
    json.dump(a, save_file, indent='\t')
```

[8]: from Accuinsight.Lifecycle.tensorflow import accuinsight

```
accu = accuinsight()

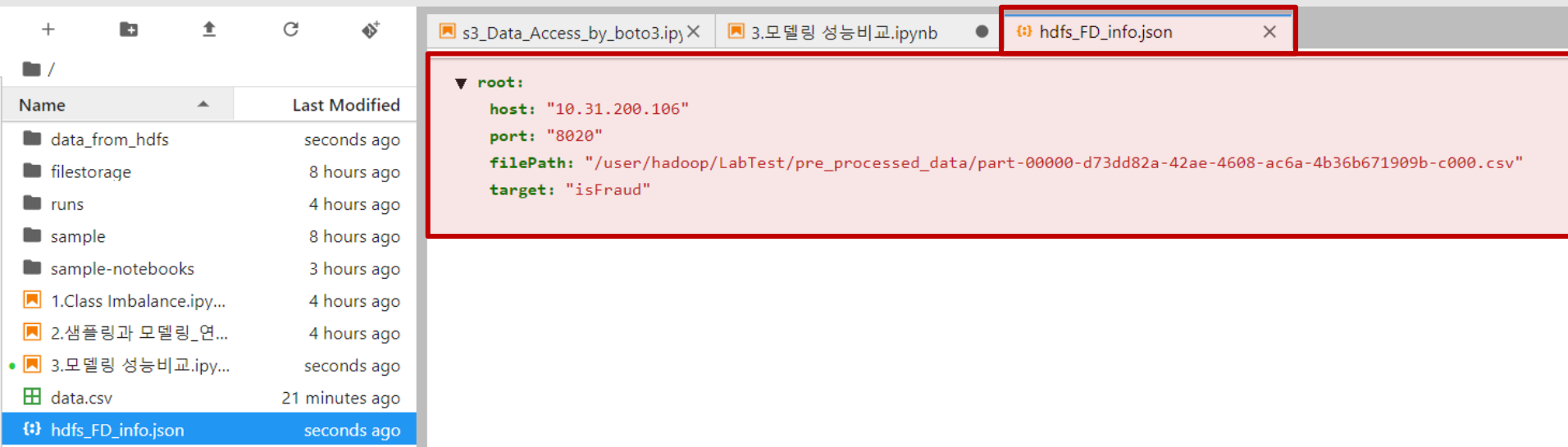
accu.get_file('/home/work/hdfs_FD_info.json')
```

Downloading file... part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv  
/home/work/data\_from\_hdfs/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000\_20210426.csv



# Modeling Using Accuinsight Modeler

✓ 어떻게 데이터를 가져올 것인가?



The screenshot displays the Accuinsight Modeler interface. On the left is a file explorer with a table of files and folders. On the right, a window titled 'hdfs\_FD\_info.json' is open, showing a JSON configuration for data access.

Name	Last Modified
data_from_hdfs	seconds ago
filestorage	8 hours ago
runs	4 hours ago
sample	8 hours ago
sample-notebooks	3 hours ago
1.Class Imbalance.ipynb	4 hours ago
2.샘플링과 모델링_연...	4 hours ago
3.모델링 성능비교.ipynb	seconds ago
data.csv	21 minutes ago
hdfs_FD_info.json	seconds ago

```
{
  "root": {
    "host": "10.31.200.106",
    "port": "8020",
    "filePath": "/user/hadoop/LabTest/pre_processed_data/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv",
    "target": "isFraud"
  }
}
```

# Modeling Using Accuinsight Modeler

## ✓ 어떻게 데이터를 가져올 것인가?

- 저장된 .json 파일을 기반으로 데이터 가져오기 : `accu.get_file`

```
storage_info_json_path = os.path.join(os.getcwd(), json_file_name)
storage_info_json_path

with open(storage_info_json_path, 'w', encoding='utf-8') as save_file:
    json.dump(a, save_file, indent='\t')

[8]: from Accuinsight.Lifecycle.tensorflow import accuinsight

accu = accuinsight()

accu.get_file('/home/work/hdfs_FD_info.json')

Downloading file... part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000.csv
/home/work/data_from_hdfs/part-00000-d73dd82a-42ae-4608-ac6a-4b36b671909b-c000_20210426.csv
```

- 파일 이름이 인식이 잘 안되는 문제가 있어서, 저장된 파일이름을 'data.csv'로 rename 합시다.

# Modeling Using Accuinsight Modeler

## ✓ 모델 저장과 성능 분석을 위한 로깅 설정

```
accu.autolog('boston-house', best_weights = True) # using model-monitor
```

```
model.fit(  
    normed_train, y_train,  
    epochs=10,  
    validation_data = (normed_valid, y_valid))
```

```
Using autolog(best_weights=True, model_monitor=True)
```

Epoch 1/10

10/10 [=====] - 1s 94ms/step - loss: 148.6967 - ma

Epoch 2/10

10/10 [=====] - 0s 34ms/step - loss: 20.8910 - mae

Epoch 3/10

< test\_ws 상태 RUNNING

요약 Experiment 작업폴더 계정폴더

run name을 입력하세요

Comparison Edit Columns 배포 저장소 이동

	Nº	Run ID	Executor	Update On	data	Version
<input type="checkbox"/>	1	tf.keras-07CE7DEC9E06483D...	trial_user03	2021-04-26 14:05:49		<a href="#">8fef655b00a0a1de6df0484705a77812f4c0465c</a>
<input type="checkbox"/>	2	tf.keras-181E29E79E6B444A...	trial_user03	2021-04-26 14:09:45		<a href="#">c1f8dc25131ce5288b5c93865f6ff2cd4117938f</a>

# 성능 평가

## ✓ Accuracy?

- Accuracy : 전체 중에서 맞춘 비율
- Class Imbalance 데이터에서 가장 주의해야 할 점!!!
- Accuracy는 무시해야 할 수도...

## ✓ 그렇다면 Precision? Recall?

- Precision 정밀도 : 사기거래라고 예측한 것들 중 맞춘 비율
- Recall 재현율 : 실제로 사기거래 중 맞춘 비율
- 무엇이 더 중요할까요?

# 성능 평가

## ✓ Confusion Matrix

Confusion Matrix

		$\hat{y}$	
		0	1
$y$	0	15052	997
	1	72	15879

$$✓ accuracy = \frac{15052 + 15879}{total}$$

## ✓ 1 입장에서...

$$▪ precision = \frac{15879}{72 + 15879}$$

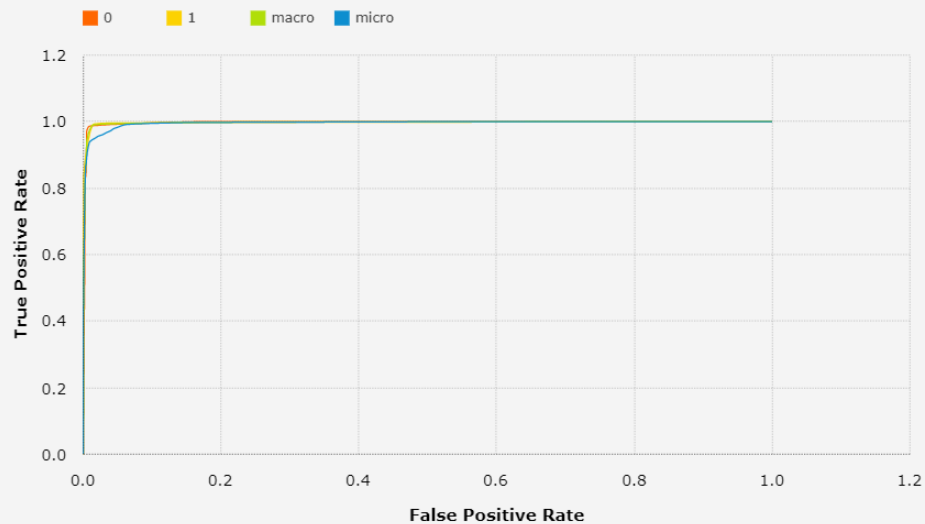
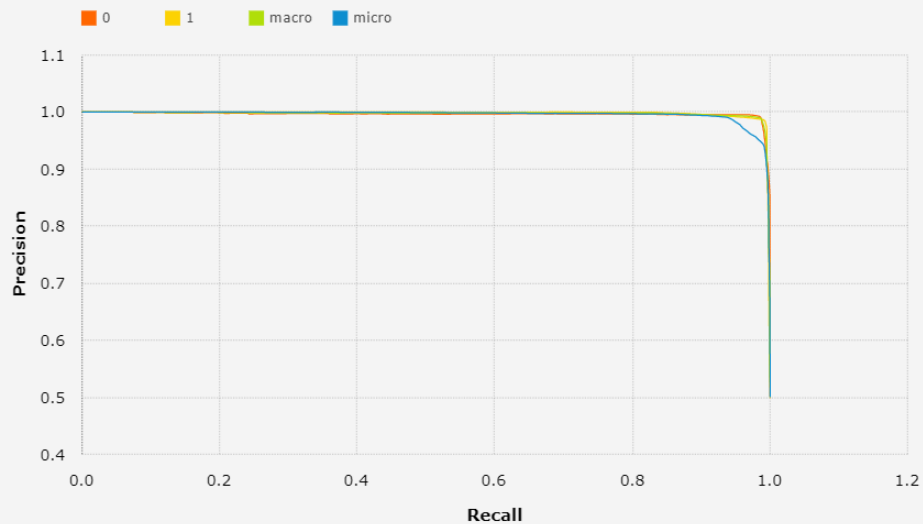
$$recall = \frac{15879}{997 + 15879}$$

# 성능 평가

## ✓ Precision-Recall Curve

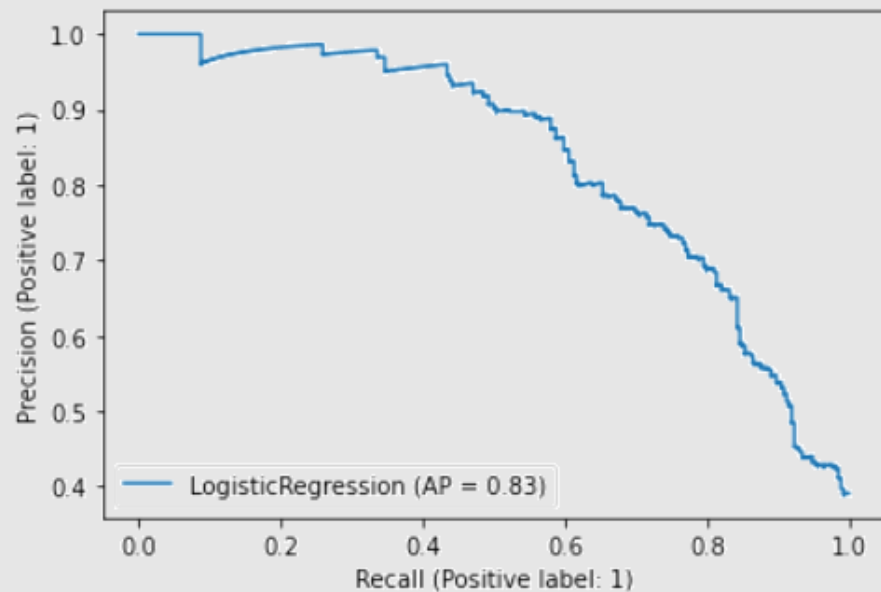
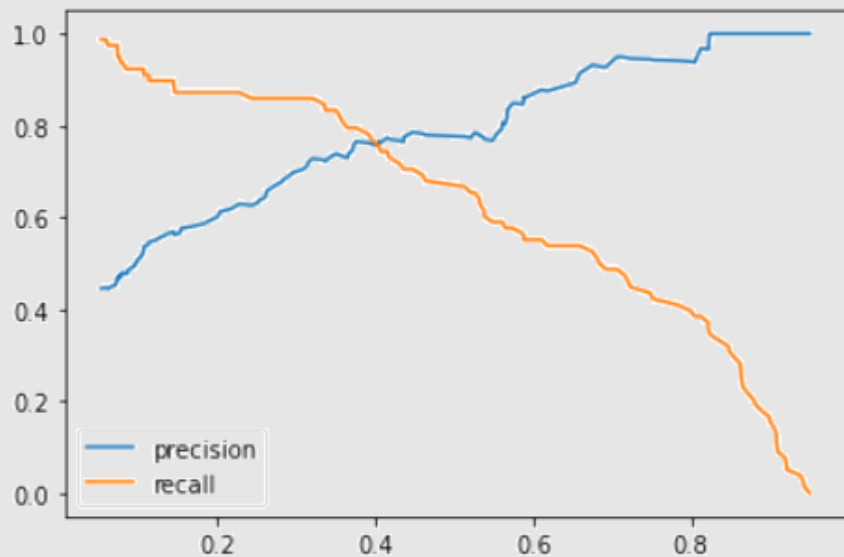
- 0 : 정상 거래 , 1 : 사기 거래
- macro : 산술평균, micro : 가중평균

Precision-Recall curve & ROC curve



# 성능 평가

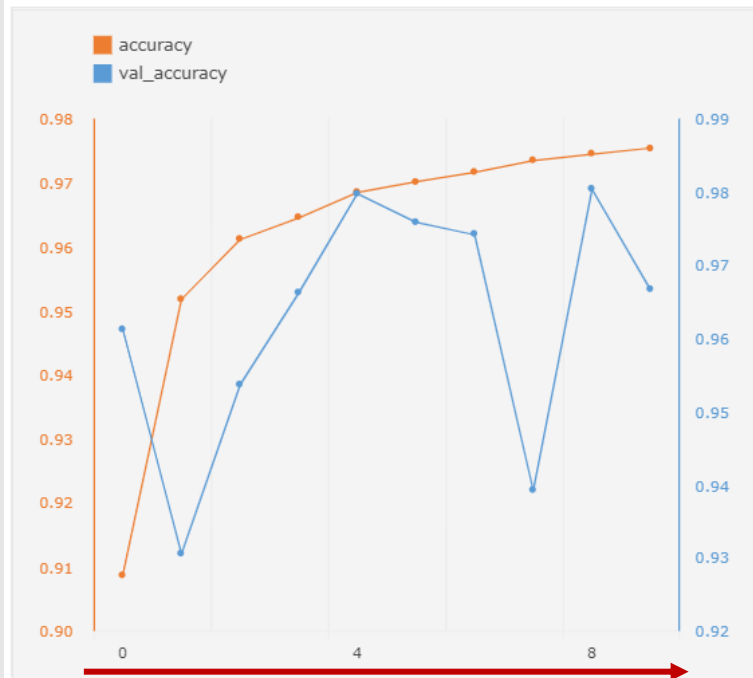
✓ Precision-Recall Curve



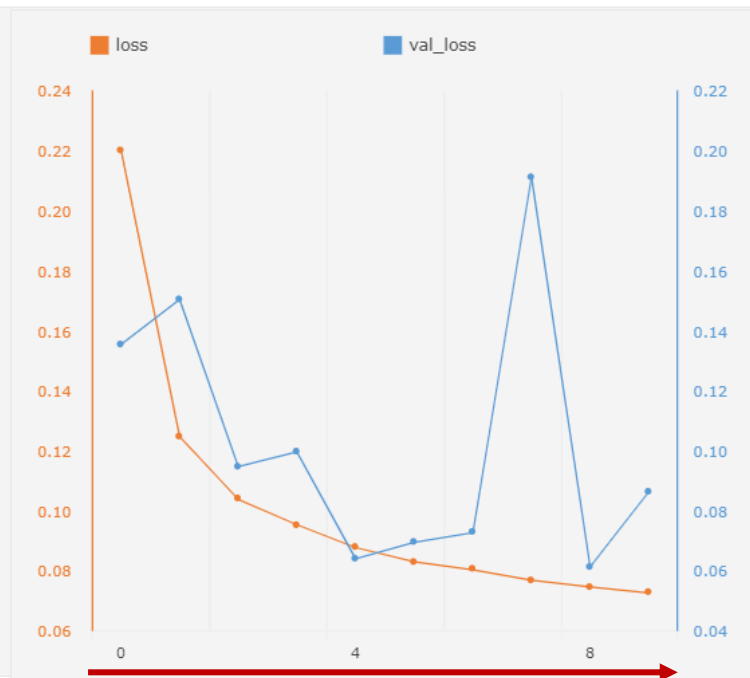
# 성능 평가

✓ Epochs 수를 더 늘리면?

accuracy / loss /



Epochs 횟수



Epochs 횟수







# 성능 평가 - 비교

✓로킹 된 모델들을 선택(최대 10개까지 가능)한 후, Comparision

홈 > 관리 > 워크스페이스 > 상

< test\_ws

상태 RUNNING




요약

Experiment

작업폴더

계정폴더

run name을 입력하세요



Comparison

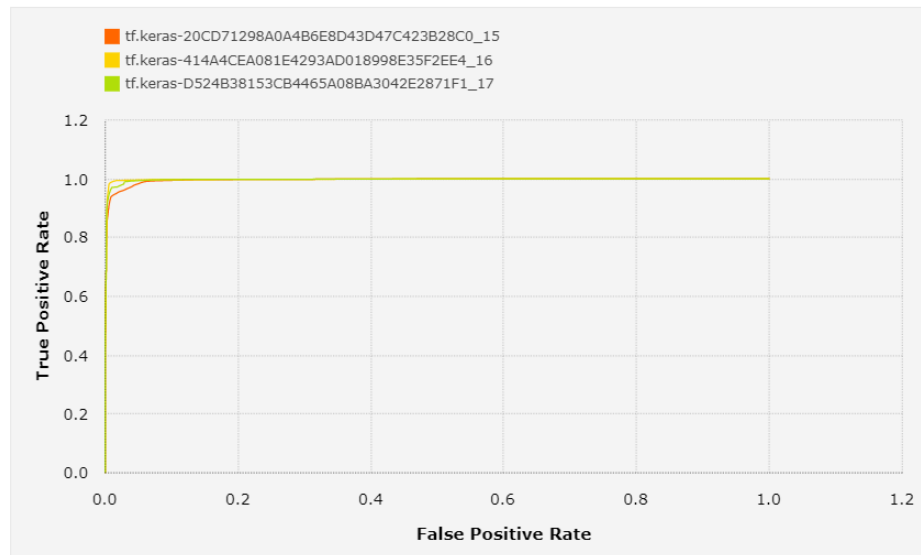
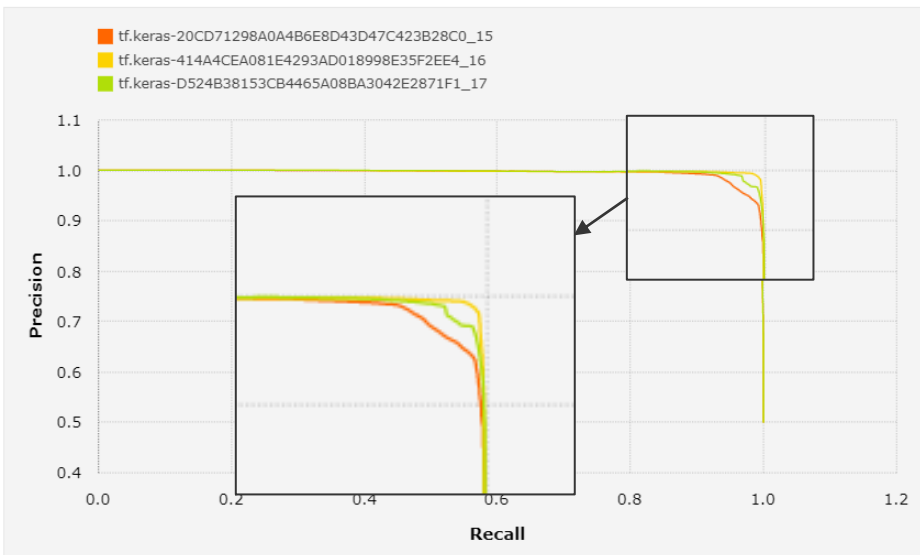
Edit Columns

배포 저장소 이동

	№	Run ID	Executor	Update On	Version	
					data	model
<input type="checkbox"/>	1	tf.keras-07CE7DEC9E06483D...	trial_user03	2021-04-26 14:05:49		<a href="#">8fef655b00a0a1de6df0484705a77812</a>
<input type="checkbox"/>	2	tf.keras-181E29E79E6B444A...	trial_user03	2021-04-26 14:09:45		<a href="#">c1f8dc25131ce5288b5c93865f6ff2cd</a>
<input checked="" type="checkbox"/>	3	tf.keras-20CD71298A0A4B6E...	trial_user03	2021-04-26 18:27:18	<a href="#">hdfs://10.31.200.106/8020/user/hadoop/LabTes...</a>	<a href="#">f5dc3b08bfc3f47a7e8ad84c025e99f1c</a>
<input checked="" type="checkbox"/>	4	tf.keras-414A4CEA081E4293...	trial_user03	2021-04-26 19:24:06	<a href="#">hdfs://10.31.200.106/8020/user/hadoop/LabTes...</a>	<a href="#">a4b8c89c17954cb54a2515ef71a17e94</a>
<input checked="" type="checkbox"/>	5	tf.keras-D524B38153CB4465...	trial_user03	2021-04-26 19:28:20	<a href="#">hdfs://10.31.200.106/8020/user/hadoop/LabTes...</a>	<a href="#">8615a827210af9c692b86e867cf4f0ae</a>

# 성능 평가 - 비교

## Model Comparison(Precision-Recall curve & ROC curve)



## Metrics

Model Version	AUC	Precision	Recall	F1-score
<a href="#">tf.keras-20CD71298A0A4B6E8D43D47C423B28C0_15</a>	0.9957880624999976	0.9958331841897203	0.96659375	0.96659375
<a href="#">tf.keras-414A4CEA081E4293AD018998E35F2EE4_16</a>	0.9976663749999992	0.9976494484769797	0.9895625	0.9895625
<a href="#">tf.keras-D524B38153CB4465A08BA3042E2871F1_17</a>	0.9973684999999988	0.9973862921872153	0.9770625	0.9770625

\* AUC : area under curve ,

\* F1-score : Precision과 Recall의 조화평균

# 실습

- ✓ 다양한 파라미터 값으로 최소 5개 이상의 모델을 생성합니다.
- ✓ Accuinsight Modeler에서 Experiment로 이동 > 모델들의 성능을 비교해 봅시다.
- ✓ 주의사항: 모델 생성(.fit) 할 때 마다, 사전에 반드시 `accu.autolog()`를 실행해 줘야 로그가 남습니다.
- ✓ 5개 이상의 모델 성능 비교에 대해서 스크린샷을 28일 18시까지 보내주셔야 합니다~!!!!

# 성능 평가

- ✓ 그러나 우리는 항상 Machine Learning Metric을 Business Metric으로 변환해서 평가할 수 있어야 한다.
- ✓ 그것이 우리가 지금 이 일을 하는 목적!