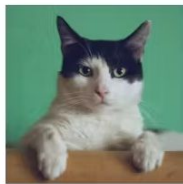


What is Imbalanced Classification



Nội dung

1. What is Imbalanced Classification?
2. Intuition for Imbalanced Classification
3. Challenge of Imbalanced Classification

Competition



MultiEarth
2023

MultiEarth 2023

MultiEarth 2022



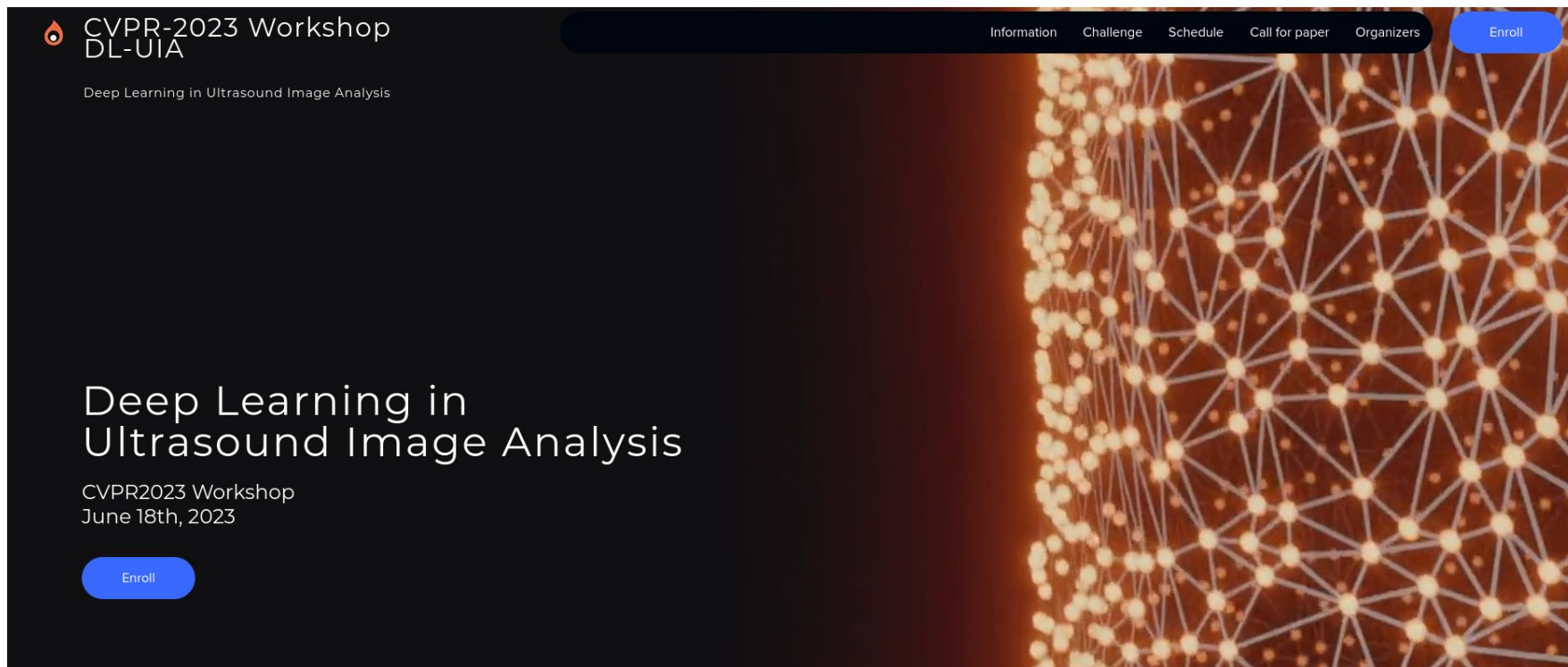
Mission of the workshop

The Multimodal Learning for Earth and Environment Workshop (MultiEarth 2023) is the second annual CVPR workshop aimed at leveraging the significant amount of remote sensing data that is continuously being collected to aid in the monitoring and analysis of the health of Earth ecosystems. The goal of the workshop is to bring together the Earth and environmental science communities as well as the multimodal representation learning communities to examine new ways to leverage technological advances in support of environmental monitoring. In addition, through a series of public challenges, the MultiEarth Workshop hopes to provide a common benchmark for remote sensing multimodal information processing. These challenges are focused on the monitoring of the Amazon rainforest and include deforestation estimation, fire detection, cross-modal image translation, and environmental change projection.

①

[MultiEarth 2023](#)

Competition



CVPR-2023 Workshop
DL-UIA

Deep Learning in Ultrasound Image Analysis

Deep Learning in
Ultrasound Image Analysis

CVPR2023 Workshop
June 18th, 2023

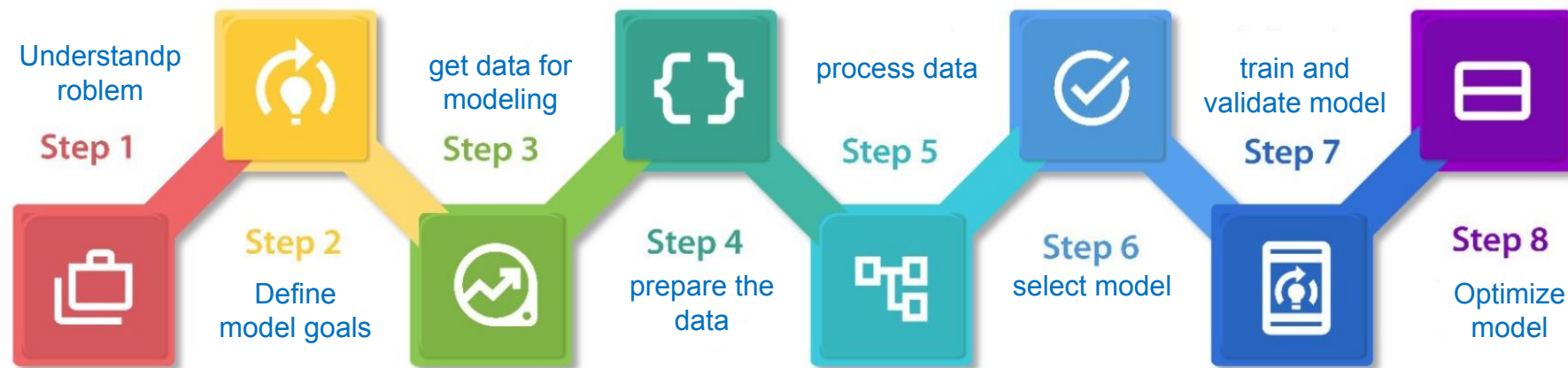
Enroll

Information Challenge Schedule Call for paper Organizers Enroll

[Deep Learning in Ultrasound Image Analysis](#)

1 - What is Imbalanced Classification

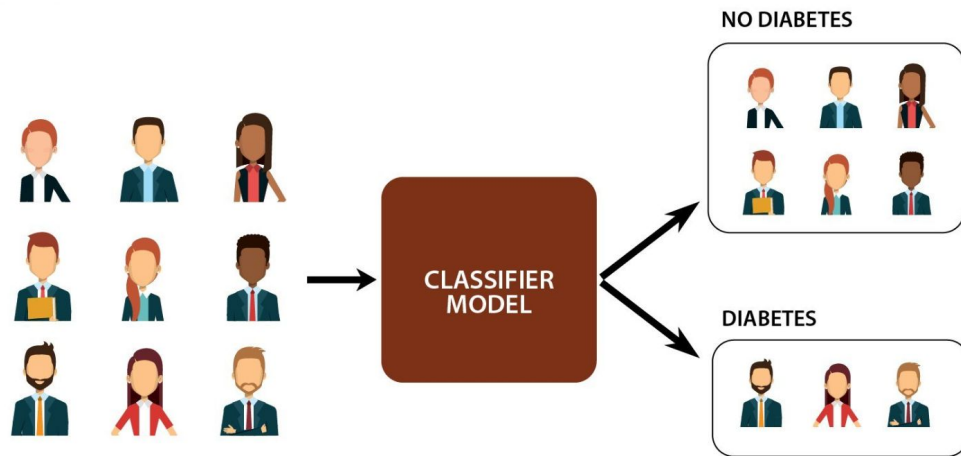
Classification Predictive Modeling



8 bước cho một Predictive Modeling Pipeline

1 - What is Imbalanced Classification

Classification Predictive Modeling

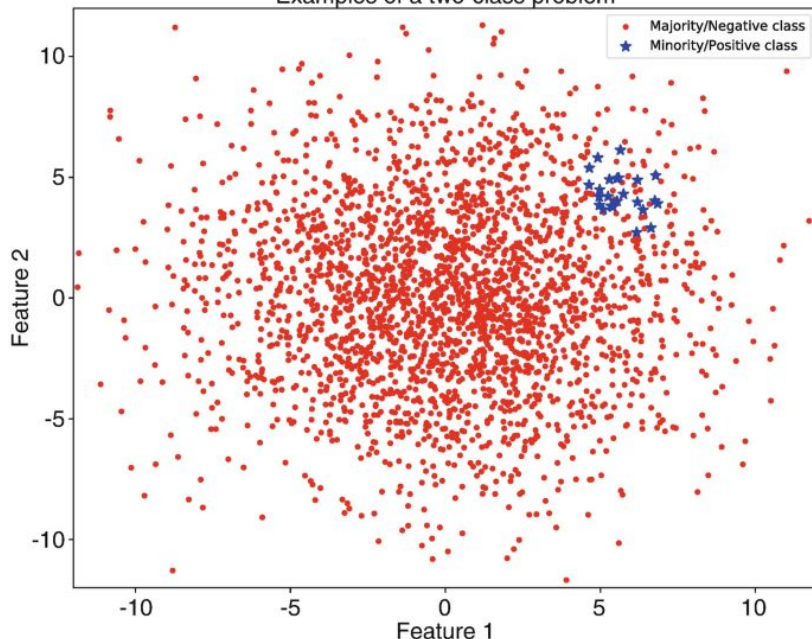


Classification là một Predictive Modeling liên quan đến việc gán nhãn cho mỗi data point/observation

1 - What is Imbalanced Classification

Imbalanced Classification Problems

Examples of a two-class problem



Imbalance data là một thách thức cho predictive modeling vì hầu hết các thuật toán sử dụng để phân loại đều được thiết kế dựa trên giả định về số lượng bằng nhau cho mỗi class. Điều này dẫn đến các mô hình có hiệu suất dự đoán kém, đặc biệt đối với những class thiểu số.

1 - What is Imbalanced Classification

Imbalanced Classification Problems

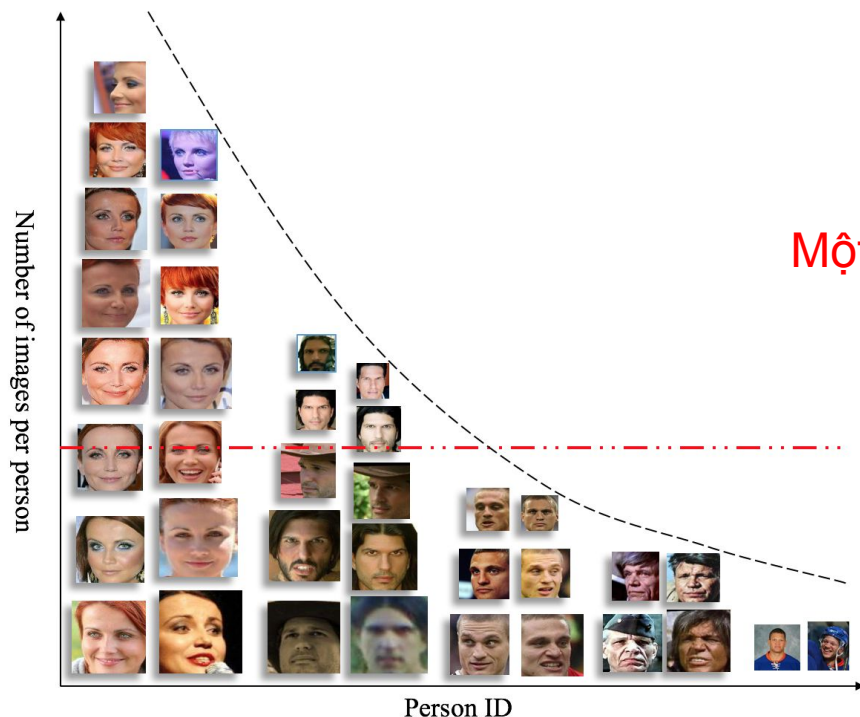
Vấn đề lệch dữ liệu giữa các class được gọi là **Imbalanced** thay vì Unbalanced là vì:

- Unbalanced đề cập đến việc class distribution đã được cân bằng và hiện không còn cân bằng nữa.
- Imbalanced đề cập đến việc class distribution vốn đã không cân bằng.

Ví dụ: Cho bài toán imbalanced binary classification với sự mất cân bằng từ 1 đến 100 (1:100) có nghĩa là cứ mỗi một sample ở một class này thì có 100 samples ở class kia

1 - What is Imbalanced Classification

Imbalanced Classification Problems



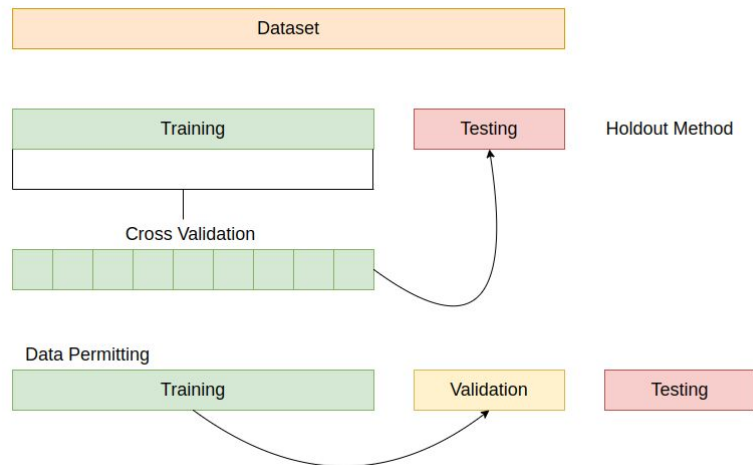
Long-tail Problem
Một loại bài toán imbalanced phức tạp hơn

1 - What is Imbalanced Classification

Causes of Class Imbalance

Hai nguyên nhân chính dẫn đến sự mất cân bằng của class distribution:

- Data sampling
- Properties of the domain



1 - What is Imbalanced Classification

Challenge of Imbalanced Classification

Imbalanced Classification có 2 dạng chính: Slight Imbalance và Severe Imabalance

- Slight Imbalance: distribution của các mẫu dữ liệu có sự phân bố không đều ở một lượng nhỏ trong dataset (ví dụ lệch 4:6)
- Severe Imbalance: distribution của các mẫu dữ liệu có sự phân bố không đều ở số lượng lớn trong dataset (ví dụ lệch 1:100 hoặc nhiều hơn)

Class có nhiều mẫu dữ liệu hơn được gọi là **major class**, ngược lại gọi là **minor class**

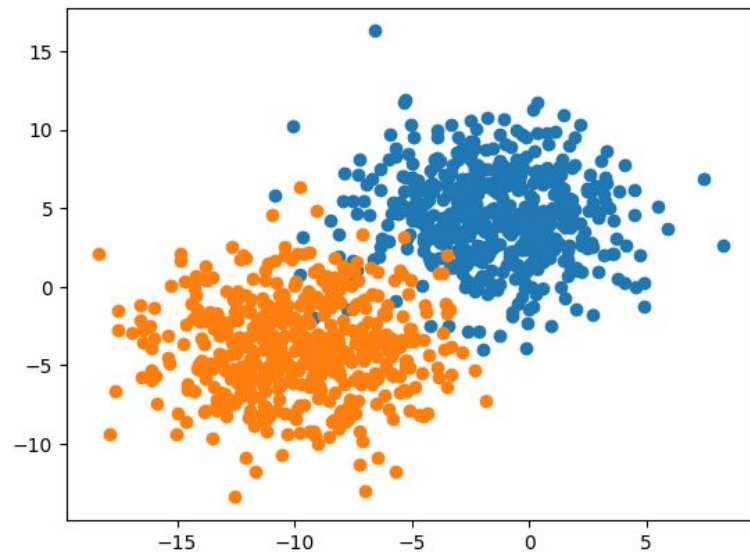
2 - Intuition for Imbalanced Classification

Create and Plot data:

```
# generate dataset
X, y = make_blobs(n_samples=1000, centers=2, n_features=2, random_state=1, cluster_std=3)

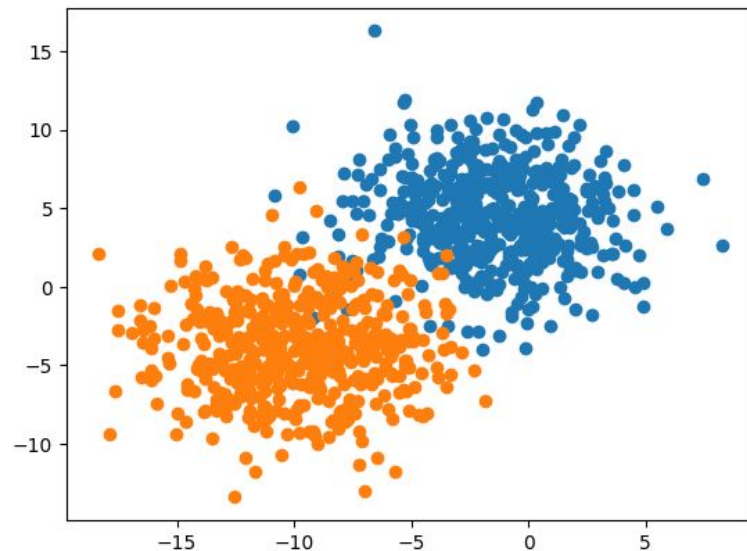
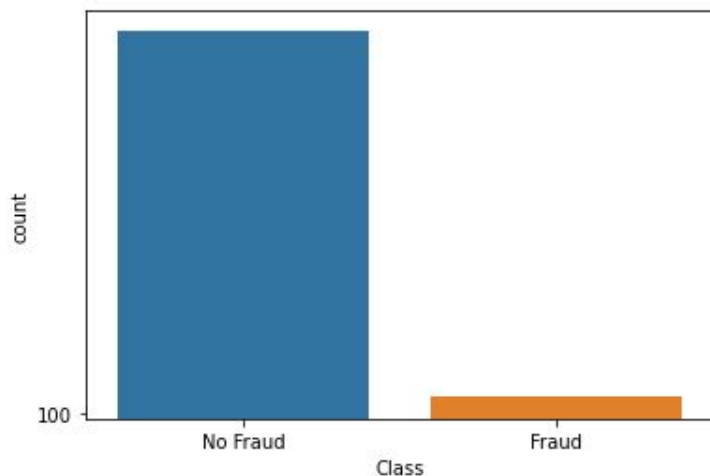
# create scatter plot for samples from each class
for class_value in range(2):
    # get row indexes for samples with this class
    row_ix = np.where(y == class_value)
    # create scatter of these samples
    plt.pyplot.scatter(X[row_ix, 0], X[row_ix, 1])

plt.pyplot.show()
```



2 - Intuition for Imbalanced Classification

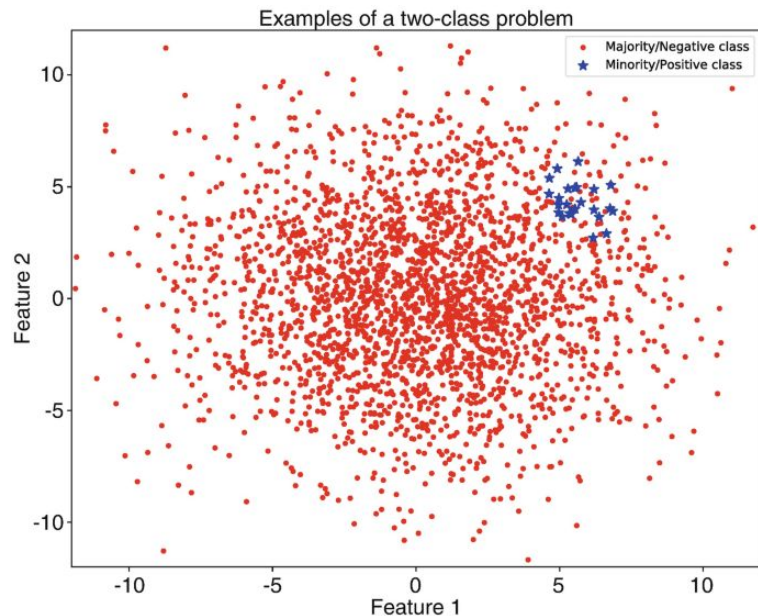
Create and Plot data:



Sự khác biệt khi kiểm tra imbalanced data giữa 2 loại biểu đồ là gì?

2 - Intuition for Imbalanced Classification

Create and Plot data:



Sử dụng scatter giúp chúng ta có cái nhìn tổng quan về phân phối data để có thể lựa chọn những giải pháp phù hợp

2 - Intuition for Imbalanced Classification

Create Synthetic Dataset with a Class Distribution

```
# create a dataset with a given class distribution
def get_dataset(proportions):
    n_classes = len(proportions)
    largest = max([v for k,v in proportions.items()])
    n_samples = largest * n_classes

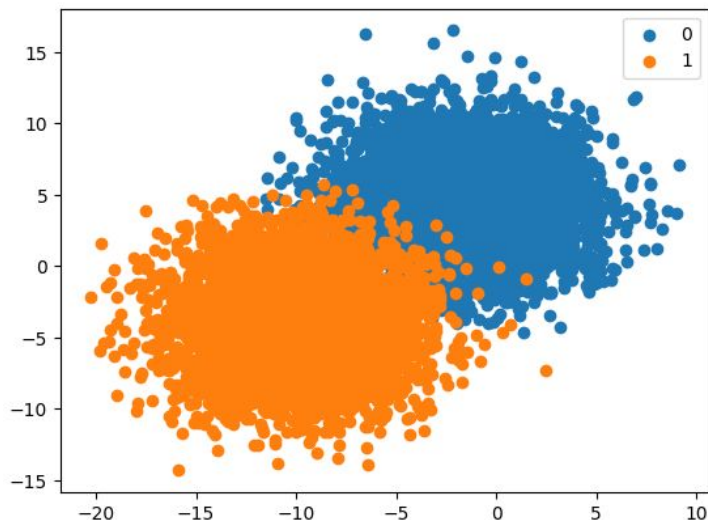
    # create dataset
    X, y = make_blobs(n_samples=n_samples, centers=n_classes, n_features=2, random_state=1, cluster_std=3)

    # collect the examples
    X_list, y_list = list(), list()
    for k,v in proportions.items():
        row_ix = np.where(y == k)[0]
        selected = row_ix[:v]
        X_list.append(X[selected, :])
        y_list.append(y[selected])

    return np.vstack(X_list), np.hstack(y_list)

def plot_dataset(X, y):
    n_classes = len(np.unique(y))
    for class_value in range(n_classes):
        row_ix = np.where(y == class_value)[0]
        # create scatter of these samples
        plt.pyplot.scatter(X[row_ix, 0], X[row_ix, 1], label=str(class_value))

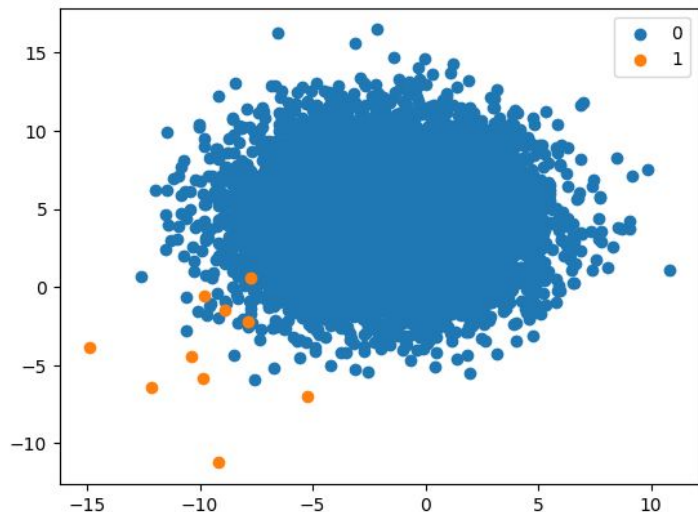
    plt.pyplot.legend()
    plt.pyplot.show()
```



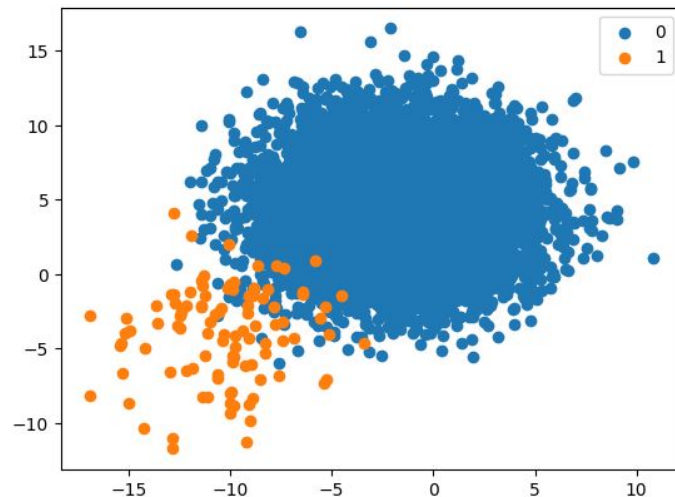
2 - Intuition for Imbalanced Classification

Effect of Skewed Class Distributions

```
# define the class distribution
proportions = {0:10000, 1:10}
# generate dataset
X, y = get_dataset(proportions)
# plot dataset
plot_dataset(X, y)
```



```
# define the class distribution
proportions = {0:10000, 1:100}
# generate dataset
X, y = get_dataset(proportions)
# plot dataset
plot_dataset(X, y)
```



3 - Challenge of Imbalanced Classification

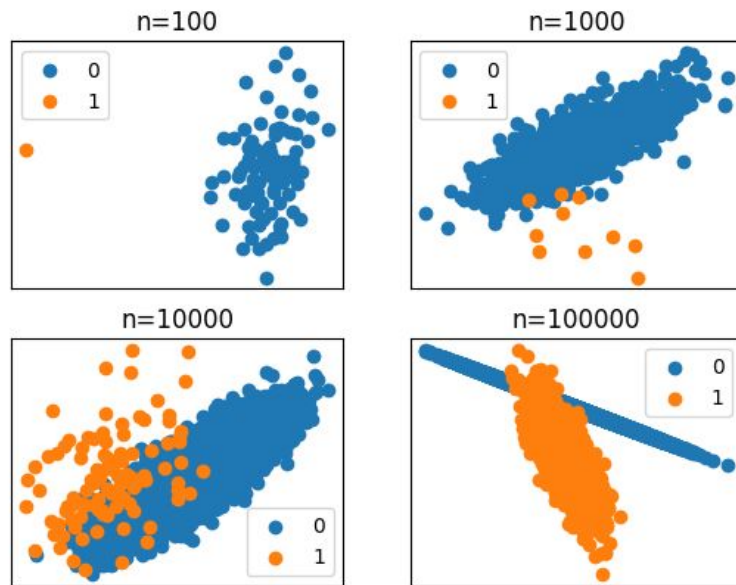
Compounding Effect of Dataset Size

```
for i in range(len(sizes)):
    n = sizes[i]

    # create the dataset
    X, y = make_classification(n_samples=n, n_features=2, n_redundant=0,
                              n_clusters_per_class=1, weights=[0.99], flip_y=0, random_state=1)

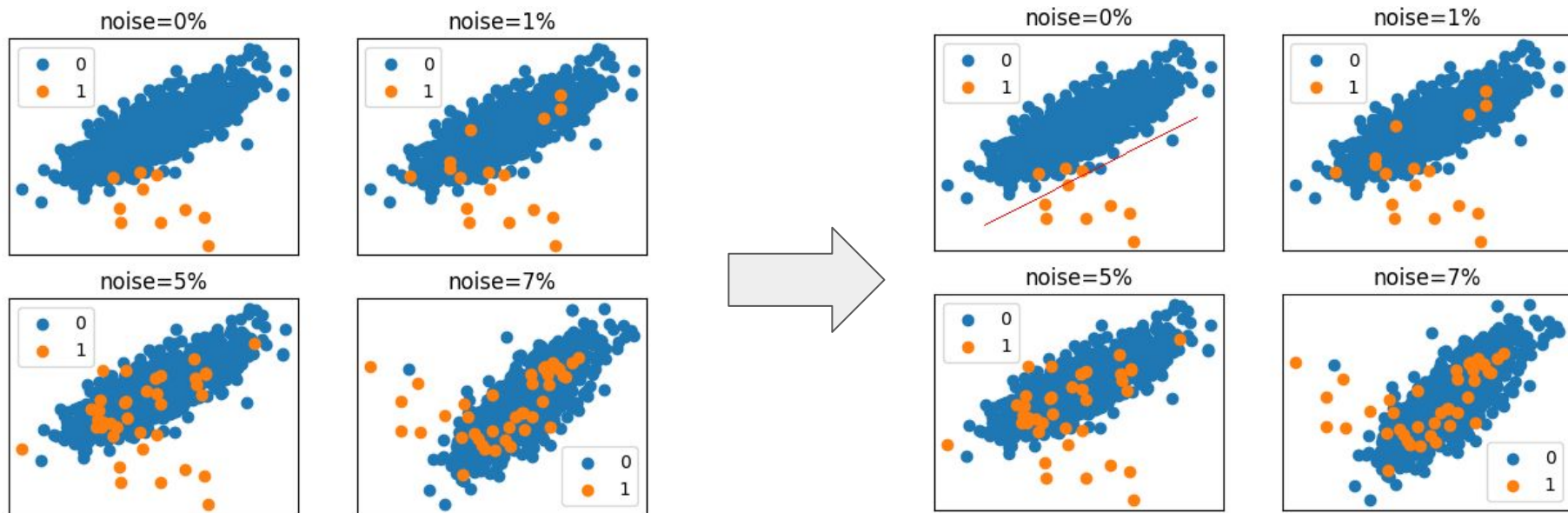
    # summarize class distribution
    counter = Counter(y)
    print('Size=%d, Ratio=%s' % (n, counter))
    plt.pyplot.subplot(2, 2, 1+i)
    plt.pyplot.title('n=%d' % n)
    plt.pyplot.xticks([])
    plt.pyplot.yticks([])

    for label, _ in counter.items():
        row_ix = np.where(y == label)[0]
        plt.pyplot.scatter(X[row_ix, 0], X[row_ix, 1], label=str(label))
    plt.pyplot.legend()
```



3 - Challenge of Imbalanced Classification

Compounding Effect of Label Noise



Việc có thêm noise sẽ làm cho việc phân tách giữa các lớp trở nên khó khăn hơn.