IPython

Setting by using Anaconda

Link download: <https://www.anaconda.com/download/#windows>

Pandas: Dataframe để lưu trữ và làm việc

Matplotlib: Vẽ sơ đồ = python

Seaborn: Thống kê

NumPy: Làm việc với số

Scikit-learn: Gói học máy cần thiết cho python

Bước 1: Trả lời câu hỏi

Bước 2: Kiểm tra dữ liệu

Bước 3: Dọn dẹp dữ liệu

Đảm bảo dữ liệu được mã hóa đúng cách

Dữ liệu nằm trong phạm vi dự kiến

Có thể thay thế hoặc xóa bỏ các dữ liệu bị thiếu

Sử dụng assert để kiểm tra trước khi làm sạch

Bước 4 : Phân tích thăm dò

Bước 5: Phân loại

Bước 6: Tái tạo

# scikit-learn Tutorials

## Machine learning: the problem setting

Xem xét một tập các mẫu dữ liệu và sau đó cố gắng dự đoán thuộc tính của một mẫu dữ liệu không xác định. Nếu mỗi mẫu dữ liệu có nhiều hơn 1 biến thì nó được cho là có một số thuộc tính( tính năng).

Học có giám sát

Phân loại: Các mẫu đã được dán nhãn và công việc cần làm là xác định nhãn cho mẫu dữ liệu chưa biết.

Hồi quy: Dự đoán đầu ra dựa vào đầu vào, vd dự đoán chiều dài của cá hồi dựa vào tuổi và chiều dài của nó

Học không giám sát

-Training set và testing set

## Loading an example dataset

from sklearn import datasets

iris = datasets.load\_iris()

digits = datasets.load\_digits()

Dữ liệu luôn luôn ở dạng mảng 2 chiều