**THỰC HÀNH 1**

1. **Nguồn lấy dữ liệu**
2. UCI: https://archive.ics.uci.edu/ml/datasets.php
3. Kaggle: <https://www.kaggle.com/datasets>
4. **Thực hành khám phá dữ liệu với MS Excel**

**NỘI DUNG 1:**

1. Download **02 bộ** dữ liệu từ các nguồn (phía trên) dưới dạng file: .csv, .txt, .data,… (ghi rõ link đến bộ dữ liệu đó)
2. Mở excel, load dữ liệu vào (data/import) (mỗi bộ dữ liệu trên một sheet)
3. Thực hành khám phá dữ liệu với Excel:

* Tên bộ dữ liệu
  + Có bao nhiêu mẫu (samples? (số hàng), Có bao nhiêu lớp (classes)?
  + Có bao nhiêu thuộc tính (feature)? (số cột)
  + Các thuộc tính của dữ liệu có những kiểu gì?
  + Dữ liệu có bị mất mát (missing value) không? Thiếu ở thuộc tính nào?

**NỘI DUNG 2: Chọn 1 trong 2 bộ dữ liệu trên và thực hiện:**

1. Các thống kê cơ bản: min, max, mean, độ lệch chuẩn (stdev),… cho tất cả các cột số
2. Vẽ biểu đồ tương quan
   * 1. Thống kê cho toàn bộ dữ liệu
     2. Thống kê cho từng lớp
3. Chuẩn hóa dữ liệu (sử dụng các cách khác nhau)
4. Đo độ tương đồng của dữ liệu: Độ đo Euclide
   * 1. Tạo trang tính mới, thực hành độ đo với 2 cột (thuộc tính) số khác nhau
5. Nếu thuộc tính là giá trị liên tục 🡪 rời rạc hóa
6. Dữ liệu thiếu: (nếu tập dl không có dl thiếu 🡪 xóa bớt 1 số giá trị) 🡪 đề xuất hướng giải quyết (điền dl thiếu???)

**Yêu cầu nộp bài:**

* **SV cần nộp ít nhất file excel trong đó trình bày các nội dung SV đã thực hiện khám phá dữ liệu. SV có thể nộp kèm theo file word trong file giải thích các công việc SV đã thực hiện.**
* **Tất cả tên file đặt theo định dạng: Mã SV\_tên SV\_TH1 (ví dụ: 661234\_Nguyễn Văn A\_TH1)**