Thực hành data mining

TH1: Khám phá một số bộ dữ liệu

1. Nguồn dữ liệu có sẵn

- UCI
  - **https://archive.ics.uci.edu/ml/datasets.php**

- Kaggle
  - **https://www.kaggle.com/datasets**

Yêu cầu 1:

1. Tìm kiếm bộ dữ liệu với các định dạng khác nhau: .csv, .txt, .data,…
2. Đọc thông tin mô tả về bộ dữ liệu
3. Tải file dữ liệu về và lưu ở các thư mục làm việc khác nhau của mình

2. Khám phá dữ liệu với Python

2.1. Python

❖Giới thiệu Python:

- https://youtu.be/HvVdgcLl9rc
- https://youtu.be/NZj6LI5a9vc

❖Hướng dẫn cài Python:

- https://www.youtube.com/watch?v=g5BdrxPhQU0&ab_channel=CodeXplore
- https://machinelearningcoban.com/faqs/

Cài đặt python và các thư viện trên Windows

- Cài đặt Python bằng Anaconda
  - Anaconda hỗ trợ rất nhiều thư viện giúp lập trình Python.
  - Để tải về Python và một số thư viện cần thiết, tải về Anaconda cho windows và cài đặt

  https://www.anaconda.com/download/success

Kiểm tra Libs

- Anaconda đã có sẵn khá là nhiều thư viện python như: Numpy, Scipy, Matplotlib, sklearn, pandas
- Để kiểm tra python của Anaconda đã có thư viện nào đó, chúng ta sẽ thử import nó trong Console.

```
>>> import numpy
>>> import sklearn
```

Cài đặt Libs bằng Anaconda

- Chúng ta mở cmd (Command Prompt) của windows gõ lệnh:
  - `conda install scikit-learn` hoặc
  - `pip install -U scikit-learn`
- Conda sẽ tự động tìm thư viện *sklearn* và cài vào đường dẫn Anaconda giúp chúng ta.

Jupyter Notebook (anaconda3)

- Thay đổi thư mục làm việc
- Cách viết, chạy chương trình
- Tham khảo:

https://www.youtube.com/watch?v=rmG994uf5m4&ab_channel=TaiChinhAcademy

2.2. Khám phá dữ liệu với Python

Tham khảo:

- https://www.youtube.com/watch?v=HPGYTWYM13s
- https://kungfupandas.lhduc.com/gi%E1%BB%9Bi-thi%E1%BB%87u-pandas.html

- **Yêu cầu 2:** thực hành trên python, sử dụng: numpy, pandas,… và ghi lại kết quả vào file word:

1. Đọc file dữ liệu, lưu vào biến **(tên biến đặt theo tên sv)**
   - Đọc các định dạng file khác nhau
   - File dữ liệu đặt ở cùng/khác thư mục với file code
2. Xem 10 dòng đầu, 10 dòng cuối của dữ liệu
3. Xem kích thước của dữ liệu
4. Liệt kê tên các cột của dữ liệu
5. Xem kiểu của các thuộc tính (các cột)
6. Xem thông tin thống kê cho các cột có kiểu dữ liệu là số
7. Xem thông tin thống kê cho tất cả các cột
8. Đếm số giá trị trong cột
9. Truy xuất vào cột/các cột dữ liệu
10. Tính các giá trị thống kê cho cột/các cột
11. Lọc dữ liệu