## **BÁO CÁO THỰC HÀNH TUẦN 7**

## **KHAI THÁC DỮ LIỆU**

*Họ và tên : Nguyễn Tiến Phong*

*MSSV : 20280071*

**Phương pháp Bottom-up Agglomerative :**

* **Đọc dữ liệu và xử lý :** 
  + Dữ liệu dùng để thực hiện thuật toán chỉ có 2 cột là :

Annual Income (k$)

Spending Score (1-100)

* **Vẽ Dendogram để xác định số cụm :**

Sử dụng thư viện *scipy* để tính toán ma trận liên kết đơn và vẽ đồ thị *dendogram*.

*(Em cũng đã thử viết ma trận liên kết đơn nhưng khi vẽ dendogram thì lại không được.)*

A picture containing text, diagram, parallel, screenshot

Description automatically generatedDựa vào *Dendogram*, em chia tập dữ liệu thành 7 cụm.

* **Các bước thực hiện phương pháp Bottom-up Agglomerative :**
  + Hàm tính khoảng cách Euclidean giữa 2 điểm dữ liệu

Tính khoảng cách Euclidean bằng hàm *norm* trong thư viện *numpy*

* + Hàm tính ma trận khoảng cách

Tính ma trận khoảng cách giữa tất cả các cặp điểm dữ liệu và biểu diễn dưới dạng ma trận 2 chiều.

* + Hàm thuật toán Bottom-up Agglomerative
    - Khởi tạo mỗi điểm trong tập dữ liệu là một cụm
    - Tìm cặp cụm có khoảng cách gần nhất trong ma trận khoảng cách và lấy vị trí
    - Gộp cặp cụm đã lấy vị trí vào cụm thứ nhất và xóa cụm thứ 2
    - Lặp lại cho đến khi số cụm bằng k
* **A picture containing screenshot, text, diagram, colorfulness

  Description automatically generatedXuất label và vẽ scatterplot và so sánh với khi sử dụng thư viện**

A picture containing text, screenshot, diagram, font

Description automatically generated

* **Nhận xét :** 
  + Kết quả tương tự nhau
  + Sử dụng Single Linkage thì có xu hướng xuất hiện các cụm kéo dài một chuỗi, các cụm rời rạc và các cụm không đồng nhất về kích thước.
  + Sử dụng các phương pháp khác như Ward thì sẽ có kết quả tốt hơn.