**BÁO CÁO PHÂN CỤM DỮ LIỆU**

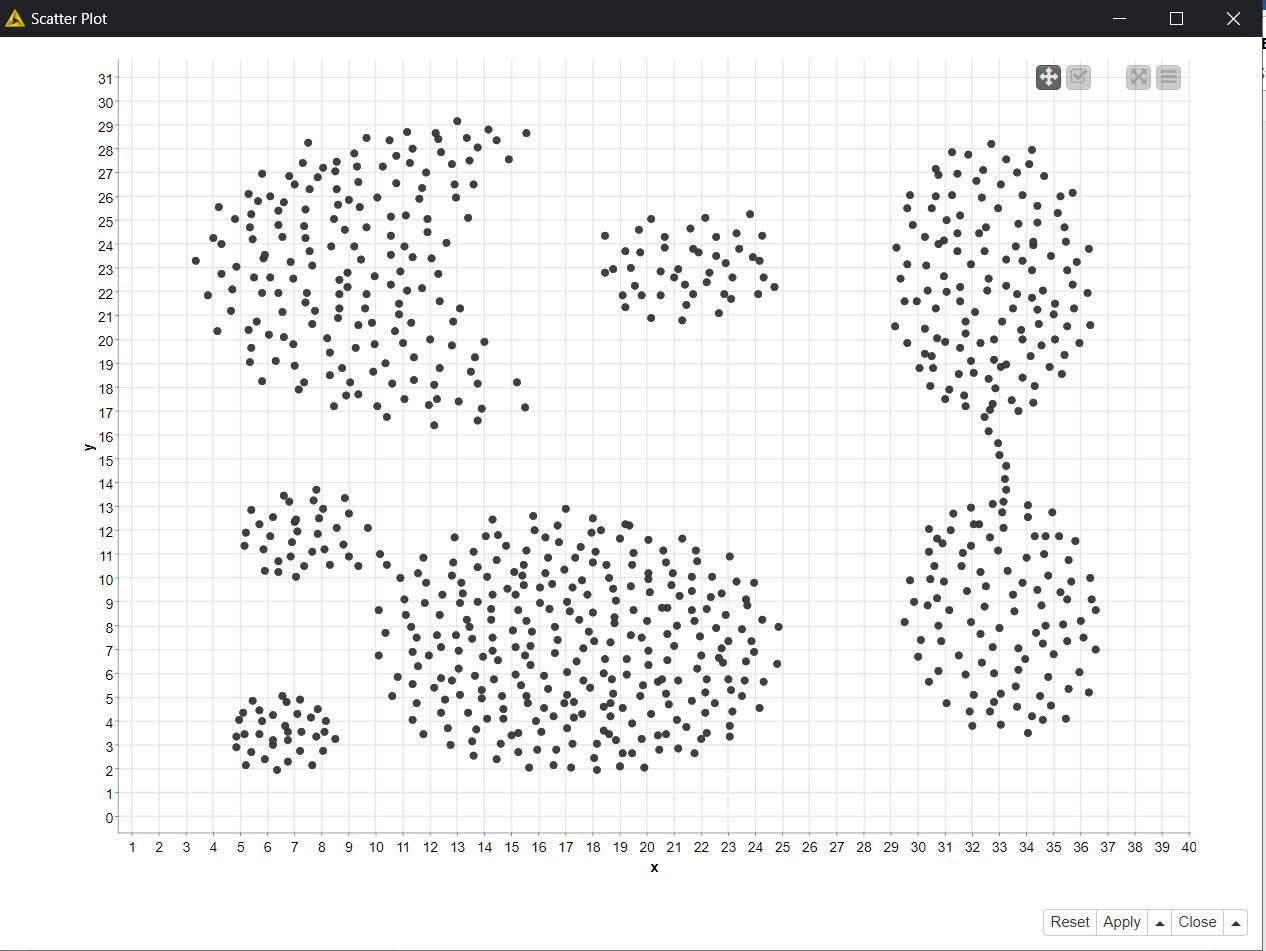
**VỚI DATASET1**

1. Khảo sát, đánh giá, nhận định tập dữ liệu
2. Khảo sát, đánh giá

- Dữ liệu là tập đối tượng (788 đối tượng) mà mỗi đối tượng có 2 giá trị (đặt là x, y) biểu diễn nó.

🡺 Vẽ biểu đồ chấm (Scatter Plot) tất cả các đối tượng trên mặt phẳng 2D.

- Biểu đồ chấm thể hiện tập dữ liệu:



- Ta thấy các đối tượng xếp thành 7 cụm khá rõ ràng.

2. Nhận định

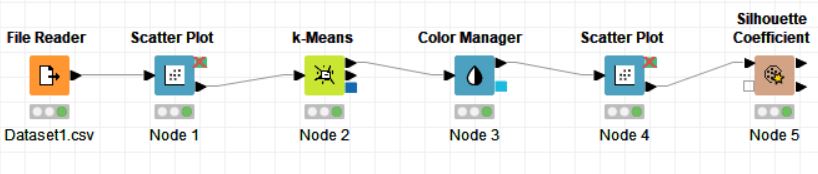
- Tập dữ liệu 1 là tập ít dữ liệu, đơn giản, có 2 chiều, đã có nhãn cụm.

II. Lựa chọn thuật toán trên Knime

Sử dụng thuật toán K-means và Hierachical Clustering.

III. Cài đặt thuật toán trên Knime

1. K-means



- File Reader: Đọc file dataset1.csv

- Node 1: Mô phỏng tập dataset1 bằng đồ thị Scatter Plot

- Node 2: Sử dụng K-means với:

+ Number of clusters: 7

+ Random initialization

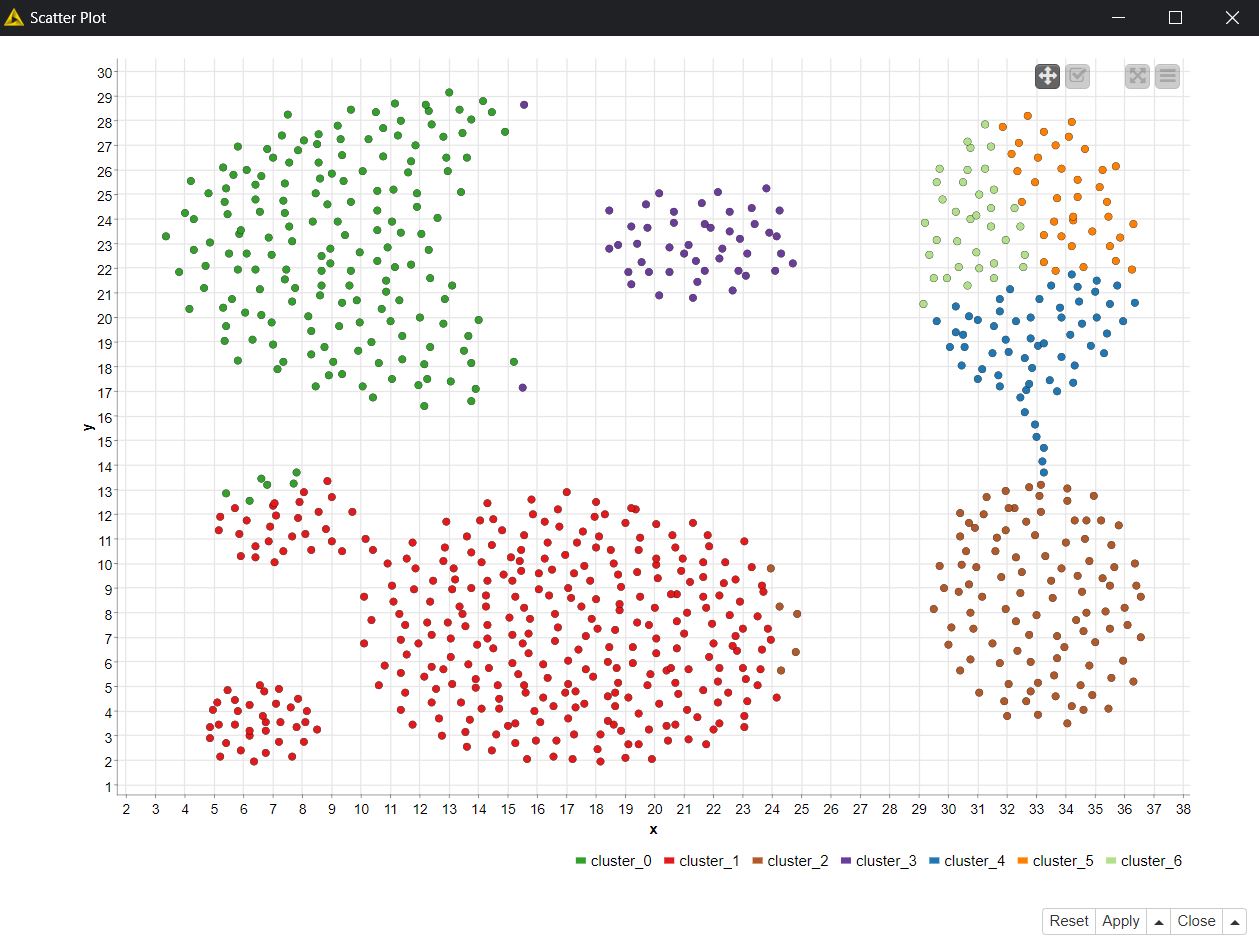
+ Column selection: x, y

- Node 3: Phân màu vào các cụm vừa chia sau Node 2

- Node 4: Mô phỏng lại tập dataset1 bằng đồ thị Scatter Plot

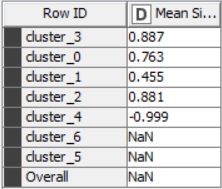
- Node 5: Đánh giá phân cụm bằng thuật toán Silhouette

🡺 kết quả:

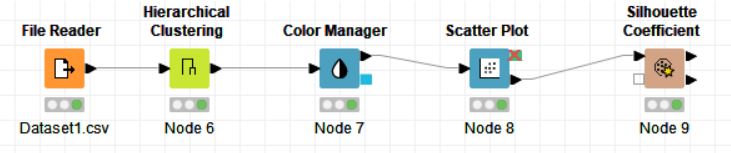


- Đánh giá:

+ Độ chính xác so với nhãn không cao



2. Hierachical Clustering



- File Reader: Đọc file dataset1.csv

- Node 6: Dùng thuật toán Hierarchical Clustering để phân cụm với:

+ Number output clusters: 7

+ Distance function: Eudidean

+ Linkage type: average

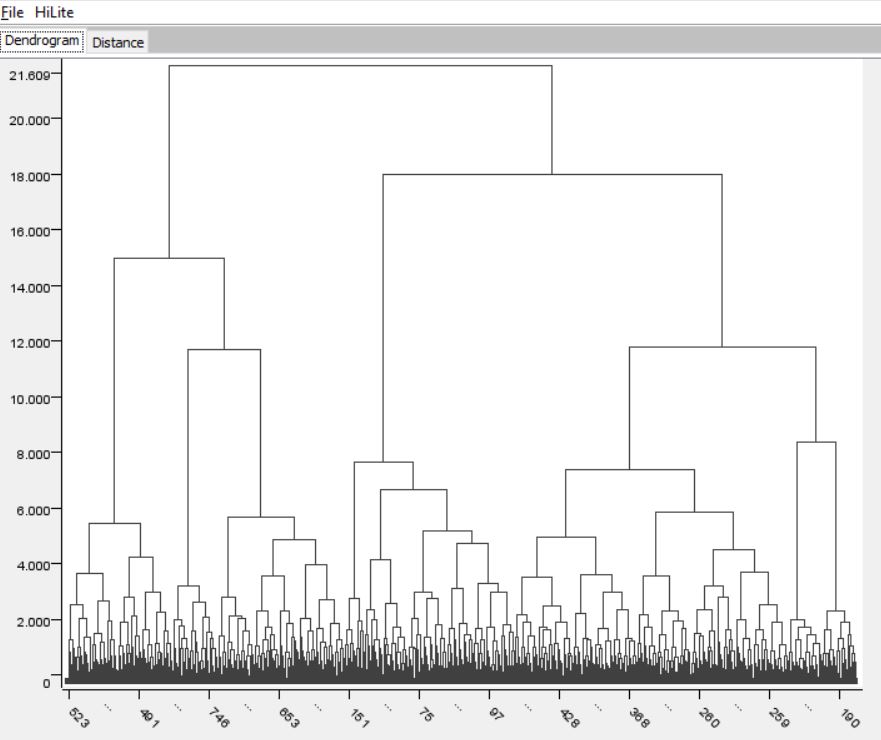
+ Indude: x, y

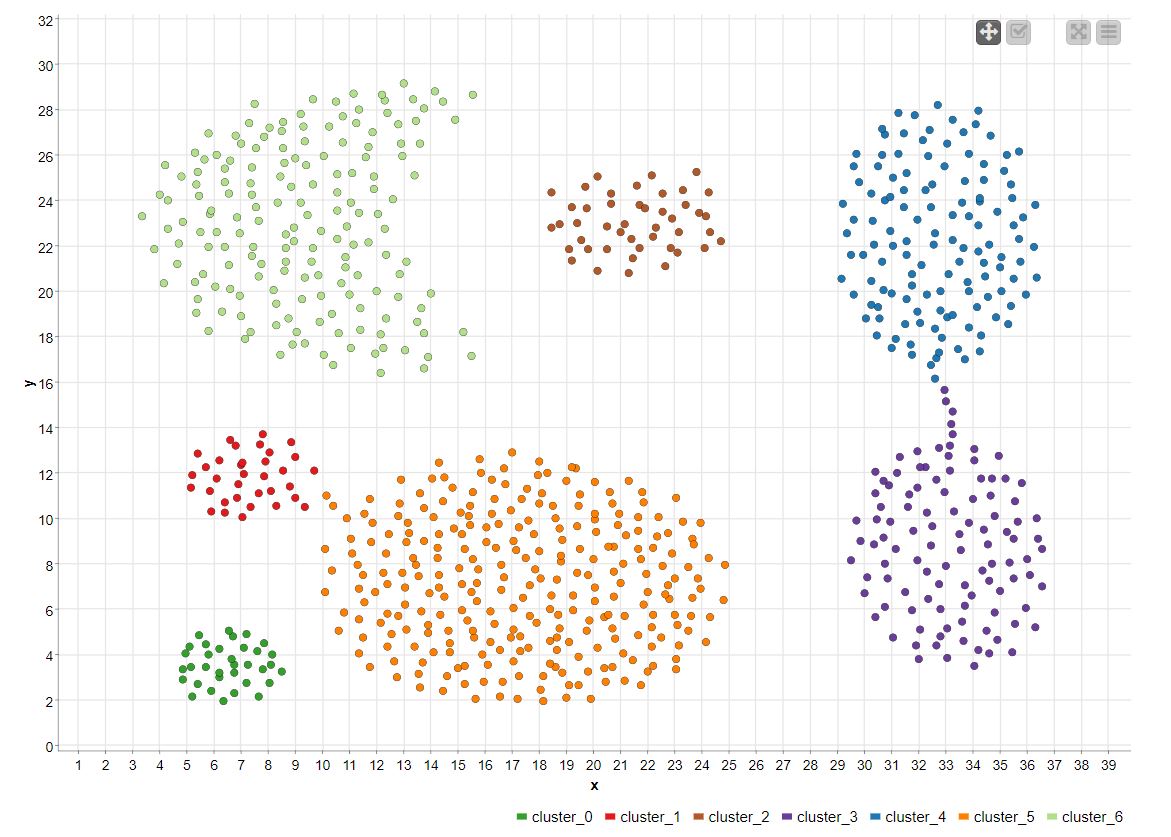
- Node 7: Phân màu vào các cụm vừa chia sau Node 6

- Node 8: Mô phỏng tập dataset1 bằng đồ thị Scatter Plot

- Node 9: Đánh giá phân cụm bằng thuật toán Silhouette

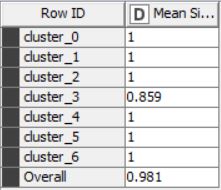
🡺 kết quả:





- Đánh giá:

+ Độ chính xác so với nhãn cao



+ Thời gian chạy chậm hơn K-means

+ Độ chính xác cao hơn K-means

**BÁO CÁO PHÂN CỤM DỮ LIỆU**

**VỚI DATASET2**

I. Khảo sát, đánh giá, nhận định tập dữ liệu

1. Khảo sát, đánh giá

- Tập dữ liệu 2 là tập dữ liệu phức tạp, nhiều dữ liệu, nhiều chiều, nhiều cụm, chưa có nhãn cụm

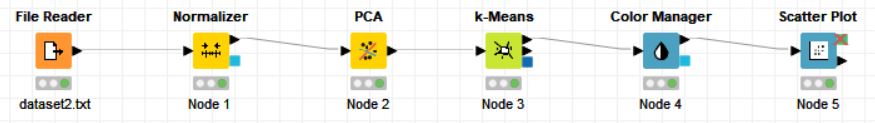
2. Nhận định

- Giảm chiều dữ liệu (xuống còn 2)

II. Lựa chọn thuật toán trên Knime

Sử dụng thuật toán K-means

III. Cài đặt thuật toán trên Knime



- File Reader: Đọc file dataset2.txt

- Node 1: Chuẩn hóa dữ liệu với:

+ Tất cả các đối tượng

+ Min-Max Normalization (min = 0.0, max = 1.0)

- Node 2: Giảm chiều dữ liệu xuống 2 chiều (PCA dimension 0, PCA dimension 1)

- Node 3: Sử dụng K-means với:

+ Number of clusters: 10

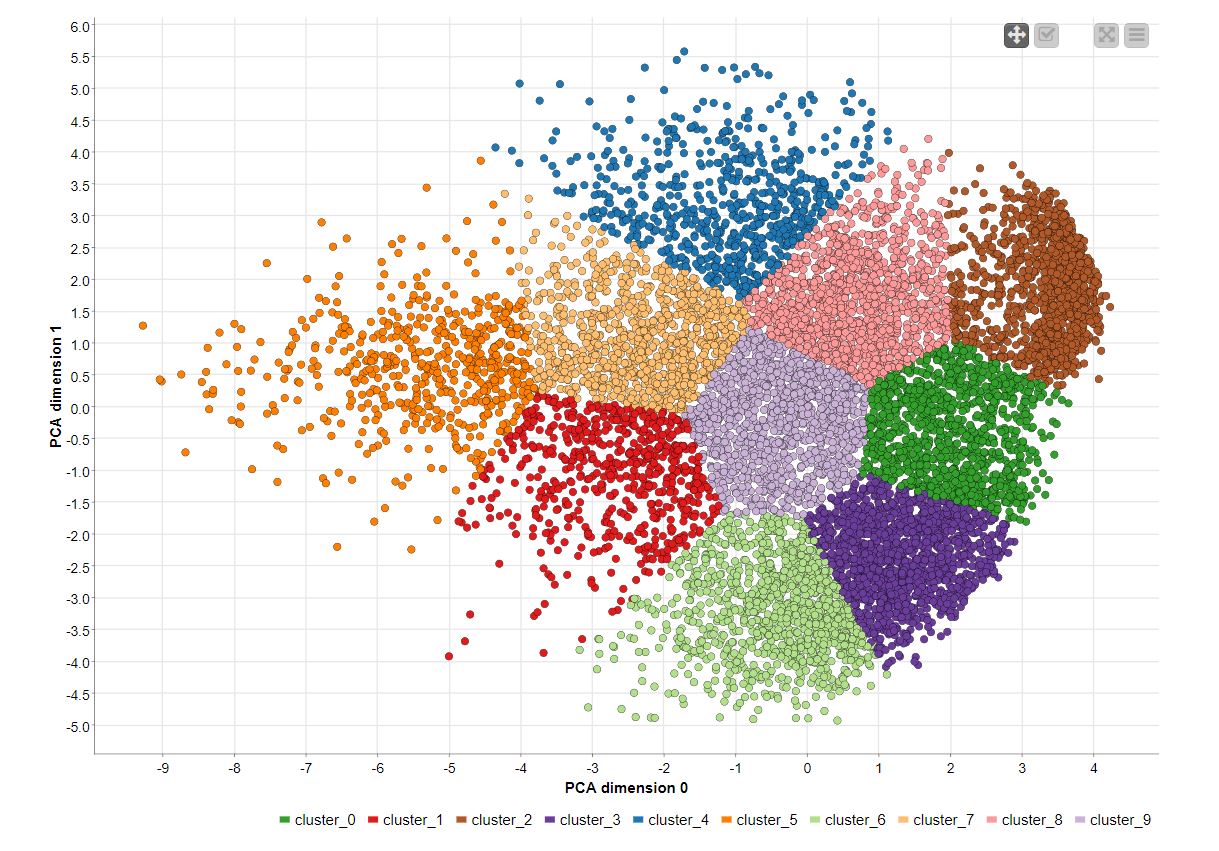
+ Random initialization

+ Column selection: PCA dimension 0, PCA dimension 1

- Node 4: Phân màu vào các cụm vừa chia sau Node 3

- Node 5: Mô phỏng tập dataset2 bằng đồ thị Scatter Plot

🡺 kết quả:



- Biểu đồ cột thể hiện số phần tử của từng cụm:

