BÀI 1: CHUẨN BỊ DỮ LIỆU

I. Mục tiêu:

Sau khi thực hành xong, sinh viên nắm được:

- Các bước làm sạch dữ liệu và tiền xử lý dữ liệu
- Sử dụng các thư viện Pandas và scikit-learn.
- Chuẩn hóa dữ liệu.
- Rời rạc hóa dữ liệu.
- PCA

II. Tóm tắt lý thuyết:

Tập dữ liệu nhiều chiều D là một tập hợp gồm n bản ghi $\overline{X_1}, \overline{X_2}, \ldots, \overline{X_n}$, sao cho mỗi $\overline{X_i}$ là một tập hợp chứa d đặc trưng được ký hiệu bởi $(x_i^1, \ldots, x_i^d)$.

1. Chuẩn hóa dữ liệu:

- Xét trường hợp thuộc tính thứ j có trung bình (mean) là μ_j và độ lệch chuẩn (standard deviation) σ_j . Khi đó, x_i^j (giá trị thuộc tính thứ j) của $\overline{X_i}$ (bản ghi thứ i) có thể được chuẩn hóa như sau:

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$

- Xấp xỉ thứ 2 sử dụng min-max scaling để ánh xạ tất cả thuộc tính thành vùng [0,1]. Đặt $\min_j$ và $\max_j$ là các giá trị nhỏ nhất và lớn nhất của thuộc tính j. Khi đó, x_i^j của $\overline{X_i}$ có thể được scale như sau:

$$y_i^j = \frac{x_i^j - \min_j}{\max_j - \min_j}$$

2. Rời rạc hóa dữ liệu:

a. Equi-width ranges: là chia các giá trị này thành các khoảng bằng nhau về độ rộng (width) hoặc bin. Cụ thể hơn là, các bin có độ rộng bằng nhau với mỗi bin