**PRACTICE #03**

**Scikit-learn with PCA & LDA**

# Thông tin cá nhân:

**MSSV**: 19127614

**Họ** **Tên**: Nguyễn Anh Tuấn

**#Lưu Ý:** Phần bonus sẽ nằm trong folder Bonus

# Dữ liệu Hoa Irish

**Link Dataset:**   
<https://www.kaggle.com/datasets/arshid/iris-flower-dataset>

**Tổng quan:**

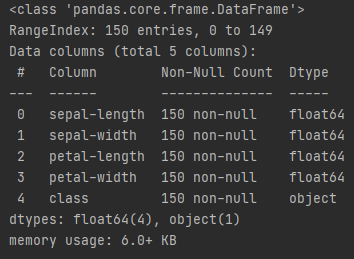
Tập dữ liệu về hoa Iris là một tập dữ liệu đa biến, lần đầu được giới thiệu bởi nhà thống kê và sinh vật học người Anh Ronald Fisher trong bài báo “The use of multiple measurements in taxonomic problems” vào năm 1936. Bộ dữ liệu bao gồm 50 mẫu từ mỗi loài trong số ba loài Iris (Iris Setosa, Iris virginica và Iris versicolor). Bốn đặc điểm được đo từ mỗi mẫu: chiều dài và chiều rộng của các đài hoa và cánh hoa, tính bằng cm.

Tập dữ liệu này đồng thời cũng có thể load từ sklearn.datasets.

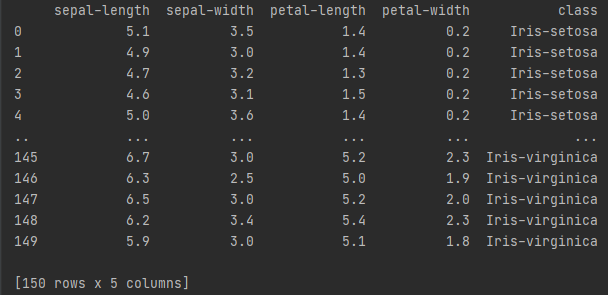
**Nội dung File CSV:**

Tập dữ liệu chứa một bộ 150 mẫu, kèm theo là 5 thuộc tính - Chiều dài cánh hoa, Chiều rộng cánh hoa, Chiều dài đài hoa, Chiều rộng đài hoa và Lớp (Loại hoa)

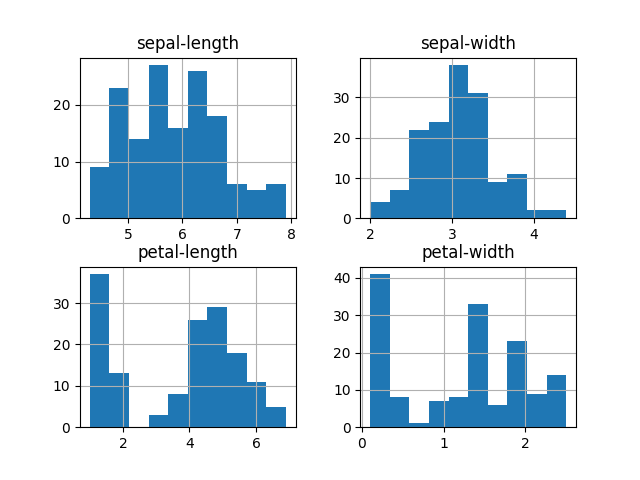
**Thông tin:**

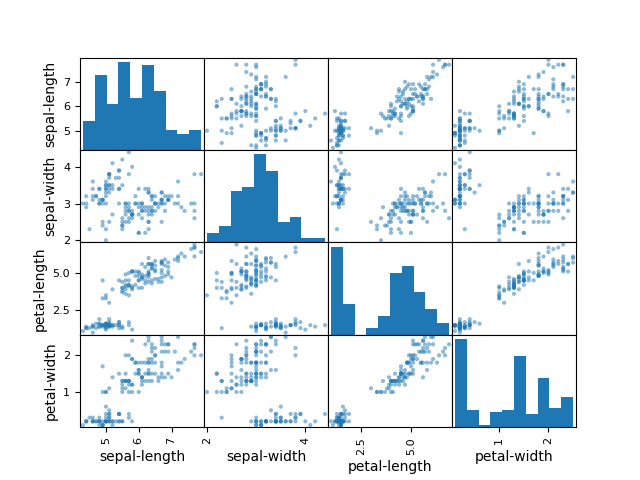
****

**Dataset:**

****

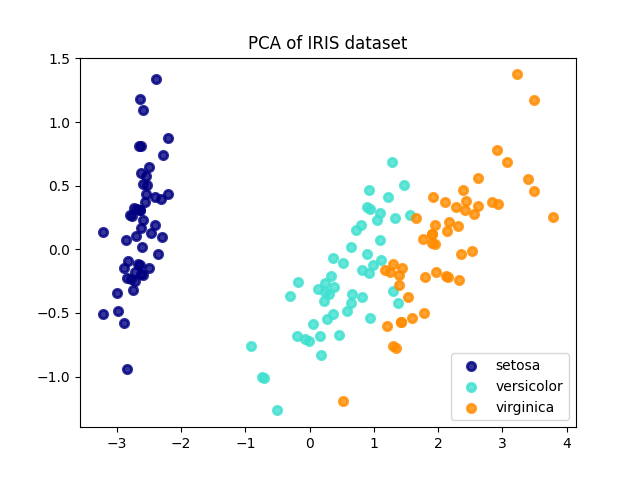
# Trực quan hóa dữ liệu

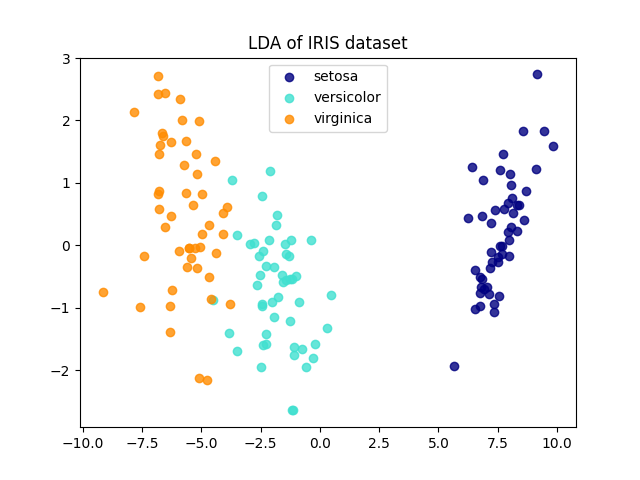




# PCA và LDA

Các dữ liệu của hoa Iris là dữ liệu 4 chiều, khó để trực quan hóa, sau khi giảm chiều bằng PCA và LDA, ta có thể trực quan hóa bằng scatter 2 chiều.





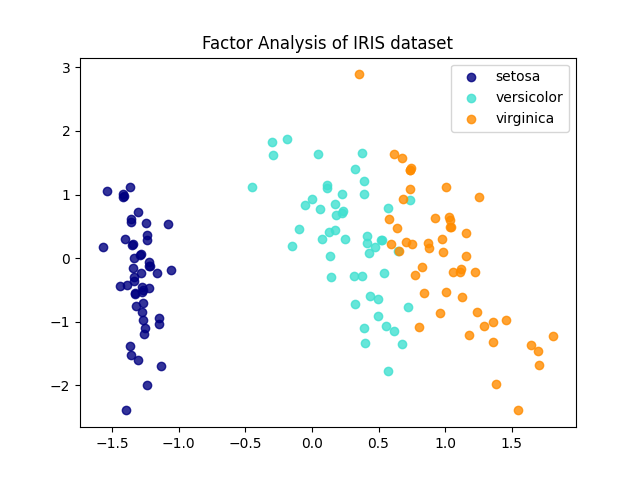
**So Sánh LDA và PCA**

|  |  |  |
| --- | --- | --- |
|  | **LDA** | **PCA** |
| **Mô hình học** | Học có giám sát - là thuật toán tiên đoán nhãn cho dữ liệu mới dựa trên tập huấn luyện (các mẫu trong tập này đều đã được gán nhãn) | Học không giám sát - là thuật toán tiên đoán nhãn cho dữ liệu mới dựa trên tập huấn luyện (các mẫu trong tập này đều chưa được gán nhãn) |
| **Xét phương sai** | Có | Không |
| **Giảm chiều** | Có | Có |
| **Số lượng thông tin** | Cần thiết | Nhiều nhất |

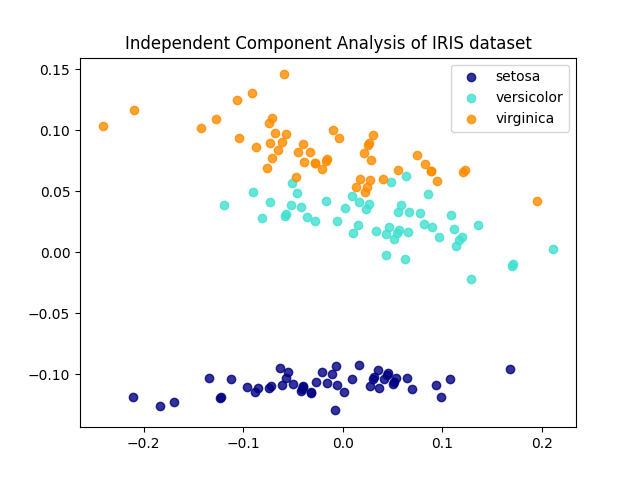
# Bonus

## Một số thuật toán phân tích đa biến khác

**Factor Analysis (FA)**

****

**Independent Component Analysis (ICA)**

****