# LẬP TRÌNH PYTHON
# Matplotlib & Ứng dụng

NGUYỄN HẢI TRIỀU[1]

[1]Bộ môn Kỹ thuật phần mềm,

Khoa Công nghệ thông tin, Trường ĐH Nha Trang

NhaTrang, February 2022

# Nội dung

# Giảm chiều dữ liệu

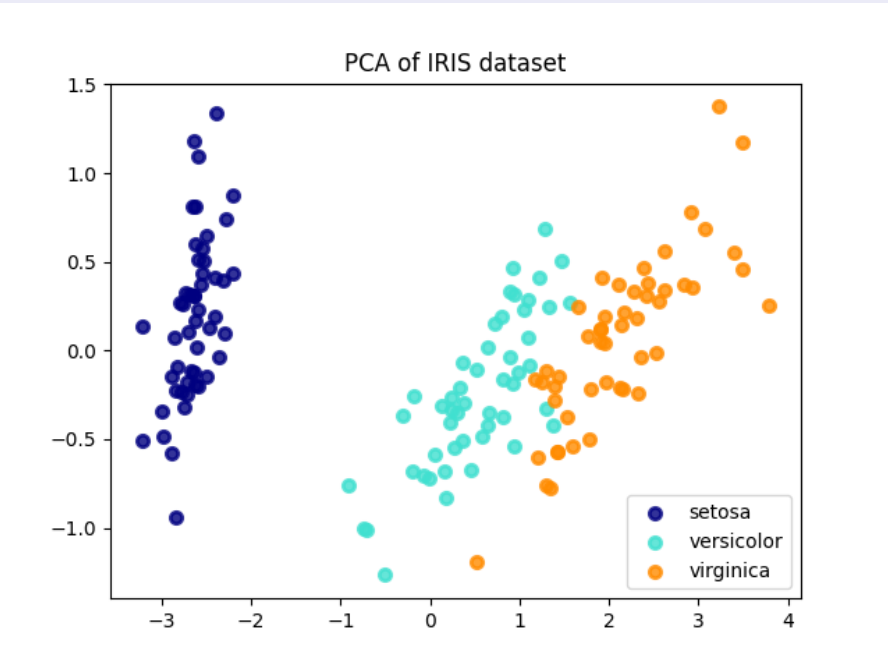
Bộ dữ liệu trong ML đôi khi có số chiều đặc trưng rất lớn làm cho việc tính toán và các tác vụ liên quan rất khó khăn $\Rightarrow$ Dẫn đến yêu cầu giảm chiều dữ liệu. Một vài thuật toán giảm chiều dữ liệu phổ biến có thể sử dụng trong scikit-learn:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- t-distributed Stochastic Neighbour Embedding (t-SNE). Code tham khảo tại đây

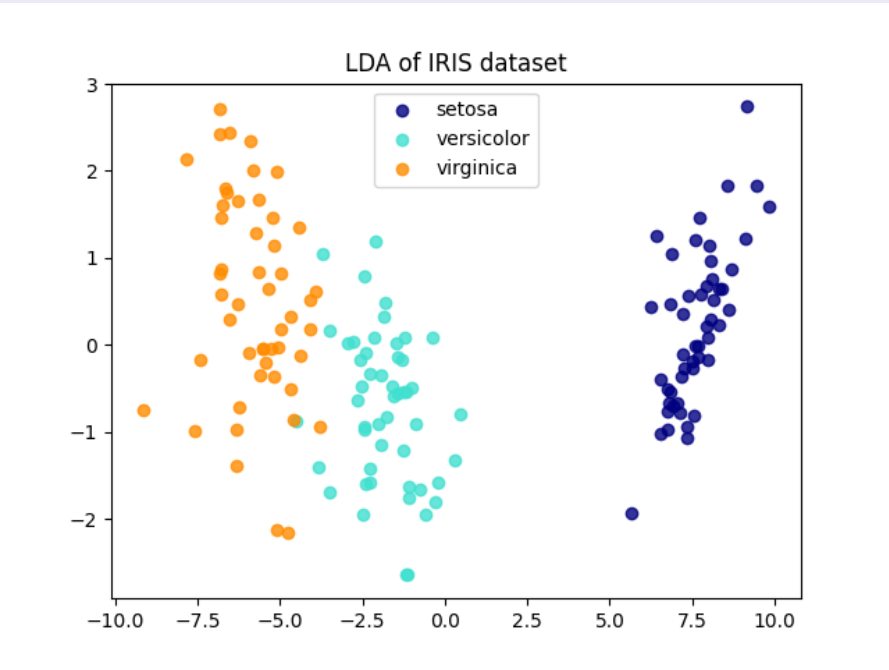
# PCA, LDA

## Ví dụ 2.1

*Sử dụng thư viện scikit-learn, áp dụng giảm chiều dữ liệu trên bộ dữ liệu iris (4 đặc trưng) về còn 2 đặc trưng và vẽ trong không gian 2D. Code tham khảo tại đây*




*(a) PCA*



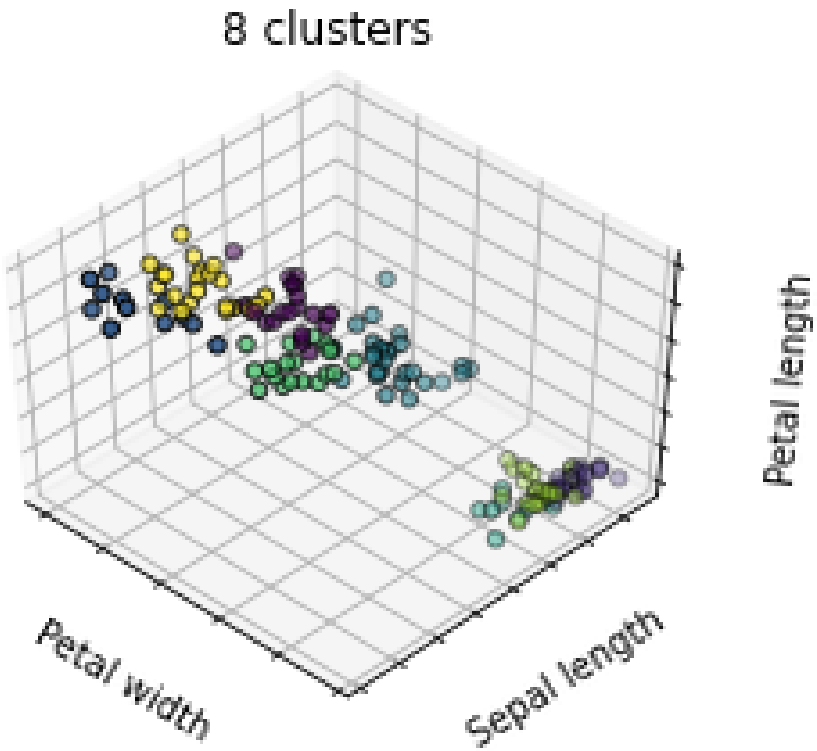

*(b) LDA*

# Phân cụm dữ liệu

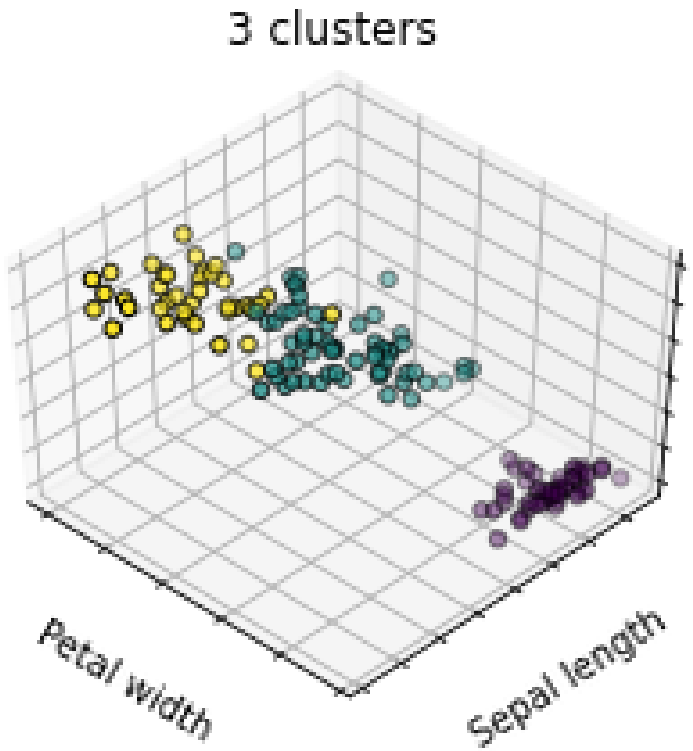
Phân cụm là bài toán thuộc nhóm Unsupervised Learning (Học không giám sát) nhằm phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Có nhiều thuật toán được sử dụng để phân cụm: K-means, DBSCAN...

# Kmeans

Áp dụng thuật toán Kmeans phân cụm cho bộ dữ liệu iris. Code tham khảo tại đây



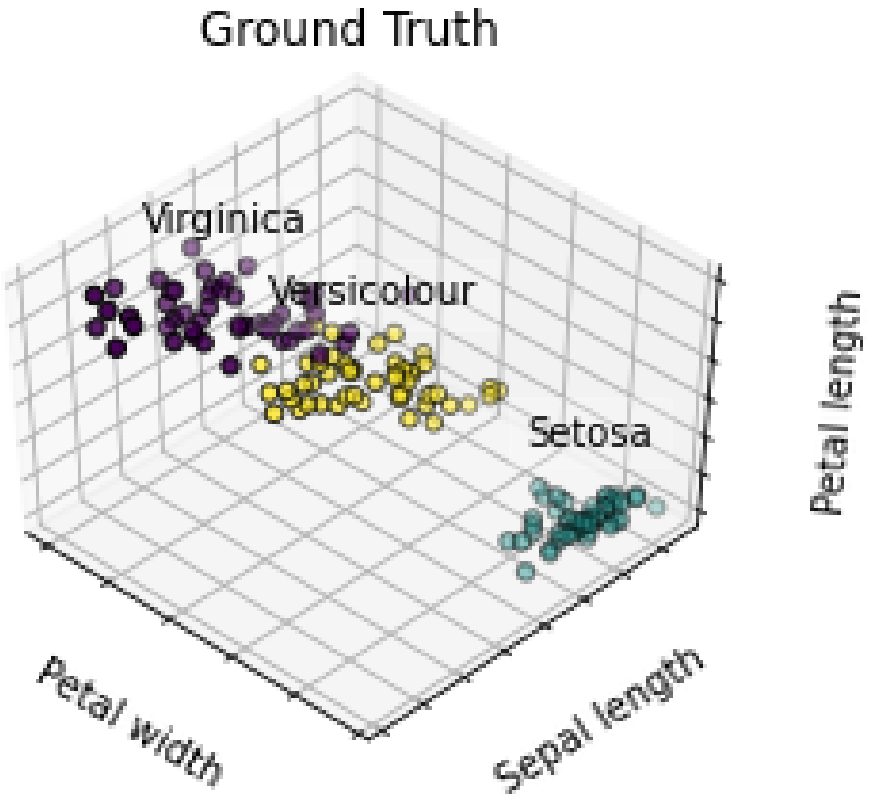

(c)




(d)

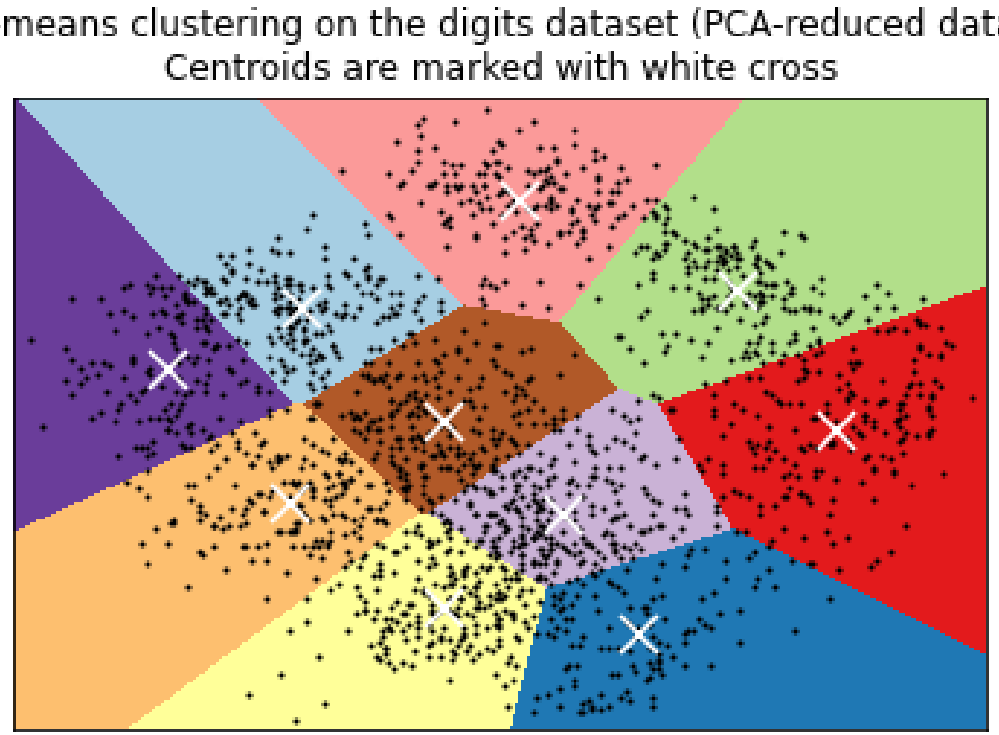
# Kmeans

Áp dụng thuật toán Kmeans phân cụm cho bộ dữ liệu iris.



# Kmeans

Áp dụng thuật toán Kmeans phân cụm cho bộ dữ liệu chữ số viết tay. Code tham khảo tại đây



# Phân loại chữ số viết tay

## Ví dụ 2.2

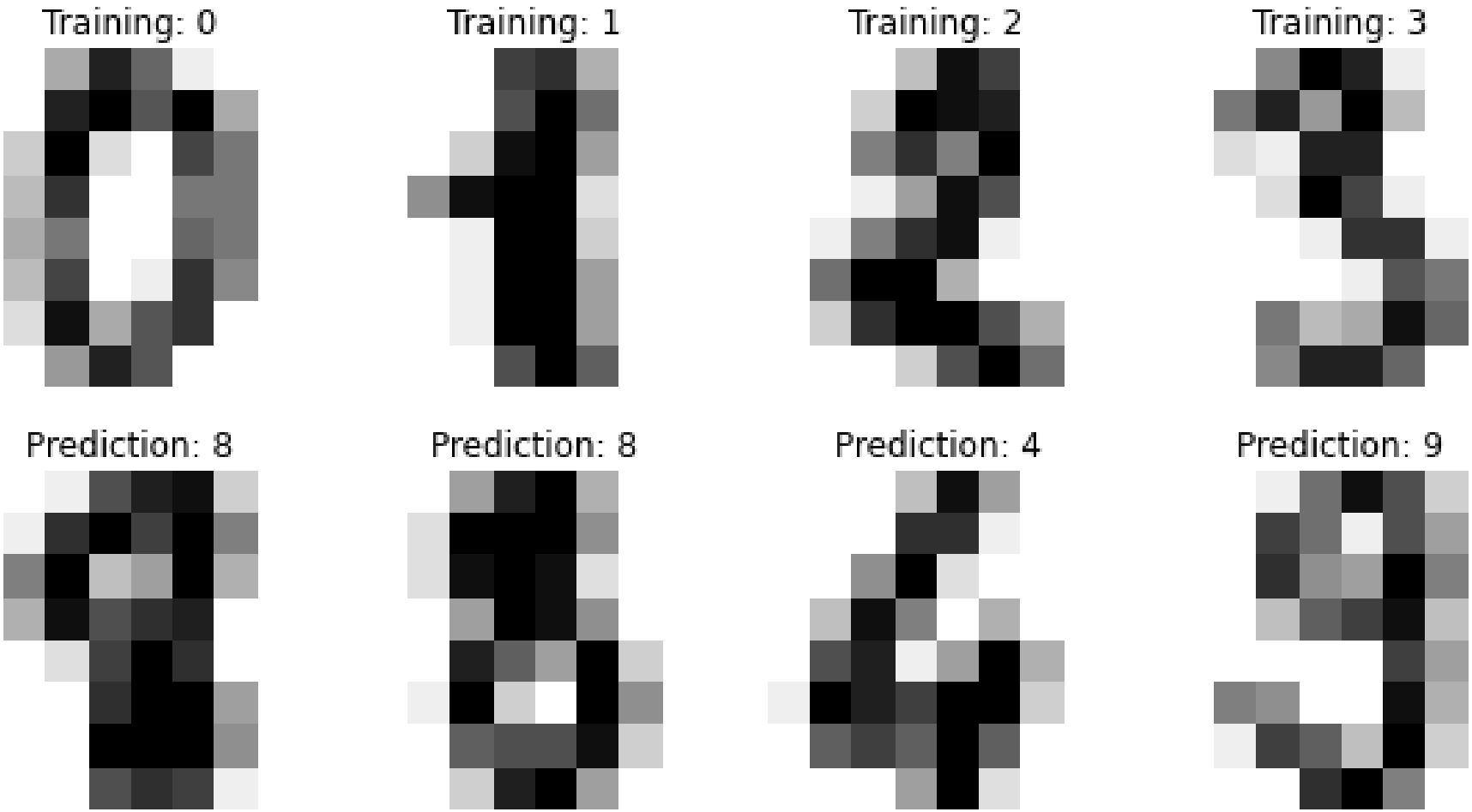
*Trong nhận dạng chữ số viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào.*

## Classification (Phân loại)

Đây là bài toán thuộc nhóm Supervised Learning (Học có giám sát) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn).

# Sử dụng SVM để phân loại chữ số viết tay

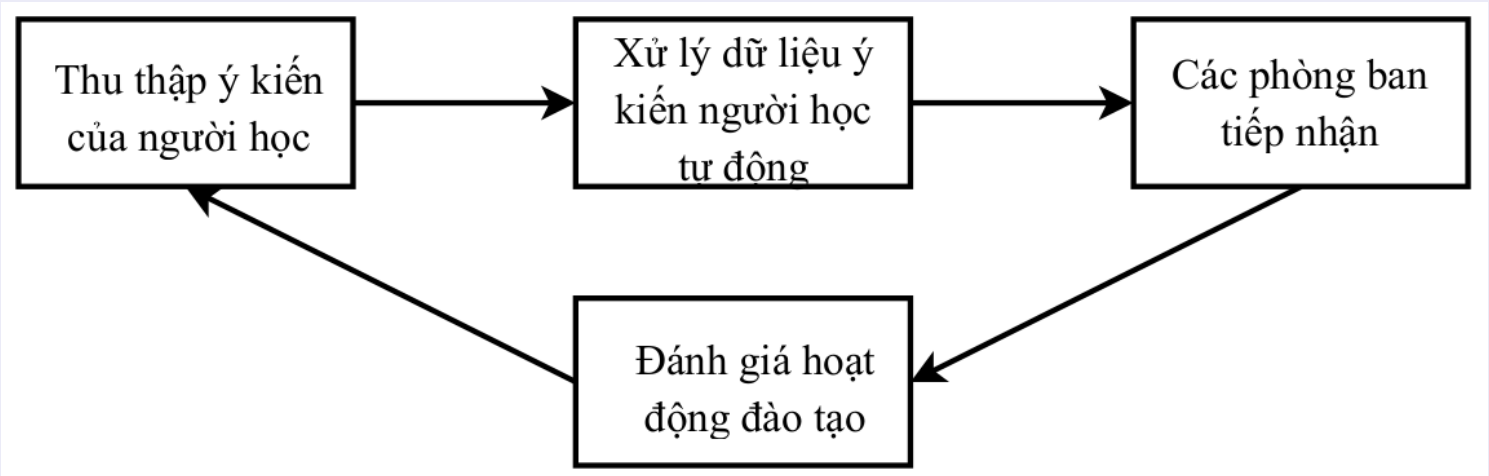
Sử dụng thư viện sklearn.svm.SVC để phân loại ảnh của các chữ số viết tay từ 0-9. Code tham khảo tại đây



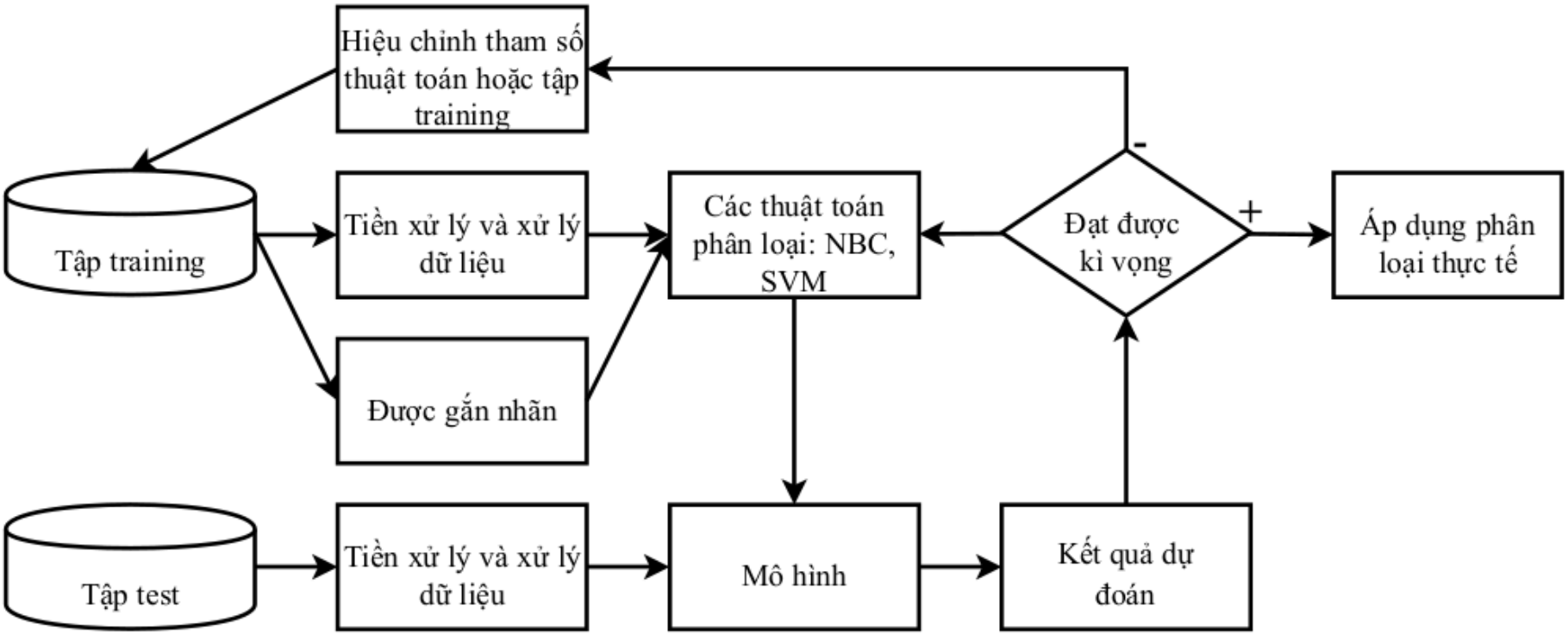
# Xây dựng ứng dụng ML

## Ví dụ 3.1 (Bài toán phân loại văn bản)

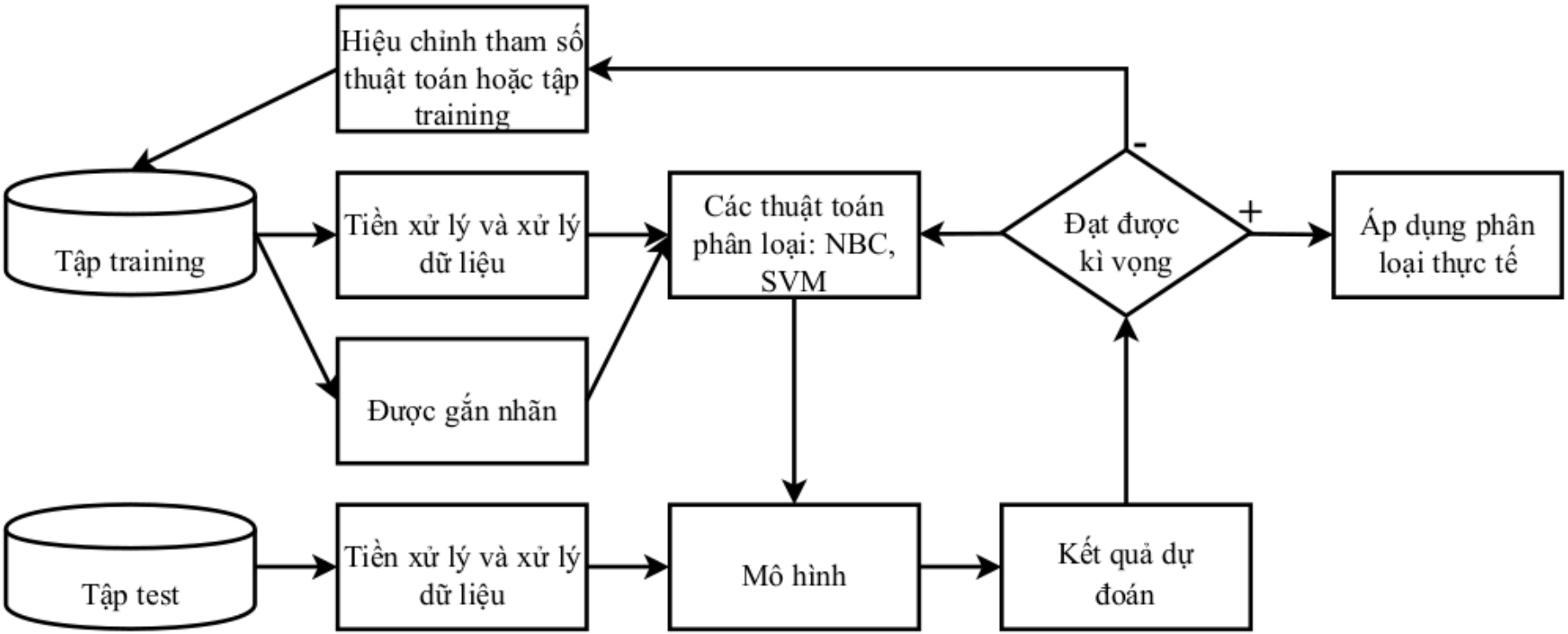
*Ví dụ thực tế về phân loại tự động ý kiến người học tại đại học Nha Trang*




Hình 1: Chu trình xử lý ý kiến người học tại ĐHNT



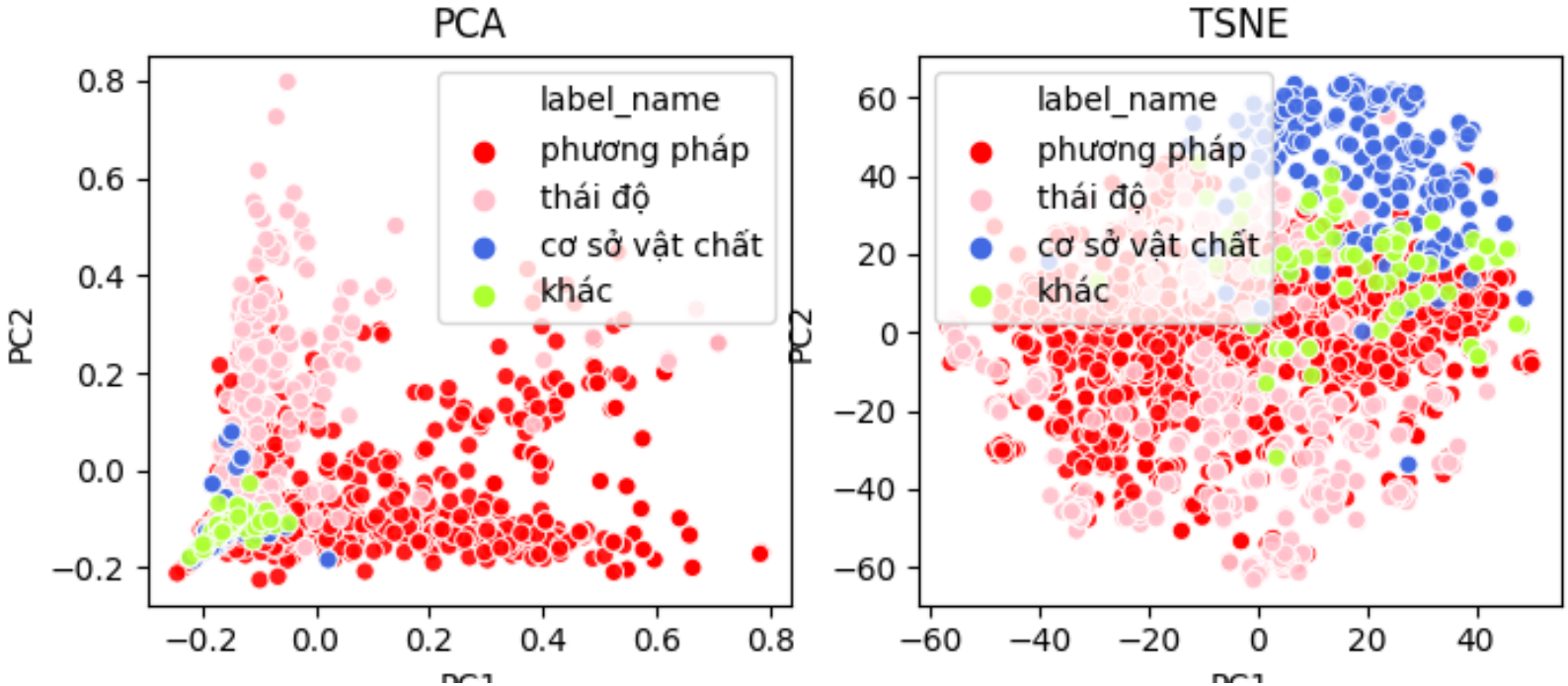

Hình 2: Minh họa quá trình phân loại văn bản




Hình 3: Minh họa quá trình phân loại văn bản

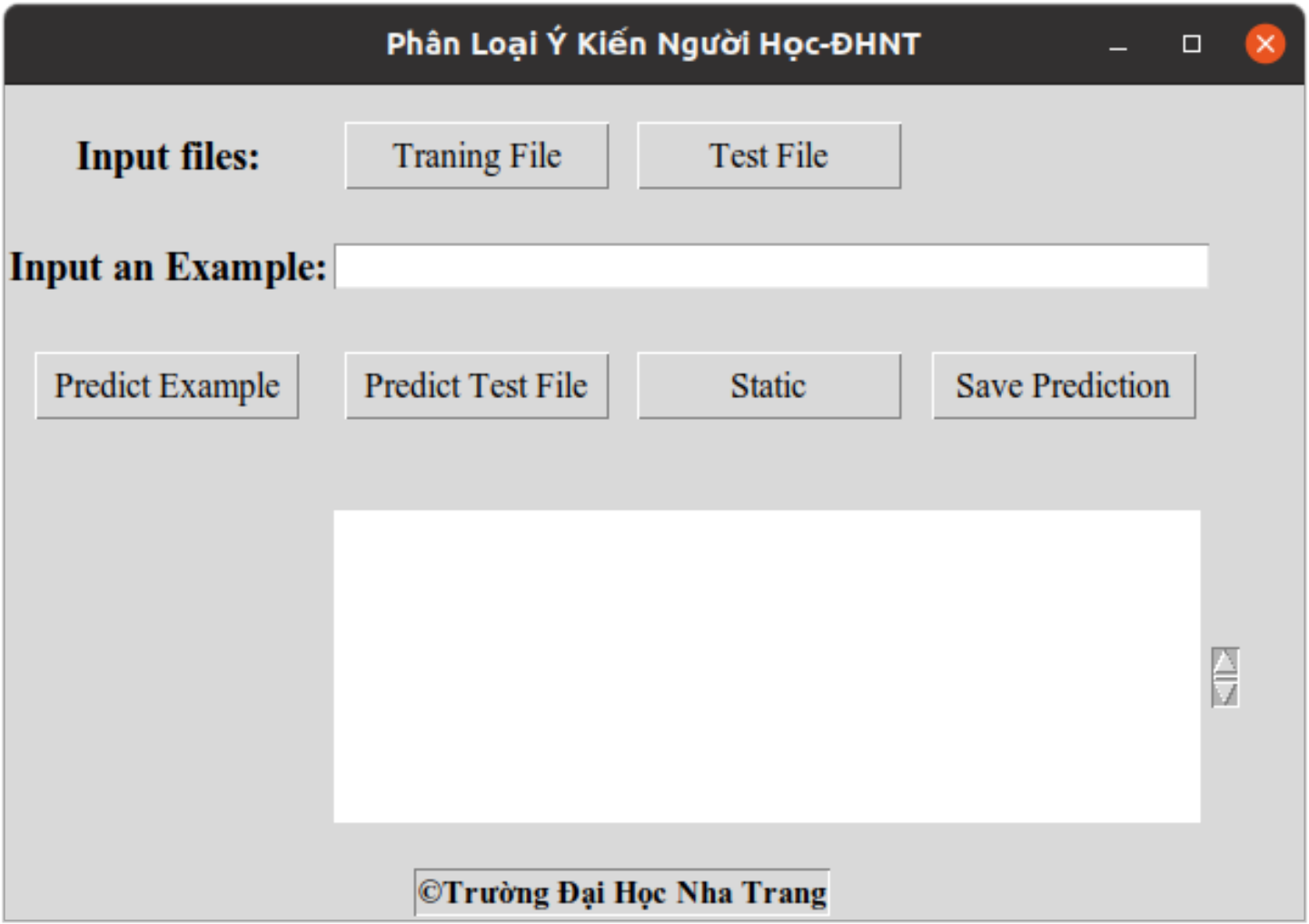
| Tên nhãn | Số lượng văn bản cho tập training | Số lượng văn bản cho tập test |
|---|---|---|
| Phương pháp giảng dạy của GV | 1099 | 469 |
| Thái độ của GV đối với người học | 518 | 222 |
| Cơ sở vật chất | 355 | 151 |
| Ý kiến khác | 92 | 47 |

Hình 4: Mô tả dữ liệu




Hình 5: Sự phân bố các điểm dữ liệu được vẽ bằng phương pháp PCA và t-SNE

# Quy trình xây dựng ứng dụng




Hình 6: Giao diện ứng dụng phân loại văn bản

# Quy trình xây dựng ứng dụng

Code tham khảo cho ứng dụng tham khảo tại đây.

- ứng dụng sử dụng thuật toán **SVM** để phần loại (xem chi tiết trong *class TextClassificationPredict(object)*)
- ứng dụng có chức năng dự đoán một câu văn bản đầu vào, một file đầu vào, thống kê độ chính xác cho file test
- sử dụng **sqllite** để lưu trữ dữ liệu
- sử dụng sklearn.externals import **joblib** để lưu trữ model huấn luyện
- sử dụng **Tkinter** để làm UI cho ứng dụng
- để đóng gói app sử dụng PyInstaller: pyinstaller app.py –onefile –windowed. Lưu ý có thể đóng gói thành file cài đặt với chương trình NSIS (Windows),

# Tài liệu tham khảo


Trung tâm tin học, Đại Học KHTN Tp.HCM
Lập trình Python nâng cao. *03/2017*.

Bernd Klein
Data Analysis: Numpy, Matplotlib and Pandas.
*bernd.klein@python-course.eu, 2021*.

Luciano Ramalho
Fluent Python (2nd Edition). *O'Reilly Media, Inc, 2021*.

Python Software Foundation
`https://docs.python.org/3/tutorial/`