**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG CÔNG NGHỆ - ĐẠI HỌC KINH TẾ QUỐC DÂN**

----------



**BÀI TẬP LỚN ỨNG DỤNG TRÍ TUỆ NHÂN TẠO**

**Đề tài:**

**Nghiên cứu và cài đặt chương trình phần mềm để thực hiện phân lớp khách hàng của ngân hàng bằng kỹ thuật Máy Vector hỗ trợ (SVM).**

**Người thực hiện : Nguyễn Minh Quang – 11225443**

**Phạm Minh Quân - 11225400**

**Môn học : Ứng dụng trí tuệ nhân tạo**

**Lớp học phần : TIHT1123(224)\_01**

**Giảng viên hướng dẫn : Ths.Phạm Thảo**

***Hà Nội, Năm 2025***

# MỤC LỤC

[MỤC LỤC 1](#_heading=h.nm83bm8b7f7k)

[DANH MỤC TỪ VIẾT TẮT 2](#_heading=h.uzkjae2ltfr4)

[LỜI MỞ ĐẦU 3](#_heading=h.2rv1fvz1cvg8)

[I. TỔNG QUAN VỀ ĐỀ TÀI 5](#_heading=h.sf4f42svfwqp)

[1.1.Phát biểu về đề tài: 5](#_heading=h.1mmcdzqlp5k3)

[1.2. Mục đích, đối tượng và phạm vi nghiên cứu: 5](#_heading=h.2an3ewtctfwx)

[1.3. Ý nghĩa khoa học và thực tiễn 5](#_heading=h.z29yangx449b)

[II. CƠ SỞ LÝ THUYẾT 7](#_heading=h.ln5996la0lul)

[2.1.Đặt vấn đề 7](#_heading=h.b3ef008ebl3b)

[2.2. Thuật toán SVM. 7](#_heading=h.ydefu98lv9q8)

[2.2.1 Khái niệm 7](#_heading=h.s5n8kowfnuw)

[2.2.2.Xây dựng bài toán: 8](#_heading=h.sq5rvtp9i224)

[III.CÀI ĐẶT THỬ NGHIỆM 11](#_heading=h.dxof28imedq2)

[3.1. Công nghệ áp dụng 11](#_heading=h.xag4m46xqm7j)

[3.1.1. Python 11](#_heading=h.65bc71rs3mu0)

[3.1.2. Các thư viện sử dụng. 11](#_heading=h.11fyttdrktlf)

[3.2. Giải thuật. 12](#_heading=h.xd78sskg7e8e)

[3.3. Xử lý dữ liệu và chọn thuộc tính . 13](#_heading=h.q4e9q1s36ljr)

[3.4.Chia dữ liệu. 14](#_heading=h.ywj4qa5jmk9g)

[3.5. Kết quả chương trình thử nghiệm. 15](#_heading=h.62jsqdwgkh73)

[IV. ĐÁNH GIÁ VÀ KẾT QUẢ NGHIÊN CỨU 20](#_heading=h.e2tcuixj1r31)

[TÀI LIỆU THAM KHẢO 21](#_heading=h.7v35wboe6pkd)

# DANH MỤC TỪ VIẾT TẮT

| STT | Từ viết tắt | Ý nghĩa |
| --- | --- | --- |
| 1 | SVM | Support Vector Machine |

**Link code:**

<https://drive.google.com/file/d/1K4gyiRdDUct-ijfS_UUthbihGSMQA_Uy/view?usp=sharing>

# LỜI MỞ ĐẦU

Ngày nay, sự phát triển của máy tính, đặc biệt là Internet đã dẫn tới sự bùng nổ thông tin. Rất nhiều dữ liệu, thông tin mới được tạo ra hằng ngày với số lượng lớn khiến cho việc phân loại thông tin trở nên khó khăn. Do đó chúng ta cần phân loại thông tin, dữ liệu để việc tìm kiếm cũng như kiểm soát thông tin dễ dàng hơn. Việc phân loại này có thể được làm thủ công đọc từng thông tin một và phân loại chúng dựa vào kinh nghiệm, cách làm này rất mát thời gian, không đáp ứng được nhu cầu và thực tế lượng dữ liệu ngày càng nhiều. Do đó cách phân loại tốt nhất là sử dụng ứng dụng trí tuệ nhân tạo, xử lí ngôn ngữ tự nhiên để phân loại. Việc xử lí ngôn ngữ tự nhiên trong Tiếng Anh đã có nhiều nghiên cứu trên thế giới, nhưng để áp dụng cho Tiếng Việt cũng có một vài điểm cần phải cái tiến ví dụ như: StopWord, Tách từ trong Tiếng Việt cũng khác so với Tiếng Anh.

Bên cạnh đó là các thuật toán phân loại, nhiều thuật toán phát triển phục vụ múc đích phân loại, phân nhóm dữ liệu. Các thuật toán như K Láng giềng gần nhất, Cây quyết định, Naïve Bayes và tiêu biểu thuật toán Support Vector Machine (SVM). Thuật toán SMV có ưu điểm đó là xử lý trong không gian số nhiều chiều, tiết kiệm tài nguyên hệ thống và có tính linh hoạt cao nhờ vào khả năng áp dụng Kernel (hàm nhân) mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính qua đó khiến cho hiệu suất phân loại cao hơn.

Với 1725 văn bản công văn nhà nước thu thập được, chia làm 10 lĩnh vực khác nhau, luân văn tiến hành xử lí số lượng văn bản. Bằng cách thêm vào các từ StopWord thuộc văn bản công văn nhà nước, tách từ Tiếng Việt, sử dụng TF-TDF để trích rút vector đặc trưng. Sau đó sử dụng thuật toán Máy Vector hỗ trợ với hàm nhân tuyến tính Linear để tạo ra Model phân loại văn bản, dữ liệu với độ chính xác trung bình cho 10 lĩnh vực là 91%, một số lĩnh vực “Công nghệ thông tin”, “Đất đai” với độ chính xác 100%. Kết quả thực nghiệm cho thấy tính khả quan việc phân lớp khách hang của ngân hang trong việc sử dụng thuật toán Máy Vector hỗ trợ và các thuật toán phân loại khác

Với những kiến thức còn hạn chế, tiểu luận của chúng em có thể sẽ có nhiều sai sót hoặc chưa được toàn diện, chúng em rất mong nhận được những lời nhận xét và góp ý từ phía thầy để có thể hoàn thiện nghiên cứu tốt hơn về sau.

*Chúng em xin chân thành cảm ơn!*

# I. TỔNG QUAN VỀ ĐỀ TÀI

## 1.1.Phát biểu về đề tài:

Vấn đề phân lớp và dự đoán là khâu rất quan trọng trong ứng dụng trí tuệ nhân tạo và trong khai phá dữ liệu, phát hiện trí thức. Kỹ thuật Support Vector Machines (SVM) được đánh giá là công cụ mạnh và tinh vi nhất hiện nay cho những bài toán phân lớp phi tuyến. Nhiều những ứng dụng đã và đang được xây dựng dựa trên kỹ thuật SVM rất hiệu quả.

## 1.2. Mục đích, đối tượng và phạm vi nghiên cứu:

Trong khuôn khổ luận văn sẽ nghiên cứu phần bài toán phân lớp quan điểm, cơ sở lý thuyết của phương pháp SVM và các vấn đề liên quan. Phân tích những giải pháp cho phép mở rộng và cải tiến để nâng cao hiệu quả ứng dụng của SVM. Đưa kỹ thuật mờ vào SVM cho phép phân chia không gian dữ liệu một cách tốt hơn, nhằm loại bỏ những vùng không được phân lớp bằng SVM thông thường.

Trình bày tổng quan về bài toán phân lớp quan điểm và cụ thể là bài toán phân lớp phân cực để phân chia các tài liệu chứa quan điểm là tích cực hay tiêu cực. Tìm hiểu dữ liệu quan điểm và viết chương trình thử nghiệm phân lớp phân cực tài liệu sử dụng SVM.

## 1.3. Ý nghĩa khoa học và thực tiễn

SVM là một phương pháp phân lớp hiện đại và hiệu quả, nắm chắc phương pháp này sẽ tạo nền tảng giúp chúng ta trong việc phát triển các giải pháp phân loại và dự đoán..., xây dựng được những ứng dụng quan trọng trong thực tế.

Ứng dụng phân lớp SVM cho bài toán phân lớp quan điểm là bài toán đã và đang được nghiên cứu và phát triển rộng rãi và có ý nghĩa cả về học thuật lẫn ứng dụng thực tế.

Ngoài phần Lời mở đầu và Tài liệu tham khảo và Phụ lục thì nghiên cứu gồm 4 chương chính:

*Chương 1: Tổng quan về đề tài*

*Chương 2: Cơ sở lý thuyết.*

*Chương 3: Cài đặt thử nghiệm*

*Chương 4: Đánh giá kết quả nghiên cứu*

# II. CƠ SỞ LÝ THUYẾT

## 2.1.Đặt vấn đề

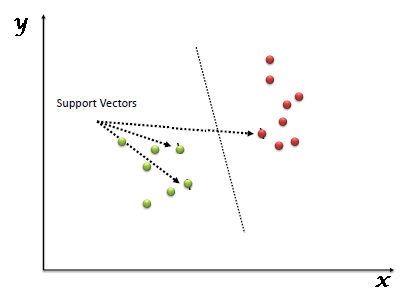
Phân loại thống kê là một nhiệm vụ phổ biến trong ứng dụng trí tuệ nhân tạo. Trong mô hình học có giám sát, thuật toán được cho trước một số điểm dữ liệu cùng với nhãn của chúng thuộc một trong hai lớp cho trước. Mục tiêu của thuật toán là xác định xem một điểm dữ liệu mới sẽ được thuộc về lớp nào. Mỗi điểm dữ liệu được biểu diễn dưới dạng một vector p-chiều, và ta muốn biết liệu có thể chia tách hai lớp dữ liệu bằng một siêu phẳng p − 1 chiều. Đây gọi là phân loại tuyến tính. Có nhiều siêu phẳng có thể phân loại được dữ liệu. Một lựa chọn hợp lý trong chúng là siêu phẳng có lề lớn nhất giữa hai lớp

.

## 2.2. Thuật toán SVM.

### 2.2.1 Khái niệm

Máy vector hỗ trợ (SVM - Support Vector Machine ) là một thuật toán giám sát giúp tìm ra một siêu phẳng phân cách tối ưu để có thể phân chia dữ liệu tuyến tính ra làm hai lớp khác nhau, được sử dụng để phân loại, hồi qui và phát hiện điểm dữ liệu bất thường. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại.

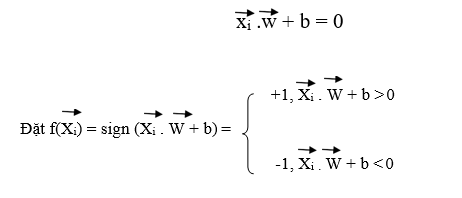
Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Vector Machine là một biên giới để chia hai lớp tốt nhất.

### 2.2.2.Xây dựng bài toán:

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

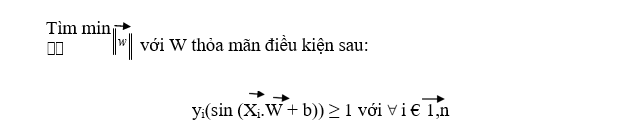
Cho tập mẫu (x1, y1), (x2, y2), … (xf, yf )} với xi ∈ Rn , thuộc vào hai lớp nhãn: yi ∈ {-1,1} là nhãn lớp tương ứng của các xi (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vectơ xi trong không gian:





Như vậy, f(Xi) biểu diễn sự phân lớp của Xi vào hai lớp như đã nêu. Ta nói yi= +1 nếu Xi € lớp I và yi = -1 nếu Xi € lớp II . Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau:



Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi về thành dạng đẳng thức. Một điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc vào các Support Vector và nó có khoảng cách đến mặt phẳng quyết định là . Cho dù các điểm khác bị xóa thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

***TÓM LẠI:*** trong trường hợp nhị phân phân tách tuyến tính, việc phân llớp được thực hiện qua hàm quyết định *f(x) = sign(<w.x> + b),* hàm này thu được bằng việc thay đổi vectơ chuẩn *w*, đây là vectơ để cực đại hóa viền chức năng

Việc mở rộng SVM để phân đa lớp hiện nay vẫn đang được đầu tư nghiên cứu. Có một phương pháp tiếp cận để giải quyết vấn để này là xây dựng và kết hợp nhiều bộ phân lớp nhị phân SVM (Chẳng hạn: trong quá trình luyện với SVM, bài toán phân m lớp có thể được biến đổi thành bài toán phân 2\*m lớp, khi đó trong mỗi hai lớp, hàm quyết định sẽ được xác định cho khả năng tổng quát hóa tối đa). Trong phương pháp này có thể đề cập tới hai cách là *một-đổi-một, một-đối-tất cả*

2.2.3.Ví dụ minh họa:

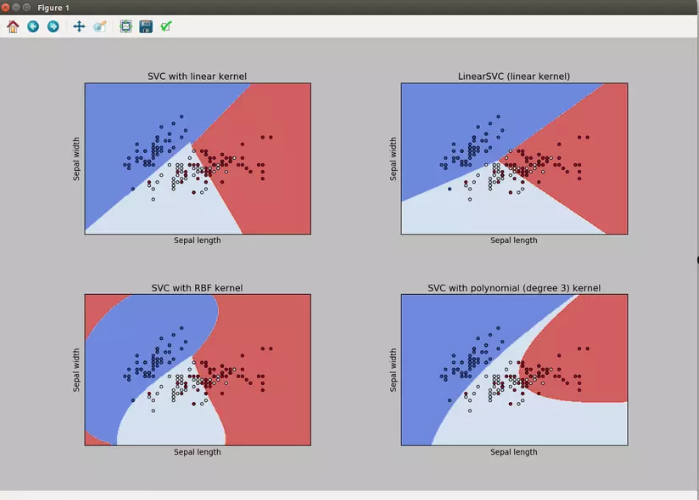
Ứng dụng Support Vector Machine trong bài toán phân loại hoa

* Tập dữ liệu Iris Flowers

Tập dữ liệu này có sẵn trong thư viện skit-learn gồm 50 mẫu về 3 loài hoa khác nhau của họ Iris là (Iris setosa, Iris virginica và Iris versicolor).

Với mỗi một mẫu hoa này có bốn thuộc tính là chiều dài và chiều rộng của đài hoa và cánh hoa với đơn vị centimet.

* Sau quá trình huấn luyện ta thu được đồ thị :



Việc phân lớp chia mặt phẳng tọa độ thành các phần khác nhau. Từ đó, khi có một điểm dữ liệu mới có thể dựa vào tọa độ của chúng để phán đoán xem nó thuộc lớp nào.

# III.CÀI ĐẶT THỬ NGHIỆM

## 3.1. Công nghệ áp dụng

Đối với phần cài đặt thử nghiệm này , nhóm đã sử dụng ngôn ngữ lập trình Python3 cùng với các thư viện liên quan đến việc huấn luyện mô hình như: Matplotlib, Scikit-learn, Pandas, Numpy.

### 3.1.1. Python

Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level)  ngữ nghĩa động (dynamic semantics). Python hỗ trợ các module và gói (packages), khuyến khích chương trình module hóa và tái sử dụng mã. Trình thông dịch Python và thư viện chuẩn mở rộng có sẵn dưới dạng mã nguồn hoặc dạng nhị phân miễn phí cho tất cả các nền tảng chính và có thể được phân phối tự do.

Python cung cấp mã ngắn gọn và dễ đọc. Trong khi các thuật toán của AI và Machine learning là hết sức phức tạp, nhưng sự đơn giản của Python cho phép các nhà phát triển tạo nên các hệ thống rất chính xác và đáng tin cậy. Ngoài ra Python còn có rất nhiều thư viện hỗ trợ cho việc huấn luyện mô hình.

### 3.1.2. Các thư viện sử dụng.

- Numpy: là một thư viện toán học phổ biến và mạnh mẽ của Python. Cho phép làm việc hiệu quả với ma trận và mảng, đặc biệt là dữ liệu ma trận và mảng lớn với tốc độ xử lý nhanh hơn nhiều lần khi chỉ sử dụng “core Python” đơn thuần.

- Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán ứng dụng trí tuệ nhân tạo được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling

- Matplotlib là một thư viện vẽ đồ thị rất mạnh mẽ hữu ích cho những người làm việc với Python và NumPy.

- Pandas là một thư viện để thực hiện phân tích và thao tác dữ liệu; bất kỳ loại xử lý, phân tích, lọc và tổng hợp dữ liệu nào. Pandas là một trong những công cụ quan trọng trong việc hỗ trợ, xử lý và phân tích dữ liệu với mã nguồn mở nhanh, mạnh, linh hoạt và dễ sử dụng,

## 3.2. Giải thuật.

**Diagram, engineering drawing

Description automatically generated**

Các bước thực hiện quá trình huấn luyện và xác định được thể hiện trong dữ liệu từ Bank-data ban đầu được xử lý , chọn ra các thuộc tính cần thiết cho huấn luyện rồi được tác thành 2 tập là tập train và tập test. Trong đó, Bank-data là tập dữ liệu được cho ban đầu (bao gồm 600 khách hàng ), tập dùng để training (huấn luyện), tập dùng để test (kiểm tra). Trong quá trình thực nghiệm, do số mẫu ít nên tập huấn luyện và tập kiểm tra được lấy ngẫu nhiên từ tập Bank-data với tỷ lệ lần lượt là 80% và 20% . Sau quá trình huấn luyện, ta cho tiến hành dự đoán kết quả dựa theo tập kiểm tra thu được kết quả dự đoán rồi so sánh kết quả dự đoán này với kết quả của tập test.

## 3.3. Xử lý dữ liệu và chọn thuộc tính .

A picture containing diagram

Description automatically generated

Xét tập dữ liệu Bank-data :Table, Excel

Description automatically generated

Hình 3.2.1. Dữ liệu của Bank-data (9 dòng đầu)

Tập dữ liệu ban đầu gồm 11 thuộc tính trong đó thuộc tính “ID ” không sử dụng trong huấn luyện mô hình, thuộc tính “pep” là kết quả và các thuộc tính còn lại được sử dụng trong việc đánh giá mô hình và kết quả dự đoán.

Xét các thuộc tính có dữ liệu dạng văn bản, để dàng cho việc tính toán nhóm đã chuyển các dữ liệu này về dạng số thực. Việc thay đổi được thực hiện theo quy tắc sau :Table

Description automatically generated

Hình 3.2.2. Chuyển đổi dữ liệu

* Thuộc tính ‘Sex’ có 2 giá trị : ‘Female’ và ‘Male’ được gán với giá trị trương ứng lần lượt là 0 và 1
* Thuộc tính ‘Region’ có 4 giá trị ‘Rural’, ‘Inner\_City’, ‘Suburban’, ‘Town’ được gán với giá trị trương ứng lần lượt là 0, 1, 2 và 3
* Các thuộc tính ‘Married’, ‘Car’, ‘Save\_Act’, ‘Current\_Act’, ‘Mortgage’, ‘Pep’ đều có 2 giá trị : ‘YES’, ‘NO’ được gán với giá trị trương ứng lần lượt là 0 và 1

Sau khi chuyển đổi ta được tập dữ liệu sau khi xử lý :Chart

Description automatically generated

Hình 3.2.3. Dữ liệu của Bank-data sau khi xử lý (9 dòng đầu)

## 3.4.Chia dữ liệu.

Sau khi xử lý dữ liệu về dạng có thể xửa lý được, nhóm tiến hành chia ngẫu nhiên tập dữ liệu này thành 2 phần là tập huấn luyện và tập kiểm tra. Tập kiểm tra chiếm 20% dữ liệu lấy vào trương đương với 120 khách hàng. Bởi vì dữ liệu cho việc huấn luyện và kiểm tra được lấy ngẫu nhiên, kết quả kiểm tra thu thập được sau mỗi lần đều có sự khác nhau.

## 3.5. Kết quả chương trình thử nghiệm.

Nhóm xin được trình bày ra kết quả của 5 lần chạy thử nghiệm để có góc nhìn tổng quan:

Lần 1: Chart, treemap chart

Description automatically generated

Kết quả:

Dự đoán đúng 70/120 ~ 58% kết quả so với tập kiểm tra trong đó:

* 16/56 ~ 29% kết quả dự đoán đúng đối với đáp án “Yes” tức giá trị 0
* 54/64 ~ 84 % kết quả dự đoán đúng đối với đáp án “No” tức giá trị 1
* Lần 2:Chart

  Description automatically generated

Kết quả:

Dự đoán đúng 69/120 ~ 58% kết quả so với tập kiểm tra trong đó:

* 14/51 ~ 27% kết quả dự đoán đúng đối với đáp án “Yes” tức giá trị 0
* 55/65 ~ 84 % kết quả dự đoán đúng đối với đáp án “No” tức giá trị 1
* Lần 3: Chart

  Description automatically generated

Kết quả:

Dự đoán đúng 66/120 ~ 56% kết quả so với tập kiểm tra trong đó:

* 9/57 ~ 16% kết quả dự đoán đúng đối với đáp án “Yes” tức giá trị 0
* 57/63 ~ 90% kết quả dự đoán đúng đối với đáp án “No” tức giá trị 1
* Lần 4Chart, treemap chart

  Description automatically generated

Kết quả:

Dự đoán đúng 79/120 ~ 67% kết quả so với tập kiểm tra trong đó:

* 20/48 ~ 42% kết quả dự đoán đúng đối với đáp án “Yes” tức giá trị 0
* 59/72 ~ 82% kết quả dự đoán đúng đối với đáp án “No” tức giá trị 1
* **Lần 5**Chart

  Description automatically generated

Kết quả:

Dự đoán đúng 66/120 ~ 56% kết quả so với tập kiểm tra trong đó:

* 15/54 ~ 28% kết quả dự đoán đúng đối với đáp án “Yes” tức giá trị 0
* 51/66 ~ 77% kết quả dự đoán đúng đối với đáp án “No” tức giá trị 1

# IV. ĐÁNH GIÁ VÀ KẾT QUẢ NGHIÊN CỨU

Qua quá trình nghiêm cứu nhóm đã ứng dụng thành công phân lớp khách hàng của ngân hàng bằng kỹ thuật ứng dụng trí tuệ nhân tạo vecto hỗ trợ (SVM). Sau nhiều lần thực hiện, hiệu quả của mô hình đã dự đoán đúng khoảng 50% - 60% so với kế quả của tập kiểm tra và tỷ lệ chính xác rất cao ở câu trả lời “No”, kết quả này là một kết quả tốt có thể dùng để dự đoán khách hàng.

Do thời gian có hạn, nên một số vấn đề vẫn chưa hoàn chỉnh. Tuy nhiên, nhóm cũng đạt được một số kết quả:

* Trình bày phương pháp SVM. Đây là một phương pháp phân lớp hiệu quả được nghiên cứu nhiều nhất trong thời gian qua.
* Cài đặt được một số công cụ giúp đỡ cho việc xây dựng chương trình như Python, các thư viện
* Cài đặt được chương trình để trích đặc trưng và tạo dữ liệu cho phân lớp SVM.

# TÀI LIỆU THAM KHẢO

<https://vietnambiz.vn/may-vector-ho-tro-support-vector-machine-svm-la-gi-20200226223210903.htm>

<https://vi.wikipedia.org/wiki/M%C3%A1y_vect%C6%A1_h%E1%BB%97_tr%E1%BB%A3#:~:text=M%C3%A1y%20vect%C6%A1%20h%E1%BB%97%20tr%E1%BB%A3%20(SVM,v%C3%A0o%20hai%20l%E1%BB%9Bp%20kh%C3%A1c%20nhau>.

<https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB>

<https://viblo.asia/p/ung-dung-support-vector-machine-trong-bai-toan-phan-loai-hoa-PdbGnLXBkyA>