|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



NGUYỄN ĐỨC HẢI

TÊN ĐỀ TÀI: PHÂN LOẠI NGUY CƠ UNG THƯ CỔ TỬ CUng BẰNG PHƯƠNG PHÁP HỌC MÁY.

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2023

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN ĐỨC HẢI

TÊN ĐỀ TÀI: PHÂN LOẠI NGUY CƠ UNG THƯ CỔ TỬ CUng BẰNG PHƯƠNG PHÁP HỌC MÁY.

|  |  |
| --- | --- |
|  | Ngành : Công nghệ thông tin |
|  | Mã số:7480201 |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | ThS. Nguyễn Đắc Hiếu |
|  |  |

HÀ NỘI, NĂM 2023

**GÁY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP**

**NGUYỄN ĐỨC HẢI ĐỒ ÁN TỐT NGHIỆP HÀ NỘI, NĂM 2023**

LỜI CAM ĐOAN

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào.Việc tham khảo các nguồn tài liệu đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

|  |  |
| --- | --- |
|  | **Tác giả ĐATN**  *Chữ ký*  Hải  **Nguyễn Đức Hải** |

LỜI CÁM ƠN

Qua hơn ba tháng tìm hiểu và thực hiện đề tài “Phân loại nguy cơ ung thư cổ tử cung bằng phương pháp học máy”, đến nay đã hoàn thành. Trong suốt quá trình thực hiện đề tài, em đã nhận được rất nhiều sự giúp đỡ và chỉ bảo nhiệt tình của thầy cô và các bạn.

Đặc biệt em muốn gửi tới thầy ThS. Nguyễn Đắc Hiếu, giảng viên trường Đại học Thuỷ Lợi người tận tình hướng dẫn em từ những ngày bắt đầu xây dựng đồ án đã có những góp ý, truyền đạt các kinh nghiệm quý báu giúp em hoàn thành được Đồ án tốt nghiệp lời cảm ơn sâu sắc nhất. Em trân thành cảm ơn ạ.

Mặc dù em đã hoàn thành đồ án, nhưng kiến thức còn hạn chế và chưa có nhiều kinh nghiệm trong thực tế nên không thể tránh khỏi những thiếu sót. Vì vậy, em kính mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để kiến thức của em trong lĩnh vực này được hoàn thiện hơn.

Em xin chân thành cảm ơn!

MỤC LỤC

[DANH MỤC CÁC HÌNH ẢNH v](#_Toc154235602)

[DANH MỤC BẢNG BIỂU vi](#_Toc154235603)

[DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ vii](#_Toc154235604)

[CHƯƠNG 1 GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN 1](#_Toc154235605)

[1.1 Lý do chọn đề tài 1](#_Toc154235606)

[1.2 Mục tiêu, phạm vi 1](#_Toc154235607)

[1.3 Đối tượng, phạm vi nghiên cứu 2](#_Toc154235608)

[1.4 Khu vực nghiên cứu 3](#_Toc154235609)

[CHƯƠNG 2 CÁC KĨ THUẬT DÙNG TRONG BÀI TOÁN 3](#_Toc154235610)

[2.1 Tổng quan về học máy 3](#_Toc154235611)

[2.1.1 Khái niệm về học máy 3](#_Toc154235612)

[2.1.2 Các phương pháp học máy 7](#_Toc154235613)

[2.2 Tiếp cận cơ sở lý thuyết 8](#_Toc154235614)

[2.3 Một số thuật toán học máy cho bài toán phân lớp 9](#_Toc154235615)

[2.3.1 Thuật toán ID3 9](#_Toc154235616)

[**2.3.1.1** **Khái niệm thuật toán ID3** 9](#_Toc154235617)

[**2.3.1.2** **Hàm số entropy** 10](#_Toc154235618)

[**2.3.1.3 Ưu nhược điểm** 11](#_Toc154235619)

[2.3.2 Thuật toán CART 11](#_Toc154235620)

[**2.3.2.1** **Khái niệm thuật toán Cart** 11](#_Toc154235621)

[**2.3.2.2** **Gini Index** 12](#_Toc154235622)

[**2.3.2.3** **Ưu nhược điểm** 13](#_Toc154235623)

[**2.3.3 Thuật toán Support Vector Machine** 14](#_Toc154235624)

[**2.3.3.1 Khái niệm thuật toán SVM** 14](#_Toc154235625)

[**2.3.3.2 Công thức** 19](#_Toc154235626)

[**2.3.3.3 Ưu nhược điểm** 20](#_Toc154235627)

[2.4 Các kĩ thuật dùng cho bài toán 21](#_Toc154235628)

[2.5 Các phương pháp đánh giá mô hình 22](#_Toc154235629)

[2.5.1 Accuracy 22](#_Toc154235630)

[2.5.2 Confusion Matrix 22](#_Toc154235631)

[2.5.3 Precision và Recall 23](#_Toc154235632)

[2.5.4 F1-score 24](#_Toc154235633)

[CHƯƠNG 3 ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH 25](#_Toc154235634)

[3.1 Phân tích chi tiết bài toán 25](#_Toc154235635)

[3.2 Các công cụ, thư viện dùng để xây dựng 28](#_Toc154235636)

[3.2.1 Sử dụng ngôn ngữ Python 28](#_Toc154235637)

[3.2.2 Các ứng dụng của python 29](#_Toc154235638)

[3.2.3 Các thư viện sử dụng 30](#_Toc154235639)

[3.3 Xây dựng mô hình 32](#_Toc154235640)

[**3.3.1 Dữ liệu thực nghiệm** 32](#_Toc154235641)

[**3.3.2** **Thực hiện xây dựng mô hình** 33](#_Toc154235642)

[CHƯƠNG 4 KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH 35](#_Toc154235643)

[4.1 Kết quả 35](#_Toc154235644)

[4.1.1 Kết quả thực hiện đối với Confusion Matrix 35](#_Toc154235645)

[4.1.2 Kết quả thực hiện đối với mô hình ID3 36](#_Toc154235646)

[4.1.3 Kết quả thực hiện đối với mô hình CART 36](#_Toc154235647)

[4.1.4 Kết quả thực hiện đối với mô hình SVM 36](#_Toc154235648)

[4.1.5 Một số kết quả thử nghiệm dự đoán khi nhập dữ liệu trực tiếp 37](#_Toc154235649)

[4.2 Đánh giá mô hình 40](#_Toc154235650)

[CHƯƠNG 5 KẾT LUẬT 40](#_Toc154235651)

[TÀI LIỆU THAM KHẢO 40](#_Toc154235652)

[PHỤ LỤC 41](#_Toc154235653)

DANH MỤC CÁC HÌNH ẢNH

[Hình 2.1 Biểu tượng (logo) của trường Đại học Thủy lợi 6](file:///F:\BOX\Box%20Sync\HD_DATN_LVThS_LATS\Huong_dan_Trinh_bay_LATS\Huong%20dan%20trinh%20bay%20DATN1.docx#_Toc414436222)

DANH MỤC BẢNG BIỂU

[Bảng 2.1 Cách định dạng lề giấy 3](file:///D:\Box\My%20Box%20Files\Quan%20Ly%20NCS\Huong%20dan%20trinh%20bay%20LATS_updated%2001-2013\Huong%20dan%20trinh%20bay%20LATS.docx#_Toc405384698)

[Bảng 2.2 Tóm tắt các kiểu định dạng (style) cho các đề mục 7](file:///D:\Box\My%20Box%20Files\Quan%20Ly%20NCS\Huong%20dan%20trinh%20bay%20LATS_updated%2001-2013\Huong%20dan%20trinh%20bay%20LATS.docx#_Toc405384699)

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

**ĐATN** Đồ án tốt nghiệp

# GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN

## Lý do chọn đề tài

Ung thư cổ tử cung là một vấn đề quan trọng trong lĩnh vực y tế. Đặc biệt là nó có liên quan đến tỷ lệ tử vong ở phụ nữ trên toàn thế giới. Ung thư cổ tử cung là một khối u ác tính xảy ra khi các tế bào mô bao phủ cổ tử cung (còn gọi là mô niêm mạc) bắt đầu phát triển và sinh sản không kiểm soát được mà không tuân theo cơ chế phân chia tế bào thích hợp. Sự phát triển không kiểm soát của tế bào này dẫn đến hình thành khối u ác tính trên cổ tử cung, và nếu không được phát hiện kịp thời, nó có thể lan ra các cơ quan và mô xung quanh. Điều này có thể gây ra nhiều biến chứng và nguy cơ nghiêm trọng cho sức khỏe của phụ nữ.

Những năm gần đây, việc ứng dụng trí tuệ nhân tạo vào các bài toán dự báo, phân loại đã được nghiên cứu và mở rộng. Đặc biệt, các mô hình học máy, học sâu đang phát triển mạnh và được ứng dụng trong nhiều lĩnh vực như: y tế, tài chính, người máy, …, bao gồm cả phân loại nguy cơ ung thư cổ tử cung. Trong lĩnh vực y tế đã có một số chuyên nghành ứng dụng AI chuẩn đoán hình ảnh nội soi… và đạt được những thành quả bước đầu rất khả quan. Với khả năng tự cập nhật trong quá trình huấn luyện, các mô hình học máy, học sâu có thể khắc phục được những hạn chế trong mô hình phân loại. Bằng cách xử lý dữ liệu lớn và phức tạp về thông tin tiền sử bệnh nhân, các kết quả xét nghiệm, dấu hiệu lâm sàng và các yếu tố nguy cơ khác, mô hình học máy có thể phân loại nguy cơ ung thư cổ tử cung với độ chính xác cao. Từ đó sẽ giúp cải thiện khả năng phát hiện sớm và đánh giá rủi ro, tiết kiệm được nhiều thời gian và chi phí. Mang lại nhiều lợi ích quan trọng trong lĩnh vực y tế và chăm sóc sức khỏe cho phụ nữ trên toàn thế giới.

Từ những vấn đề thực tế, cùng với sự phát triển của công nghệ hiện đại và ứng dụng các phương pháp học máy. Vì vậy, em lựa chọn đề tài: “**Phân loại nguy cơ ung thư cổ tử cung bằng phương pháp học máy**”.

## Mục tiêu, phạm vi

* Mục tiêu

+Chuẩn đoán và dự đoán: phát triển mô hình học máy để chuẩn đoán ung thư cổ tử cung và dự đoán nguy cơ mắc bệnh.

* Phạm vi

+Có thể mở rộng nghiên cứu đối với nhiều quốc gia, tùy thuộc vào khả năng thu thập dữ liệu và mức độ phổ biến của bệnh.

+Dữ liệu: thu thập dữ liệu về lối sống chi tiết, về các yếu tố nguy cơ, lịch sử bệnh, và kết quả kiểm tra xét nghiệm của những người phụ nữ.

+Thuật toán: Áp dụng và đánh giá các thuật toán học máy như SVM, Decision Trees để xây dựng mô hình.

+Đánh giá: Đánh giá hiệu suất của các mô hình trên qua các độ đo như: Precision, Recall, F1 Score.

## Đối tượng, phạm vi nghiên cứu

* Đối tượng nghiên cứu

+Tiêu chuẩn lựa chọn:

+Giới tính: Chỉ chọn phụ nữ vì ung thư cổ tử cung liên quan chặt chẽ đến hệ sinh sản ở

nữ giới.

+Độ tuổi: Chọn phụ nữ ở các độ tuổi có nguy cơ cao về ung thư cổ tử cung. Điều này

có thể là phụ nữ trưởng thành có thể bắt đầu quan hệ tình dục và có nguy cơ nhiễm HPV,

một trong những yếu tố gây ra ung thư cổ tử cung.

+Số lượng đối tác tình dục: Những người có số lượng đối tác tình dục nhiều có thể có

nguy cơ cao hơn.

+Có thông tin y tế đầy đủ.

+Tiêu chuẩn loại trừ:

+Không phải là phụ nữ: Nếu là nam giới thì loại trừ hết vì vấn đề nghiên cứu này chỉ

liên quan đến phụ nữ.

+Dữ liệu thiếu: Nếu thiếu các dữ liệu thông tin quan trọng như tuổi, số lượng đối tác

tình dục, hoặc các chỉ số liên quan đến ý tế, xét nghiệm, có thể loại trừ.

* Phạm vi nghiên cứu

+Sử dụng dữ liệu y tế và lối sống chi tiết để xây dựng mô hình.

+Dữ liệu có thể thu thập được từ các cơ sở y tế, cuộc khảo sát và các nguồn thông tin y

tế khác.

## Khu vực nghiên cứu

Khía cạnh địa lý: Nghiên cứu sẽ chọn một số khu vực địa lý quan trọng có tình trạng

ung thư cổ tử cung đáng chú ý. Các khu vực này có thể được xác định dựa trên mức độ

phổ biến của bệnh, độ phức tạp của hệ thống ý tế, và các yếu tố xã hội địa phương. Điều

này giúp tập trung nghiên cứu vào những vùng có nhu cầu cấp bách về việc sớm phát

hiện và quản lý ung thư cổ tử cung.

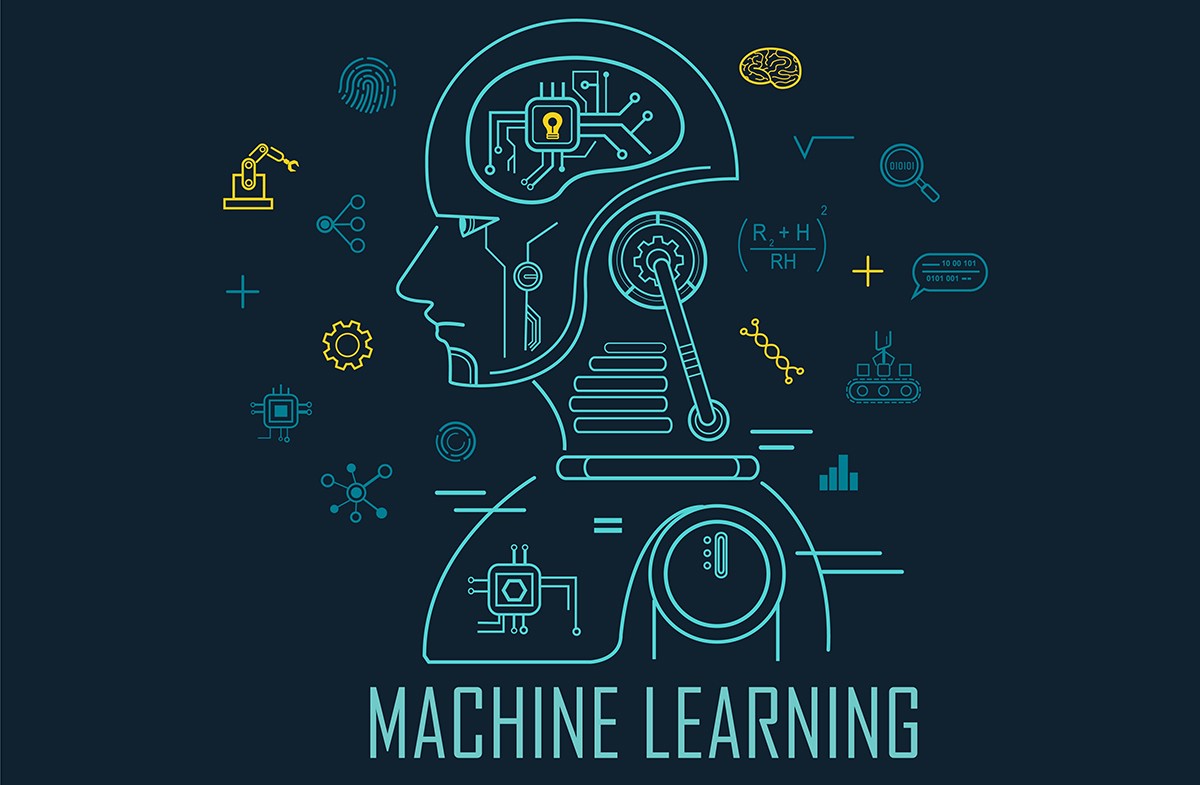
# CÁC KĨ THUẬT DÙNG TRONG BÀI TOÁN

## Tổng quan về học máy

### Khái niệm về học máy

Học máy hay machine learning là một lĩnh vực của trí tuệ nhân tạo, nó nghiên cứu về cách xây dựng các mô hình và thuật toán tự động học từ dữ liệu và cải thiện hiệu suất của chúng theo thời gian. Trong học máy, các thuật toán được thiết kế để cho máy tính học từ dữ liệu một cách tự động mà không cần được lập trình bằng tay. Các mô hình học máy có thể học từ các tập dữ liệu mẫu đầu vào để dự đoán kết quả trong tương lai. Đầu vào ở đây đề cập đến tất cả các biến dự đoán hoặc biến độc lập dùng để đi vào mô hình và đầu ra thường là các giá trị dự đoán (có thể là giá trị phân loại hoặc số,...).

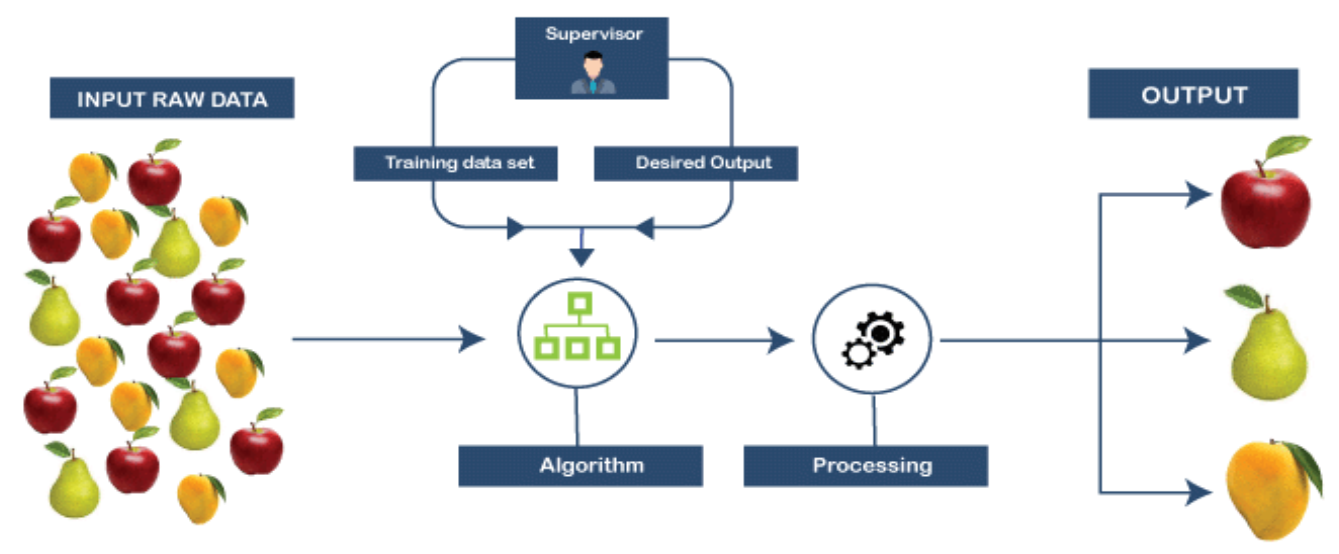
Quá trình Machine Learning bắt đầu với việc các dữ liệu (số liệu, hình ảnh hoặc văn bản, âm thanh,...) được thu thập từ các nguồn. Dữ liệu đó sẽ được sử dụng để đào tạo hoặc cung cấp thông tin cho mô hình học máy. Càng nhiều dữ liệu được cung cấp sẽ càng tốt. Lập trình viên sau khi chọn mô hình học máy sử dụng sẽ tiến hành huấn luyện mô hình bằng cách cung cấp dữ liệu và để mô hình máy tính tự học từ dữ liệu đó nhằm tìm ra các mẫu hoặc đưa ra dự đoán. Theo thời gian, lập trình viên cũng có thể điều chỉnh và thêm các dữ liệu mới để cải thiện độ chính xác của mô hình.



Hình . Minh họa học máy (Machine Learning)

Học máy (Machine Learning) đóng vai trò quan trọng trong nhiều lĩnh vực khác nhau, đặc biệt là trong lĩnh vực công nghệ thông tin, khoa học dữ liệu và trí tuệ nhân tạo. Các ứng dụng của học máy bao gồm:

+ Dự đoán và phân loại: Học máy có thể dự đoán và phân loại các đối tượng dựa trên dữ liệu đầu vào, ví dụ như phân loại email là thư rác hay thư quan trọng, dự đoán giá của một căn nhà dựa vào các thông tin như vị trí, diện tích, ... hay xác định một bệnh nhân có bị bệnh ung thư hay không từ các triệu chứng quan sát và ghi chép được.

 Hình . Ứng dụng học máy cho dự đoán và phân loại

+Xử lý ngôn ngữ tự nhiên: là một lĩnh vực trong học máy và trí tuệ nhân tạo mà nói đến khả năng của máy tính hiểu, hiểu và tương tác với ngôn ngữ tự nhiên của con người, ví dụ như: dịch máy, nhận dạng giọng nói…



Hình .3 Ứng dụng học máy cho nhận dạng giọng nói

+Thị giác máy tính: là một lĩnh vực trong trí tuệ nhân tạo và học máy tập trung vào việc giúp máy tính hiểu và hiểu rõ hình ảnh, video, giống như các con người thị giác. Ví dụ như: nhận diện phân loại đối tượng, nhận diện khuôn mặt và điểm ảnh, nhận diện vật thể và giám sát…



Hình .4 Ứng dụng học máy cho nhận diện khuôn mặt

+Robot được tích hợp trí tuệ nhân tạo và học máy để có khả năng tự động hóa quá trình quyết định, học hỏi từ dữ liệu và thích ứng với môi trường xung quanh. Có nhiều loại như: robot dịch vụ, robot giáo dục, robot giao thông, robot y tế…

 Hình .5 Ứng dụng học máy cho robot

### Các phương pháp học máy

Để phân loại machine learning có rất nhiều cách nhưng thông thường thì machine  
learning được phân làm bốn loại chính sau:

**Học có giám sát (Supervised learning):** Học có giám sát có thể xây dựng một mô hình dự đoán có khả năng dự đoán nhãn của các mẫu dữ liệu mới mà nó chưa từng thấy. Là hoạt động mà máy tính học với dữ liệu đã được gán nhãn (label), nói dễ hiểu hơn thì với học có giám sát khi đưa mỗi một đầu vào X, thì sẽ thu được một nhãn Y tương ứng.

+ Các thuật toán trong học có giám sát được chia thành hai nhóm phân lớp hay cũng có thể gọi là phân loại (classification) và hồi quy (regression).

+ Phân loại có một số thuật toán nổi bật như: Decision Trees, Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN),...

+ Còn hồi quy có thể kể đến một số thuật toán như: Linear Regression, Neural Network, Support Vector Regression,...

**Học không giám sát (Unsupervised learning):** Là hoạt động mà máy tính học trên dữ  
liệu mà không được gán nhãn, khi đó các thuật toán học máy sẽ tìm ra sự tương quan giữa các dữ liệu hoặc mô hình hóa dữ liệu đó. Từ đó, có thể phân loại các dữ liệu về thành các nhóm, các lớp giống nhau mà chúng đã được học hoặc giảm số chiều dữ liệu (dimension reduction). Học không giám sát được biết đến với các thuật toán phân cụm tiêu biểu như: K-Means Clustering, Gaussian Mixture, DBSCAN Clustering,...

**Học tăng cường (Reinforcement learning):** Là hoạt động mà máy tính học tập trung vào việc học và ra quyết định trong môi trường có tính tương tác. Trong Unsupervised learning, “agent” được coi là một đại diện học tập tương tác qua lại với một môi trường và tìm cách tối đa hóa một reward function (hàm phần thưởng) trong suốt quá trình ấy. Agent không được cung cấp dữ liệu có nhãn mà nó phải tự tìm hiểu, tạo ra các hành động để tìm thấy mục đích. Thử và sai là cách agent thử nghiệm các hành động trong môi trường rồi vẽ ra tiếp các hành động sau đó. Học tăng cường có các thuật toán ra quyết định (Decision Making) như: Q-Learning, R-Learning, TD Learning,...

**Học bán giám sát (Semi-supervised learning):** Là một phương pháp học máy mà mô hình được đào tạo trên một tập dữ liệu lớn, trong đó chỉ một phần nhỏ được gán nhãn. Phương pháp này kết hợp giữa học giám sát và học không giám sát nhằm tận dụng cả thông tin có nhãn và không nhãn để tạo ra mô hình có khả năng dự báo tốt trên dữ liệu mới. Học bán giám sát có một số thuật toán hay sử dụng như: co-training, self-training, semi-supervised SVM

## Tiếp cận cơ sở lý thuyết

* Cây quyết định (Decision tree): là một mô hình có giám sát (supervised learning), có

thể được áp dụng cho cả bài toán phân lớp (classification) và hồi quy (regression).

Các thuộc tính của cây quyết định là: Thuộc tính rời rạc và không có thứ tự và thuộc tính liên tục.

* Cây quyết định:

+ Có cấu trúc cây giống như biểu đồ luồng.

+ Mỗi nút trong thể hiện một sự kiểm tra trên một thuộc tính.

+ Mỗi nhánh đại diện cho một kết quả của sự kiểm tra.

+ Các nút lá đại diện cho các nhãn lớp hoặc phân khối lớp

* Thuật toán cơ bản (thuật toán tham lam)

+ Cây quyết định được xây dựng theo cách chia để trị từ trên xuống (top-down).

+ Tại các vị trí khởi tạo, tất cả các mẫu thuộc nút gốc.

+ Các thuộc tính được phân loại (nếu giá trị của thuộc tính là liên tục, thì phải được rời rạc hóa trước).

+ Các mẫu được phân chia đệ quy dựa vào các thuộc tính được chọn.

+ Thuộc tính kiểm tra được lựa chọn vào kinh nghiệm hoặc độ đo thống kê.

* Điều kiện dừng phân chia

+ Tất cả các mẫu của nút xem xét thuộc tính cùng một lớp.

+ Không có thuộc tính nào để phân chia – biểu quyết đa số được sử dụng để gán nhãn phân loại cho lá.

+ Không còn mẫu nào.

* Cây quyết định bao gồm một số thuật toán để xây dựng cây và thực hiện các quyết

định:

+ ID3 (Iterative Dichotomiser 3): Thuật toán này sử dụng độ do Entropy để tìm cách chia dữ liệu hiệu quả nhất. Nó chia cây theo cách tốt nhất để giảm độ không chắc chắn trong quyết định.

+ CART (Classification and Regression Trees): Sử dụng Gini Index để đo lường độ không thuần khiết của dữ liệu. Nó giảm Gini Index bằng cách chọn cách chia tốt nhất.

* Support Vector Machine (SVM): là một thuật toán học máy thuộc lớp các thuật toán

phân loại và hồi quy. Mục tiêu của SVM là tạo ra một đường ranh giới (decision boundary) sao cho khoảng cách từ mỗi điểm dữ liệu đến đường ranh giới là lớn nhất. SVM thường được sử dụng trong bài toán phân loại nhưng cũng có thể áp dụng vào bài toán hồi quy.

* Thuật toán SVM thường hiệu quả trong việc xử lý các dữ liệu tuyến tính và phi tuyến

tính và nó đã được áp dụng trong nhiều lĩnh vực như thị giác máy tính, dự đoán chuỗi thời gian, phân loại y tế và phân loại văn bản.

## Một số thuật toán học máy cho bài toán phân lớp

### Thuật toán ID3

#### **Khái niệm thuật toán ID3**

Trước hết ta cần tìm hiểu về Decison Tree. Decison Tree hay cây quyết định là mô hình thuộc nhóm học có giám sát. Nó có các nút trong cây là các biến đầu vào, cành là các quy tắc quyết định còn lá là biến mục tiêu hay kết quả. Nó phân chia dữ liệu dựa trên các thuộc tính đầu vào.

Thuật toán ID3 (Iterative Dichotomiser 3) là thuật toán để xây dựng cây quyết định, áp dụng cho các bài toán phân lớp. Nó được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen, Stone trong thống kê. Thuật toán ID3 có thể giải thích theo ba bước:

Bước 1: Đầu tiên phải xác định được các thuộc tính tốt nhất để phân chia dữ liệu, tính độ đo của mỗi thuộc tính trong tập dữ liệu đầu vào (thường sử dụng hàm số entropy) và chọn thuộc tính có độ đo nhỏ nhất để phân chia dữ liệu.

Bước 2: Tiếp theo là xây dựng nên cây quyết định bằng cách tạo nút quyết định cho thuộc tính được chọn trước đó, tạo ra các nhánh con của nút quyết định dựa trên các giá trị riêng biệt của thuộc tính phân chia, sau đó chia tập dữ liệu đầu thành tập con tương ứng các nhánh con. Qúa trình này cứ lặp đi lặp lại cho đến khi đạt được điều kiện dừng.

Bước 3: Cuối cùng là đánh nhãn cho lá bằng cách chọn nhãn phổ biến nhất trong tập dữ liệu tương ứng.

#### **Hàm số entropy**

Xét một bài toán với C class khác nhau. Giả sử ta đang làm việc với một *non-leaf node* (nút quyết định: các nút khác nút lá) với các điểm dữ liệu tạo thành một tập S với số phần tử là |S|=N. Giả sử thêm rằng trong số N điểm dữ liệu này điểm thuộc vào class. Xác suất để mỗi điểm dữ liệu rơi vào một class được xấp xỉ bằng . Như vậy, entropy tại node này được tính bởi:

(2- )

Tiếp theo, giả sử thuộc tính được chọn là x. Dựa trên x, các điểm dữ liệu trong S được phân ra thành K child node với số điểm trong mỗi child node lần lượt là . Ta có công thức tổng trọng số entropy của mỗi child node:

(2- )

Tiếp theo, ta định nghĩa information gain dựa trên thuộc tính x:

(2- )

Điều kiện dừng, để tránh quá khớp (overfitting), một trong số các phương pháp sau có thể được sử dụng. Tại một nút (node), nếu một trong số các điều kiện sau đây xảy ra, ta không tiếp tục phân chia node đó và coi nó là một nút lá (leaf node):

+ Nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một lớp.

+ Nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.

+ Nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của tree này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.

+ Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

+ Nếu tổng số nút lá vượt quá một ngưỡng nào đó.

+ Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

**2.3.1.3 Ưu nhược điểm**

**Ưu điểm:**

Thuật toán này khá dễ hiểu và giải thích bằng cách sử dụng các quy tắc quyết định. Nó có khả năng làm việc với dữ liệu không hoàn hảo (bị khuyết, nhiễu) mà không cần tới tiền xử lý.

**Nhược điểm:**

ID3 không xử lý được thuộc tính liên tục mà chỉ phù hợp với các dữ liệu có tính rời rạc. Quá khớp (overfitting): ID3 có xu hướng tạo ra cây quyết định phức tạp và quá khớp dữ liệu huấn luyện. Tức là cây quyết định có thể tạo ra quyết định chi tiết và phân lớp chính xác dữ liệu huấn luyện nhưng lại hoạt động không được tốt lắm trên dữ liệu kiểm tra và dữ liệu mới. Nó thường xảy ra trong trường hợp cây quyết định phân chia quá chi tiết để tương thích với nhiễu hoặc biến động không cần thiết trong dữ liệu huấn luyện.

### Thuật toán CART

#### **Khái niệm thuật toán Cart**

Thuật toán CART là một thuật toán cây quyết định được sử dụng cho cả bài toán phân loại và hồi quy. Nó được giới thiệu bởi Leo Breiman, Jerome Fridman, Richard Olshen, và Charles Stone vào năm 1984. CART tập trung vào việc chia nhóm dữ liệu thành các thành phần con, mỗi phần con tương ứng bởi một quy tắc quyết định.

+Cách hoạt động của thuật toán:

1. Chọn thuộc tính tốt nhất: Tìm thuộc tính và giá trị ngưỡng tốt nhất để chia dữ liệu thành hai phần con sao cho hiệu suất của mô hình là tốt nhất. Các phương pháp đánh giá có thể bao gồm độ thuần khiết của các phần con.
2. Tạo nút: Tạo một nút quyết định trong cây với quy tắc chia được chọn. Nút này biểu diễn một quyết định dựa trên thuộc tính và giá trị ngưỡng
3. Phân loại hoặc hồi quy: Nếu nút là lá (không thể phân chia tiếp ), nó sẽ được gán nhãn (phân loại) hoặc dự báo giá trị (hồi quy) dựa trên giá trị trung bình của các mẫu trong nút.
4. Lặp lại cho cả hai phần con: Quy trình trên được lặp lại cho cả hai phần con được tạo ra bởi quy tắc chia.
5. Tiêu chí dừng: Quy trình sẽ tiếp tục đến khi một điều kiện dừng được đáp ứng, chẳng hạn như đạt đến độ sâu tối đa của cây. Số lượng mẫu tối thiểu trong mỗi lá, hoặc độ thuần khiết không đạt đến ngưỡng nhất định

#### **Gini Index**

Gini =1-

Trong đó

C: số lớp cần phân loại

pi= ni/N

ni là số phần tử ở lớp thứ i

N là số phần tử ở node đó

Gini\_index = gini(p)-

Trong đó

Gini(p): chỉ số gini ở node cha

K: số node con được tách ra

Gini(): chỉ số gini ở node con thứ k

M: số phần tử ở node p

: là số phần tử ở node con thứ i

=M

Gini split: chọn thuộc tính có chỉ số nhỏ

=

#### **Ưu nhược điểm**

**Ưu điểm:**

-Dễ hiểu và triển khai: Thuật toán CART sử dụng nguyên tắc đơn giản và dễ hiểu, giúp người ta có thể giải thích mô hình mội cách dễ dàng.

-Khả năng xử lý cả dữ liệu số và dữ liệu rời rạc: CART có thể được sử dụng cho cả bài toán phân loại và dự đoán giá trị số.

-Khả năng xử lý dữ liệu thiếu: Cây quyết định có khả năng xử lý tốt dữ liệu, tức là dữ liệu mà bị thiếu ở một số thuộc tính

-Khả năng xử lý biến độc lập tương quan: CART không bị ảnh hưởng lớn bởi các biến độc lập tương quan, nghĩa là nó có khả năng xử lý tốt khi có sự tương quan giữa các thuộc tính

**Nhược điểm:**

-Dễ bị quá mức phức tạp: Cây quyết định có thể trở nên quá mức phức tạp và quá mức fit dữ liệu huấn luyện nếu không kiểm soát được. Điều này dẫn đến hiện tượng quá mức (overfitting), khi mô hình hoạt động tốt trên dữ liệu huấn luyện nhưng kém hiệu suất trên dữ liệu mới.

- Khả năng bị chi phối bởi dữ liệu nhiễu: CART có thể nhạy cảm với dữ liệu nhiễu, tức là các ngoại lệ hoặc giá trị sai lệch có thể ảnh hưởng đến quá trình xây dựng cây.

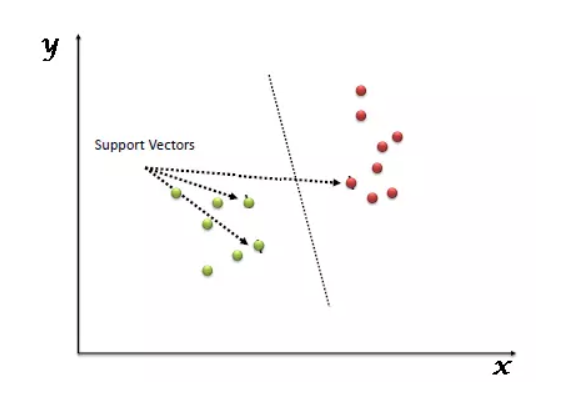
- Tính không ổn định: Những thay đổi nhỏ trong dữ liệu đầu vào có thể dẫn đến các cây quyết định khác nhau, tạo ra tính không ổn định.

- Có thể chịu ảnh hưởng của biến có số lượng lớp lớn: Khi một biến có số lượng lớp lớn, nó có thể trở thành ưu tiên trong quá trình chia nhánh, làm giảm khả năng quan sát các biến khác.

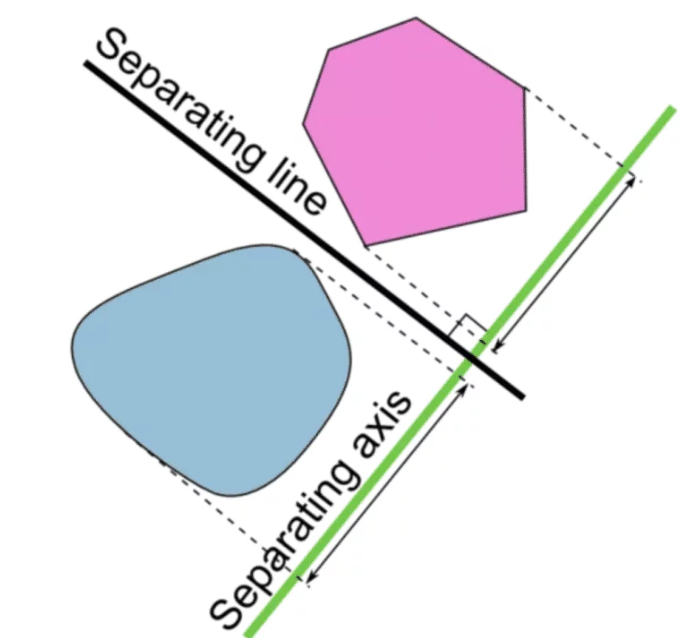
**2.3.3 Thuật toán Support Vector Machine**

**2.3.3.1 Khái niệm thuật toán SVM**

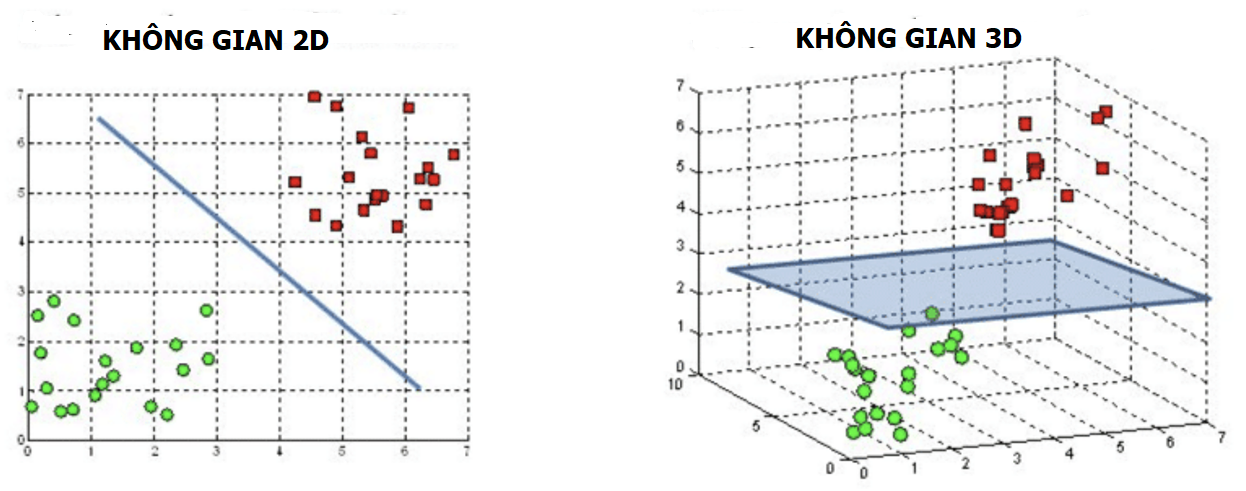
SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.



Mục tiêu của thuật toán SVM là tìm ra một siêu phẳng trong không gian N chiều (tương ứng với N đặc trưng) để chia dữ liệu thành hai phần, mỗi phần thuộc về một lớp khác nhau. Trong không gian N chiều, siêu phẳng là một không gian con có kích thước N-1 chiều. Đồng nghĩa với việc trong mặt phẳng 2 chiều, siêu phẳng là một đường thẳng còn trong một không gian 3 chiều thì siêu phẳng là một mặt phẳng. Theo ngôn ngữ của đại số tuyến tính, siêu phẳng này phải có lề cực đại (lề tối đa) và phân chia hai bao lồi cách đều với chúng.

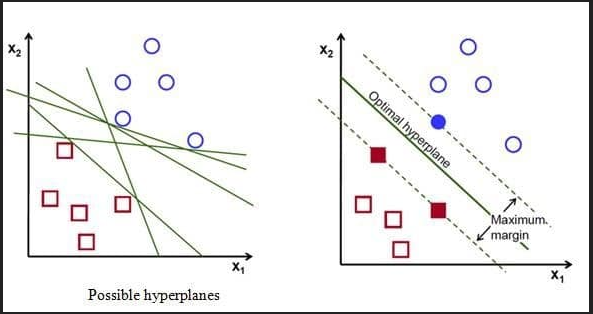


Nếu bạn không quen với đại số tuyến tính, bạn chỉ cần nhớ rằng trong không gian N chiều, một siêu phẳng là một không gian con có kích thước N-1 chiều. Một cách trực quan, trong một mặt phẳng (2 chiều) thì siêu phẳng là một đường thẳng, trong một không gian 3 chiều thì siêu phẳng là một mặt phẳng. Có thể bạn đã biết kiến thức này ở cấp 3 nếu buổi học hôm đó bạn không ngủ gật. Siêu phẳng tạo ra một biên giới phân chia 2 lớp của dữ liệu.



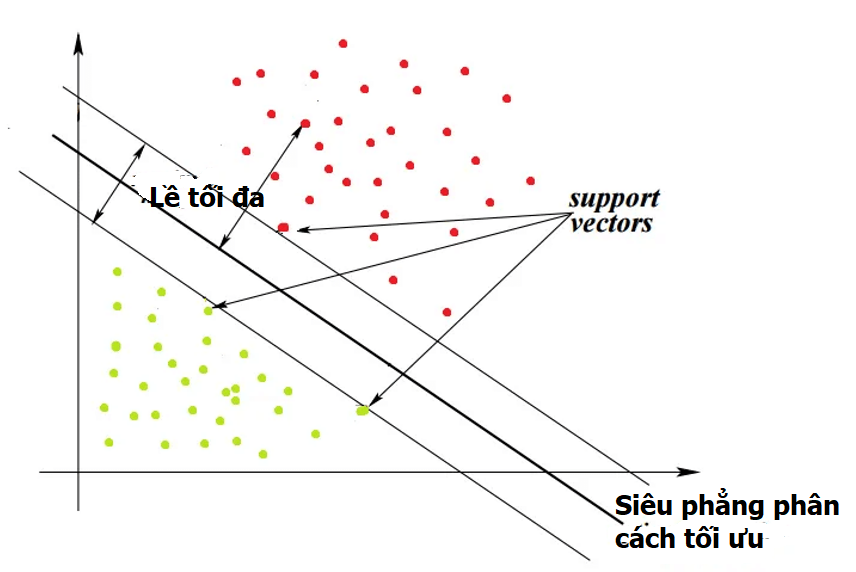
Hình . Siêu phẳng trong không gian 2D với 3D

Để phân chia hai lớp dữ liệu, rõ ràng là có rất nhiều siêu phẳng có thể làm được điều này. Mặc dù vậy, mục tiêu của chúng ta là tìm ra siêu phẳng có lề rộng nhất tức là có khoảng cách tới các điểm của hai lớp là lớn nhất. Hình dưới đây là một ví dụ trực quan về điều đó.



Hình . Siêu phẳng tối ưu có lề cực đại

Mỗi điểm trong không gian vector có thể được xem như một vector từ gốc tọa độ đến điểm đó. Các điểm dữ liệu nằm trên hoặc gần siêu phẳng được gọi là vector hỗ trợ và chúng có tác động trực tiếp đến vị trí và hướng của siêu phẳng. Các vector hỗ trợ này được sử dụng để tối ưu lề siêu phẳng. Nếu các vector hỗ trợ bị loại bỏ, vị trí của siêu phẳng sẽ thay đổi, và đặc biệt là các vector hỗ trợ phải cách đều siêu phẳng.



Hình . Các đường phân cách hai lớp

Với những bài toán có dữ liệu có thể phân chia tuyến tính thì SVM chỉ cần xác định siêu phẳng này dựa trên 2 tiêu chí đó sau:

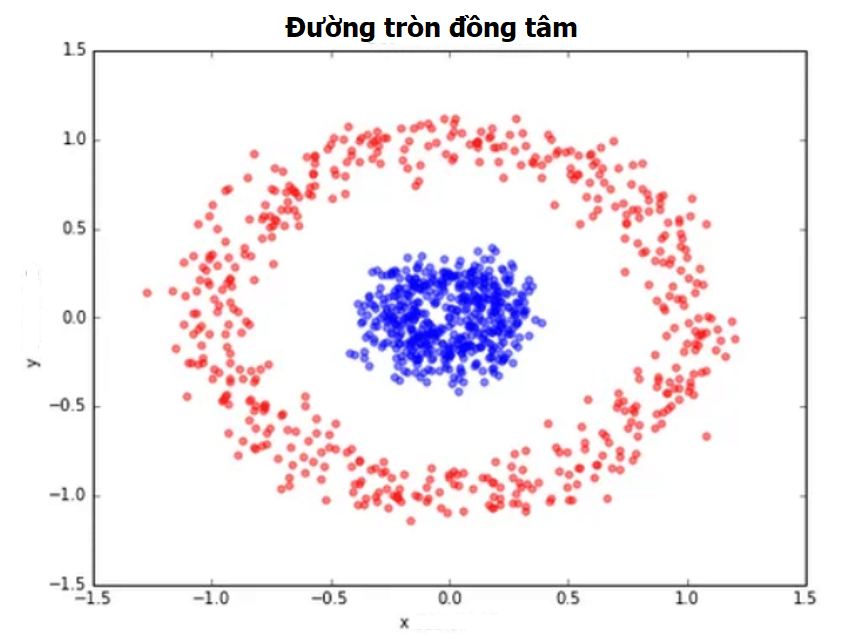
+ Một siêu phẳng có thể phân chia được hai lớp.

+ Siêu phẳng này phải cách đều hai lớp với khoảng cách lề đến điểm gần nhất với siêu phẳng của hai lớp đó là lớn nhất.

Tuy thế đối với bài toán có tập dữ liệu không thể phân chia tuyến tính được thì rõ ràng là không thể vẽ một đường thẳng để phân loại các điểm dữ liệu xanh và đỏ. Để giải quyết vấn đề này, có hai giải pháp đó là:

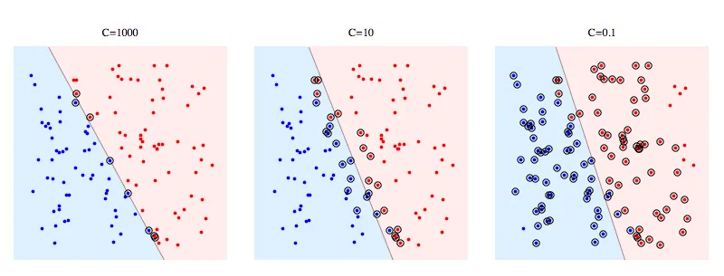
+ Lề mềm (Soft Margin)

+ Thủ thuật kernel (Kernel Tricks)



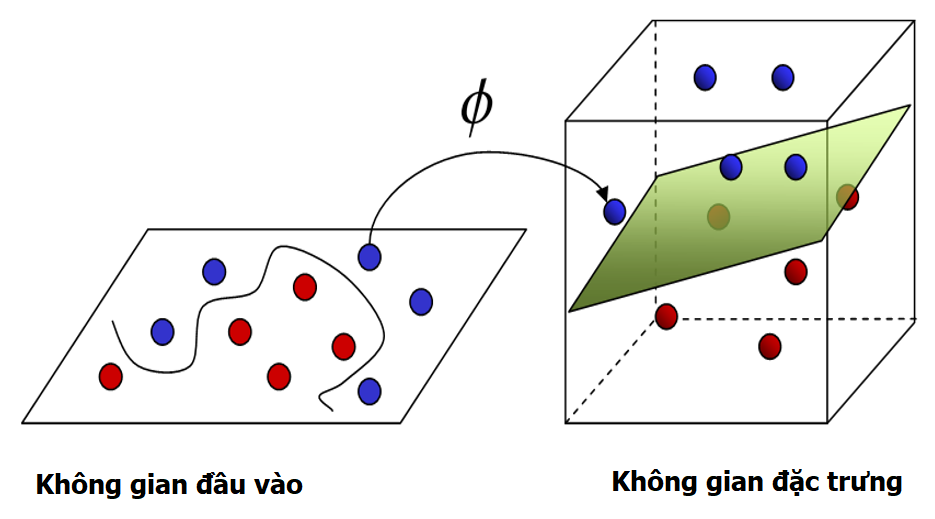
Hình . Trường hợp hai lớp không phân chia tuyến tính

Trong SVM với lề mềm, thuật toán cho phép tồn tại một số lỗi nhất định và tối đa hóa lề bằng cách giữ cho lề càng rộng càng tốt. Nó giúp cân bằng giữa việc phân loại sai với tối đa hóa lề. Lề mềm áp dụng cho các trường hợp khi dữ liệu nằm gần lề hoặc phân loại sai. Trong SVM mức độ chấp nhận lỗi là một siêu số quan trọng. Mức độ chấp nhận lỗi được coi như một tham số phạt (C). Tham số phạt càng lớn có nghĩa là SVM càng bị phạt nặng khi thực hiện phân loại sai. Do đó, lề càng hẹp và càng ít vector hỗ trợ được sử dụng.



Hình . SVM và mức độ chấp nhận lỗi khác nhau

Bằng thủ thuật Kernel, sẽ có một hàm ánh xạ dữ liệu từ không gian ít chiều hơn sang không gian nhiều chiều hơn, từ đó ta tìm được siêu phẳng phân tách dữ liệu. Nhận xét một cách trực quan, kỹ thuật này giống như việc bạn gập tờ giấy lại để có thể dùng kéo cắt một lỗ tròn trên nó. Các kiểu kernel có thể là tuyến tính, đa thức, sigmoid,…



Hình . Thủ thuật kernel ánh xạ dữ liệu sang không gian đa chiều hơn

**2.3.3.2 Công thức**

Trong không gian 2 chiều, ta biết rằng khoảng cách từ một điểm có toạ độ ( , ) tới đường thẳng có phương trình + + b = 0 được xác định bởi:

Trong không gian ba chiều, khoảng cách từ một điểm có toạ độ ( , ,) tới một mặt phẳng có phương trình + ++ b = 0 được xác định bởi:

Hơn nữa, nếu ta bỏ dấu trị tuyệt đối ở tử số, chúng ta có thể xác định được điểm đó nằm về phía nào của đường thẳng hay mặt phẳng đang xét. Những điểm làm cho biểu thức trong dấu giá trị tuyệt đối mang dấu dương nằm về cùng 1 phía (tôi tạm gọi đây là phía dương của đường thẳng), những điểm làm cho biểu thức trong dấu giá trị tuyệt đối mang dấu âm nằm về phía còn lại (tôi gọ là phía âm). Những điểm nằm trên đường thẳng/măt phẳng sẽ làm cho tử số có giá trị bằng 0, tức khoảng cách bằng 0.

Việc này có thể được tổng quát lên không gian nhiều chiều: Khoảng cách từ một điểm (vector) có toạ độ tới siêu mặt phẳng (hyperplane) có phương trình + b = 0 được xác định bởi:

Với = với d là số chiều của không gian.

**2.3.3.3 Ưu nhược điểm**

**Ưu điểm**

Hiệu suất tốt: SVM là một giải thuật dành cho bài toán phân loại vô cùng hiệu quả, trong các trường hợp các lớp dữ liệu có sự tách biệt rõ ràng hoặc khi số lượng đặc trưng lớn nó đặc biệt hiệu quả.

Xử lý dữ liệu phi tuyến tính: SVM sử dụng các hàm kernel để ánh xạ dữ liệu từ không gian đặc trưng ban đầu sang cao hơn giúp phân loại các lớp dữ liệu phi tuyến tính một cách hiệu quả.

**Nhược điểm**

Bài toán số chiều cao: Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả khá tồi.

Đòi hỏi lượng tính toán lớn: Đặc biệt khi sử dụng các hàm kernel phức tạp kèm tập dữ liệu lớn thì việc huấn luyện SVM sẽ ngốn rất nhiều thời gian và tài nguyên tính toán.

Khó xử lý với tập dữ liệu lớn: Việc lưu trữ các vector hỗ trợ có thể tốn nhiều bộ nhớ trong trường hợp các tập dữ liệu lớn.

## Các kĩ thuật dùng cho bài toán

* Thu thập dữ liệu: Gom dữ liệu của các bệnh viện, cơ sở y tế hoặc các nguồn khác liên

quan đến thông tin người bệnh như Kaggle.

* Tiền xử lý dữ liệu: Loại bỏ dữ liệu thiếu, chuyển đổi các biến không phải số thành

dạng số, xử lý nhiễu, và chuẩn hóa dữ liệu để tạo điều kiện cho mô hình học máy.

* Chọn đặc trưng: Lựa chọn các trường quan trọng để xây dựng mô hình. Các trường

này có thể liên quan đến tuổi, số lần mang thai, số lần hút thuốc và các thông tin về y tế.

* Chia dữ liệu: Phân chia dữ liệu thành tập huấn luyện (training) và tập kiểm thử

(testing) để đào tạo đánh giá mô hình.

* Chọn mô hình: Sử dụng các thuật toán học máy như SVM, cây quyết định để xây

dựng mô hình.

* Đào tạo mô hình: Mô hình được đào tạo trên tập huấn luyện để học cách dự đoán

nguy cơ ung thư cổ tử cung dựa trên các đặc trưng đã chọn.

* Đánh giá mô hình: Sử dụng tập kiểm thử để đánh giá hiệu suất của mô hình. Các độ

đo như độ chính xác, độ nhạy (recall), độ đặc hiệu (precision) và F1-Score.

* Điều chỉnh tham số: Sau khi đã huấn luyện và đánh giá được mô hình học máy của

mình, chúng ta có thể điều chỉnh lại tham số nếu cần để tăng độ chính xác cho hệ thống.

* Đưa ra dự đoán: Cuối cùng, chúng ta sẽ sử dụng mô hình mà chúng ta đã tạo để phân

tích thông tin và đưa ra dự đoán một cách chính xác.

## Các phương pháp đánh giá mô hình

### Accuracy

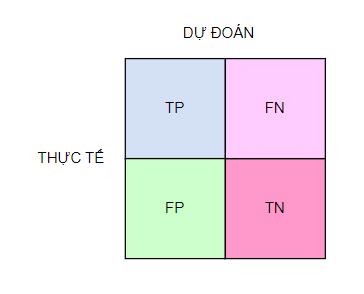
Đây là một độ đo của bài toán phân lớp đơn giản nhất. Accuracy được tính bằng cách lấy số dự đoán đúng chia cho toàn bộ các dự đoán.

Ví dụ dễ hiểu: nếu đếm được có 8 điểm dữ liệu được dự đoán đúng trên tổng số 10 điểm. Vậy ta kết luận độ chính xác của mô hình là 0.8 (hay 80%).

Nhược điểm: độ đo này chỉ cho ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất hay dữ liệu của lớp nào thường bị phân loại nhầm nhất vào các lớp khác hơn nó hoàn toàn không chỉ ra được.

### Confusion Matrix

Confusion Matrix hiểu đơn giản là tổng kết giữa nhãn thực tế của dữ liệu cần dự đoán với nhãn của dữ liệu được dự đoán ra bằng thuật toán học máy.



Hình . Minh họa độ đo Confusion Matrix

*Một confusion matrix đại diện điểm trên ma trận 2 chiều bao gồm Thực tế và Dự đoán với 4 chỉ số sau:*

True Positive: Các giá trị thực sự Positive và được dự đoán là Positive.

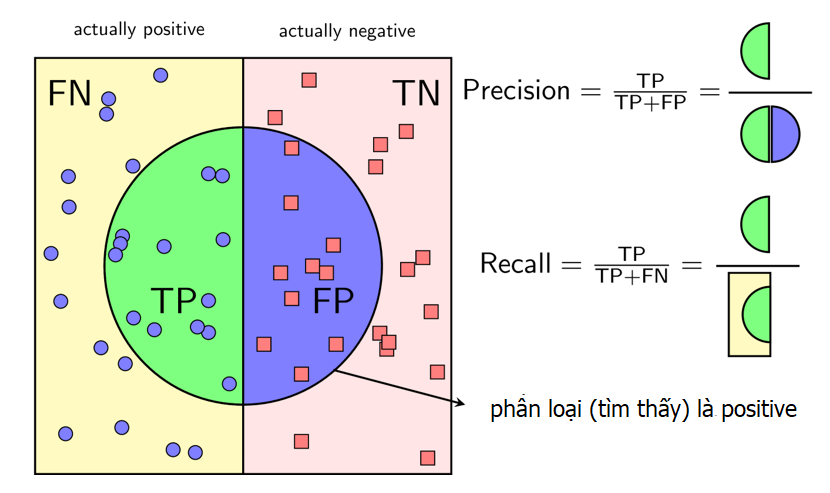
False Negative (FN): Các giá trị thực sự là Positive nhưng được dự đoán sai là Negative.

False Positive: Các giá trị thực sự là Negative nhưng được dự đoán sai là Positive.

True Negative (TN): Các giá trị thực sự Negative và được dự đoán là Negative.

Ví dụ dễ hiểu như: có 99 người thực tế là ung thư trong đó 98 người được đoán là ung thư (TP), 1 người bị dự đoán là không ung thư (FN). 10 người không bị ung thư trong thực tế trong đó 6 người được dự đoán ung thư (FP), 4 người dự đoán là không ung thư (TN).

### Precision và Recall



Hình . Minh họa độ đo Precision, Recall

Precision (độ chuẩn xác) là tỷ lệ giữa các điểm true positive trên tổng điểm được phân thành positive. Như ở ví dụ có 99 người thực tế là ung thư trong đó 98 người được đoán là ung thư (TP), 1 người bị dự đoán là không ung thư (FN), 10 người không bị ung thư trong thực tế trong đó 6 người được dự đoán ung thư (FP), 4 người dự đoán là không ung thư (TN). Thì tỷ lệ tính theo độ đo này sẽ là: 98/(98+6)= 98/104

Recall (độ nhạy) có ý nghĩa là tỷ lệ giữa các điểm true positive trên tổng điểm positive thực sự. Như ở ví dụ có 99 người thực tế là ung thư trong đó 98 người được đoán là ung thư (TP), 1 người bị dự đoán là không ung thư (FN) thì ta có tỷ số là 98/(98+1)=98/99

### F1-score

F1 Score là sự trung bình điều hòa giữa độ đo precision và độ đo recall. Do đó nó đại diện hơn trong việc đánh giá độ chính xác trên đồng thời precision và recall. Công thức của F1 Score được tính toán như sau:

có giá trị nằm trong khoảng . càng cao thì mô hình phân lớp càng tốt.

Bảng . Bảng giá trị F1 - score

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 1 | 1 |
| 0.1 | 0.1 | 0.1 |
| 0.4 | 0.4 | 0.4 |
| 1 | 0.5 | 0.67 |
| 0.4 | 0.9 | 0.55 |

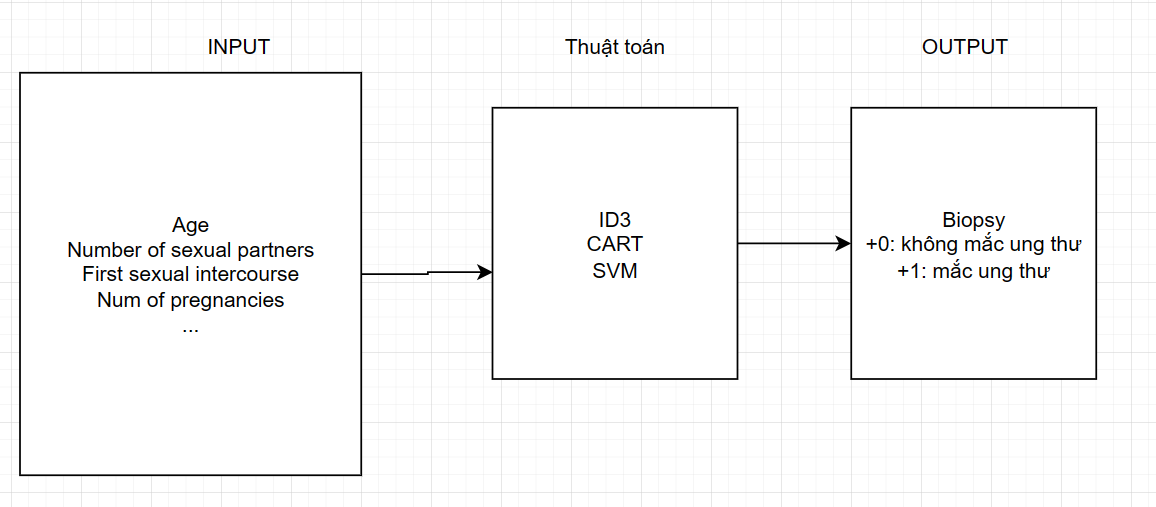
Một mô hình có chỉ số cao chỉ khi cả 2 chỉ số và để cao. Một trong 2 chỉ số này thấp đều sẽ kéo điểm xuống. Trường hợp xấu nhất khi 1 trong hai chỉ số và bằng 0 sẽ kéo điểm về 0. Trường hợp tốt nhất khi cả điểm chỉ số đều đạt giá trị bằng 1, khi đó điểm sẽ là 1.

Qua việc sử dụng chỉ số , ta đã có một thước đo đáng tin cậy về hiệu năng của mô hình trong các bài toán phân loại, đặc biệt khi dữ liệu về một lớp lớn hơn gấp nhiều lần so với dữ liệu về lớp còn lại.

# ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH

## Phân tích chi tiết bài toán

Bài toán đặt ra là phân loại nguy cơ ung thư cổ tử cung bằng cách sử dụng hồ sơ lâm sàng và kết quả xét nghiệm trong phòng thí nghiệm của họ. Dự báo các sự kiện liên quan đến ung thư cổ tử cung trong thực hành lâm sàng có xu hướng khá không chính xác và hay thay đổi. Vì thế tôi muốn xây dựng một mô hình dự đoán chính xác hơn các bệnh nhân có nguy cơ ung thư cổ tử cung dựa vào phương pháp học máy. Điều này có thể giúp các bác sĩ lâm sàng đưa ra quyết định sáng suốt về cường độ điều trị cho từng bệnh nhân mắc ung thư. Bộ dữ liệu sử dụng bao gồm đầu vào là 35 thuộc tính đặc trưng về các chỉ số y khoa của người bệnh và 1 thuộc tính nhãn lớp (gọi là nhãn kết quả). Nhãn kết quả có hai giá trị cụ thể là 0 và 1, tức là ung thư và không ung thư. Để giải quyết bài toán này tôi sử dụng các thuật toán phân loại thuộc nhóm học có giám sát. Cụ thể là thuật toán ID3 (Iterative Dichotomiser 3), CART và SVM (Support Vector Machine).

 Hình . Sơ đồ biến đầu vào (input), thuật toán, biến đầu ra (output)

**Các thông tin đầu vào của tập dữ liệu bao gồm:**

+Age: tuổi

+Number of sexual partners: Số lượng đối tác tình dục

+First sexual intercourse: Quan hệ tình dục lần đầu

+Num of pregnancies: Số lần mang thai

+Smokes: Hút thuốc

+Smokes (years): Số năm hút thuốc

+Smokes (packs/year): số gói thuốc theo năm

+Hormonal Contraceptives: Thuốc tránh thai nội tiết tố

+Hormonal Contraceptives (years): Số năm đã sử dụng các phương pháp tránh thai nội tiết

+IUD: vòng tránh thai

+IUD (years): Số năm sử dụng vòng tránh thai

+STDs: Có bị lây truyền qua đường tình dục k

+STDs (number): số lượng bệnh lây qua đường tình dục

+STDs:condylomatosis: bị nhiễm vi khuẩn gây mụn có đầu tròn

+STDs:cervical condylomatosis: nhiễm vi khuẩn gây mụn cổ tử cung

+STDs:vaginal condylomatosis: nhiễm vi khuẩn gây mụn âm đạo

+STDs:vulvo-perineal condylomatosis: nhiễm vi khuẩn gây mụn ở vùng âm đạo và xung quanh

+STDs:syphilis: bị nhiễm sừng

+STDs:pelvic inflammatory disease: bị nhiễm trùng nội mạc tử cung

+STDs:genital herpes: bị nhiễm herpes sinh dục

+STDs:molluscum contagiosum: bị nhiễm molluscum contagiosum

+STDs:AIDS: bị nhiễm bệnh AIDS

+STDs:HIV: bị nhiễm HIV

+STDs:Hepatitis B: bị nhiễm HepatitisB

+STDs:HPV: bị nhiễm HPV (Human Papillomavirus) là một loại virut thường gây ra các bệnh lây truyền qua đường tình dục và có thể dẫn đến ung thư

+STDs: Number of diagnosis: số lượng chuẩn đoán bệnh lây qua đường tình dục

+STDs: Time since first diagnosis: thời gian kể từ lần chuẩn đoán bệnh lây truyền qua đường tình dục đầu tiên

+STDs: Time since last diagnosis: thời gian kể từ lần chuẩn đoán bệnh lây truyền qua đường tình dục gần nhất

+Dx:Cancer: biểu thị sự xuất hiện hoặc vắng mặt của ung thư

+Dx:CIN: Biểu thị sự xuất hiện hoặc vắng mặt của bệnh CIN (Cervical Intraepithelial Neoplasia), một bệnh có thể dẫn đến ung thư cổ tử cung.

+Dx:HPV: Biểu thị sự xuất hiện hoặc vắng mặt của vi rút HPV trong việc chẩn đoán.

+Dx: Biểu thị sự xuất hiện hoặc vắng mặt của bất kỳ bệnh lý nào được chẩn đoán. +Hinselmann: Kết quả của bài kiểm tra Hinselmann để xác định có dấu hiệu của ung thư cổ tử cung hay không.

+Schiller: Kết quả của bài kiểm tra Schiller để đánh giá sự biến đổi của cổ tử cung

+Citology: Kết quả của bài kiểm tra Citology để kiểm tra tế bào tử cung

Thông tin đầu ra là: Biopsy kết quả bệnh nhân có bị ung thư cổ tử cung hay không (0 hoặc 1).

## Các công cụ, thư viện dùng để xây dựng

### Sử dụng ngôn ngữ Python

Python là ngôn ngữ lập trình được sáng tạo bởi Guido Van Rossum – một chuyên gia lập trình người Hà Lan vào năm 1990. Người sáng lập muốn Python là một ngôn ngữ dễ hiểu và tương đồng với tiếng anh. Nhờ vậy, nó rất thân thiện với người dùng, không khó để người mới làm quen. Tất cả các câu lệnh, cú pháp đều rất trực quan và rõ ràng. Kết quả từ khảo sát của Stack Overflow cho thấy Python là ngôn ngữ lập trình được yêu thích nhất. Python đã cho ra mắt với rất nhiều phiên bản khác nhau, phiên bản gần đây nhất là Python 3. Trên windows bạn hoàn toàn có thể tải và cài đặt python dễ dàng với vài thao tác thông qua đường link dưới đây:

<https://www.python.org/downloads/windows/>

**Đặc điểm của python:**

Python là một ngôn ngữ thông dịch, nó trực tiếp chạy từng dòng mã. Nếu có lỗi trong mã chương trình, nó sẽ ngừng chạy. Do đó, lập trình viên có thể nhanh chóng tìm ra lỗi trong đoạn mã.

Python là một ngôn ngữ đễ sử dụng, từ ngữ giống trong tiếng Anh. Không giống như các ngôn ngữ lập trình khác, Python không sử dụng dấu ngoặc ôm. Thay vào đó, ngôn ngữ này sử dụng thụt đầu dòng.

Python là ngôn ngữ linh hoạt, cao cấp gần gũi với ngôn ngữ con người hơn các ngôn ngữ lập trình khác.

Python là ngôn ngữ lập trình hướng đối tượng, nhưng ngôn ngữ này cũng hỗ trợ các phương thức lập trình khác như lập trình hàm và lập trình cấu trúc.

**Ưu điểm:**

Là một ngôn ngữ có cấu trúc rõ ràng, cú pháp ngắn gọn. Có trên tất cả các nền tảng hệ điều hành từ UNIX, MS – DOS, Mac OS, Windows và Linux và các OS khác thuộc họ Unix.

Tương thích mạnh mẽ với Unix, hardware, third-party software với số lượng thư viện khổng lồ (400 triệu người sử dụng)

Python với tốc độ xử lý cực nhanh, python có thể tạo ra những chương trình từ những script siêu nhỏ tới những phần mềm cực lớn như Blender 3D.

**Nhược điểm:**

Python không có các thuộc tính như: protected, private hay public, không có vòng lặp do…while và switch….case.

Mặc dù tốc độ xử lý của Python nhanh hơn PHP nhưng không bằng JAVA và C++.



Hình . Tổng quan về ngôn ngữ python [5]

### Các ứng dụng của python

+ Python có thể sử dụng trong phát triển web: python có nhiều framework mạnh như Django và Flask, chúng giúp phát triển ứng dụng web nhanh chóng và dễ dàng.

+ Khoa học dữ liệu và trí tuệ nhân tạo: trong lĩnh vực này python được sử dụng rất phổ biến nhờ vào các thư viện hỗ trợ như numpy, pandas, sklearn.

+ Khai phá dữ liệu và phân tích: các thư viện như pandas và matplotlib có thể được sử dụng để phân tích và trực quan hóa dữ liệu.

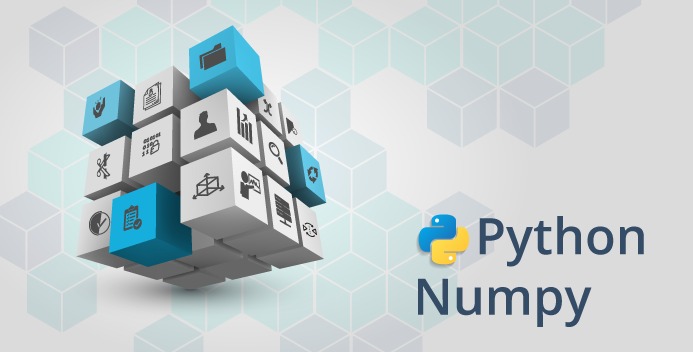
+ Đồ họa, trò chơi điện tử: python có thư viện opencv để xử lý ảnh, video và thư viện pygame để phát triển các trò chơi.

+ Ngoài ra python còn được sử dụng trong việc phát triển các công cụ quản lý mạng và bảo mật, kiểm thử.

### Các thư viện sử dụng

*Thư viện numpy:*

Xử lý mảng đa chiều, ma trận. Là một gói xử lý (Processing Package) phổ biến của Python. NumPy làm phong phú ngôn ngữ lập trình Python với các cấu trúc dữ liệu mạnh mẽ để tính toán hiệu quả các mảng và ma trận đa chiều. Numpy không chỉ là một gói mô-đun để xử lý mảng mà nó còn cung cấp khả năng quản lý mảng cực kỳ vượt trội.



Hình . Minh họa thư viện Numpy [6]

*Thư viện pandas:*

**Một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ, linh hoạt và mang hàm ý. Tên thư viện được bắt nguồn từ panel data (bảng dữ liệu). Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều) và dữ liệu chuỗi thời gian.**



Hình 3.4 Minh họa thư viện Pandas [7]

*Thư viện sklearn:*

Là thư viện mạnh mẽ nhất dành cho các thuật toán học máy 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 gồm: classification, regression, clustering, và dimensionality reduction.



Hình . Minh họa thư viện Sklearn [8]

*Thư viện Tkinter:*

Là thư viện đồ họa để tạo nên giao diện cho bài toán. Thư viện tkinter cho phép người dùng tạo ra các cửa sổ giao diện với textbox, label, combobox, button,...



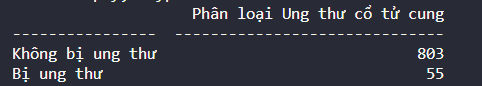
Hình . Minh họa thư viện Tkinter [9]

## Xây dựng mô hình

**3.3.1 Dữ liệu thực nghiệm**

Bộ dữ liệu sử dụng được lấy từ kaggle chứa hồ sơ bệnh án của 858 bệnh nhân. Trong số 858 bệnh nhân có 803 bệnh nhân không mắc ung thư cổ tử cung và 55 bệnh nhân mắc ung thư cổ tử cung trong quá trình nghiên cứu. Đường link thu thập dữ liệu:

[Cervical Cancer Risk Classification (kaggle.com)](https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification)



Bộ dữ liệu này có 35 thuộc tính đặc trưng về các chỉ số y khoa và 1 thuộc tính nhãn lớp. Nhưng chỉ lấy 10 thuộc tính quan trọng để sử dụng trong mô hình bao gồm:

Input:

Age (Tuổi): Tuổi thường là một yếu tố quan trọng vì nguy cơ mắc bệnh thường tăng theo tuổi.

Number of sexual partners (Số lượng đối tác tình dục): Thông tin này có thể liên quan đến rủi ro nhiễm bệnh qua đường tình dục.

First sexual intercourse (Quan hệ tình dục lần đầu): Thời điểm này có thể liên quan đến nguy cơ mắc bệnh.

Num of pregnancies (Số lần mang thai): Thông tin về số lần mang thai có thể ảnh hưởng đến rủi ro.

Smoking (Hút thuốc): Thông tin về hút thuốc có thể là một yếu tố nguy cơ.

Hormonal Contraceptives (Thuốc tránh thai nội tiết tố): Sử dụng thuốc tránh thai có thể ảnh hưởng đến nguy cơ.

IUD (Vòng tránh thai): Việc sử dụng vòng tránh thai cũng là một yếu tố quan trọng.

STDs (Bệnh lây truyền qua đường tình dục): Thông tin về các bệnh lây truyền qua đường tình dục, đặc biệt là HPV, có thể quan trọng.

STDs (Number of diagnoses) (Số lượng chuẩn đoán bệnh lây truyền qua đường tình dục): Số lượng chuẩn đoán có thể liên quan đến nguy cơ.

STDs:HPV (Vi rút HPV): Thông tin về HPV có thể là một yếu tố nguy cơ chính.

Output:

Biopsy(Kết quả của quá trình lấy mẫu sinh thiết): Đây có thể là yếu tố quan trọng nhất vì nó thường được sử dụng để xác định chính xác có ung thư hay không. Nhãn kết quả (mắc ung thư hay không mắc).

Bộ dữ liệu này khi giải quyết bài toán sẽ dùng 80% để huấn luyện mô hình và 20% để kiểm tra và đánh giá.

**3.3.2** **Thực hiện xây dựng mô hình**

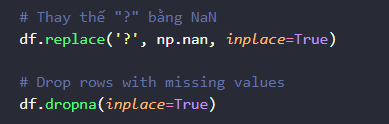
**Bước 1:** Thu thập dữ liệu.

Như đã trình bày trên mục 3.1 ở mục các thông tin đầu vào, dữ liệu được sử dụng trong bài toán được tải trực tiếp từ trang web kaggle.

**Bước 2:** Tiền xử lý dữ liệu.

Loại bỏ dữ liệu thiếu, chuyển đổi các biến không phải số thành dạng số, xử lý nhiễu, và

chuẩn hóa dữ liệu để tạo điều kiện cho mô hình học máy.

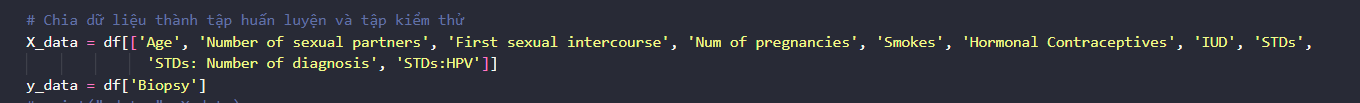


df.replace(“?”, np.nan, inplace=True): thay thế “?” bằng Nan.

df = df.dropna(inplace=true): Sau khi đã chuyển đổi sang Nan thì bản ghi có chứa Nan sẽ bị loại bỏ.

**Bước 3: Chọn các trường đặc trưng.**

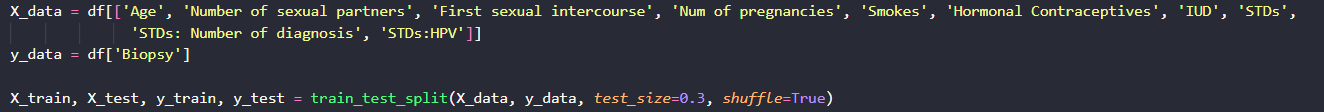
Như đã trình bày trên mục 3.3.1 dữ liệu được sử dụng để xây dựng mô hình

****

**Bước 4: Chia dữ liệu.**

Phân chia dữ liệu thành tập huấn luyện (training) và tập kiểm thử (testing) để đào tạo

đánh giá mô hình.



**Bước 5: Chọn mô hình**

Chọn mô hình: Sử dụng các thuật toán học máy như SVM, ID3, CART để xây

dựng mô hình.

**Bước 6:** **Đào tạo mô hình**

Mô hình được đào tạo trên tập huấn luyện để học cách dự đoán nguy cơ ung thư cổ tử

cung dựa trên các đặc trưng đã chọn

**Bước 7: Đánh giá mô hình**

Sử dụng tập kiểm thử để đánh giá hiệu suất của mô hình. Các độ đo như độ chính xác

(Accuracy), độ nhạy (recall), độ đặc hiệu (precision) và F1-Score.

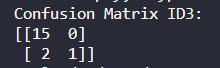
**Bước 8: Kết quả**

# KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

## Kết quả

### Kết quả thực hiện đối với Confusion Matrix

**ID3:**



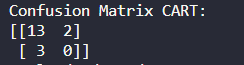
Ta có thể thấy với thuật toán ID3 trong 18 người tập dữ liệu test thì có 15 người không bị ung thư cổ tử cung được dự đoán đúng, 0 người không có bệnh bị dự đoán thành có bệnh, 2 người có bệnh được dự đoán thành không có bệnh và 1 người có bệnh được dự đoán đúng.

**SVM:**



Ta có thể thấy với thuật toán SVM trong 18 người tập dữ liệu test thì có 12 người không bị ung thư cổ tử cung được dự đoán đúng, 2 người không có bệnh bị dự đoán thành có bệnh, 4 người có bệnh được dự đoán thành không có bệnh và 0 người có bệnh được dự đoán đúng.

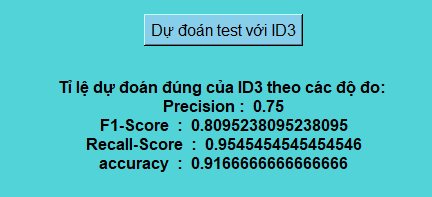
**CART:**



Ta có thể thấy với thuật toán SVM trong 18 người tập dữ liệu test thì có 13 người không bị ung thư cổ tử cung được dự đoán đúng, 2 người không có bệnh bị dự đoán thành có bệnh, 3 người có bệnh được dự đoán thành không có bệnh và 0 người có bệnh được dự đoán đúng.

### Kết quả thực hiện đối với mô hình ID3

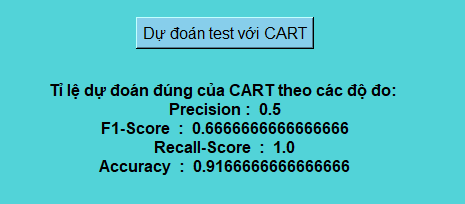
Kết quả trên tập dữ liệu test với mô hình ID3 theo độ đo Accuracy, Precision, Recall và F1-score có tiền xử lý dữ liệu:



Hình . Kết quả dự đoán test với id3 (sử dụng tiền xử lý dữ liệu)

### Kết quả thực hiện đối với mô hình CART

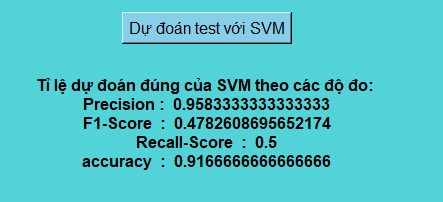
Kết quả trên tập dữ liệu test với mô hình CART theo độ đo Accuracy, Precision, Recall và F1-score có tiền xử lý dữ liệu:



Hình .10 Kết quả dự đoán test với CART (sử dụng tiền xử lý dữ liệu)

### Kết quả thực hiện đối với mô hình SVM

Kết quả trên tập dữ liệu test với mô hình SVM theo độ đo Accuracy, Precision, Recall và F1-score có tiền xử lý dữ liệu:

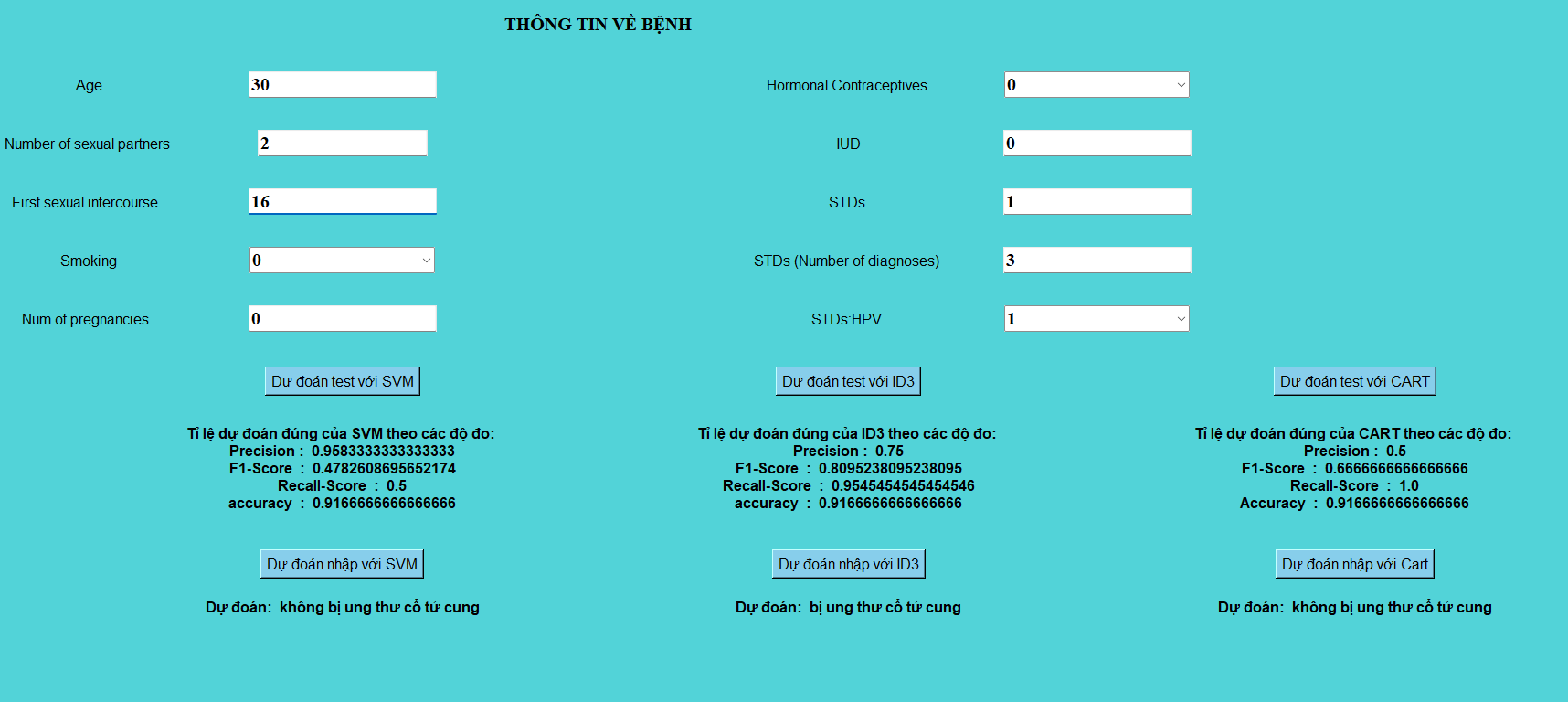


Hình .11 Kết quả dự đoán test với SVM (sử dụng tiền xử lý dữ liệu)

### Một số kết quả thử nghiệm dự đoán khi nhập dữ liệu trực tiếp

VD1: Age: 24, Number of sexual partners: 2, Firstsexual intercourse: 16, Smoking: 0, Num of pregnancies: 0, Hormonal Contraceptives: 0, IUD: 0, STDs: 1, STDs(Number of diagnoses): 3, STDs:HPV: 1

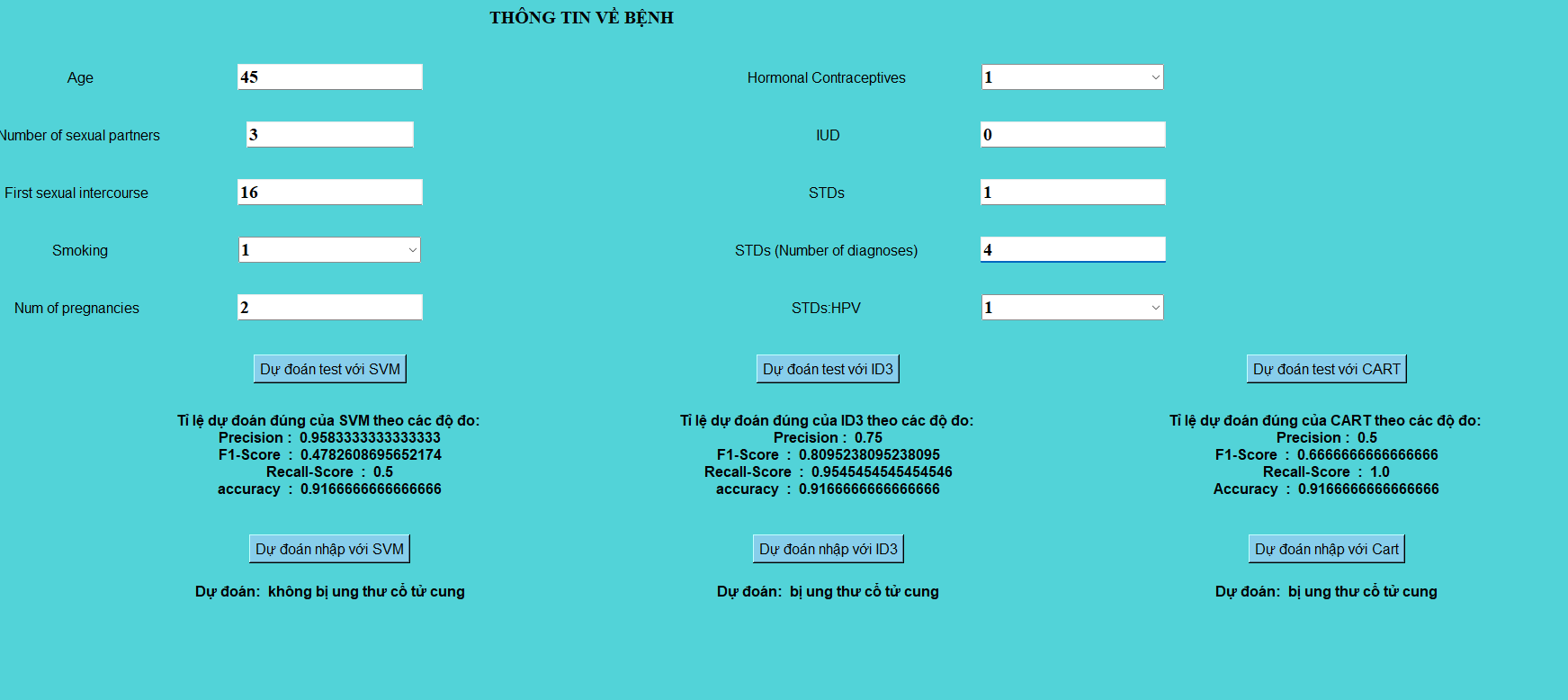
Kết quả trả về: Thuật toán ID3 dự báo mặc ung thư cổ tử cung còn hai thuật toán SVM và CART thì trả về không mắc ung thư.



Hình 4.5 Ví dụ một kết quả dự đoán từ nhập dữ liệu

VD2: Age: 45, Number of sexual partners: 3, Firstsexual intercourse: 16, Smoking: 1, Num of pregnancies: 2, Hormonal Contraceptives: 1, IUD: 0, STDs: 1, STDs(Number of diagnoses): 4, STDs:HPV: 1

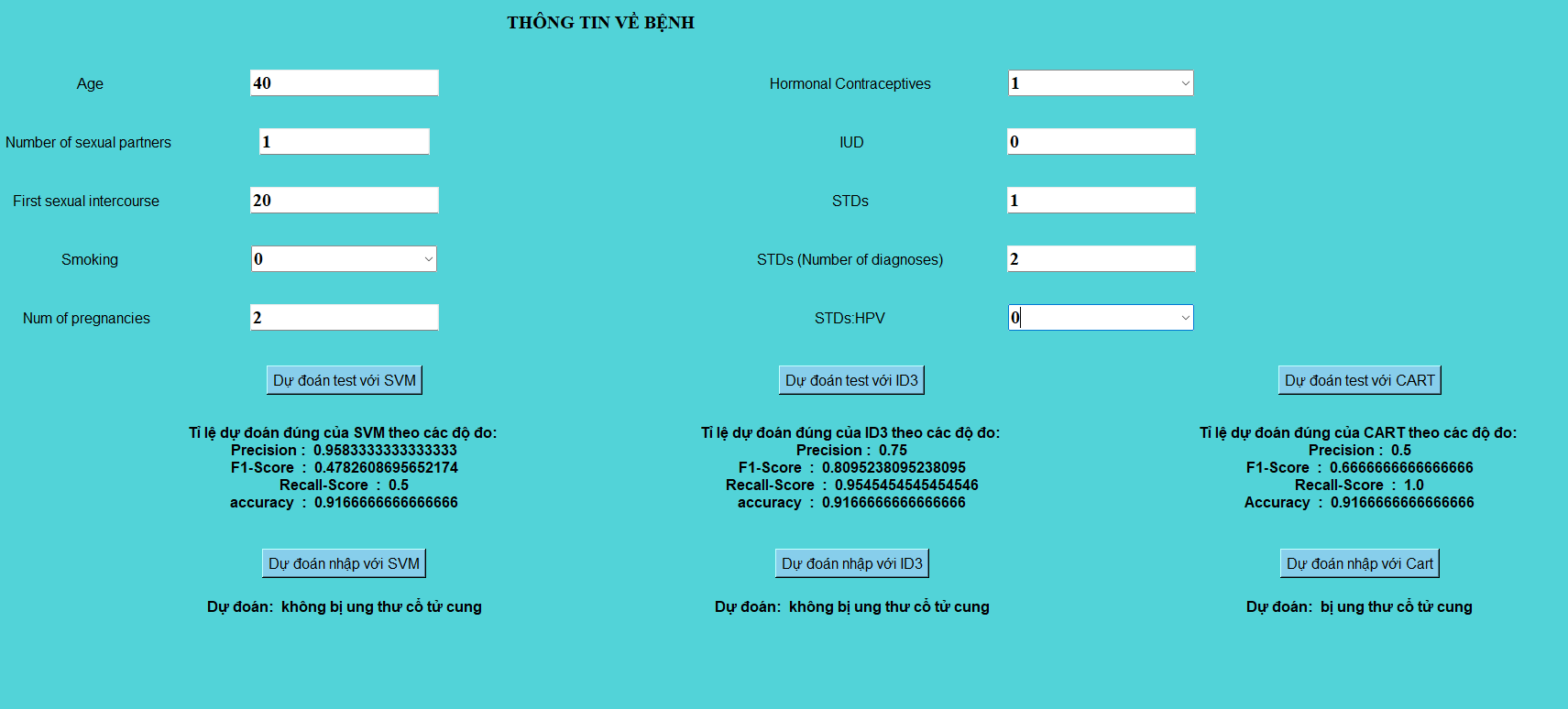
Kết quả trả về: Thuật toán ID3 và CART dự đoán mắc ung thư cổ tử cung còn thuật toán SVM dự báo không mắc ung thư.



Hình 4.5 Ví dụ một kết quả dự đoán từ nhập dữ liệu

VD3: Age: 40, Number of sexual partners: 1, Firstsexual intercourse: 20, Smoking: 0, Num of pregnancies: 2, Hormonal Contraceptives: 1, IUD: 0, STDs: 1, STDs(Number of diagnoses): 2, STDs:HPV: 0

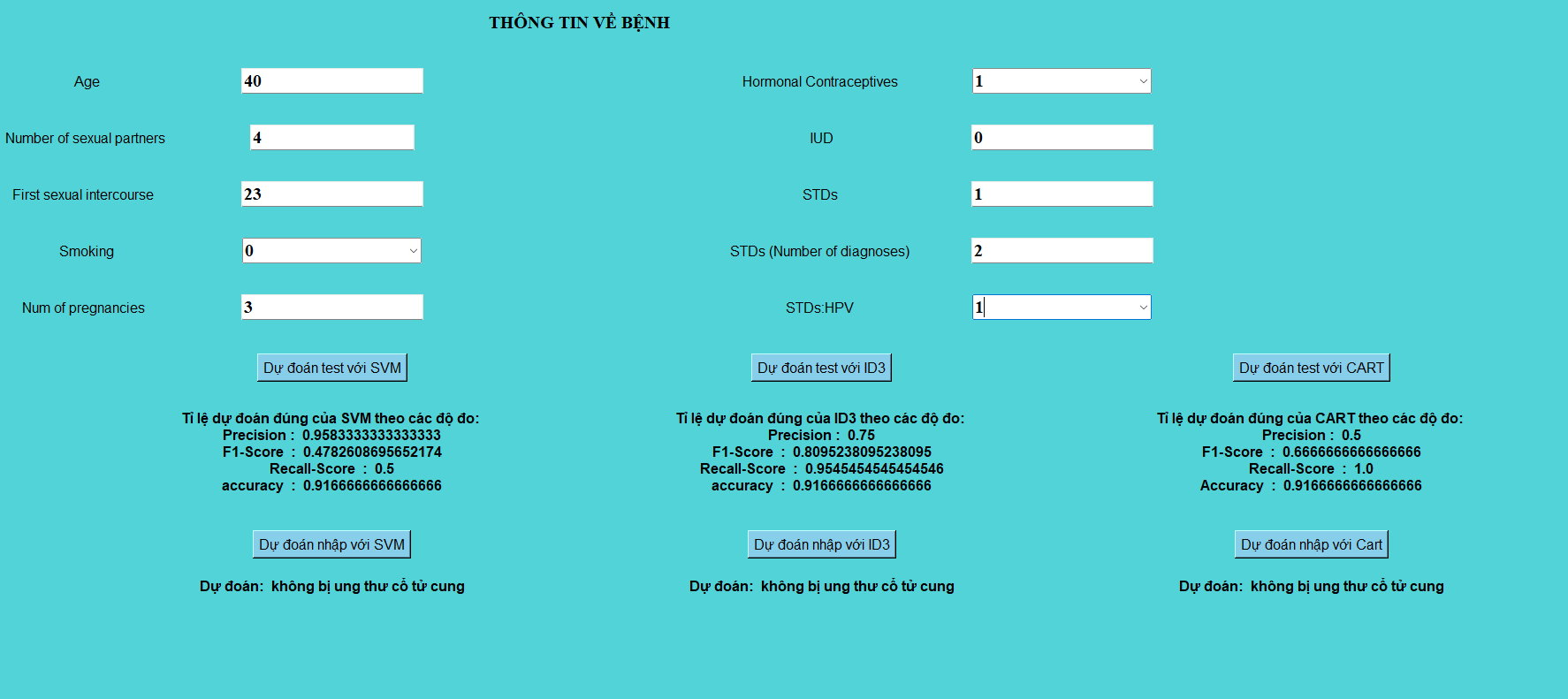
Kết quả trả về: Thuật toán CART dự đoán bị ung thư cổ tử cung còn 2 thuật toán ID3 và SVM cho dự đoán không mắc ung thư.



Hình 4.5 Ví dụ một kết quả dự đoán từ nhập dữ liệu

VD4: Age: 40, Number of sexual partners: 4, Firstsexual intercourse: 23, Smoking: 0, Num of pregnancies: 3, Hormonal Contraceptives: 1, IUD: 0, STDs: 1, STDs(Number of diagnoses): 2, STDs:HPV: 1

Kết quả trả về: cả 3 thuật toán SVM, ID3, CART đều dự đoán là không mắc ung thư.



Hình 4.5 Ví dụ một kết quả dự đoán từ nhập dữ liệu

## Đánh giá mô hình

# KẾT LUẬT

Với đề tài: “Phân loại nguy cơ ung thư cổ tử cung bằng phương pháp học máy” em đã hoàn thành trong thời gian được giao và đạt được kết quả như sau:

- Tìm hiểu kĩ hơn về học máy, các thuật toán phân lớp.

- Tìm hiểu về bệnh ung thư cổ tử cung, thu thập được dữ liệu bệnh nhân mắc bệnh.

- Tìm hiểu được các thuật toán cụ thể như ID3, CART, SVM và áp dụng chúng giải quyết bài toán dự đoán khả năng mắc bệnh ung thư cổ tử cung.

- Xây dựng được mô hình dự đoán với thuật toán ID3 và SVM, thử nghiệm trên bộ dữ liệu thu thập được.

Tuy vậy, do kinh nghiệm và kiến thức còn hạn chế, dữ liệu thu thập được vẫn chưa thực sự nhiều, độ chính xác của mô hình không được cao như mong đợi. Em sẽ tiếp tục học hỏi, tích lũy kiến thức, và mong các thầy cô đưa ra những góp ý để đề tài của em được hoàn thiện hơn. Em xin chân thành cảm ơn.

Hướng phát triển tiếp theo trong tương lai em sẽ cố gắng xây dựng được một bộ dữ liệu đầy đủ hơn, với bài toán dự đoán nguy cơ ung thư cổ tử cung có thể sử dụng thêm một số thuật toán học sâu để cải thiện hơn chức năng dự đoán.

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | [DOANVANNGOC-LA.pdf (hmu.edu.vn)](https://sdh.hmu.edu.vn/images/DOANVANNGOC-LA.pdf) |

PHỤ LỤC