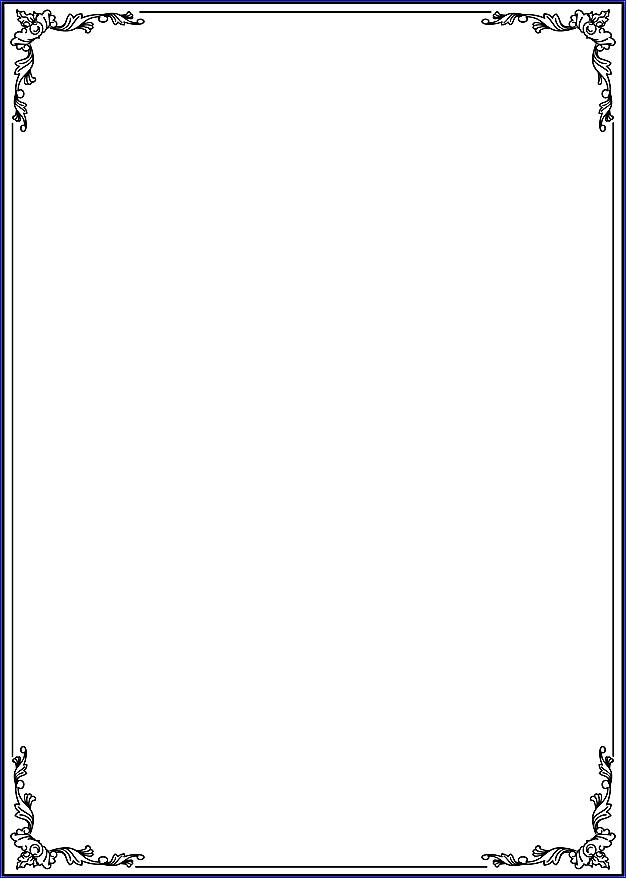
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH**



**BÁO CÁO CUỐI KÌ**

**MÔN: CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM**

**Đề tài: NGHIÊN CỨU VÀ ÁP DỤNG THUẬT TOÁN PHÂN LOẠI.  
SO SÁNH ĐỘ CHÍNH XÁC CỦA CÁC THUẬT TOÁN TRONG VIỆC NHẬN DIỆN CÁC KHỐI U ÁC TÍNH VÀ TRUNG TÍNH.**

**GVHD: Th.s Trần Phong Nhã**

***Nhóm sinh viên thực hiện:***

* Huỳnh Trọng Nhân MSV: 5951071066
* Nguyễn Quang Phước MSV: 5951071080
* Nguyễn Anh Tuấn MSV: 5951071117
* Đinh Hữu Thanh Nguyên MSV: 5951071065

TPHCM, ngày 9 tháng 09 năm 2021

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ**

ĐỀ TÀI: NGHIÊN CỨU VÀ ÁP DỤNG THUẬT TOÁN CÁC THUẬT TOÁN PHÂN LOẠI TRONG VIỆC NHẬN DIỆN CÁC KHỐI U ÁC TÍNH VÀ TRUNG TINH.

Sinh viên thực hiện: HUỲNH TRỌNG NHÂN  
NGUYỄN QUANG PHƯỚC

NGUYỄN ANH TUẤN

ĐINH HỮU THANH NGUYÊN

TP. Hồ Chí Minh, năm 2021 tháng 09 ngày 9

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI **CỘNG HÒA XÃ HỘI CHỦ NGHIÃ VIỆT NAM**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH** Độc lập – Tự do – Hạnh phúc

# THIẾT KẾ TỔNG QUAN ĐỀ TÀI

**Mã sinh viên:**5951071066 **Họ tên SV:** Huỳnh Trọng Nhân

5951071080 Nguyễn Quang Phước

5951071117 Nguyễn Anh Tuấn

5951071065 Đinh Hữu Thanh Nguyên

**Khóa:** 59 **Lớp:** CQ.59.CNTT

1. **Tên đề tài:**

Nghiên cứu và áp dụng thuật toán các thuật toán phân lớp trong việc nhận diện các khối u ác tính và trung tính.

1. **Mục đích, yêu cầu:**

Giúp việc chuẩn đoán và phát hiện sẽ nhanh hơn quá trình xét nghiệm và chính xác hơn quá trình chuẩn đoán chủ quan của bác sĩ

1. **Nội dung và phạm vi đề tài:**

Khối u trong bệnh ung thư còn gọi là các khối u ác tính, không chỉ lớn lên về kích thước mà còn tạo thành các rễ cắm vào khu vực xung quanh, phá hoại các vùng này. Người ta gọi đây là tính chất xâm lấn của ung thư. Hình ảnh con cua với càng, chân bám rộng xung quanh mô tả tính chất này đã trở thành biểu tượng của bệnh ung thư.

Các tế bào bị ung thư còn gọi là các tế bào ác tính có khả năng tách ra khỏi khối u ban đầu, trôi dạt đến các nơi khác trong cơ thể, sinh sôi tiếp ở đó tựa như ong tách khỏi đàn và đến nơi khác tạo một tổ ong mới. Tính chất này gọi là di căn.

Nếu các tế bào của khối u không phải là ung thư thì khối u đó là lành tính. Khối u lành tính có tính chất không xâm lấn các mô gần đó hoặc lan sang các khu vực khác của cơ thể (di căn). Một khối u lành tính ít đáng lo ngại trừ khi nó chèn ép vào các mô, dây thần kinh hoặc mạch máu gần đó và gây tổn thương. Ví dụ u xơ trong u xơ tử cung hoặc lipomas là ví dụ điển hình của khối u lành tính.

**Công nghệ, công cụ và ngôn ngữ lập trình:**

Sử dụng ngôn ngữ lập trình chủ yếu là R trong việc thống kê dữ liệu cho việc phân loại, dự đoán và Python trong việc làm cú pháp trở nên trực quan hơn. Ngoài ra còn sử dụng khá nhiều thư viện đi kèm trong việc xây dựng mô hình.

1. **Các kết quả chính dự kiến sẽ đạt được và ứng dụng:**

Mô hình có thể đạt được dự đoán chính xác là 92% cho việc chuẩn đoán khối u cho bệnh nhân

1. **Giảng viên và cán bộ hướng dẫn**

Họ tên: Trần Phong Nhã

Đơn vị công tác: Trường Đại Học Giao Thông Vận Tải Cơ Sở 2

Điện thoại: Email:

|  |  |
| --- | --- |
|  | ***Ngày 9 tháng 09 năm 2021***  ***Giảng viên hướng dẫn*** |
|  |  |

**Mục Lục**

[THIẾT KẾ TỔNG QUAN ĐỀ TÀI 3](#_Toc85615758)

[LỜI CẢM ƠN 7](#_Toc85615759)

[GIỚI THIỆU VẤN ĐỀ 9](#_Toc85615760)

[CHƯƠNG I: MÔ TẢ BÀI TOÁN VÀ GIẢI PHÁP 11](#_Toc85615761)

[**1.** **Mô tả bài toán:** 11](#_Toc85615762)

[**1.2** **Mục tiêu nghiên cứu.** 11](#_Toc85615763)

[**1.3** **Các thuật toán Classification được áp dụng trong báo cáo** 11](#_Toc85615764)

[**1.2.1 Thuật toán SVM (Super Vector Machine)** 11](#_Toc85615765)

[**1.2.2 Thuật toán KNN (K-Nearest Neighbor)** 11](#_Toc85615766)

[**1.2.3 Thuật toán RandomForest** 11](#_Toc85615767)

[**1.4** **Supervised Learning (Học có giám sát):** 11](#_Toc85615768)

# LỜI CẢM ƠN

Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý Thầy Cô ở Khoa CNTT, Đăc biệt là thầy Trần Phong Nhã đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Nhờ có những lời hướng dẫn, dạy bảo của các thầy cô nên đề tài nghiên cứu của em mới có thể hoàn thiện tốt đẹp.

Một lần nữa, em xin chân thành cảm ơn thầy/cô – người đã trực tiếp giúp đỡ, quan tâm, hướng dẫn em hoàn thành tốt bài báo cáo này trong thời gian qua.

Bài báo cáo thực tập thực hiện trong khoảng thời gian gần 5 tuần. Bước đầu đi vào thực tế của em còn hạn chế và còn nhiều bỡ ngỡ nên không tránh khỏi những thiếu sót , em rất 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 đồng thời có điều kiện bổ sung, nâng cao ý thức của mình.

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

|  |
| --- |
| ***Tp. Hồ Chí Minh, ngày 09 tháng 09 năm 2021***  ***Giảng viên hướng dẫn*** |
|  |

# GIỚI THIỆU VẤN ĐỀ

Một trong những căn bệnh hiểm nghèo mà cho đến nay sự tác động của y học tuy có tiến bộ nhưng vẫn không mang lại nhiều sự khả quan, đặc biệt là ở các nước chưa phát triển và các nước đang phát triển, đó là bệnh ung thư. Theo thống kê của Tổ chức y tế thế giới (WHO), hàng năm có trên dưới 9 triệu người tử vong do căn bệnh này và sang thế kỷ XXI, người ta đã tiên đoán ung thư sẽ là nguyên nhân gây tử vong cao nhất. Ở Việt Nam, ung thư được nghiên cứu và phát triển khá sớm nhưng lại nằm rải rác trong các chuyên ngành như Khoa ngoại, Khoa tai mũi họng … Nay một chuyên ngành mới được ra đời như một một giải pháp cho việc chuẩn đoán đó là chuyên ngành tin sinh học. Được ra đời dựa nhu cầu của xã hội về phát hiện chuẩn chính xác và nhanh chong việc phát hiện các khối U ác tích trong cơ cũng như một giải pháp để xử bệnh ngay từ lúc đầu, giúp cải thiện về sức khỏe, tài chính cũng quy trình xử lý bệnh của bệnh nhân.

Một trong căn bệnh ung thư nguy hiểm nhất mọi thời đại có thể nhắc đến trong mọi trang báo y học, đứng đầu trong khả năng gây tử vong hàng loạt ở phụ nữ. Đó là ung thư vú.

Ung thư vú (UTV) là bệnh ung thư phổ biến nhất ở nữ giới. Theo GLOBOCAN, ước tính năm 2018 trên thế giới có khoảng hơn 2 triệu ca ung thư vú mới mắc ở nữ được chuẩn đoán chiếm 25% tổng số các loại ung thư và 627.000 ca tử vong. Tại Việt Nam, ung thư vú đứng hàng đầu trong những loại ung thư ở phụ nữ. Tỷ lệ mắc ung thư vú trên thế giới nói chung và Việt Nam nói riêng có xu hướng tăng với tỷ lệ mắc bệnh đạt sấp xỉ 23/100000 trong những năm gần đây. Ung thư vú là tình trạng tế bào ung thư khởi phát tại vú. Các tế bào ung thư ác tính tập hợp thành khối u ác tính, chúng sinh sôi rất nhanh và di căn sang các bộ phận khác trong cơ thể. Ung thư vú hầu hết xuất hiện ở nữ giới. Tuy nhiên, trong một số ít trường hợp, nam giới cũng mắc bệnh này.

Việc phát hiện và chuẩn đoán khối u thông qua các thông số chi tiết cụ thể của khối u là một rất cần thiết để tiến hành đề xuất ra các giải pháp thiết thực, thứ giúp bệnh nhân và các bác sĩ xử lý tình hình bệnh một cách nhanh chóng.

# CHƯƠNG I: MÔ TẢ BÀI TOÁN VÀ GIẢI PHÁP

## **Mô tả bài toán:**

Quay trở lại bài toán trên, có thể đưa ra rất nhiều mô hình để dự đoán khối u cho bệnh nhân nhưng có lẽ thứ ta cần nhất là độ chính xác và tính tường minh của mô hình. Trong bài báo này, ta sẽ không tiến hành đề cập vào tính tường minh của mô hình vì nói về tính tường minh sẽ liên quan mạnh đến chuyên ngành ngoài lĩnh vực và cũng không được đề cập trong mục tiêu của bài tập lớn. Thế nên trong báo này sẽ tiến hành xây nhiều mô hình để dự đoán cho một bài toán với những thuật toán khác nhau từ đó đưa ra các nhận xét về ưu nhược điểm của các thuật toán xây dựng mô hình

## **Mục tiêu nghiên cứu.**

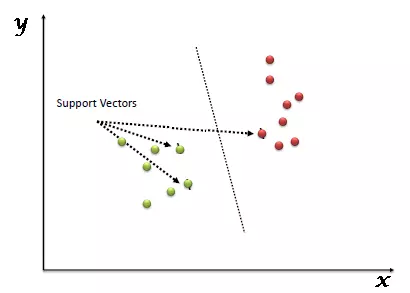
Nghiên cứu nhằm đánh giá tổng quát về việc áp dụng thuật toán trong y học cũng như việc diễn giải tính tường minh của chúng giúp chúng có thể áp dụng một cách tối ưu hóa nhất trong y học. Ngoài ra, việc áp dụng mô hình thuật toán vào trong y học giúp tối đa hóa lợi ích, thời gian và nhân lực cho nền y học nước nhà

* 1. **Các thuật toán Classification được áp dụng trong báo cáo**

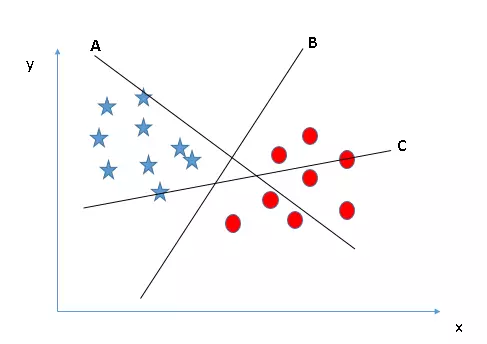
**1.2.1 Thuật toán SVM (Super Vector Machine)**

*Bỏ qua những kiến thức toán học trong thuật toán, tiến hành đi vào ý tưởng chính để thuật toán SVM có thể hoạt động*

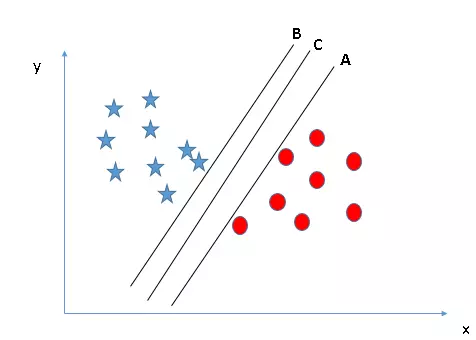
* SVM là một thuật toán supervised learning, nó có thể sử dụng cho cả việc phân loại và hồi quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại (Đối ngược với thuật toán Logistics Regression). Thuật toán này, ta có đồ thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn 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. 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.



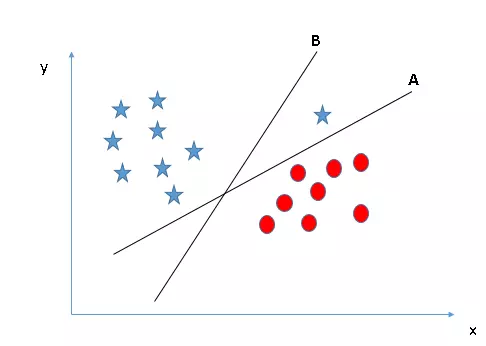
* SVM làm việc như thế nào? Ở trên, chúng ta đã thấy được việc chia hyper-plane. Bấy giờ làm thế nào chúng ta có thể xác định "Làm sao để vẽ-xác định đúng hyper-plane". Chúng ta sẽ theo các tiêu chí sau:
  + Identify the right hyper-plane (Scenario-1):Ở đây, có 3 đường hyper-lane (A,B and C). Bây giờ đường nào là hyper-lane đúng cho nhóm ngôi sao và hình tròn. Quy tắc số một để chọn 1 hyper-lane, chọn một hyper-plane để phân chia hai lớp tốt nhất. Trong ví dụ này chính là đường B.



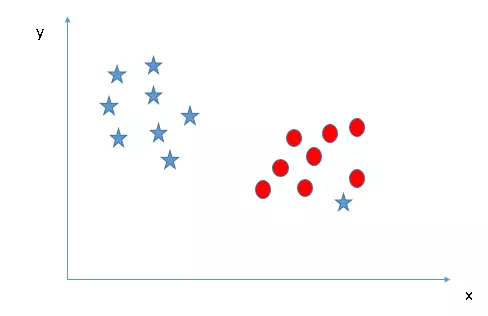
* + Identify the right hyper-plane (Scenario-2):Ở đây chúng ta cũng có 3 đường hyper-plane (A,B và C), theo quy tắc số 1, chúng đều thỏa mãn. Quy tắc thứ hai chính là xác định khoảng cách lớn nhất từ điểu gần nhất của một lớp nào đó đến đường hyper-plane. Khoảng cách này được gọi là "Margin", Hãy nhìn hình bên dưới, trong đấy có thể nhìn thấy khoảng cách margin lớn nhất đấy là đường C. Cần nhớ nếu chọn lầm hyper-lane có margin thấp hơn thì sau này khi dữ liệu tăng lên thì sẽ sinh ra nguy cơ cao về việc xác định nhầm lớp cho dữ liệu.



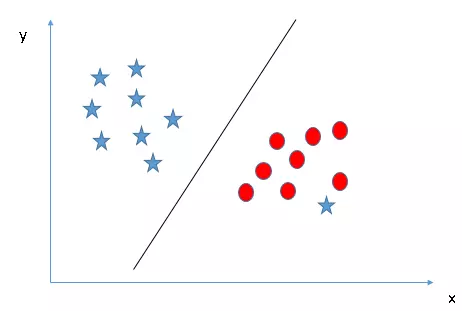
* + Identify the right hyper-plane (Scenario-3): Sử dụng các nguyên tắc đã nêu trên để chọn ra hyper-plane cho trường hợp sau: Có thể có một vài người sẽ chọn đường B bởi vì nó có margin cao hơn đường A, nhưng đấy sẽ không đúng bởi vì nguyên tắc đầu tiên sẽ là nguyên tắc số 1, chúng ta cần chọn hyper-plane để phân chia các lớp thành riêng biệt. Vì vậy đường A mới là lựa chọn chính xác.



* + Can we classify two classes (Scenario-4)? Tiếp the hãy xem hình bên dưới, không thể chia thành hai lớp riêng biệt với 1 đường thẳng, để tạo 1 phần chỉ có các ngôi sao và một vùng chỉ chứa các điểm tròn. Ở đây sẽ chấp nhận, một ngôi sao ở bên ngoài cuối được xem như một ngôi sao phía ngoài hơn, SVM có tính năng cho phép bỏ qua các ngoại lệ và tìm ra hyper-plane có biên giới tối đa . Do đó có thể nói, SVM có khả năng mạnh trong việc chấp nhận ngoại lệ.

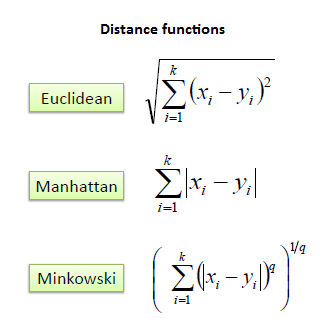


* + Find the hyper-plane to segregate to classes (Scenario-5): Trong trường hợp dưới đây, không thể tìm ra 1 đường hyper-plane tương đối để chia các lớp, vậy làm thế nào để SVM phân tách dữ liệu thành hai lớp riêng biệt? Cho đến bây giờ chúng ta chỉ nhìn vào các đường tuyến tính hyper-plane. SVM có thể giải quyết vấn đề này, Khá đơn giản, nó sẽ được giải quyết bằng việc thêm một tính năng, Ở đây chúng ta sẽ thêm tính năng z = x^2+ y^2. Bây giờ dữ liệu sẽ được biến đổi theo trục x và z như sau

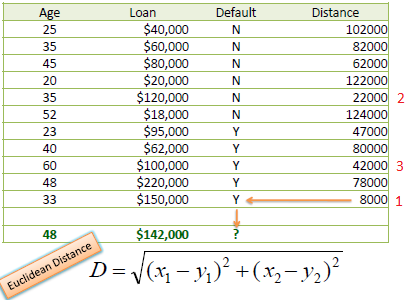


**1.2.2 Thuật toán KNN (K-Nearest Neighbor)**

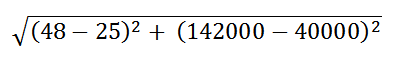
Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại **gần nhau** trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp.



Ví dụ chúng ta có dữ liệu là tuổi, khoản vay và khả năng vở nợ như hình:

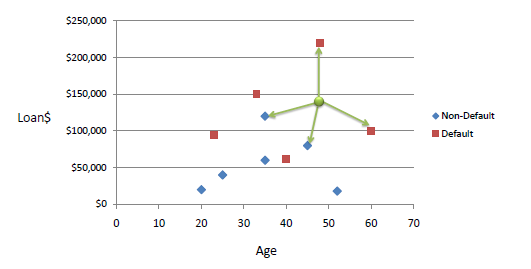


Dữ liệu cần phân loại của chúng ta là {age: 48, loan: 142000}. Đây dữ liệu 2 chiều và chúng ta cần dự đoán người này thuộc nguy cơ vở nợ hay không. Chúng ta sẽ dùng một cách khá phổ biến để tính khoảng cách là Euclidean. Ví dụ ở hàng đầu tiên khoảng cách sẽ được tính:



Thực hiện tương tự, ta sẽ tính được khoảng cách ở cột Distance, từ đó chọn ra k = 3 khoảng cách nhỏ nhất (gần với dữ liệu vào nhất). Với 3 khoảng cách này chúng ra nhận được 3 label là (Yes, No, Yes). Trong 3 label này Yes xuất hiện nhiều hơn nên chúng ta sẽ đưa ra dự đoán người này có khả năng vở nợ.

Vì đây là dử liệu 2 chiều nên chúng ta củng có thể biểu diễn dữ liệu trong hệ tọa độ như hình:



Trên hệ tọa độ này chúng ta thể dễ dàng nhận thấy cách chúng ta chọn k điểm gần nhất. Nhưng với dữ liệu lớn, nhiều chiều thì việc biểu diễn dữ liệu trên một không gian là không hề dễ dàng.

**1.2.3 Thuật toán RandomForest**

Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. Random forests có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng. Nó có thể được sử dụng để phân loại các ứng viên cho vay trung thành, xác định hoạt động gian lận và dự đoán các bệnh. Nó nằm ở cơ sở của thuật toán Boruta, chọn các tính năng quan trọng trong tập dữ liệu.

Thuật toán Random Forests

Giả sử bạn muốn đi trên một chuyến đi và bạn muốn đi đến một nơi mà bạn sẽ thích.

Vậy bạn sẽ làm gì để tìm một nơi mà bạn sẽ thích? Bạn có thể tìm kiếm trực tuyến, đọc các bài đánh giá trên blog và các cổng thông tin du lịch hoặc bạn cũng có thể hỏi bạn bè của mình.

Giả sử bạn đã quyết định hỏi bạn bè và nói chuyện với họ về trải nghiệm du lịch trong quá khứ của họ đến những nơi khác nhau. Bạn sẽ nhận được một số khuyến nghị từ tất cả các bạn. Bây giờ bạn phải tạo danh sách các địa điểm được đề xuất. Sau đó, bạn yêu cầu họ bỏ phiếu (hoặc chọn địa điểm tốt nhất cho chuyến đi) từ danh sách các địa điểm được đề xuất bạn đã thực hiện. Địa điểm có số phiếu bầu cao nhất sẽ là lựa chọn cuối cùng của bạn cho chuyến đi.

Trong quá trình quyết định ở trên, có hai phần. Trước tiên, hãy hỏi bạn bè về trải nghiệm du lịch cá nhân của họ và nhận được đề xuất từ nhiều nơi họ đã ghé thăm. Điều này cũng giống như sử dụng thuật toán cây quyết định. Ở đây, mỗi người trong số các bạn chọn những nơi mà họ đã ghé thăm cho đến nay. Phần thứ hai, sau khi thu thập tất cả các khuyến nghị, là thủ tục bỏ phiếu để chọn địa điểm tốt nhất trong danh sách các khuyến nghị. Toàn bộ quá trình nhận được khuyến nghị từ bạn bè và bỏ phiếu cho họ để tìm ra nơi tốt nhất được gọi là thuật toán rừng ngẫu nhiên.

Về mặt kỹ thuật, nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin, tỷ lệ tăng và chỉ số Gini cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và lớp phổ biến nhất được chọn là kết quả cuối cùng. Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác.

Thuật toán hoạt động như thế nào?

Nó hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

Ưu điểm: Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Lý do chính là nó mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến. Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy. Random forests cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu. Bạn có thể nhận được tầm quan trọng của tính năng tương đối, giúp chọn các tính năng đóng góp nhiều nhất cho trình phân loại. Nhược điểm: Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

Các tính năng quan trọng

Random forests cũng cung cấp một chỉ số lựa chọn tính năng tốt. Scikit-learn cung cấp thêm một biến với mô hình, cho thấy tầm quan trọng hoặc đóng góp tương đối của từng tính năng trong dự đoán. Nó tự động tính toán điểm liên quan của từng tính năng trong giai đoạn đào tạo. Sau đó, nó cân đối mức độ liên quan xuống sao cho tổng của tất cả các điểm là 1.

Điểm số này sẽ giúp bạn chọn các tính năng quan trọng nhất và thả các tính năng quan trọng nhất để xây dựng mô hình.

Random forests sử dụng tầm quan trọng của gini hoặc giảm tạp chất trung bình (MDI) để tính toán tầm quan trọng của từng tính năng. Gini tầm quan trọng còn được gọi là tổng giảm trong tạp chất nút. Đây là mức độ phù hợp hoặc độ chính xác của mô hình giảm khi bạn thả biến. Độ lớn càng lớn thì biến số càng có ý nghĩa. Ở đây, giảm trung bình là một tham số quan trọng cho việc lựa chọn biến. Chỉ số Gini có thể mô tả sức mạnh giải thích tổng thể của các biến. Random Forests và cây quyết định Random Forests là một tập hợp của nhiều cây quyết định. Cây quyết định sâu có thể bị ảnh hưởng quá mức, nhưng Random forests ngăn cản việc lấp đầy bằng cách tạo cây trên các tập con ngẫu nhiên. Cây quyết định nhanh hơn tính toán. Random forests khó giải thích, trong khi cây quyết định có thể diễn giải dễ dàng và có thể chuyển đổi thành quy tắc.

## **Supervised Learning (Học có giám sát):**

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN}X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN}Y={y1,y2,…,yN}, trong đó xi,yixi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y(xi,yi)∈X×Y được gọi là tập training data (dữ liệu huấn luyện). Từ tập training data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập XX sang một phần tử (xấp xỉ) tương ứng của tập YY:

yi≈f(xi),  ∀i=1,2,…,Nyi≈f(xi),  ∀i=1,2,…,NMục đích là xấp xỉ hàm số ff thật tốt để khi có một dữ liệu xx mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x)y=f(x).

**Ví dụ 1:** trong nhận dạng chữ 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.



[MNIST](http://yann.lecun.com/exdb/mnist/): bộ cơ sở dữ liệu của chữ số viết tay.  
(Nguồn: [Simple Neural Network implementation in Ruby)](http://www.rubylab.io/2015/03/18/simple-neural-network-implenentation-in-ruby/)

Ví dụ này khá giống với cách học của con người khi còn nhỏ. Ta đưa bảng chữ cái cho một đứa trẻ và chỉ cho chúng đây là chữ A, đây là chữ B. Sau một vài lần được dạy thì trẻ có thể nhận biết được đâu là chữ A, đâu là chữ B trong một cuốn sách mà chúng chưa nhìn thấy bao giờ.

**Ví dụ 2:** Thuật toán dò các khuôn mặt trong một bức ảnh đã được phát triển từ rất lâu. Thời gian đầu, facebook sử dụng thuật toán này để chỉ ra các khuôn mặt trong một bức ảnh và yêu cầu người dùng tag friends - tức gán nhãn cho mỗi khuôn mặt. Số lượng cặp dữ liệu (khuôn mặt, tên người) càng lớn, độ chính xác ở những lần tự động tag tiếp theo sẽ càng lớn.

**Ví dụ 3:** Bản thân thuật toán dò tìm các khuôn mặt trong 1 bức ảnh cũng là một thuật toán Supervised learning với training data (dữ liệu học) là hàng ngàn cặp (ảnh, mặt người) và (ảnh, không phải mặt người) được đưa vào. Chú ý là dữ liệu này chỉ phân biệt mặt người và không phải mặt người mà không phân biệt khuôn mặt của những người khác nhau.

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

#### **Classification (Phân loại)**

Một bài toán được gọi là classification nếu các label của input data được chia thành một số hữu hạn nhóm. Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này.

#### **Regression (Hồi quy)**

(tiếng Việt dịch là Hồi quy, tôi không thích cách dịch này vì bản thân không hiểu nó nghĩa là gì)

Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu?

Gần đây [Microsoft có một ứng dụng dự đoán giới tính và tuổi dựa trên khuôn mặt](https://how-old.net/). Phần dự đoán giới tính có thể coi là thuật toán **Classification**, phần dự đoán tuổi có thể coi là thuật toán **Regression**. Chú ý rằng phần dự đoán tuổi cũng có thể coi là ***Classification*** nếu ta coi tuổi là một số nguyên dương không lớn hơn 150, chúng ta sẽ có 150 class (lớp) khác nhau.

Chúng ta có thể hình Machine Learning chính là một dạng của đa thức nội suy Lagrange chúng ta có (input và outcome). Việc chúng ta cần làm là tìm ra phương trình sao đáp ứng tương đối sấp xỉ với kết quả đầu ra.

Nói về đa thức nội suy Lagrange:

Ví dụ ta có:  P(x) là một đa thức bậc hai và chúng ta có thể tính được

P(1)=2,  P(2)=5,  P(3)=12.

Từ đó ta phải tìm ra được phương trình sao cho đáp ứng được kết quả đầu ra từ đó có thể thử nghiệm trên tập dữ liệu chưa được thấy bao giờ như là lấy phương trình trên để tìm kết quả của P(5) = ?.

Đây có thể được coi là một định nghĩa, cấu hình chung của phương thức Supervised learning. Nhưng bản thân phương thức học này lại chia ra làm 2 loại là Classification và Regression. Vậy điểm khác nhau giữa loại này trong phương thức học tập này ?

Với Classification, đây là một nhánh trong phương thức học máy dự đoán outcome nằm trong khối lượng nhãn có hạn nghĩa là nó sẽ trả kết quả dự đoán này trong khung giới hạn quy định. Lấy ví dụ bên trên khi tạo một học máy kiểm tra Emails spam thì kết trả về phải là True(1) or False(0) không thể nào xuất hiện kết quả thứ 3 xuất hiện và [0,1] chính là không gian giới hạn outcome của dữ liệu này.

Thế khi nói về