**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH**

🖎🕮✍



**ĐỒ ÁN MÔN HỌC**

PHÂN TÍCH DỮ LIỆU CÁC CÔNG TY GIAN LẬN CỦA MỘT VĂN PHÒNG KIỂM TOÁN ẤN ĐỘ BẰNG PHẦN MỀM ORANGE

**Học phần**: Khoa Học Dữ Liệu

**Nhóm Sinh Viên**:

**Chuyên Ngành**: KẾ TOÁN

**Khóa**: K48

**Giảng Viên**: TS. Đặng Ngọc Hoàng Thành

**TP. Hồ Chí Minh, Ngày 06 tháng 08 năm 2023**

**Lời cảm ơn**

Lời đầu tiên em xin gửi lời cảm ơn đến Đại học Kinh tế TP.HCM, đặc biệt là thầy Đặng Ngọc Hoàng Thành, giảng viên đảm nhiệm bộ môn Khoa học dữ liệu của nhóm chúng em. Tuy thời gian thầy trò tiếp xúc chỉ vỏn vẹn có sáu buổi học, nhưng cũng đủ để chúng em cảm nhận được sự nhiệt tình, tận tâm của thầy trong quá trình giảng dạy. Môn Khoa học dữ liệu đối với chúng em khá là khô khan, ít thú vị hơn so với những môn học khác, tuy nhiên, nó đã đem lại cho chúng em những kiến thức mới và những kiến thức này hoàn toàn thực tiễn và chúng em có thể áp dụng nó vào nhiều vấn đề nghiên cứu cũng như công việc.

Đối với chúng em, Khoa học dữ liệu là một môn học khó vì vậy để có thể thực hiện được đồ án nhóm cuối kì như thế này thật sự không dễ dàng, chúng em đã dành rất nhiều thời gian, tâm huyết để tự nghiên cứu và cũng nhờ có sự hướng dẫn tận tình của thầy nên mới có thể hoàn thành được. Trong quá trình làm bài có thể có những sai sót mà chúng em mắc phải, mong thầy có thể đưa ra nhận xét và một số phương hướng để chúng em giải quyết sai lầm để có thể hoàn thiện hơn về mặt tri thức trong tương lai. Chúc thầy thật nhiều sức khỏe, hạnh phúc và sẽ mãi giữ được sự nhiệt huyết để truyền đạt kiến thức cho những khóa sau. Nhóm em xin cảm ơn thầy rất nhiều ạ!

# MỤC LỤC

[MỤC LỤC 2](#_Toc147364800)

[DANH MỤC HÌNH 3](#_Toc147364801)

[CHƯƠNG 1. TỔNG QUAN 5](#_Toc147364802)

[1.1. Tổng Quan Về Bài Toán Phân Lớp Dữ Liệu 5](#_Toc147364803)

[1.2. Giới Thiệu Về Phần Mềm Orange 5](#_Toc147364804)

[1.3. Lý Do Chọn Lựa Đề Tài 5](#_Toc147364805)

[CHƯƠNG 2. CÁC MÔ HÌNH PHÂN LỚP DỮ LIỆU 6](#_Toc147364806)

[2.1. Các Mô Hình Phân Lớp Dữ Liệu 6](#_Toc147364807)

[2.1.1. Mô Hình Logistic Regression 6](#_Toc147364808)

[2.1.2. Mô Hình Support Vector Machine 7](#_Toc147364809)

[2.1.3. Mô Hình Decision Tree 8](#_Toc147364810)

[2.2. Quy Trình Phân Lớp Dữ Liệu 9](#_Toc147364811)

[2.2.1 Quá trình phân lớp dữ liệu 9](#_Toc147364812)

[2.2.1. Tiền Xử Lý Dữ Liệu 11](#_Toc147364813)

[2.2.2. Phân Lớp Dữ Liệu 11](#_Toc147364814)

[2.2.3. Đánh Giá Tính Hiệu Quả 12](#_Toc147364815)

[CHƯƠNG 3. CÁC KẾT QUẢ THỰC NGHIỆM 15](#_Toc147364816)

[3.1. Bộ dữ liệu 15](#_Toc147364817)

[Tiền xử lý dữ liệu: 15](#_Toc147364818)

[Bảng mô tả dữ liệu 15](#_Toc147364819)

[3.2. Các Kết Quả Thực Nghiệm 16](#_Toc147364820)

[3.3. Phân Tích và Đánh Giá 17](#_Toc147364821)

[3.3. Kết quả dự báo 21](#_Toc147364822)

[3.4 Kiến thức chuyên ngành 21](#_Toc147364823)

[CHƯƠNG 4. KẾT LUẬN 23](#_Toc147364824)

[4.1. Các Kết Quả Đạt Được 23](#_Toc147364825)

[4.2. Những Hạn Chế và Hướng Phát Triển 23](#_Toc147364826)

[4.2.1 Hạn chế 23](#_Toc147364827)

[4.2.2 Hướng phát triển 23](#_Toc147364828)

[TÀI LIỆU THAM KHẢO 25](#_Toc147364829)

# 

# DANH MỤC HÌNH

[Hình 1: Minh họa mô hình Hồi quy Logistic 6](file:///C:\Users\hoan\Downloads\DoanKHDL%20(3).docx#_Toc147352528)

[Hình 2: Minh họa Mô hình SVM. 7](file:///C:\Users\hoan\Downloads\DoanKHDL%20(3).docx#_Toc147352533)

[Hình 3: Xây dựng mô hình phân lớp 9](#_Toc147352540)

[Hình 4: Đánh giá mô hình phân lớp. 10](#_Toc147352546)

[Hình 5: Quá trình phân lớp dữ liệu mới. 11](#_Toc147352552)

[Hình 6: Minh họa phương pháp ma trận nhầm lẫn. 12](#_Toc147363647)

[Hình 7: Minh họa phương pháp ROC 14](#_Toc147363658)

[Hình 8: Minh họa phương pháp AUC 14](#_Toc147363663)

[Hình 9: Phân tách dữ liệu. 16](#_Toc147363668)

[Hình 10. Mô hình phân lớp. 17](#_Toc147363674)

[Hình 11: Kết quả Test & Score 17](#_Toc147363683)

[Hình 12 : Kết quả Confusion Matrix của phương pháp Logistic Regession 18](#_Toc147363737)

[Hình 13: Kết quả Confusion Matrix của phương pháp Decision tree 19](#_Toc147363744)

[Hình 14: Kết quả Confusion Matrix của phương pháp SVM 19](#_Toc147363749)

[Hình 15.1: Kết quả phân lớp theo mô hình Decision tree 20](#_Toc147363755)

[Hình 15.2: Kết quả phân lớp theo mô hình Decision tree (skip Score) 21](#_Toc147363768)

[Hình 16: Kết quả dự báo của Predeictions 21](#_Toc147363774)

# CHƯƠNG 1. TỔNG QUAN

## 1.1. Tổng Quan Về Bài Toán Phân Lớp Dữ Liệu

Phân lớp dữ liệu là một chức năng khai thác dữ liệu chỉ định các mục trong một tập hợp cho các danh mục hoặc lớp mục tiêu. Mục tiêu của phân lớp dữ liệu là dự đoán chính xác lớp mục tiêu cho từng trường hợp trong dữ liệu.

Bài toán phân lớp của nhóm: Dự báo nguy cơ gian lận/ không gian lận của các công ty dựa vào khả năng xảy ra rủi ro ở các báo cáo thu thập (Bài toán phân lớp).

## 1.2. Giới Thiệu Về Phần Mềm Orange

Giới thiệu về Phần Mềm Orange: Orange là một công cụ khai phá dữ liệu và học máy nguồn mở, được viết bằng Python. Orange cung cấp một môi trường trực quan và tương tác để phân tích dữ liệu và xây dựng các mô hình học máy.

Orange có một số tính năng chính, bao gồm:

* Trực quan hóa dữ liệu: Orange cung cấp một loạt các công cụ để trực quan hóa dữ liệu, bao gồm biểu đồ, đồ thị, và bản đồ nhiệt.
* Khai thác dữ liệu: Orange cung cấp một loạt các thuật toán khai thác dữ liệu, bao gồm phân loại, hồi quy, và clustering.
* Học máy: Orange cung cấp một loạt các mô hình học máy, bao gồm các mô hình dựa trên cây, các mô hình dựa trên sự hỗ trợ vector, và các mô hình dựa trên mạng nơ-ron.

Orange là một công cụ mạnh mẽ cho phân tích dữ liệu và học máy. Nó là một lựa chọn tốt cho các nhà khoa học dữ liệu, kỹ sư máy học, và sinh viên.

## 1.3. Lý Do Chọn Lựa Đề Tài

Trong thời đại công nghệ và thông tin bùng nổ như hiện nay, khoa học dữ liệu là một ngành khoa học liên quan đến việc thu thập, xử lý và phân tích dữ liệu lớn để tìm ra các mẫu và xu hướng ẩn. Ngành này sử dụng các kỹ thuật và công nghệ từ các lĩnh vực khác nhau, bao gồm toán học, thống kê, máy học, trí tuệ nhân tạo và khoa học máy tính.

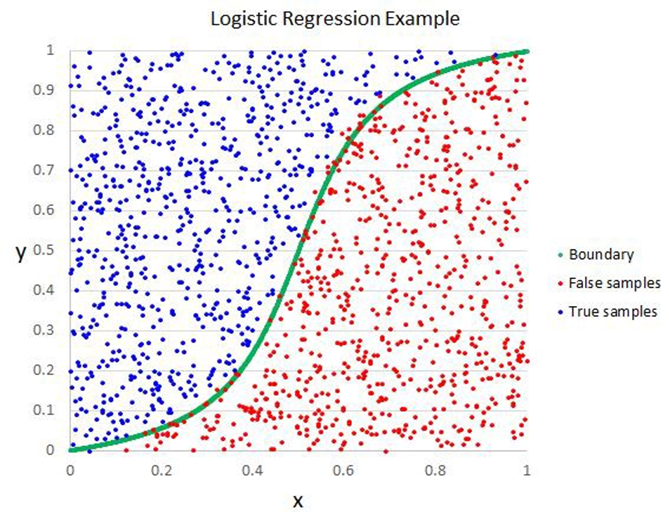
Tình hình gian lận báo cáo tài chính của các công ty hiện nay vẫn đang là một vấn đề đáng lo ngại. Theo một báo cáo của Hiệp hội Kiểm toán và Thẩm định Quốc tế (IFAC), gian lận trong báo cáo tài chính đang gia tăng trên toàn thế giới. Báo cáo cho thấy rằng số vụ gian lận được phát hiện đã tăng 10% trong năm 2022. Có một số nguyên nhân dẫn đến tình trạng này, bao gồm: sự phức tạp ngày càng tăng của các báo cáo tài chính, sự phát triển của công nghệ: Công nghệ có thể được sử dụng để thực hiện gian lận một cách tinh vi hơn; … Điều này tạo nên thách thức lớn đối với người quản lý công ty cũng như các công ty kiểm toán và các kiểm toán viên. Từithực tế đó, các nhà lãnh đạo và các công ty kiểm toán mong muốn có được một công cụ nhằm lưu trữ, truy xuất và có khả năng xử lý dữ liệu tốt để hỗ trợ các kiểm toán viên. Đáp ứng được các yêu cầu trên, các phần mềm về khoa học dữ liệu đã và đang được ứng dụng nhiều trong các công ty kiểm toán. Nhằm nghiên cứu và tìm hiểu rõ hơn về việc đánh giá được các gian lận nhờ vào các công cụ khai phá dữ liệu cũng như để ứng dụng những kiến thức được học tại trường trong bộ môn Khoa học dữ liệu, nhóm chúng em đã quyết định thực hiện đề tài “Phân tích dữ liệu các công ty gian lận của một văn phòng kiểm toán Ấn Độ”

# CHƯƠNG 2. CÁC MÔ HÌNH PHÂN LỚP DỮ LIỆU

## 2.1. Các Mô Hình Phân Lớp Dữ Liệu

Định nghĩa: Là quá trình phân một đối tượng dữ liệu vào một hay nhiều lớp (loại) đã cho trước nhờ một mô hình phân lớp. Mô hình này được xây dựng dựa trên một tập dữ liệu đã được gán nhãn trước đó (thuộc về lớp nào). Kỹ thuật này dùng để rút trích các thông tin cần thiết từ kho dữ liệu có sẵn. Vì thế, đối với kỹ thuật này, chúng ta sẽ áp dụng các thuật toán khác nhau tùy thuộc vào mục tiêu sử dụng. Đây cũng là kỹ thuật có vai trò quan trọng trong việc dự báo các quy luật, xu hướng, … bằng cách mô tả các thuộc tính liên quan để đối tượng được phân loại vào một lớp cụ thể.

### 2.1.1. Mô Hình Logistic Regression

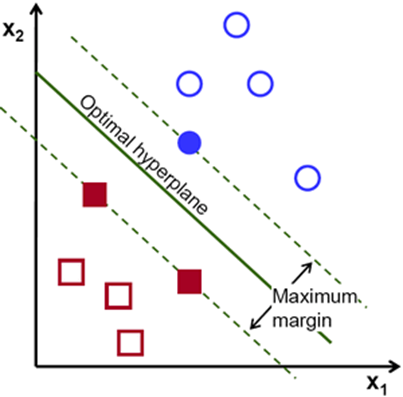
**Khái niệm:** là mô hình dùng để dự đoan xác suất của một sự kiện xảy ra dựa trên một hoặc nhiều biến độc lập. Hồi quy logistic sử dụng một hàm logistic để biến đổi giá trị của biến phụ thuộc nhị phân thành một giá trị liên tục nằm trong khoảng từ 0 đến 1. Giá trị này có thể được hiểu như là xác suất để sự kiện xảy ra khi biết các biến độc lập.

Hình 1: Minh họa mô hình Hồi quy Logistic

Các loại hồi quy logistic:

* Hồi quy logistic nhị phân: Đây là loại hồi quy logistic khi biến phụ thuộc chỉ có hai kết quả có thể, chẳng hạn như có hay không, đạt hay không đạt, bệnh hay khỏe, v.v. Hồi quy logistic nhị phân sử dụng một hàm logistic để biến đổi giá trị của biến phụ thuộc nhị phân thành một giá trị liên tục nằm trong khoảng từ 0 đến 1. [Giá trị này có thể được hiểu như là xác suất để sự kiện xảy ra khi biết các biến độc lập](https://bing.com/search?q=c%c3%a1c+lo%e1%ba%a1i+h%e1%bb%93i+quy+logistic).
* Hồi quy logistic đa thức: Đây là loại hồi quy logistic khi biến phụ thuộc có hai hoặc nhiều kết quả có thể mà không cần sắp xếp thứ tự. Ví dụ: dự đoán chất lượng thực phẩm, loại hoa, màu sắc, v.v. [Hồi quy logistic đa thức sử dụng một hàm softmax để biến đổi giá trị của biến phụ thuộc thành một vector xác suất cho mỗi lớp](https://aws.amazon.com/vi/what-is/logistic-regression/).
* Hồi quy logistic thứ tự: Đây là loại hồi quy logistic khi biến phụ thuộc có hai hoặc nhiều kết quả có thể mà có sắp xếp thứ tự. Ví dụ: đánh giá sản phẩm, mức độ hài lòng, mức độ khó khăn, v.v. Hồi quy logistic thứ tự sử dụng một hàm logit để biến đổi giá trị của biến phụ thuộc thành một giá trị liên tục nằm trong khoảng từ -∞ đến +∞.

### 2.1.2. Mô Hình Support Vector Machine

Ứng dụng thực tế trong các lĩnh vực như phát hiện ung thư, nhận diện khuôn mặt, nhận dạng chữ viết tay,....

Hình 2: Minh họa Mô hình SVM.

Khái niệm: là một thuật toán học có giám sát để phân loại và hồi quy. SVM tìm ra một siêu phẳng (hyperplane) để phân chia dữ liệu thành hai hoặc nhiều lớp khác nhau sao cho khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất là lớn nhất. SVM có nhiều ứngdụng thực tế trong các lĩnh vực như phát hiện ung thư, nhận diện khuôn mặt, nhận dạng chữ viết tay,...

**Ưu điểm:**

* [SVM có khả năng tạo ra các mô hình phân loại chính xác và ổn định, đặc biệt khi số lượng biến độc lập lớn hơn số lượng quan sát](https://viblo.asia/p/support-vector-machine-trong-hoc-may-mot-cai-nhin-don-gian-hon-XQZkxoQmewA).
* [SVM có thể giảm thiểu việc trang bị quá mức (overfitting) bằng cách sử dụng tham số C để kiểm soát độ lớn của lề](https://viblo.asia/p/support-vector-machine-trong-hoc-may-mot-cai-nhin-don-gian-hon-XQZkxoQmewA).
* [SVM có thể xử lý được các bài toán phi tuyến bằng cách sử dụng các hàm hạt nhân khác nhau, chẳng hạn như tuyến tính, đa thức, RBF, sigmoid, ....](https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB)
* [SVM có khả năng chấp nhận ngoại lệ và tìm ra siêu phẳng có biên giới tối đa](https://qastack.vn/stats/24437/advantages-and-disadvantages-of-svm).

**Nhược điểm:**

* SVM khó khăn trong việc xác định các tham số tối ưu cho mô hình, chẳng hạn như tham số C, gamma, epsilon, v.v. [Các tham số này ảnh hưởng đến hiệu suất của mô hình và cần được tìm kiếm bằng các phương pháp như tìm kiếm lưới (grid search) hoặc tìm kiếm ngẫu nhiên (random search)](https://vi.wikipedia.org/wiki/M%C3%A1y_vect%C6%A1_h%E1%BB%97_tr%E1%BB%A3).
* SVM khó khăn trong việc diễn giải ý nghĩa của mô hình, đặc biệt khi sử dụng các hàm hạt nhân phi tuyến. [Không có cách nào để biết được các biến độc lập nào quan trọng nhất hoặc làm thế nào chúng ảnh hưởng đến kết quả](https://vi.wikipedia.org/wiki/M%C3%A1y_vect%C6%A1_h%E1%BB%97_tr%E1%BB%A3).

### 2.1.3. Mô Hình Decision Tree

Mô hình Decision Tree là một mô hình học máy phân loại, được sử dụng để dự đoán giá trị mục tiêu của một biến phụ thuộc dựa trên các biến độc lập. Mô hình này hoạt động bằng cách chia dữ liệu thành các nhóm dựa trên các giá trị của các biến độc lập. Mỗi nhóm được gán một giá trị mục tiêu.

Cấu trúc của mô hình Decision Tree là một cây phân nhánh, với mỗi nút đại diện cho một quyết định. Các nút lá của cây đại diện cho các giá trị mục tiêu có thể có.

**Cách xây dựng mô hình Decision Tree**

Mô hình Decision Tree đượci xây dựng bằng cách sử dụng một thuật toán học máy. Thuật toán này sẽ bắt đầu từ một nút gốc, đại diện cho toàn bộ tập dữ liệu. Sau đó, thuật toán sẽ phân chia tập dữ liệu thành hai nhóm dựa trên một biến độc lập. Nhóm nào có độ đồng nhất cao hơn sẽ được chọn làm nhóm gốc cho nút con tiếp theo. Quái trình này sẽ tiếp tục cho đến khi tất cả dữ liệu được phân loại thành các nhóm riêng biệt.

**Ưu điểm**

* Mô hình Decision Tree có thể đượci hiểu và giải thích một cách dễ dàng.
* Mô hình Decision Tree có thể được sử dụng để giải quyết các vấn đề phân loại phức tạp.
* Mô hình Decision Tree có thể được sử dụng để xử lý dữ liệu lớn.

**Nhược điểm**

* Mô hình Decision Tree có thể gặp vấn đề overfitting.
* Mô hình Decision Tree có thể không hiệu quả đối với các vấn đề dự đoán liên tục.

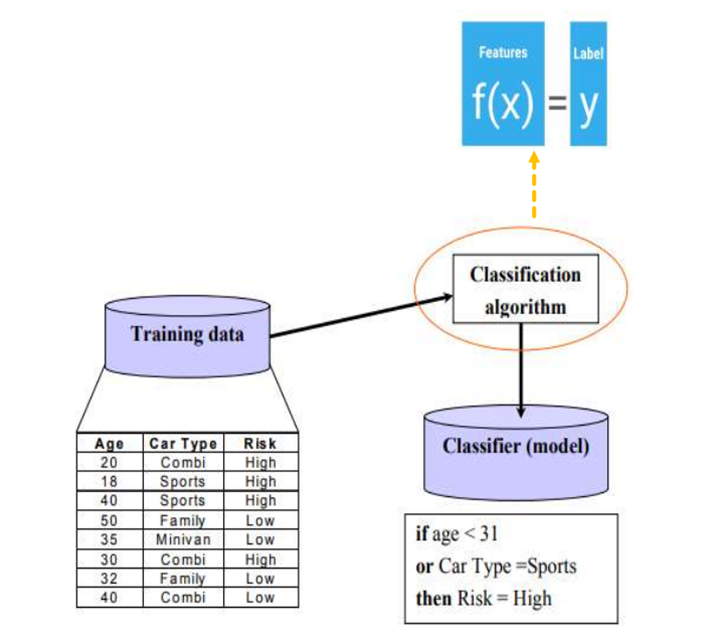
## 2.2. Quy Trình Phân Lớp Dữ Liệu

### 2.2.1 Quá trình phân lớp dữ liệu

Quá trình phân lớp dữ liệu gồm có hai bước chính:

***Bước 1:******Xây dựng mô hình phân lớp (hay còn gọi là giai đoạn “học’’ hoặc “huấn luyện’’)***

Quá trình huấn luyện này nhằm mục đích xây dựng một mô hình mô tả dữ liệu đã có sẵn. Đầu vào của quá trình này là một tập dữ liệu mẫu được gán nhãn và tiền xử lý, mỗi phần tử của dữ liệu đã được giả định thuộc về một phân lớp trước, lớp ở câu nói này có nghĩa là giá trị của một thuộc tính được chọn làm thuộc tính gán nhãn hoặc phân lớp. Mỗi bộ giá trị được gọi chung là mỗi phần tử dữ liệu (data Hình 1: Xây dựng mô hình phân lớp 1 tuple), cũng có thể là các mẫu, ví dụ, đối tượng hay các trường hợp khác. Từ đó cho ra kết quả là mô hình phân lớp đã được huấn luyện.

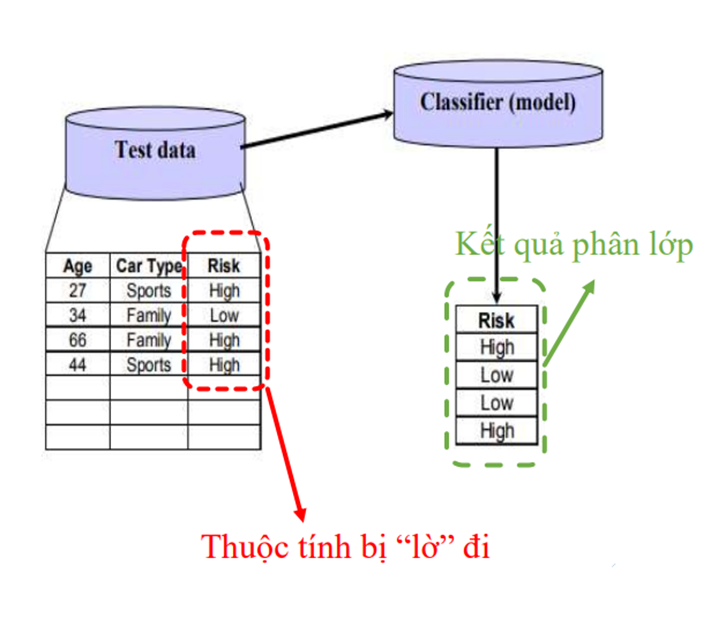


Hình 3: Xây dựng mô hình phân lớp

· ***Bước 2: Sử dụng mô hình, trong đó chia thành 2 bước nhỏ:***

· ***Bước 2.1: Đánh giá mô hình (kiểm tra tính đúng đắn của mô hình)***

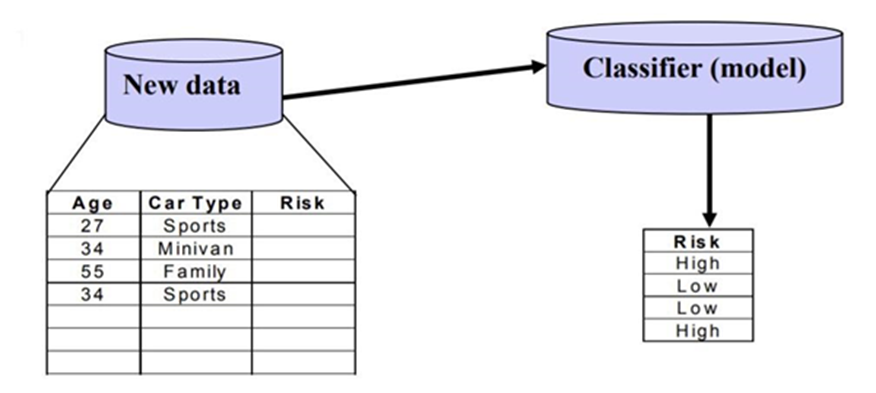
Dữ liệu đầu vào: là một tập dữ liệu mẫu khác đã được gán nhãn và tiền xử lý. Tuy nhiên lúc đưa vào mô hình phân lớp, ta “lờ” đi thuộc tính đã được gán nhãn. Bước này dùng mô hình đã xây dựng ở bước đầu tiên để phân lớp dữ liệu mới. Trước tiến độ chính xác mang tính chất dự đoán của mô hình phâni lớp vừa tạo ra được ước lượng. Holdout là một kỹ thuật đơn giản để ước lượng tính đúng đắn này dựa trên tỷ lệ phần trăm các mẫu trong tập dữ liệu kiểm tra được mô hình phân lớp đúng (so với thực tế). Tính đúng đắn của mô hình sẽ đượci xác định bằng cách so sánh thuộc tínhi gán nhãn của dữ liệu đầu vào và kết quả phân lớp của mô hình.



Hình 4: Đánh giá mô hình phân lớp.

· ***Bước 2.2: Phân lớp dữ liệu mới***

Dữ liệu đầu vào: là dữ liệu “khuyết” thuộc tính cần dự đoán lớp (nhãn). Mô hình sẽ tự độngi phân lớp (gán nhãn) cho các đối tượngi dữ liệu này dựa vào những gì được huấn luyện ở bước 1.



Hình 5: Quá trình phân lớp dữ liệu mới.

### 2.2.1. Tiền Xử Lý Dữ Liệu

Tiền xử lý dữ liệu là quá trình chuẩn bị dữ liệu để sử dụng cho các mục đích phân tích. Dữ liệu thô thường không phù hợp để phân tích trực tiếp, vì nó có thể chứa các lỗi, thiếu sót và bất thường. Tiền xử lý dữ liệu giúp loại bỏ các lỗi và bất thường này, đồng thời chuẩn hóa dữ liệu để phù hợp với các thuật toán phân tích.

Tiền xử lý dữ liệu bao gồm một số bước cơ bản sau:

* Loại bỏ các giá trị bị thiếu: Nếu một dữ liệu có giá trị bị thiếu, chúng có thể được thay thế bằng giá trị trung bình hoặc giá trị của một số dữ liệu tương tự.
* Sửa chữa các lỗi dữ liệu: Các lỗi dữ liệu có thể được sửa chữa bằng cách sử dụng các kỹ thuật thống kê hoặc các thuật toán học máy.
* Chuẩn hóa dữ liệu: Dữ liệu có thể được chuẩn hóa bằng cách chuyển đổi chúng sang một thang đo chung, chẳng hạn như thang đo từ 0 đến 1 hoặc thang đo từ -1 đến 1.
* Giảm thiểu dữ liệu: Dữ liệu có thể được giảm thiểu bằng cách sử dụng các kỹ thuật như lấy mẫu hoặc giảm kích thước.
* Tiền xử lý dữ liệu hình ảnh: Dữ liệu hình ảnh có thể được tiền xử lý bằng các kỹ thuật như loại bỏ nhiễu, tăng cường độ tương phản và chuyển đổi màu sắc.

### 2.2.2. Phân Lớp Dữ Liệu

Là quá trình phân loại một đối tượng dữ liệu thành một hoặc nhiều lớp (loại) nhất định bằng mô hình phân lớp. Mô hình này được xây dựng dựa trên một tập dữ liệu đã được gắn nhãn trước đó.

Kỹ thuật này dùng để trích xuất các thông tin cần thiết từ kho dữ liệu có sẵn. Do đó, đối với kỹ thuật này, chúng ta sẽ linh hoạt áp dụng các thuật toán khác nhau tùy thuộc vào mục đích sử dụng.

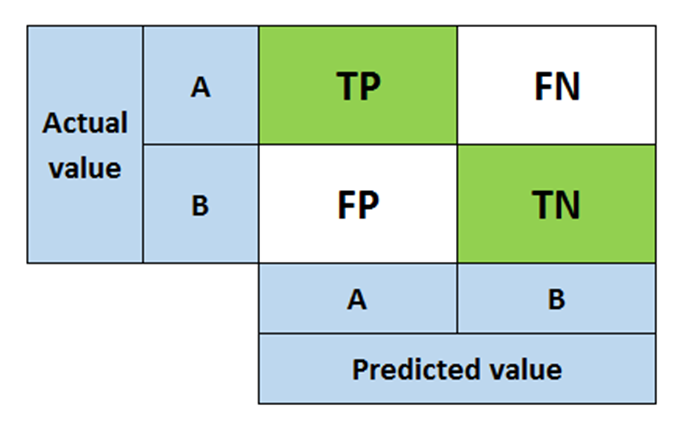
Bên cạnh đó, đây cũng được coi là một kỹ thuật quan trọng trong việc dự đoán các quy luật, xu hướng, ... bằng cách mô tả các thuộc tính liên quan đến các đối tượng được phân loại vào một lớp cụ thể.

### 2.2.3. Đánh Giá Tính Hiệu Quả

Đánh giá tính hiệu quả của mô hình phân lớp dữ liệu là một quá trình quan trọng để đảm bảo rằng mô hình đang hoạt động tốt và có thể được sử dụng để đưa ra các quyết định chính xác. Có nhiều cách khác nhau để đánh giá hiệu quả của mô hình phân lớp dữ liệu.

* Ma trận nhầm lẫn (Confusion Matrix)

 Ma trận nhầm lẫn là một bảng hiển thị số lượng dữ liệu được phân loại chính xác và số lượng dữ liệu được phân loại sai. Ma trận nhầm lẫn có thể được sử dụng để tính toán các chỉ số hiệu suất khác nhau, chẳng hạn như độ chính xác, độ nhạy và độ đặc hiệu.



Hình 6: Minh họa phương pháp ma trận nhầm lẫn.

Lấy ví dụ về bài toán phân nhóm công ty gian lận. Trong đó lớp A là nhóm công ty không gan lận, B là nhóm công ty gian lận. Các thuật ngữ chính của Confusion matrix trong trường hợp này như sau:

Giả sử lớp A là lớp công ty không gian lận và lớp B là lớp công ty gian lận. Các thuật ngữ chính của ma trận nhầm lẫn như sau:

* + TP (True Positive): Số lượng dự đoán chính xác về nhóm công ty không gian lận.
  + TN (True Negative): Số lượng dự đoán chính xác về nhóm công ty gian lận
  + FP (False Positive): Số lượng dự đoán sai lệch về công ty gian lận. Là khi mô hình dự đoán công ty không gian lận nhưng thực tế là công ty gian lận.
  + FN (False Negative): Số lượng dự đoán sai lệch về công ty không gian lận. Là khi mô hình dự đoán công ty gian lận nhưng thực tế công ty không gian lận.
* Độ chính xác (Actuary)

Độ chính xác là tỷ lệ giữa số lượng dữ liệu được phân loại chính xác và tổng số dữ liệu được phân loại. Độ chính xác là một chỉ số hiệu suất phổ biến, nhưng nó có thể bị ảnh hưởng bởi sự cân bằng của các lớp trong tập dữ liệu.

* Độ nhạy (Recall)

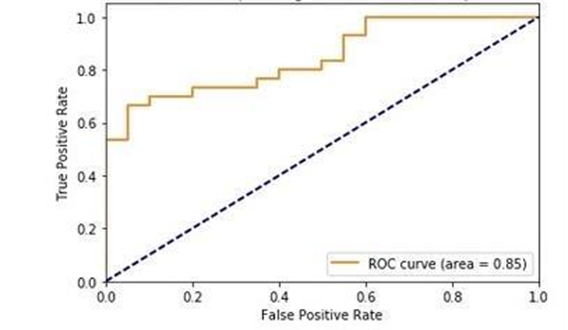
 Độ nhạy là tỷ lệ giữa số lượng dữ liệu thực sự thuộc lớp dương được phân loại chính xác và tổng số dữ liệu thực sự thuộc lớp dương. Độ nhạy là một chỉ số hiệu suất quan trọng khi cần phát hiện tất cả các dữ liệu thuộc lớp dương.

* Precision (độ chính xác) cho biết tỉ lệ số điểm TP (True positive) trong số những điểm được phân loại (dự đoán) là positive (TP + FP).
* F1- score là giá trị trung bình điều hòa của hai độ đo Precision và Recall.

=> F1 có giá trị gần với giá trị nhỏ hơn giữa 2 giá trị Precision và Recall và F1 càng lớn nếu cả 2 giá trị Precision và Recall đều lớn cho thấy độ tin cậy của mô hình cao hơn.

* ROC (Receiver Operating Characteristic)

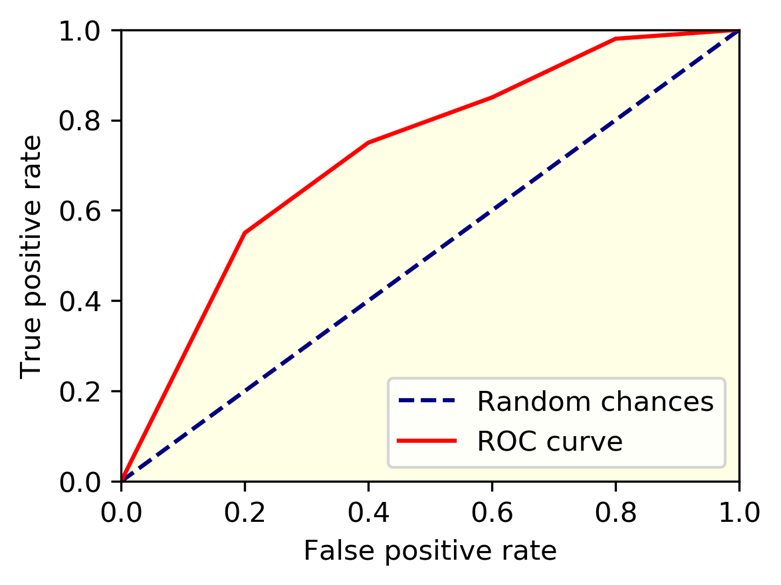
Là đồ thị được sử dụng phổ biến trong đánh giá các mô hình phân loại nhị phân. Đường cong này được sinh ra từ việc biểu diễn tỷ lệ dự báo true positive rate (TPR) dựa trên tỷ lệ dự báo false positive rate (FPR) tại các ngưỡng khác nhau. Một mô hình được gọi là hiệu quả khi có ROC càng tiệm cận với điểm (0;1) hay có TPR cao và FPR thấp thì mô hình càng phù hợp.



Hình 7: Minh họa phương pháp ROC

* AUC (Area Under the Curve)

Phần diện tích nằm dưới đường cong ROC và có giá trị là một số dương nhỏ hơn hoặc bằng 1. Giá trị này càng lớn thì độ chính xác của mô hình càng cao.



Hình 8: Minh họa phương pháp AUC

# CHƯƠNG 3. CÁC KẾT QUẢ THỰC NGHIỆM

## 3.1. Bộ dữ liệu

### Tiền xử lý dữ liệu:

Đồ án sử dụng bộ dữ liệu được lấy từ nguồn kaggle.com đã được xác minh nên không có dữ liệu thiếu, và bộ dữ liệu này đầy đủ thông tin để tiến hành làm đồ án nên không cần phải thực hiện tiền xử lí dữ liệu.

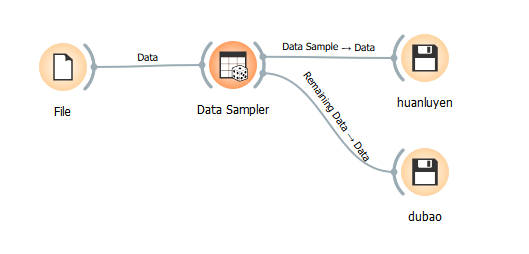
Nguồn dữ liệu: <https://www.kaggle.com/code/tejkirandm/audit-risk-prediction/notebook>

### Bảng mô tả dữ liệu

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Ý nghĩa | Mô tả |
| Sector\_score | Giá trị điểm rủi ro lịch sử của đơn vị mục tiêu sử dụng quy trình phân tích | Số thực |
| Para\_A | đại diện cho sự khác biệt giữa chi tiêu thực tế kiểm tra và chi tiêu dự kiến trong báo cáo tóm tắt A (tính bằng Rs). | Số thực |
| Score\_A | Điểm của A | Số thực |
| Para\_B | đại diện cho sự khác biệt giữa chi tiêu thực tế kiểm tra và chi tiêu dự kiến trong báo cáo tóm tắt B (tính bằng Rs). | Số thực |
| Score\_B | Điểm của B | Số thực |
| Total | Tổng số chênh lệch được tìm thấy trong các báo cáo khác là (tính bằng Rs). | Số thực |
| Numbers | Điểm chênh lệch lịch sử | Số thực |
| Money\_value | Gía trị tiền | Số thực |
| Loss\_score: | Điểm tổn thất | Số nguyên |
| History\_score | điểm lịch sử của tổ chức | Số nguyên |
| Score | Điểm qua quá trình thực hiện | Số nguyên |
| Risk | Rủi ro (0= gian lận, 1= không gian lận) | Số nguyên |

Trong đó: "Rs" là ký hiệu của "rupee", đồng tiền của Ấn Độ. Bộ dữ liệu không bí mật đầy đủ trong một năm từ 2015 đến 2016 của các công ty được thu thập từ Văn phòng Kiểm toán Ấn Độ để xây dựng công cụ dự đoán nhằm phân loại các công ty đáng ngờ. Bộ dữ liệu gồm 777 đối tượng, để phù hợp với mục đích nghiên cứu và cấu hình laptop, nhóm em xin được rút gọn bộ còn từ 18 cột thuộc tính còn 12 thuộc tính.

Nhóm phân tách dữ liệu từ dữ liệu gốc “trial\_risk”, sử dụng 70% dữ liệu ban đầu để làm dữ liệu huấn luyện mô hình phân lớp dữ liệu (huanluyen.tab). Và sử dụng 30% dữ liệu còn lại để làm dữ liệu dự báo cho nghiên cứu (dubao.tab). Trong đó Risk là biến phụ thuộc, các biến còn lại là biển độc lập.



Hình 9: Phân tách dữ liệu.

## 3.2. Các Kết Quả Thực Nghiệm

**Các bước thực hiện mô hình phân lớp dữ liệu**

Bước 1: Chọn dữ liệu File “trial.xlsx” và chọn cột “Risk” làm Target.

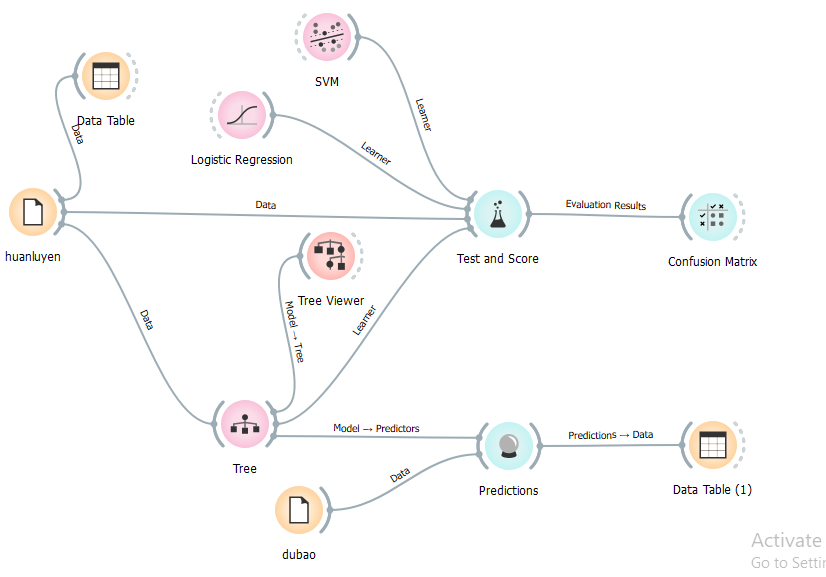
Bước 2: Dùng Data Sampler chia dữ liệu làm 2 phần theo tỉ lệ 70% đặt tên File là “huanluyen” và 30% với tên “Dubao”.

Bước 3: Mở File “Data” → Nối File và 3 phương pháp Logistic Regression, SVM và Decision Tree với Test and Score → Nối Test and Score với Confusion Matrix để thực hiện đánh giá kết quả và đánh giá ma trận nhầm lẫn.

Bước 4: Liên kết phương pháp tốt nhất và File “Dubao” với Predictions để đánh giá và phân loại dữ liệu đầu vào.

Bước 5: Xuất kết quả dự báo qua Data Table.

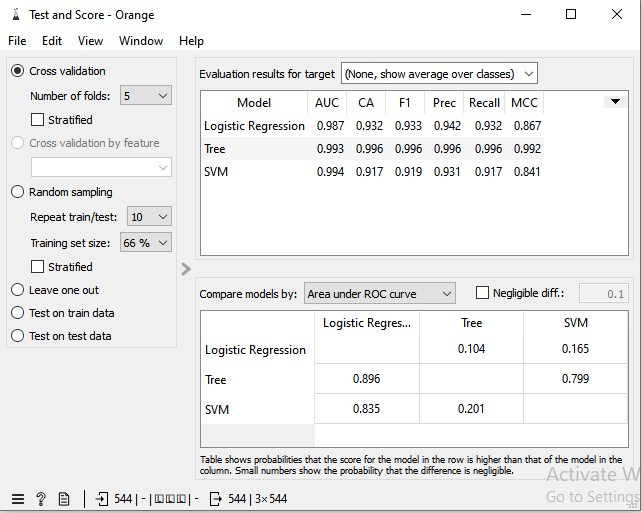
**­­­Mô hình phân lớp**



Hình 10. Mô hình phân lớp.

## 3.3. Phân Tích và Đánh Giá

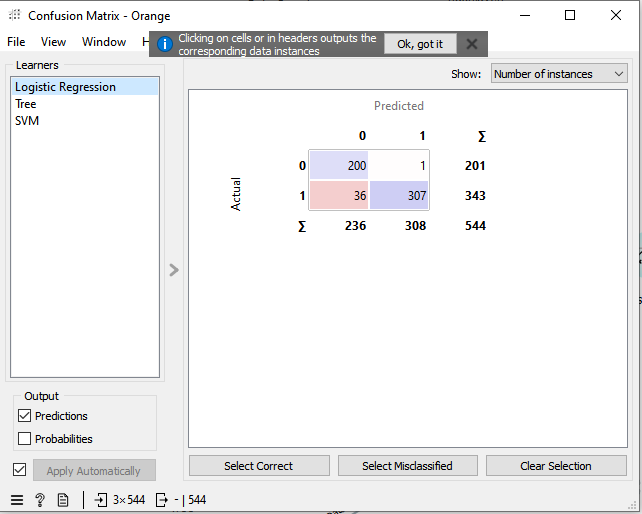
**3.3.1. Đánh giá mô hình dựa trên kết quả của Test and Score**



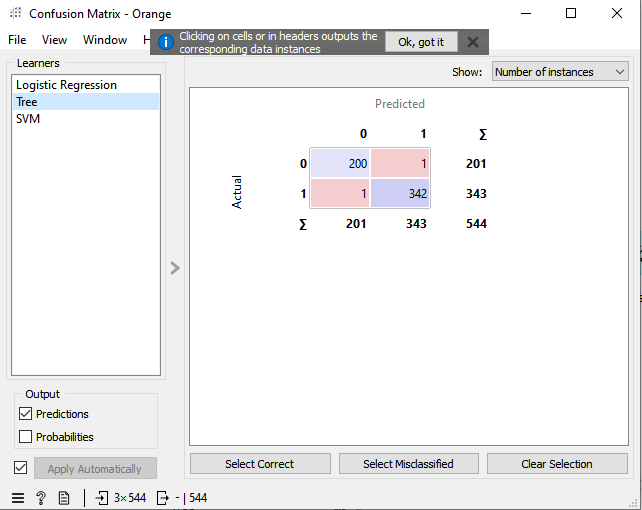
Hình 11: Kết quả Test & Score

Kết quả cho thấy rằng: AUC (Accuracy: tính chính xác) của SVM là lớn nhất (0.994) > Decision tree =0.993 >Logistic Regression (0.987). Tuy nhiên, các chỉ số như, Precision (Độ chính xác), Recall (Độ phủ) và F1 (Giá trị trung bình điều hòa giữa Precision và Recall) của Decision tree (0,996) lại lớn hơn hai phương pháp còn lại (trong khoảng từ 0.841 đến 0.933) Do đó chọn phương pháp Decision tree.

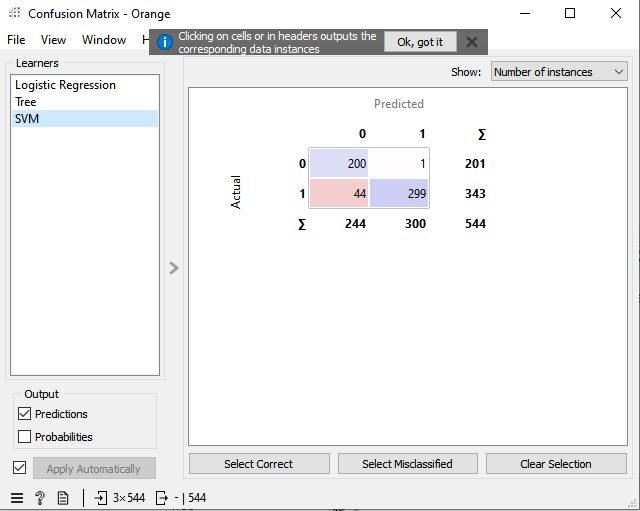
**2.2.3.2 Đánh giá mô hình dựa trên kết quả của Confusion Matrix:**



Hình 12: Kết quả Confusion Matrix của phương pháp Logistic Regression



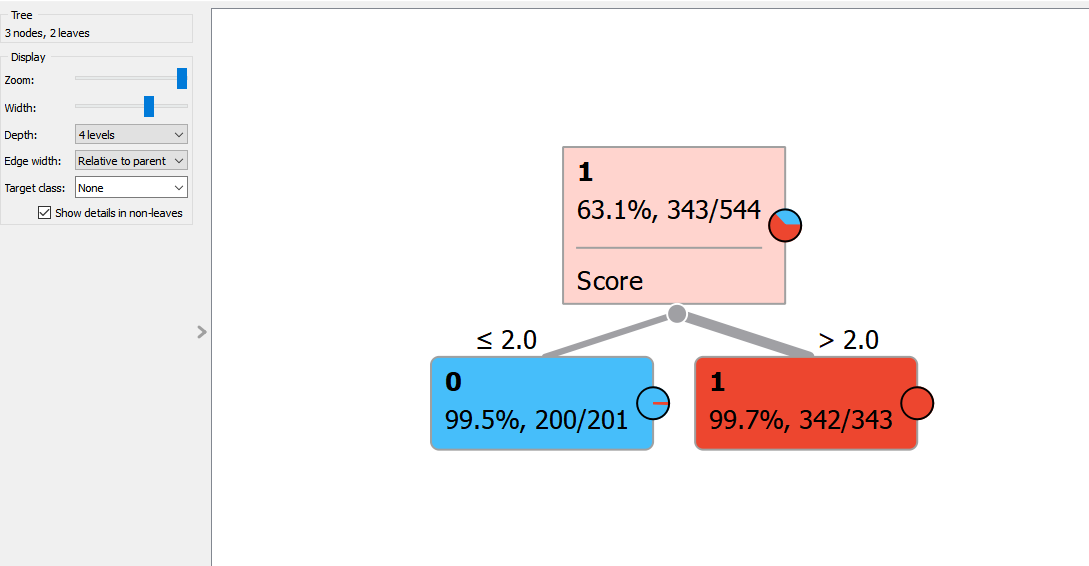
Hình 13: Kết quả Confusion Matrix của phương pháp Decision tree



Hình 14: Kết quả Confusion Matrix của phương pháp SVM

Sai lầm loại 2: Dự đoán là không gian lận nhưng thực tế là gian lận là sai lầm nghiêm trọng. Nếu sai lầm này càng nhiều thì việc dự đoán công ty nào gian lận sẽ không còn ý nghĩa. Decision tree = 2 < Logistic Regression =36 < SVM= 44. Nhận thấy sai lầm loại 2 của phương pháp Decision tree là nhỏ nhất nên nhóm sử dụng phương pháp Decision Tree để dự báo và không cần xét Test and Score.

**Mô hình Decision tree**



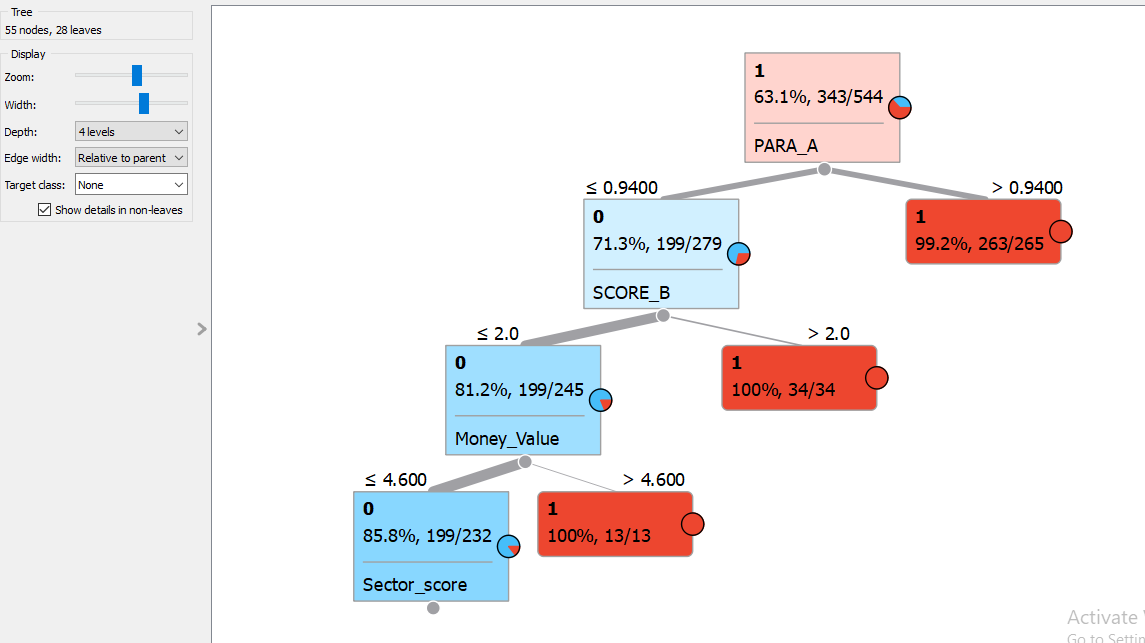
Hình 15.1: Kết quả phân lớp theo mô hình Decision Tree

Kết quả hình trên cho thấy

* Nếu Score bé hơn bằng 2 thì công ty được xếp vào nhóm không gian lận với độ chính xác 99.5%. Ngược lại, nếu điểm Score của công ty lớn hơn 2 thì công ty được xếp vào nhóm gian lận với độ chính xác 99.7%. Nhóm tiếp tục xử dụng các biểu đồ để xem xét mối quan hệ giữa các biến độc lập còn lại lên biến phụ thuộc Risk.

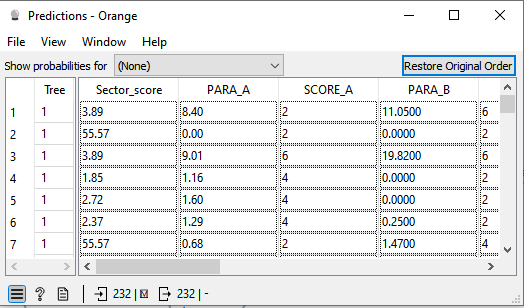
Tiếp tục sử dụng mô hình Decision tree quan sát mô hình phân lớp công ty gian lận và không gian lận của các biến phụ thuộc khác lên biến độc lập Risk bằng cách skip biến Score, nhóm em nhận thấy:

* Nếu Para\_A lớn hơn 0.94 thì công ty được xếp loại gian lận với độ chính xác 99.2%
* Nếu Para\_A bé hơn bằng 0.94
  + Nếu Score\_B bé hơn bằng 2:
    - Money\_value, bé hơn bằng 4.6 thì công ty được xết loại không gian lận với độ chính xác 85.8%
    - Money\_value, lớn hơn 4.6 thì công ty được xết loại gian lận với độ chính xác tuyệt đối.
  + Nếu Score\_B lớn hơn 2, công ty được xếp loại là gian lận với độ chính xác tuyệt đối.



Hình 15.2: Kết quả phân lớp theo mô hình Decision tree (skip Score)

## 3.3. Kết quả dự báo

****

Hình 16: Kết quả dự báo của Predeictions

## 3.4 Kiến thức chuyên ngành

Trong báo cáo tài chính, score là một chỉ số được sử dụng để đánh giá khả năng gian lận báo cáo tài chính của một doanh nghiệp. Chỉ số này được tính toán dựa trên 8 biến số tài chính, bao gồm:

* Tỷ lệ lợi nhuận trên doanh thu (ROS)
* Tỷ lệ lợi nhuận trên vốn chủ sở hữu (ROE)
* Tỷ số thanh toán hiện hành (Current ratio)
* Tỷ số nợ trên vốn chủ sở hữu (Debt to equity ratio)
* Tỷ số vòng quay tổng tài sản (Total asset turnover ratio)
* Tỷ lệ tăng trưởng doanh thu (Sales growth rate)
* Tỷ lệ tăng trưởng lợi nhuận (Profit growth rate)
* Tỷ lệ vòng quay hàng tồn kho (Inventory turnover ratio)

Ý nghĩa của biến Para trong các bảng cáo cáo:

* Trong bảng cân đối kế toán, chỉ số Para có thể được sử dụng để thể hiện Tỷ số nợ trên vốn chủ sở hữu (Debt to equity ratio) hoặc Tỷ số vòng quay tổng tài sản (Total asset turnover ratio)
* Trong bảng báo cáo kết quả hoạt động kinh doanh, chỉ số Para có thể được sử dụng để thể hiện tỷ lệ lợi nhuận trên doanh thu (ROS) hoặc tỷ lệ lợi nhuận trên vốn chủ sở hữu (ROE).

Para có thể tác động đến score theo hướng tích cực hoặc tiêu cực. Ví dụ, tỷ lệ lợi nhuận trên doanh thu (ROS) là một chỉ số para, được tính bằng cách chia lợi nhuận sau thuế cho doanh thu thuần. Nếu ROS tăng, thì score của doanh nghiệp cũng sẽ tăng theo. Ngược lại, nếu ROS giảm, thì score của doanh nghiệp cũng sẽ giảm theo.

Ý nghĩa của biến Loss\_score và Money\_value trong các bảng cáo cáo:

* Loss có thể tác động đến score theo hướng tiêu cực. Ví dụ, lỗ thuần là một chỉ số loss, được tính bằng cách trừ lợi nhuận sau thuế cho chi phí. Nếu lỗ thuần tăng, thì score của doanh nghiệp cũng sẽ giảm theo.
* Money\_value có thể tác động đến score theo hướng tích cực hoặc tiêu cực. Ví dụ, doanh thu thuần là một chỉ số money\_value, được tính bằng tổng giá trị của tất cả các sản phẩm hoặc dịch vụ đã được bán trong một kỳ. Nếu doanh thu thuần tăng, thì score của doanh nghiệp cũng sẽ tăng theo. Ngược lại, nếu doanh thu thuần giảm, thì score của doanh nghiệp cũng sẽ giảm theo.

# CHƯƠNG 4. KẾT LUẬN

## 4.1. Các Kết Quả Đạt Được

* Dựa vào đánh giá từ 2 phương pháp đánh giá Test & Score và Confusion Matrix, phương pháp tốt nhất chính là Decision tree với tính chính xác chính xác cao và các chỉ số rất cao Precision, Recall và F1 (bằng 0.996 và 0.992), so với 2 phương pháp còn lại, và Sai lầm loại 2 chỉ có (1) nhỏ nhất. Với kết quả dự báo có tính chính xác tối đa và mức độ sai lầm tối thiểu, do vậy, nhóm xin đề xuất sử dụng phương pháp Decision tree. Ngày nay, các báo cáo tài chính ngày càng trở nên phức tạp hơn, khiến chúng dễ bị gian lận hơn, cùng với việc phát triển của công nghệ, các báo cáo có thể gian lận một cách tinh vi hơn.
* Đồ án giúp phân tích các đặc điểm của các công ty gian lận và không gian lận và dự báo khả năng gian lận trong báo cáo tài chính của công ty. Khi kiểm toán viên tiến hành kiểm tra, họ sẽ sử dụng mô hình này để hỗ trợ cho công việc kiểm kê của mình, hạn chế tối đa rủi ro trong quá trình làm việc dựa vào các công thức và mô hình đã được xây dựng sẵn trong bài nghiên cứu.

## 4.2. Những Hạn Chế và Hướng Phát Triển

### 4.2.1 Hạn chế

* Nhóm em sử dụng bộ dữ liệu trong quá khứ (2015-2016) nên kết quả nghiên cứu sẽ có độ chính xác chưa sát với tình hình thực tế.
* Nhóm em gặp khó khăn trong việc chọn đề tài mang tính chuyên ngành, vì vậy phạm vi của bài nghiên cứu bị hạn chế trong hiểu biết của nhóm.

### 4.2.2 Hướng phát triển

* Các doanh nghiệp nên đầu tư, cập nhật trang thiết bị công nghệ tiên tiến, những phần mềm mới nhất để kiểm toán viên có thể vận dụng công nghệ vào kiểm toán. Thực hiện đào tạo, có những chính sách khuyến khích kiểm toán viên áp dụng công nghệ thông tin nhằm giúp kiểm toán viên có thể hiểu được rõ hơn về dữ liệu khách hàng, dữ liệu trong quá khứ và có thể cải thiện khả năng tìm kiếm thông tin chuyên sâu để tìm hiểu được các thông tin kiểm toán hữu ích. Hoặc khuyến khích các kiểm toán viên sử dụng công nghệ thông tin để có thể phân tích tính đúng đắn và độ tin cậy của quá trình xử lý dữ liệu kiểm toán, đơn giản hóa quá trình kiểm toán và cải thiện hiệu quả kiểm toán. Tuy việc đầu tư vào phần mềm kiểm toán hay công nghệ để kịp xu thế khiến doanh nghiệp bỏ ra với các khoản chi phí không nhỏ. Nhưng đổi lại, hệ thống sẽ được bảo mật hơn, phần mềm sẽ cập nhật và ứng dụng cao hơn nhằm phục vụ công việc kiểm toán tốt hơn nữa trong tương lai.
* Kiểm toán viên cũng phải tự nâng cao trình độ, cập nhật các kiến thức cần thiết đặc biệt là về công nghệ thông tin nhằm phục vụ công việc kiểm toán. Bởi việc sử dụng công nghệ sẽ làm giảm thời gian công việc của kiểm toán viên cụ thể: công nghệ thông tin làm giảm thời gian lập báo cáo kiểm toán, giảm thời gian chọn mẫu. Một số phần mềm kiểm toán còn giúp kiểm toán viên giảm thời gian thực hiện các thử nghiệm kiểm soát và thử nghiệm cơ bản. Giúp kiểm toán viên có thể chọn mẫu nhanh và chính xác hơn, giảm thời gian tạo ra giấy tờ làm việc của kiểm toán viên.

# TÀI LIỆU THAM KHẢO

1. Batóg, B., Wawrzyniak, K. (2022). *Comparison of Influence of Various Proposals of Transforming Nominants into Stimulants on Linear Ordering and Grouping of Listed Companies*. In: Jajuga, K., Dehnel, G., Walesiak, M. (eds) **Modern Classification and Data Analysis**. Springer, Cham.

2. Speller, J., Staerk, C., Gude, F. et al. (2023). *Robust gradient boosting for generalized additive models for location, scale and shape*. **Adv Data Anal Classif**.

3. Zeyuan Wang (2023). Based on the Big Data Audit Risk Model Audit Process Research. Frontiers in Business, Economics and Management.

4. Trần Mạnh Dũng (2023). ỨNG DỤNG CÔNG NGHỆ THÔNG TIN TRONG KIỂM TOÁN NỘI BỘ TẠI CÁC DOANH NGHIỆP CỔ PHẦN TRÊN ĐỊA BÀN THÀNH PHỐ HÀ NỘI. Tạp chí Kinh tế & Phát triển số 308 tháng 02/2023