# Đồ án Khoa học Dữ Liệu - Thầy Nguyễn Mạnh Tuấn

Data Science (Trường Đại học Kinh tế Thành phố Hồ Chí Minh)

**ĐẠI HỌC UEH**
**TRƯỜNG KINH DOANH**
**KHOA KẾ TOÁN**

# DỰ ÁN KẾT THÚC HỌC PHẦN

## Môn: KHOA HỌC DỮ LIỆU

Giảng viên hướng dẫn: Thầy Nguyễn Mạnh Tuấn
Mã lớp học phần: 22C1INF50905974
Nhóm nghiên cứu: 4

Tp. Hồ Chí Minh, ngày 18 tháng 12 năm 2022

**BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

| Danh sách thành viên | MSSV | Mức độ tham gia |
|---|---|---|
| Dương Kim Ngân | 31211022552 | 100% |
| Đào Tuyết Nhung | 31211023057 | 100% |
| Đoàn Thị Ngọc Trâm | 31211022170 | 100% |
| Nguyễn Đỗ Thảo My | 31211022124 | 100% |
| Phạm Tô Minh Vỹ | 31211022666 | 100% |

*Đối với đồ án nhóm chọn thực hiện bài toán liên quan đúng chuyên ngành mà nhóm đang theo học - Kiểm toán. Lý do là vì việc gian lận trên báo cáo tài chính hiện nay xảy ra ngày càng nhiều. Gian lận trên Báo cáo tài chính là những sai sót mang tính trọng yếu mà các Kiểm toán viên phải thực sự cẩn trọng trong việc thực thi kiểm kê và đánh giá Báo cáo tài chính. Việc đánh giá rủi ro gian lận tốt, thì sẽ tạo điều kiện cho các Kiểm toán viên nhìn nhận những sai sót một cách sáng suốt với những chiêu trò của các doanh nghiệp. Hơn thế nữa, với những đánh giá chính xác, sẽ là những minh chứng để Kiểm toán viên đối chứng với doanh nghiệp khi có những ý kiến khi từ chối các báo cáo tài chính không minh bạc. Và là một sinh viên chuyên ngành Kiểm toán của trường Đại học UEH, đây là cách để nhóm có thể nâng cao tính chuyên nghiệp*

# DANH MỤC BẢNG BIỂU, HÌNH ẢNH MINH HỌA

## BẢNG BIỂU



## HÌNH ẢNH

# MỤC LỤC

# CHƯƠNG 1: KHÁI QUÁT ĐỒ ÁN

## 1. Khái quát về Kiểm toán và Doanh nghiệp

### 1.1. Khái niệm Báo cáo Tài chính

Theo chuẩn mực kế toán quốc tế (IAS 01) của Ủy ban chuẩn mực kế toán quốc tế (IASB), báo cáo tài chính là các thông tin kinh tế được kế toán viên trình bày dưới dạng bảng biểu, cung cấp thông tin về tình hình tài chính, tình hình kinh doanh và các luồng tiền của doanh nghiệp đáp ứng nhu cầu cho người sử dụng thông tin trong việc ra quyết định kinh tế.

Hệ thống báo cáo tài chính doanh nghiệp hiện nay bao gồm bốn báo cáo chính: bảng cân đối kế toán, báo cáo kết quả kinh doanh, báo cáo lưu chuyển tiền tệ và thuyết minh báo cáo tài chính. Mỗi báo cáo tài chính cung cấp thông tin về các khía cạnh khác nhau trong tình hình tài chính của doanh nghiệp nhằm giúp các đối tượng sử dụng thông tin có thể đánh giá toàn diện về tình hình tài chính đó, từ đó đưa ra quyết đinh hợp lý. Cụ thể:

+ Bảng cân đối kế toán: cung cấp thông tin về tình trạng tài chính (giá trị tài sản, nợ phải trả và vốn chủ sở hữu) của doanh nghiệp tại một thời điểm.

+ Báo cáo kết quả kinh doanh: cung cấp thông tin về doanh thu, chi phí và kết quả lãi hoặc lỗ trong một kỳ kế toán của doanh nghiệp.

+ Báo cáo lưu chuyển tiền tệ cung cấp thông tin về các khoản tiền thu, chi trong một kỳ hoạt động của doanh nghiệp.

+ Thuyết minh báo cáo tài chính là bản giải trình giúp các đối tượng sử dụng thông tin hiểu rõ hơn về các con số trên bảng cân đối kế toán, báo cáo kết quả kinh doanh, báo cáo lưu chuyển tiền tệ. Thuyết minh báo cáo tài chính 10 thường bao gồm bốn nội dung cơ bản: chính sách kế toán áp dụng tại doanh nghiệp; các thông tin bổ sung cho các khoản mục trên báo cáo tài chính; biến động vốn chủ sở hữu và các thông tin khác.

Các báo cáo tài chính trong doanh nghiệp có mối quan hệ mật thiết với nhau, mỗi sự thay đổi của một chỉ tiêu trong báo cáo này trực tiếp hay gián tiếp ảnh hưởng đến các báo cáo kia. Qua đó, người sử dụng thông tin nhận biết được và tập trung vào các chỉ tiêu tài chính liên quan trực tiếp tới mục tiêu phân tích của họ.

## 1.2. Nghiệp vụ Kiểm toán

Thuật ngữ kiểm toán thường đề cập đến cuộc kiểm tra, đánh giá, kết luận và xác nhận tính đầy đủ, hợp pháp của số liệu, tài liệu kế toán, báo cáo tài chính của các cơ quan, tổ chức. Hay nói một cách dễ hiểu, kiểm toán là hoạt động kiểm tra lại các thông tin tài chính được cung cấp bởi kế toán nhằm xác định và đối chiếu mức độ phù hợp giữa thông tin đó với các chuẩn mực đã được thiết lập. Kiểm toán thường dành cho các đối tượng có niềm đam mê đến tình hình tài chính của một tổ chức nào đó nhưng lại không có nghiệp vụ về tài chính, kế toán. Vì thế, các doanh nghiệp thường cần đến những người kiểm toán viên để đưa ra những đánh giá đúng đắn cho doanh nghiệp của họ.

Kiểm toán ngày càng khẳng định được vai trò vô cùng quan trọng trong nền kinh tế mới hiện nay. Nó thể hiện được chuẩn mực của kế toán trong các hoạt động kinh tế, quản trị kinh doanh do kế toán cung cấp. Bên cạnh đó, kiểm toán viên cũng giữ một vai trò quan trọng không kém đối với cơ quan nhà nước, doanh nghiệp hay tổ chức. Kiểm toán viên giúp kiểm soát ngân quỹ nhà nước và sự vận động của toàn bộ ngân quỹ và tài sản quốc gia. Không những thế, nó còn giúp cơ quan nhà nước đưa ra được những chính sách hiệu quả dựa trên kết quả thu nhận được. Ngoài ra, kiểm toán viên còn giúp cho doanh nghiệp kiểm tra, đánh giá các thông tin tài chính - kế toán và đưa ra những quyết định kinh doanh kịp thời.

Nếu phát hiện có hành vi không tuân thủ pháp luật và các quy định, kiểm toán viên phải báo cáo những hành vi này với các cơ quan nhà nước có thẩm quyền, thông

báo với đại diện chủ sở hữu của đơn vị được kiểm toán và các đối tượng bên ngoài đơn vị được kiểm toán. Nếu nghi ngờ có hành vi không tuân thủ pháp luật và quy định, kiểm toán viên phải thực hiện các thủ tục kiểm toán bổ sung để làm rõ những nghi ngờ này. Đây là một trong những trách nhiệm của kiểm toán viên đối với gian lận và sai sót.

Những hình thức gian lận trên Báo cáo tài chính :

+ Che dấu công nợ và chi phí: Che dấu công nợ đưa đến giảm chi phí là một trong những kỹ thuật gian lận phổ biến tên Báo cáo tài chính nhằm mục đích khai khống lợi nhuận. Khi đó, lợi nhuận trước thuế sẽ tăng tương ứng với số chi phí hay công nợ bị che dấu. Đây là phương pháp dễ thực hiện và khó bị phát hiện vì thường không để lại dấu vết. Có ba phương pháp chính thực hiện giấu gian lận và chi phí như: Không ghi nhận công nợ và chi phí, đặc biệt không lập đầy đủ các khoản dự phòng; vốn hoá chi phí; không ghi nhận hàng bán trả lại – các khoản giản trừ và không trích trước chi phí bảo hành.

+ Ghi nhận doanh thu không có thật hay khai cao doanh thu: Là việc ghi nhận sổ sách một nghiệp vụ bán hàng hay cung cấp dịch vụ không có thực. Kỹ thuật thường sử dụng là tạo ra các khách hàng giả mạo thông qua lập chứng từ giả mạo danh nhưng hàng hoá không được giao và đầu niên độ sau sẽ lập bút toán hàng bán bị trả lại. Khai cao doanh thu còn được thực hiện thông qua việc cố ý ghi tăng các nhân tố trên hoá đơn như số lượng, giá bán… hoặc ghi nhận doanh thu khi các điều kiện giao hàng chưa hoàn tất, chưa chuyển quyền sở hữu và chuyển rủi ro đối với hàng hoá, dịch vụ được bán

+ Định giá sai tài sản: Việc định giá sai tài sản được thực hiện thông qua việc không ghi giảm giá trị hàng tồn kho khi hàng đã hư hỏng, không còn sử dụng hay không lập đầy đủ dự phòng giảm giá hàng tồn kho, nợ phải thu khó đòi, các khoản đầu tư ngắn, dài hạn. Các tài sản thường bị định giá sai như tài sản mua qua hợp nhất kinh doanh, tài sản cố định, không vốn hoá đầy đủ các chi phí mô hình, phân loại không đúng tài sản.

+ Ghi nhận sai niên độ: Doanh thu hay chi phí được ghi nhận không đúng với thời kỳ mà nó phát sinh. Doanh thu hoặc chi phí của kỳ này có thể chuyển sang kỳ kế tiếp hay ngược lại để làm tăng hoặc giảm thu nhập theo mong muốn.

+ Không khai báo đầy đủ thông tin: Việc không khai báo đầy đủ thông tin nhằm hạn chế khả năng phân tích của người sử dụng Báo cáo tài chính. Các thông tin thường không được khai báo đầy đủ trong thuyết minh như: Nợ tiềm tàng, các sự kiện phát sinh sau ngày khoá sổ kế toán, thông tin về bên có liên quan, các những thay đổi về chính sách kế toán…

Những dấu hiệu nhận biết gian lận trên Báo cáo tài chính:

+ Doanh nghiệp có cơ cấu sở hữu phức tạp, sở hữu chéo, thành lập nhiều công ty con

+ Lợi nhuận vượt trội trong lĩnh vực hoạt động thông thường, không đặc sắc

+ Dòng tiền từ hoạt động kinh doanh liên tục âm, được bù đắp từ dòng tiền từ hoạt động tài chính (tăng vốn)

+ Lợi nhuận cao bất thường trước các đợt tăng vốn, lợi nhuận không đến từ hoạt động kinh doanh chính mà đến từ việc bán tài sản, thông qua các hợp đồng hợp tác kinh doanh, hợp đồng uỷ thác, giao dịch với các bên liên quan

+ Hay thay đổi người đại diện theo pháp luật, Kế toán trưởng

### 1.3. Thực trạng gian lận của các doanh nghiệp hiện nay

Hiện nay, việc gian lận Báo cáo tài chính trên thế giới nói chung và Việt Nam nói riêng ngày càng gia tăng và dần trở thành vấn đề đáng quan tâm đối với các doanh nghiệp, chính phủ và các nhà đầu tư. Và trong bối cảnh ngày nay, quá trình sử dụng thông tin đăng tải trên báo cáo tài chính đã thể hiện vai trò quan trọng trọng quản lý và đầu tư. Ở Việt Nam, hiện tượng chênh lệch giữa các báo cáo tài chính trước và sau kiểm toán đã tạo nên tâm lý nghi ngại. Đặc biệt những gian lận báo cáo tài chính gần đây của công ty Cổ phần tập đoàn FLC và công ty Cổ phần NTACO và một loạt các công ty niêm yết khác trên sàn chứng khoán đã bị phát hiện đã gây ra tâm lý nghi ngại cho nhà đầu tư, ảnh hưởng rất nhiều đến hoạt động của thị trường vốn . Các doanh nghiệp ngày càng sử dụng nhiều thủ thuật gian lận báo cáo tài chính tinh vi, như là tăng vốn ảo thông qua sử dụng các công ty con còn gọi tắt là SPE, điều chỉnh doanh thu, lợi nhuận thông qua SPE, thực hiện các giao dịch khống để rút tiền vay ngân hàng thông qua SPE và còn rất nhiều hình thức gian lận khác. Đây là một thực trạng đáng quan

ngại mà chúng ta cần ngăn chặn ngay từ bây giờ.

### 1.4. Yêu cầu của ngành

- Giỏi tính toán, yêu thích những con số: Kiểm toán là một trong các ngành liên quan mật thiết đến toán học, với việc kiểm tra, rà soát thông tin tính toán trên Báo cáo tài chính

- Kỹ năng diễn đạt ngắn gọn và thuyết phục cao: Hoạt động hiệu quả của kiểm toán phụ thuộc rất nhiều vào sự tin tưởng của đối tượng sử dụng dịch vụ. Chính vì vậy không phải người nào cũng có thể lắng nghe và dễ dàng đồng ý với những nhận định mà kiểm toán viên đưa ra, ngay cả khi đã có những bằng chứng cụ thể và xác thực. Do vậy, để thành công trong ngành kiểm toán thì cần phải có khả năng diễn giải và thuyết phục cao. Bên cạnh kiến thức chuyên môn thì kỹ năng này sẽ giúp kiểm toán viên dễ dàng hơn trong việc thuyết phục người nghe tiếp nhận vấn đề một cách chính xác và nhanh chóng.

- Biết quản lý thời gian và chịu được áp lực công việc: Có thể nói kiểm toán là một công việc đòi hỏi khả năng chịu áp lực công việc cao cho nên người làm công việc này phải là người có sức khỏe và tinh thần tốt. Song song đó, phải biết cách sắp xếp thời gian hợp lý để có thể hoàn thành công việc đúng tiến độ đã được đề ra.

- Phải có tư duy phân tích cao và óc quan sát: Đặc thù của nghề kiểm toán đòi hỏi chúng ta phải có tư duy phân tích rất cao. Đồng thời, đặc điểm khác biệt lớn nhất giữa nghề kế toán và kiểm toán chính là kỹ năng nhận diện vấn đề nhanh chóng. Ngoài ra, áp lực hoàn thành công việc trong giới hạn thời gian nhất định cũng đòi hỏi kiểm toán viên phải nắm bắt vấn đề thật nhanh để tìm ra những điểm sai lệch trong các bản báo cáo tài chính.

- Tính độc lập, khách quan: Độc lập về tư tưởng và độc lập về hình thức là những yêu cầu cần thiết để kiểm toán viên hành nghề đưa ra kết luận hoặc được coi là đưa ra kết luận một cách không thiên vị, không mâu thuẫn về lợi ích hoặc không bị ảnh hưởng một cách bất hợp lý từ người khác.

## 2. Lý do chọn đề tài

Với điều kiện nền kinh tế thị trường phát triển như hiện nay, báo cáo tài chính của các doanh nghiệp là một trong những yếu tố mang đến thông tin rất hữu ích và quan trọng cho các nhà đầu tư, cơ quan có thẩm quyền như: cục thuế,.. chủ nợ. Bên cạnh đó, báo cáo

tài chính còn hỗ trợ ra quyết định đối với những nhà tài trợ về nguồn vốn cho doanh nghiệp như các chủ nợ hoặc các nhà đầu tư tiềm năng, hỗ trợ các nhà quản lý công ty trong việc đưa ra những phương án quan trọng cho công ty. Và thậm chí, đối với một doanh nghiệp có báo cáo tài chính "đẹp" sẽ để lại ấn tượng tốt và tạo ra cho công ty một lợi thế cạnh tranh nhất định. Tóm lại, chính vì tầm quan trọng của báo cáo tài chính, nó thật sự cần thiết được trình bày một cách chính xác các giao dịch kinh tế, trình bay trung thực, phản ánh đầy đủ và không bao gồm những sai sót mang tính trọng yếu. Thế nhưng, tình trạng gian lận trên báo cáo tài chính hiện nay trở nên đáng lo ngại, khi mà ngày càng nhiều doanh nghiệp dùng chiêu trò này để gây tổn hại tới tất cả những chủ thể sử dụng báo cáo tài chính bằng cách trình bày các thông tin sai lệch mang tính trọng yếu. Theo VAS 33: " Trọng yếu là thuật ngữ dùng để thể hiện tầm quan trọng của một thông tin (một số liệu kế toán) trong báo cáo tài chính. Thông tin được coi là trọng yếu có nghĩa là nếu thiếu thông tin đó hay thiếu chính xác của thông tin đó sẽ ảnh hưởng đến các quyết định của người sử dụng báo cáo tài chính. Mức độ trọng yếu tùy thuộc vào tầm quan trọng của thông tin và tính chất của thông tin hay của sai sót được đánh giá trong hoàn cảnh cụ thể. Mức trọng yếu là một ngưỡng, một điểm chia cắt, chứ không phải là nội dung của thông tin cần phải có".

Xét đến những tác hại mà việc gian lận gây ra đối với những chủ nợ, nhà đầu tư hoặc các cơ quan quản lý là vô cùng nghiêm trọng. Với các chủ nợ (hay những người đi vay), trong nhiều trường hợp, doanh nghiệp chọn gian lận báo cáo tài chính nhằm che đi đi tình trạng mất khả năng thanh toán nợ. Điều này, công ty sẽ trục lợi lừa thêm tiền những người cho vay, và làm tăng các khoản nợ xấu đối với ngân hàng, các tổ chức tín dụng và chủ nợ. Với các nhà đầu tư, đại đa số khi họ đưa ra một quyết định hoặc một đánh giá một công ty, sẽ dẫn đến tình trạng đánh giá quá cao so với giá trị doanh nghiệp. Như vậy, các nhà đầu tư có thể gây ảnh hưởng đến vốn và khả năng sinh lợi của họ. Với các cơ quan thuế, việc gian lận trên báo cáo tài chính sẽ làm doanh nghiệp trốn được một khoản thuế phải nộp chẳng hạn: thuế thu nhập doanh nghiệp, thuế giá trị gia tăng,.. bằng việc giảm đi doanh thu và tăng chi phí. Khi đó, việc mất đi một khoản thuế làm giảm đi nguồn ngân sách của nhà nước.

Nhìn chung, gian lận trên báo cáo tài chính luôn là những việc đáng bị lên án vì nó mang đến những ảnh hưởng rất nghiêm trọng đối với rất nhiều người và thậm chí là ảnh hưởng liên quan đến tình hình kinh tế đất nước. Vì vậy, Kiểm toán viên cần phải tìm ra những sai sót trọng yếu của báo cáo tài chính thông qua thiết lập những bài toán nhằm dự báo và phát hiện gian lận.

## 3. Mục tiêu đồ án

Mục tiêu của đồ án này liên quan đến ba bài toán:

+ Bài toán 1: Dự báo khả năng doanh nghiệp lợi dụng các khoản thanh toán của khách hàng nhằm che giấu công nợ để gian lận trên Báo cáo tài chính

+ Bài tóa 2: Phát hiện công ty gian lận trong các nhóm công ty có cùng tính chất

+ Bài toán 3: Dự báo khả năng gian lận trên BCTC

## 4. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đồ án là dự trên các mặt sai sót mang tính trọng yếu (gian lận) trong báo cáo tình của của các doanh nghiệp có cùng tính chất.

## 5. Mô tả dữ liệu và cấu trúc dữ liệu

Dữ liệu trong đồ án được tổng hợp từ nguồn sau:

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

https://archive.ics.uci.edu/ml/datasets/Audit+Data

Đây là những nguồn được đánh giá là có những bộ dữ liệu đáng tin cậy để phục vụ cho việc nghiên cứu vấn đề rủi ro gian lận trên Báo cáo tài chính của các doanh nghiệp, và sự ảnh hưởng lớn của các khoản phải thu (khoản khách hàng vay) đến việc gian lận của các công ty

- Default of credit card clients dataset

| Thuộc tính | Ý nghĩa | Mô tả |
|---|---|---|
| ID | Mã số Khách hàng | Số nguyên |
| LIMIT_BAL | Số dư tín dụng dùng | Số thực |
| SEX | Giới tính<br>1: nam<br>2: nữ | Nam hoặc nữ |

| | | |
|---|---|---|
| EDUCATION | Học vấn được hiện thị theo 6 cách sau:<br>1 : đã tốt nghiệp<br>2: đại học<br>3: cấp ba<br>4: khác<br>5,6 : không rõ | Số nguyên |
| MARRIAGE | Có 3 tình trạng hôn nhân của tệp khách hàng này.<br>1: kết hôn<br>2: độc thân<br>3: khác | Số nguyên |
| AGE | Độ tuổi của khách hàng | Số nguyên |
| PAY_X | Tình trạng trả nợ trong tháng X | Số thực |
| BILL_AMTX | Sao kê hóa đơn | Số thực |
| PAY_AMTX | Số tiền thanh toán trước trong tháng X | Số thực |
| Default payment of next month | Khả năng thanh toán tháng tới với lần lượt theo trạng thái:<br>1: Trả được c tiền<br>0: Không trả tiền | Có hoặc không |

*Bảng 1.1 Bảng mô tả cấu trúc của bộ dữ liệu về Default credit clients payment*

- Audit Risk

| Thuộc tính | Ý nghĩa | Mô tả |
|---|---|---|
| Sector_score | Giá trị điểm rủi ro lịch sử của đơn vị mục tiêu bằng thủ tục phân tích | Số không nguyên |
| LOCATION_ID | ID duy nhất của công ty | Số thực |
| PARA_A | Sự khác biệt được tìm thấy trong kế hoạch kiểm tra và báo cáo tóm tắt của A | Số không nguyên |
| Score_A | Điểm của A | Số không |

| | | |
|---|---|---|
| | | nguyên |
| Risk_A | Rủi ro của A | Số không nguyên |
| PARA_B | Sự khác biệt được tìm thấy trong kế hoạch kiểm tra và báo cáo tóm tắt của B | Số không nguyên |
| Score_B | Điểm của B | Số không nguyên |
| Risk_B | Rủi ro của B | Số không nguyên |
| TOTAL | Tổng số lượng chênh lệch được tìm thấy trong các báo cáo khác | Số không nguyên |
| numbers | Những khác biệt trong lịch sử | Số thực |
| Score_B | Điểm của B | Số không nguyên |
| Risk_C | Rủi ro của C | Số không nguyên |
| Money_Value | Gía trị tiền | Số không nguyên |
| Score_MV | Điểm của MV | Số không nguyên |
| Risk_D | Rủi ro của D | Số không nguyên |
| District_Loss | Dữ liệu của quận bị mất | Số thực |
| PROB | Vấn đề | Số không |

| | | |
|---|---|---|
| | | nguyên |
| RiSk_E | Rủi ro của E | Số không nguyên |
| History | Lịch sử | Có hoặc không |
| Prob | Vấn đề | Số không nguyên |
| Risk_F | Rủi ro của F | Số không nguyên |
| Score | Điểm số | Số không nguyên |
| Inherent_Risk | Rủi ro tiềm tàng | Số không nguyên |
| CONTROL_RISK | Rủi ro kiểm soát | Số không nguyên |
| Detection_Risk | Rủi ro phát hiện | Số không nguyên |
| Audit_Risk | Rủi ro kiểm toán | Số không nguyên |
| Risk | Rủi ro | Có hoặc không |
| Audit Risk = Inherent Risk x Control Risk x Detection Risk (%) | | |

## CHƯƠNG II. QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ

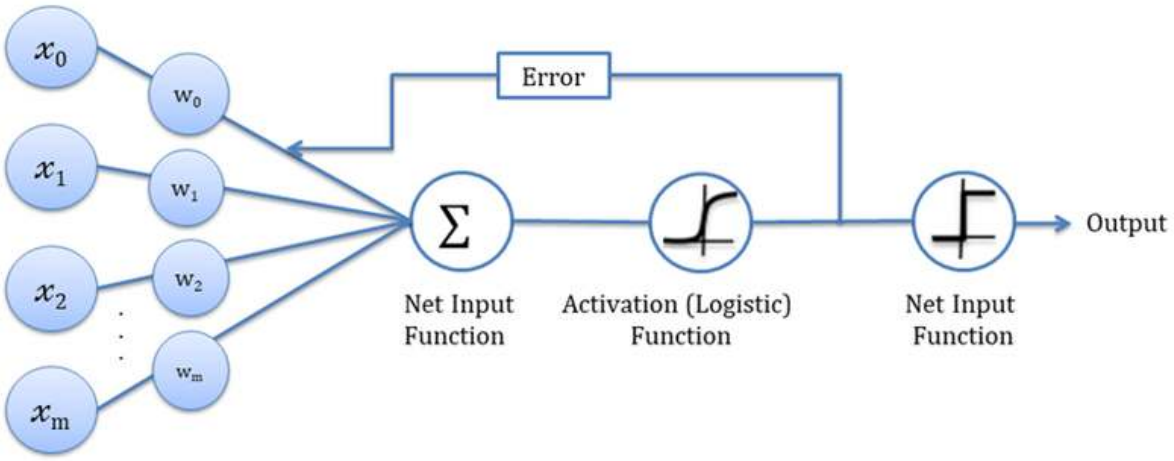
## 1. Các phương pháp dự đoán và quy trình cụ thể

### 1.1. Phân lớp dữ liệu

**- Phân lớp dữ liệu:** là cách dùng để khai thác dữ liệu của các mục được chỉ định trong một tập hợp lớn các danh mục hoặc lớp. Mục tiêu để dự đoán chính xác các lớp mục tiêu cho mỗi trường hợp trong tập hợp dữ liệu.

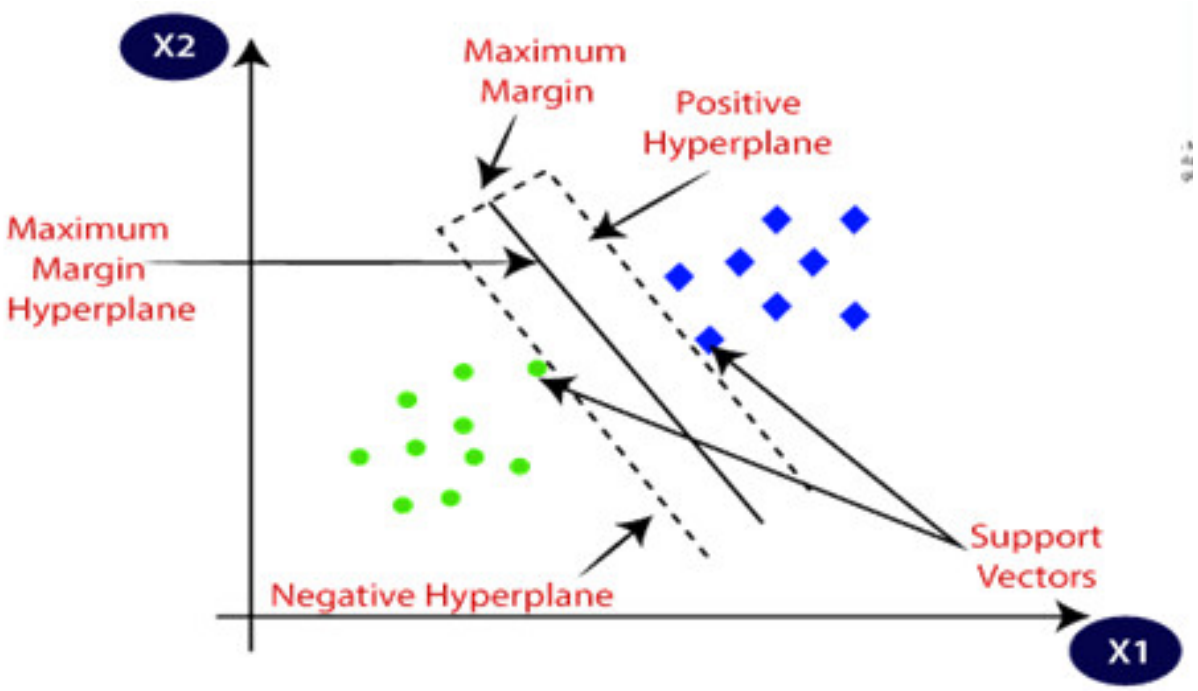
Ví dụ, một mô hình phân loại được sử dụng để xác định những người xin vay là rủi ro tín dụng thấp, trung bình hoặc cao.

**- Hồi quy logistic (Logistic Regression):** là phương pháp phân lớp dựa trên xác suất. Một mô hình đơn giản (dễ cài đặt, dễ diễn giải kết quả, huấn luyện đơn giản), không cần thông tin để phân phối của các lớp trong không gian đặc trưng, phân lớp nhanh. Tuy nhiên chỉ áp dụng với biến phụ thuộc rời rạc.



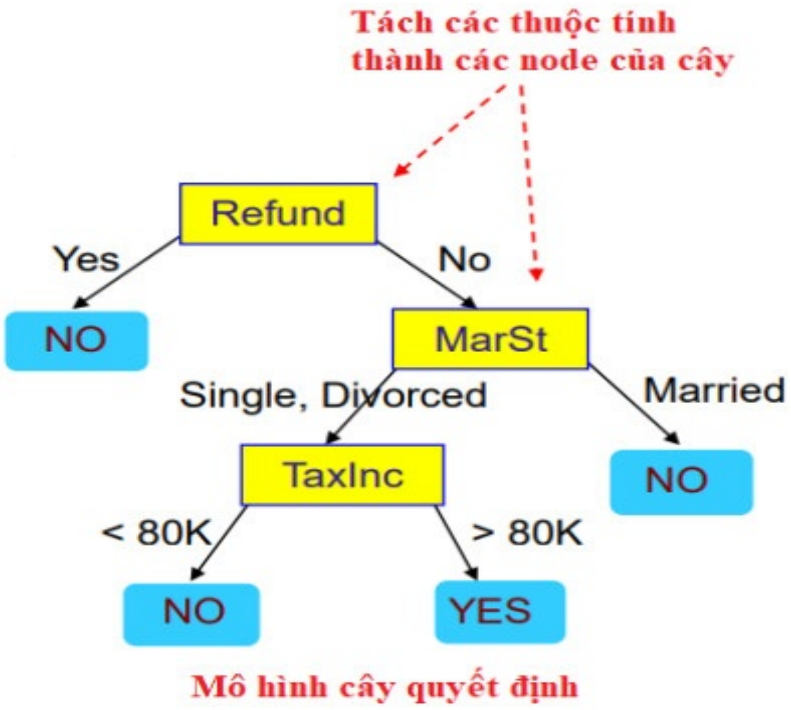

*Hình. Minh họa phương pháp hồi quy logistic*

**- Phương pháp SVM (Support Vector Machine):** là một thuật toán có giám sát, mô hình nhận dữ liệu đầu vào và xem chúng như những vector trong không gian sau đó phân chia chúng vào các lớp khác nhau bằng cách xây dựng siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu. Để có được kết quả phân lớp tối ưu thì phải xác định siêu phẳng (hyperplane) có khoảng cách đến các điểm dữ liệu (margin) của tất cả các lớp xa nhất có thể. SVM có khả năng phân lớp nhanh và tiết kiệm bộ nhớ. Tuy nhiên đối mặt với kho dữ liệu lớn hay số chiều lớn hơn số mẫu dữ liệu huấn luyện thì trở nên kém hiệu quả, nhạy cảm với nhiễu hoặc thiếu thông tin xác suất phân lớp.




*Hình. Minh họa phương pháp SVM (Support Vector Machine)*

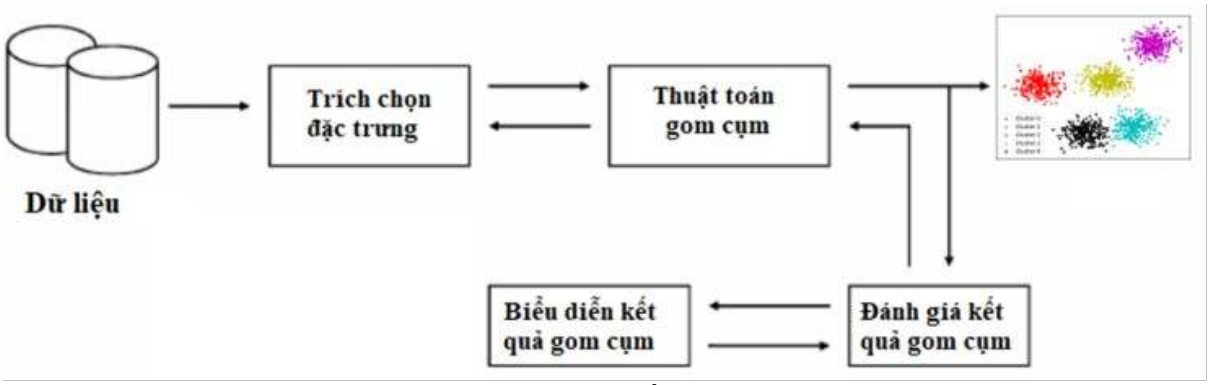
**- Phương pháp Cây quyết định (Decision Tree):** trong lĩnh vực quản trị, cây quyết định là đồ thị các quyết định đi kèm với các kết quả khả dĩ nhằm giúp quá trình ra quyết định. Trong khai thác dữ liệu, cây quyết định là phương pháp mô tả, phân loại và tổng quát hóa dữ liệu cho trước. Với hình thức dễ hiểu và không đòi hỏi việc chuẩn hóa dữ liệu, nó có thể xử lý trên nhiều kiểu kiến thức khác nhau và xử lý hiệu quả một lượng lớn dữ liệu trong một thời gian ngắn. Bên cạnh vẫn còn hạn chế tỏng việc xử lý tình huống với dữ liệu phụ thuộc thời gian và chi phí xây dựng mô hình cao.




*Hình. Minh họa phương pháp Cây quyết định (Decision Tree)*

## 1.2. Phân cụm dữ liệu

- **Phân cụm dữ liệu** là quá trình gom cụm/ nhóm các đối tượng/dữ liệu có đặc điểm tương đồng vào các nhóm/ cụm tương ứng. Trong đó, tương đồng giữa những phần tử trong cùng cụm; khác biệt với với những phần tử trong các cụm khác. (Slide thầy)




*Hình Minh hoạ về phân cụm dữ liệu*

**- Đặc điểm**

+ Số cụm dữ liệu không được biết trước vì vậy việc phân cụm dữ liệu thuộc nhóm học không giám sát (unsupervised learning)

+ Có nhiều cách tiếp cận, mỗi cách lại có vài kỹ thuật

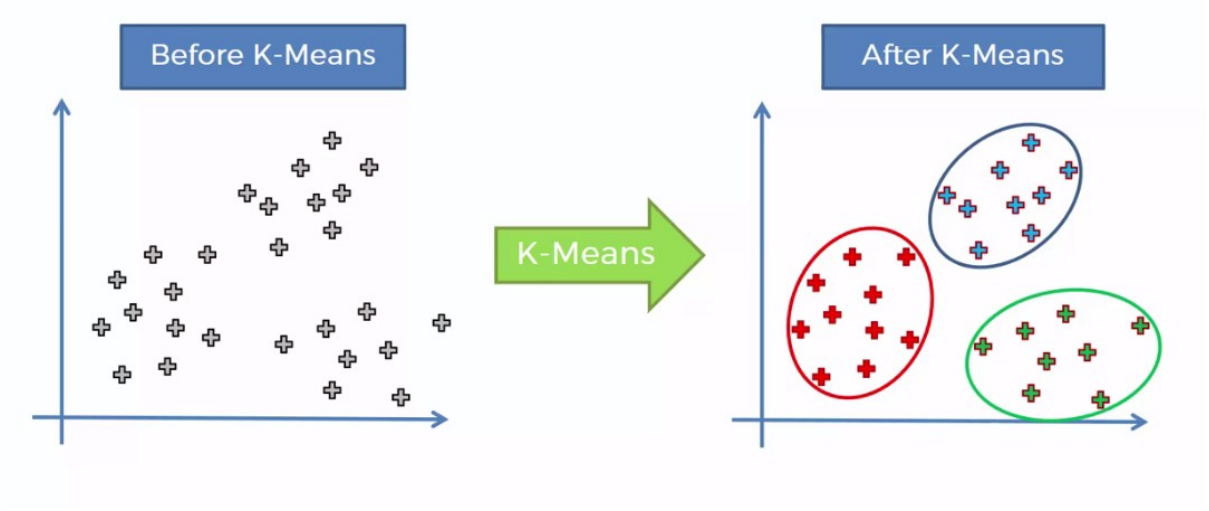
+ Các kỹ thuật khác nhau thường mang lại kết quả khác nhau.

**- Đánh giá mô hình phân cụm:**

+ Ưu điểm của phân cụm phân lớp là không phải xác định trước số lượng cụm điều này khá vượt trội so với K-Means. Tuy nhiên, nó không hoạt động tốt với lượng dữ liệu khổng lồ.

+ Thuật toán phân cụm phân lớp có thể được sử dụng để xác định, dự đoán số cụm trước khi thực hiện thuật toán K-Means.

*Thuật toán K-Means: thuộc nhóm thuật toán phân cụm dựa trên phân hoạch




*Hình Minh họa về thuật toán K-Means*

-**Ý tưởng chính**: ta xem mỗi đối tượng trong tập dữ liệu là một điểm trong không gian d chiều (với d là số lượng thuộc tính của đối tượng)

+ Bước 1: Loại bỏ các hàng dữ liệu bị khuyết. Thuộc bước Tiền xử lý dữ liệu

+Bước 2: Chọn k điểm bất kỳ làm trung tâm ban đầu của k cụm.

+Bước 3: Phân mỗi điểm dữ liệu cào cụm có trung tâm gần nó nhất. Nếu các điểm dữ liệu ở từng cụm vừa được phân chia không thay đổi so với kết quả của lần phân chia trước nó thì ta dừng lại thuật toán.

+Bước 4: Cập nhật lại tình hình cho từng cụm: Lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cụm đó sau khi phân chia ở bước 2

+Bước 5: Quay lại bước 2.

**-Các bước quy trình của SVM, LR và Tree:**

+ Bước 1: Nhập dữ liệu cần huấn luyện vào Orange

+ Bước 2: Nối widget dữ liệu huấn luyện và SVM, Tree, LR với Test and score, sau đó nối widget vào Confusion Matrix để thực hiện đánh giá kết quả và đánh giá ma trận nhầm lẫn.

+ Bước 3: Sau khi chọn được phương pháp dự báo tốt nhất, nối dữ liệu huấn luyện vào SVM, hoặc Tree, hoặc LR. Đồng thời nhập dữ liệu dùng để dự báo vào Orange.

+ Bước 4: Liên kết phương pháp dự báo tốt nhất và dữ liệu dự báo với Predictions để đánh giá và phân loại dữ liệu đầu vào.

+ Bước 5: Xuất kết quả dự báo bằng Data Table

## 2. Tìm hiểu về dữ liệu

### 2.1. Phân tích dữ liệu

**- Với bộ dữ liệu Audit Risk:**

+ Dữ liệu thô chứa …. đối tượng (hàng) và …. thuộc tính (cột)

+ Trong mỗi đối tượng sẽ là một đại diện cho mỗi doanh nghiệp được chọn để cho vào rà soát về những rủi ro gian lận, mỗi thuộc tính đại diện những đặc trưng của từng đối tượng doanh nghiệp

+ Từ bộ dữ liệu, khi xét đến các cột dữ liệu hiện thị những đặc trưng, có 5 thuộc tính chính được tích hợp từ những thuộc tính còn lại và có sự tác động không nhỏ đối việc phân tích và phát hiện các nguy cơ gian lận của một số doanh nghiệp. Năm thuộc tính bao gồm:

- Audit Risk (rủi ro kiểm toán)
- Dectection_Risk (rủi ro phát hiện)
- Control_Risk (rủi ro kiểm soát)
- Inhenrent_Risk (rủi ro tiềm tàng)
- Risk (Rủi ro gian lận)

Nhóm cho rằng Audit Risk là một biến phụ thuộc để xét dữ liệu của biến Risk, khi Risk có giá trị là "1" nghĩa là doanh nghiệp có nguy cơ gian lận và Risk có giá trị là "0" là nguy cơ doanh nghiệp không gian lận. Bởi vì, nhóm nhận thấy mối quan hệ giữa biến Audit Risk với các biến còn lại như sau: Audit_Risk = Inhenrent Risk x Control Risk x Detection Risk.

→ Qua việc đánh giá bộ dữ liệu, nhóm có một vài kết luận như sau: những doanh nghiệp có biến Audit Risk với giá trị nhỏ hơn 1, thì nhìn chung hầu hết các biến Risk là "0" - nghĩa là doanh nghiệp không có nguy cơ gian lận. Tuy nhiên, ở chiều ngược lại, chỉ số của biến Audit lớn hơn 1 thì nhìn chung sẽ cho kết quả biến Risk là "1" _ có nguy cơ gian lận ở các nhóm doanh nghiệp này

**- Với bộ dữ liệu Default credit clients payment:**

+ Dữ liệu thô bao gồm …. hàng (thuộc tính) và 14 cột (đặc trưng)

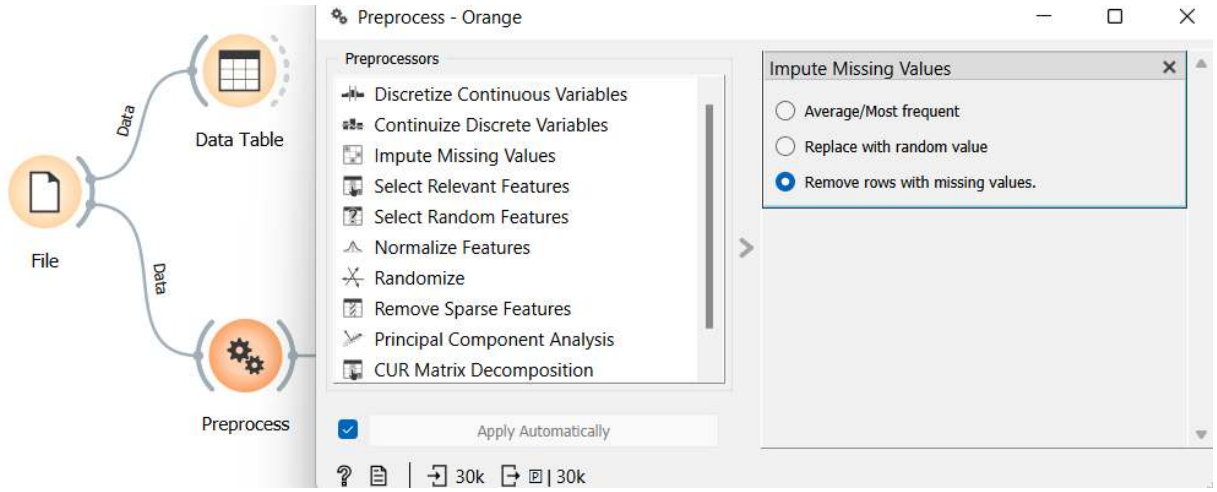
+ Mỗi một thuộc tính là đại diện cho một khách hàng, số tiền vay và mỗi thuộc tính chứa các đặc điểm của các khách hàng đó

+ Mỗi đơn vị chứa các loại thuộc tính sau: person_age, person_income, person_home_ownership, person_emp_length, loan_intent, loan_grade, loan_amnt, loan_int_rate, loan_status, loan_percent_income, cb_person_default_on_file, cb_person_cred_hist_lenght sẽ thể hiện được thông tin liên quan một cách cụ thể về khách hàng, các khoản vay, tình trạng thanh toán và có sự tác động lớn đến việc kiểm soát các khoản phải thu được ghi nhận cùng với những gian lận có liên qua của doanh nghiệp

+ Nhóm nhận định bộ dữ liệu này có biến phụ thuộc là loan_status, với dữ liệu "1" là khách hàng có thể trả tiền, và dữ liệu "0" là khách hàng không thể trả tiền.

## 2.2. Tiền xử lý dữ liệu

Hai bộ dữ liệu mà nhóm chọn được tổng hợp và chọn lọc từ trang https://www.kaggle.com/datasets. Bộ dữ liệu Audit risk mà nhóm chọn để nghiên cứu đã có đầy đủ thông tin, dữ liệu cần thiết cũng như không bị thiếu hoặc mất dữ liệu nên nhóm quyết định bỏ qua bước tiền xử lý dữ liệu đối với bộ dữ liệu này. Đối với bộ dữ liệu Credit risk vì dữ liệu quá lớn và có một số thông tin bị thiếu sót nên nhóm đã dùng preprocessing để xóa đi các hàng có chứa dữ liệu bị mất.



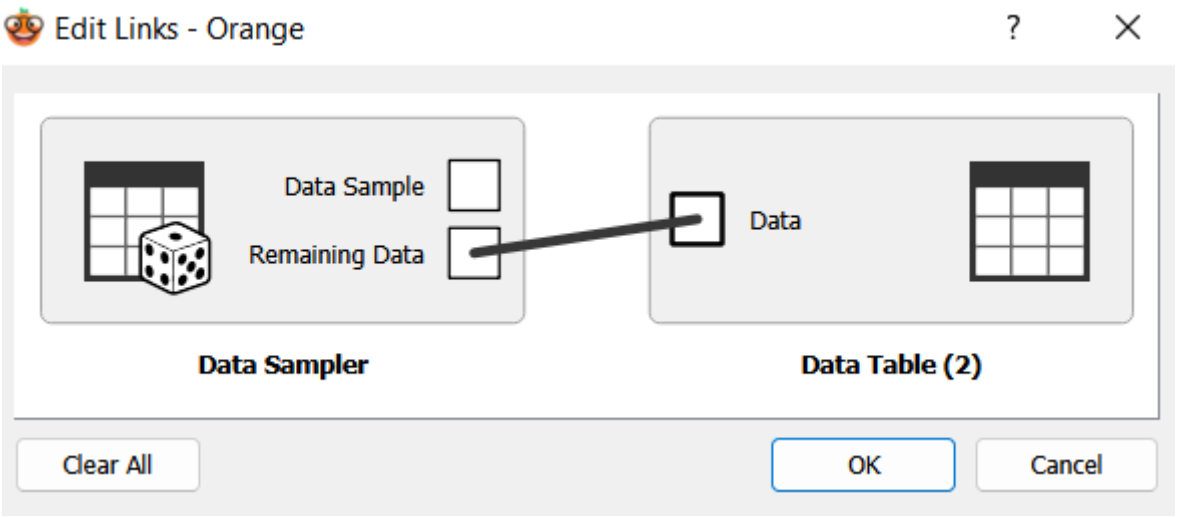
*Hình. Mô tả xử lý dữ liệu bị mất*

### 2.3. Phân tách dữ liệu

Trong bài nghiên cứu của nhóm, hai bộ dữ liệu được sử dụng để phân tích đều được tách ra thành 2 file dữ liệu riêng biệt:

+ 70% của mỗi bộ dữ liệu được dùng để làm dữ liệu mẫu cho mô hình phân lớp dữ liệu

+ 30% dữ liệu còn lại của mỗi bộ được dùng để dự báo




*Hình. Mô tả phân tách bộ dữ liệu*

## 3. Thực nghiệm

### 3.1. Kiến thức chuyên ngành

***Một là***, vận dụng các kiến thức về kế toán, kiểm toán và đặc biệt là kiến thức liên quan đến hạch toán khoản phải thu. Nếu dự báo khách hàng không trả tiền, doanh nghiệp phải thực hiện xóa nợ bằng cách ghi tăng chi phí quản lý doanh nghiệp (Nợ TK 642), ghi giảm khoản dự phòng khó đòi (Nợ TK 229) và ghi giảm khoản phải thu (Có TK 131).

***Hai là***, vận dụng phương trình kế toán:

Tài sản = Nợ phải trả + Vốn chủ sở hữu

Và, công thức tính lợi nhuận như sau:

Lợi nhuận = Doanh thu - Chi phí

***Ba là***, vận dụng các kiến thức về kế toán, kiểm toán về các khả năng gian lận của các công ty thông qua các thủ thuật làm sai lệch báo cáo tài chính (che giấu khoản tiền thu được…) và nhận biết các rủi ro kiểm toán theo công thức:

$$AR = IR \times CR \times DR$$

Trong đó:

- Rủi ro kiểm toán (Audit Risk - AR) là khi kiểm toán viên và các công ty kiểm toán đưa ra những ý kiến, nhận xét sai lầm vì những sai sót trọng yếu trên báo cáo tài chính, làm ảnh hưởng đến việc đưa ra quyết định của những người sử dụng báo cáo tài chính đó.
- Rủi ro kiểm toán được xác định dựa vào 3 yếu tố:

+ Rủi ro tiềm tàng (Inherent risk - IR): Là sự tồn tại sai sót trọng yếu trong bản thân đối tượng kiểm toán (tức tồn tại ngay trong chức năng hoạt động và môi trường quản lý của doanh nghiệp).

+ Rủi ro kiểm soát (Control risk - CR): Là sự tồn tại sai sót trọng yếu mà hệ thống kiểm soát nội bộ không phát hiện và ngăn chặn kịp thời.

+ Rủi ro phát hiện (Detection risk - DR): Là sự tồn tại sai sót trọng yếu mà hệ thống kiểm toán hay chuyên gia kiểm toán không phát hiện ra được.

Bài toán 1: + Bài toán 1: Dự báo khả năng doanh nghiệp lợi dụng các khoản thanh toán của khách hàng nhằm che giấu công nợ để gian lận trên Báo cáo tài chính

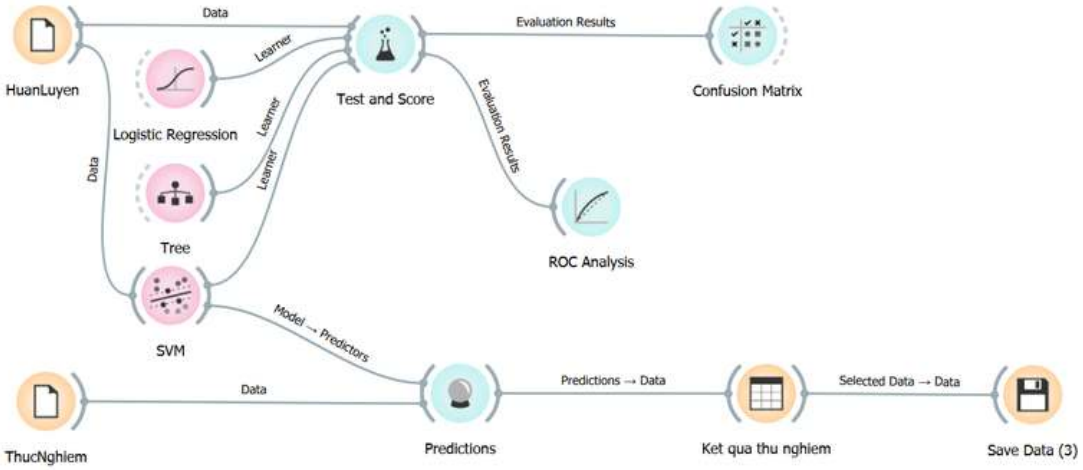
## 3.2. Bài toán 1: Dự báo khả năng doanh nghiệp lợi dụng các khoản thanh toán của khách hàng nhằm che giấu công nợ để gian lận trên Báo cáo tài chính

### 3.2.1. Mô tả bài toán

Gian lận là hình thức một số doanh nghiệp cố ý tạo ra để trực lợi cho bản thân. Một số hình thức gian lận thường xuyên được sử dụng như: sửa đổi hoặc làm giả các hóa đơn, chứng từ liên quan đến báo cáo tài chính; cố tình che giấu, bỏ sót, ghi chép sai sự thật một số thông tin liên quan đến nghiệp vụ kế toán. Trong số đó, hành vi gian lận thường xuất hiện khi khai sai số tiền thu được từ khách hàng, ảnh hưởng lớn đến doanh thu, chi phí và lợi nhuận của doanh nghiệp. Điều này dẫn đến việc doanh nghiệp sẽ bị đánh giá sai lệch về khả năng tài chính, khiến cho những người sử dụng thông tin từ báo cáo tài chính này đưa ra những quyết định sai lầm về đầu tư hoặc cho vay.

### 3.2.2. Chạy mô hình và kết quả

+ Xây dựng mô hình

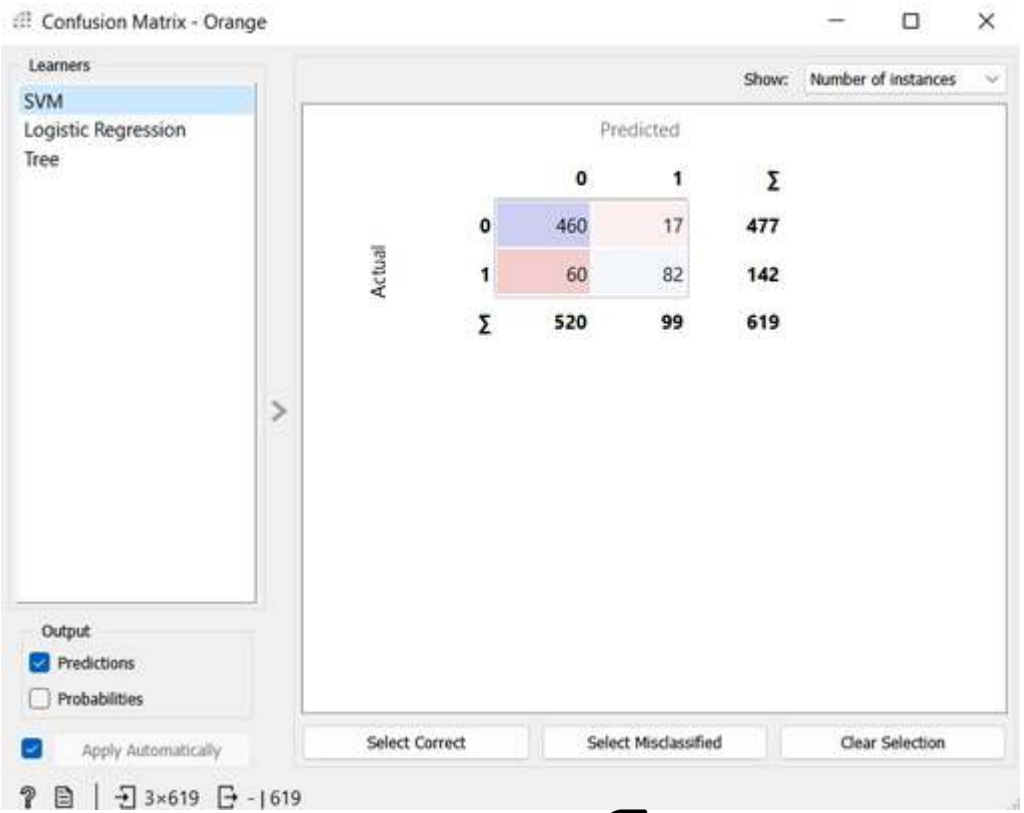



*Hình. Mô hình xây dựng bài toán 2*

### 3.2.3. Kết quả và đánh giá

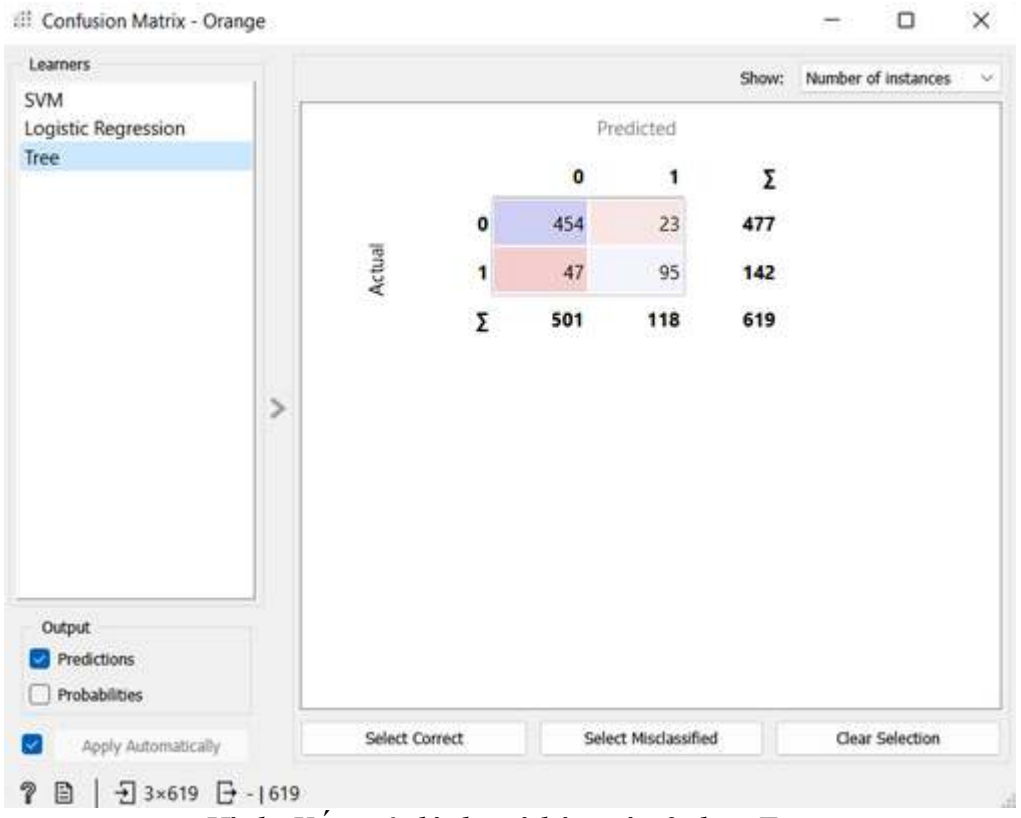
**+ Theo ma trận nhầm lẫn, ta có**

- 0: Không thu được tiền
- 1: Có thu được tiền




*Hình. Kết quả đánh giá bài toán 2 theo SVM*


*Hình. Kết quả đánh giá bài toán 2 theo Logistic Regression*




*Hình. Kết quả đánh giá bài toán 2 theo Tree*

+ Sai lầm loại 2: Công ty dự đoán không thu được tiền nhưng thực tế có thu được tiền. Vì khi công ty dự báo là không thu được tiền nhưng thực tế có thu được tiền nên số tiền thực tế mà doanh nghiệp đang có sẽ nhiều hơn số tiền đã được dự đoán hay ghi trên sổ sách.

⇨ Có ăn chặn hoặc gian lận

⇨ Rủi ro cao

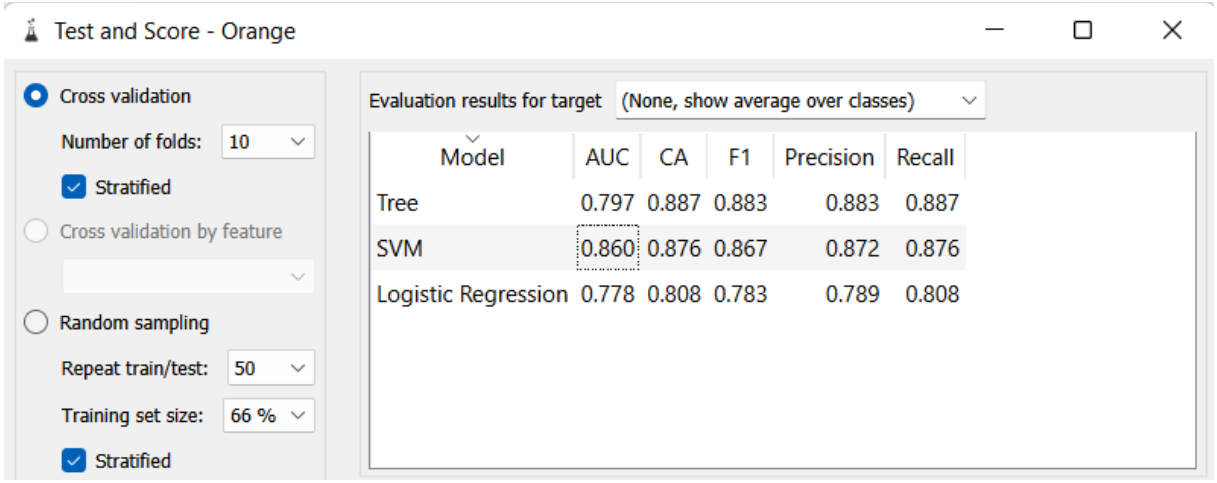
+ Xét ma trận nhầm lẫn ta có sai lầm loại 2:

LR (= 94) > SVM (= 60) > Tree (= 47)

⇨ LR lớn nhất

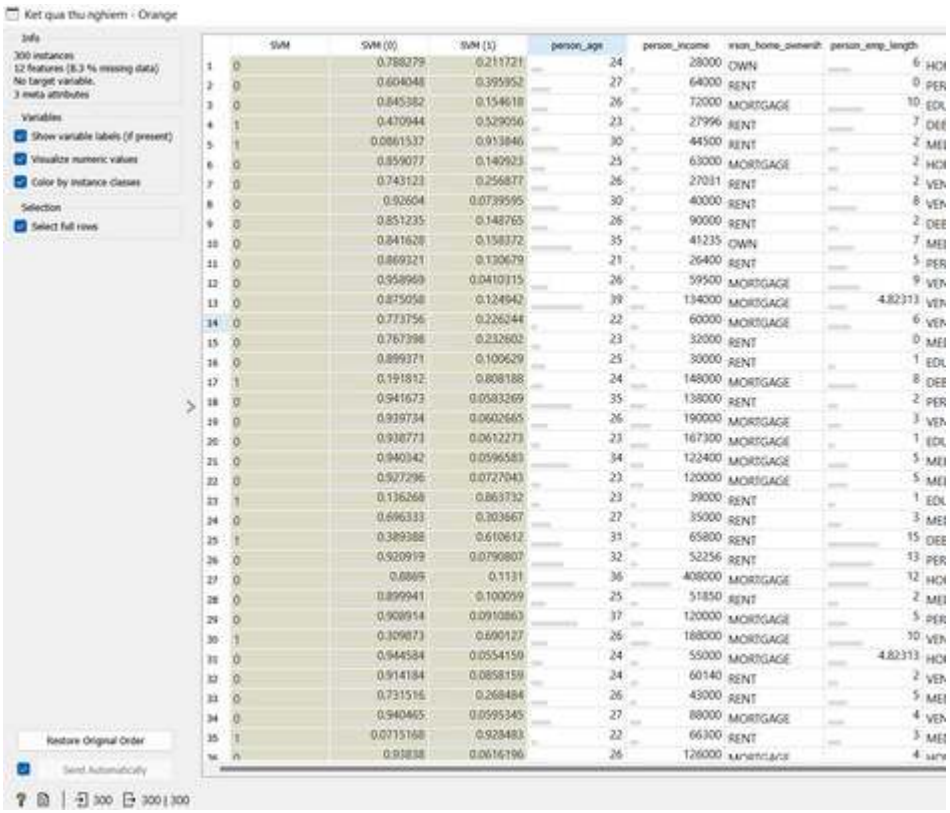
⇨ Loại phương pháp LR (Logistic Regression)

+ Theo Test and Score, xét AUC thì SVM = 0.860 lớn nhất nên chọn SVM làm mô hình dự báo.




*Hình. Kết quả đánh giá bài toán 2 và quyết định chọn phương pháp nghiên cứu*

+ Kết quả đánh giá:




*Hình. Kết quả nghiên cứu bài toán 2*

## 3.3. Bài toán 2: Phát hiện công ty gian lận trong nhóm các công ty cùng tính chất

### 3.3.1. Mô tả bài toán

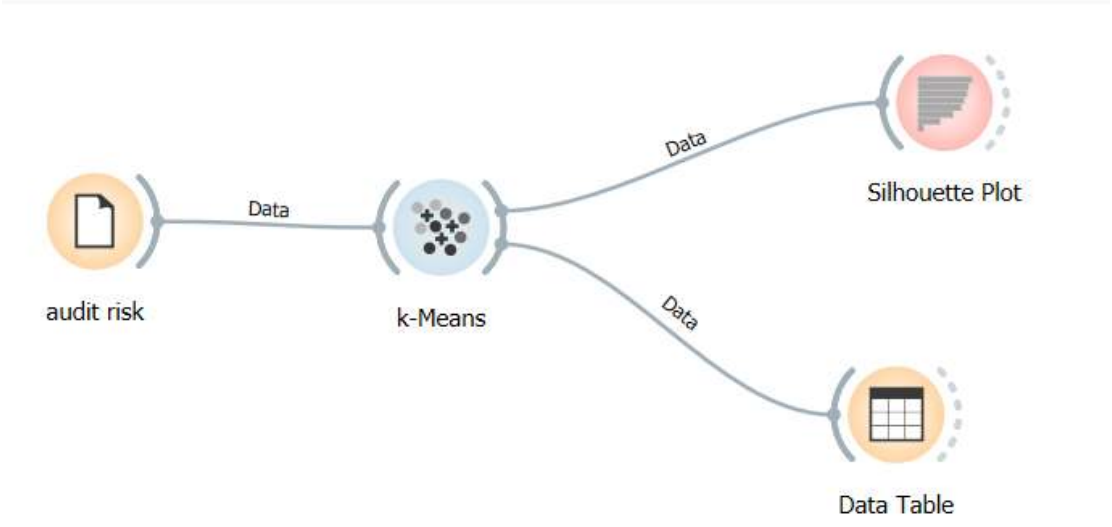
Dựa trên những điểm bất hợp lí trong các Báo Cáo Tài Chính là một cách phổ biến, để phát hiện các loại gian lận trong doanh nghiệp có cùng bản chất với nhau. Ví dụ như, liên quan đến các khoản phải thu cao hơn hoặc thấp hơn một các bất thường của một công ty dược Cửu Long với các công ty khác cùng ngành. Khi thực hiện kiểm toán các báo cáo tài chính này, kiểm toán viên sẽ phát hiện được các sự khác thường mang tính trọng yếu. Từ đó, họ sẽ tiến hành rà soát và thực hiện đánh giá lại các khoản phải thu cũng như số lượng tiền mặt trong công ty, chi phí và các giao dịch liên quan của doanh nghiệp nhằm xác minh tính đúng đắn và minh bạch của doanh nghiệp đó.

Ta sẽ dùng Orange để tiến hành chạy dữ liệu với mục đích là phân nhóm các

công ty và so sánh các chỉ tiêu được trình bày trên báo cáo tài chính.

### 3.3.2. Chạy mô hình và kết quả đánh giá

+ Xây dựng mô hình




*Hình 1 Mô hình xây dựng bài toán 2*

**k-Means** Tue Nov 29 22, 13:42:32

**Number of clusters:** 2
**Optimization:** random initialization, 10 re-runs limited to 300 steps

**Data**

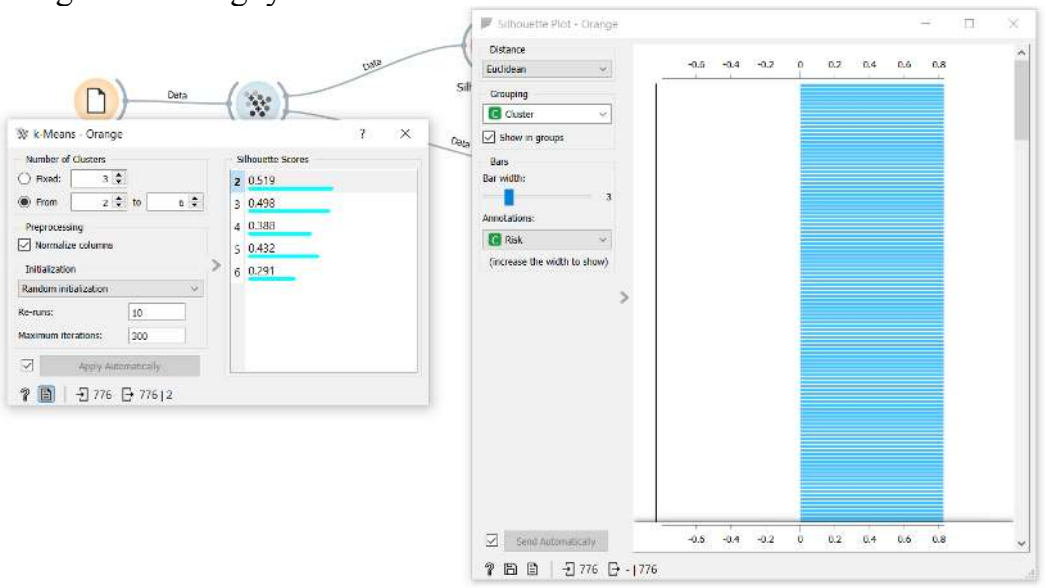
**Data instances:** 776
**Features:** Sector_score, PARA_A, Score_A, Risk_A, PARA_B, Score_B (1), Risk_B, TOTAL, numbers, Score_B (2), Risk_C, Money_Value, Score_MV, Risk_D, District_Loss, PROB, RiSk_E, History, Prob, Risk_F, Score, Inherent_Risk, CONTROL_RISK, Detection_Risk, Audit_Risk, Risk (total: 26 features)
**Meta attributes:** LOCATION_ID

**Silhouette scores for different numbers of clusters**

| | |
|---|---|
| **2** | 0.519 |
| **3** | 0.498 |
| **4** | 0.388 |
| **5** | 0.432 |
| **6** | 0.291 |

+ Kết quả và đánh giá

Kết quả phân loại K-Means chạy từ 2-6 cụm, chọn phân thành 2 cụm tương ứng với điểm Silhouette cao nhất là 0,519. Vậy ta chọn chia thành 2 nhóm để so sánh giữa các công ty.



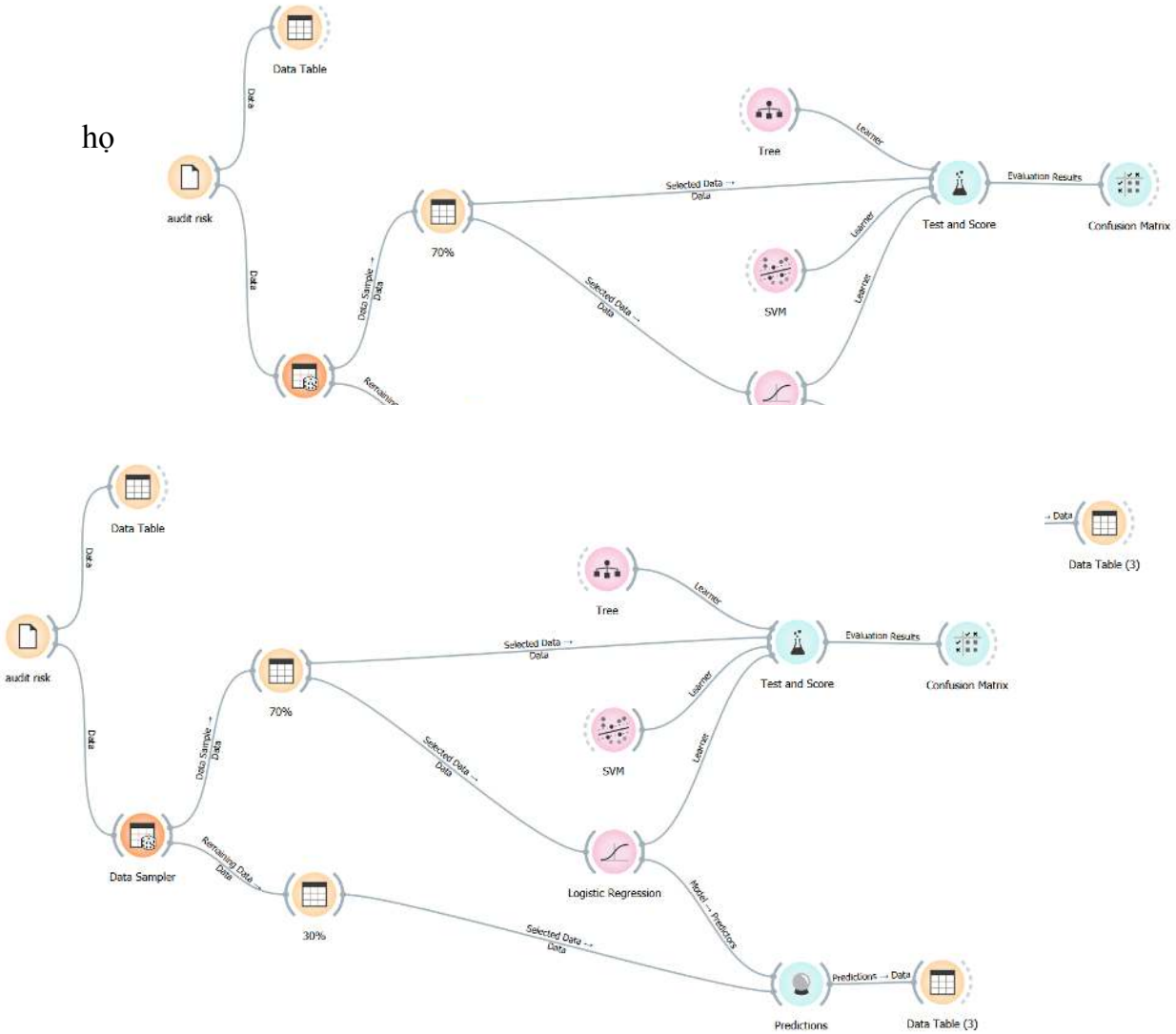

*Minh hoạ kết quả phân cụm*

## 3.4. Bài toán 3: Dự báo khả năng gian lận trên Báo cáo tài chính

### 3.4.1. Mô tả bài toán

Cho đến tận ngày nay, khi thời đại công nghệ số càng ngày càng phát triển thì việc minh bạch trong Báo cáo tài chính vẫn là một vấn đề nan giải. Chỉ vì để "làm đẹp" cho Báo cáo tài chính để lôi kéo đầu tư mà các công ty sẵn sàng có những hành vi giận lận cũng như "lách" khỏi các quy định và chuẩn mực kế toán. Mà những sự "điều chỉnh" đầy tinh vi đó thì hiện tại vẫn chưa có mô hình công nghệ nào đủ hoàn chỉnh để có thể phát hiện ra toàn bộ. Chính vì thế mà các kiểm toán viên càng mang trên mình trách nhiệm quan trọng và lớn lao để phát hiện những điều bất thường ấy,

họ




phải giữ được cái đầu lạnh và cái nhìn khách quan khi đọc Báo cáo tài chính để đánh giá một cách chính xác và công tâm nhất. Và vì thế, bài toán này lập ra để dự báo khả năng gian lận và phát hiện gian lận trên Báo cáo tài chính

### *3.4.2. Xây dựng mô hình*

*HÌnh 2 Mô hình xây dựng bài toán 3*

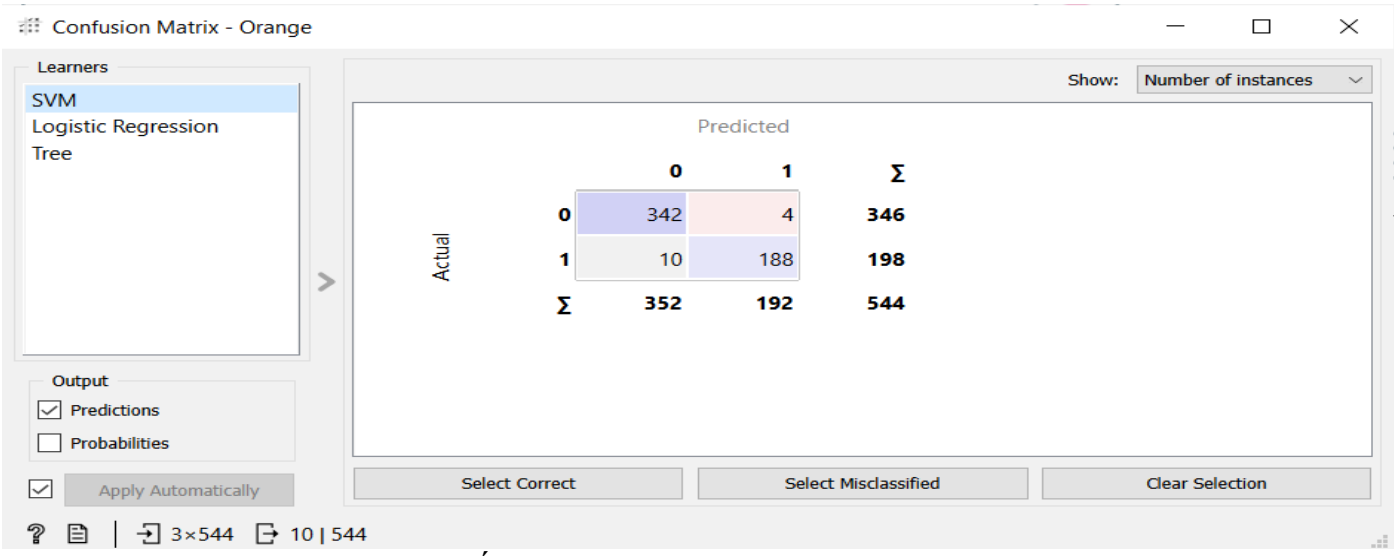
### *3.4.3. Kết quả và đánh giá*

+ Theo ma trận nhầm lẫn, ta có

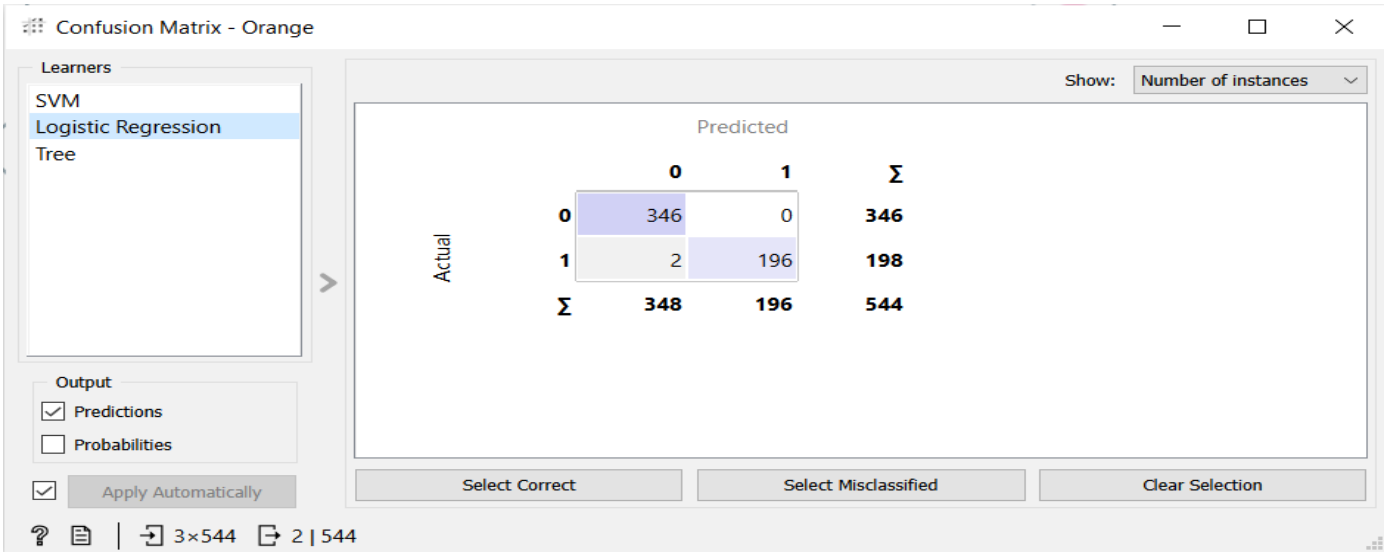
- 0: không gian lận
- 1: có gian lận

+ Sai lầm loại 2: Dự báo không gian lận nhưng trên thực tế là có sự gian lận nên điều này làm ảnh hưởng đến nhận xét và đánh giá của kiểm toán viên đối với báo cáo tài chính của công ty, làm cho những người đọc báo cáo tài chính này đưa ra

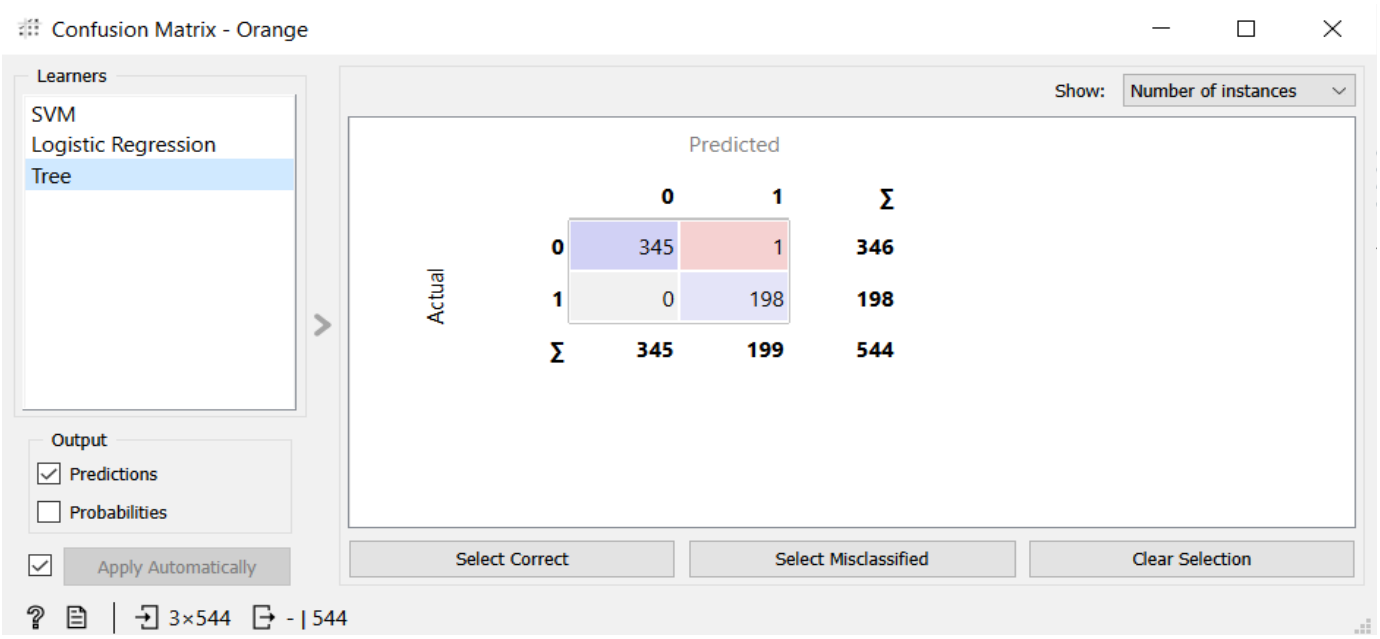
những quyết định không chính xác.




*Hình. Kết quả đánh giá bài toán 3 theo SVM*



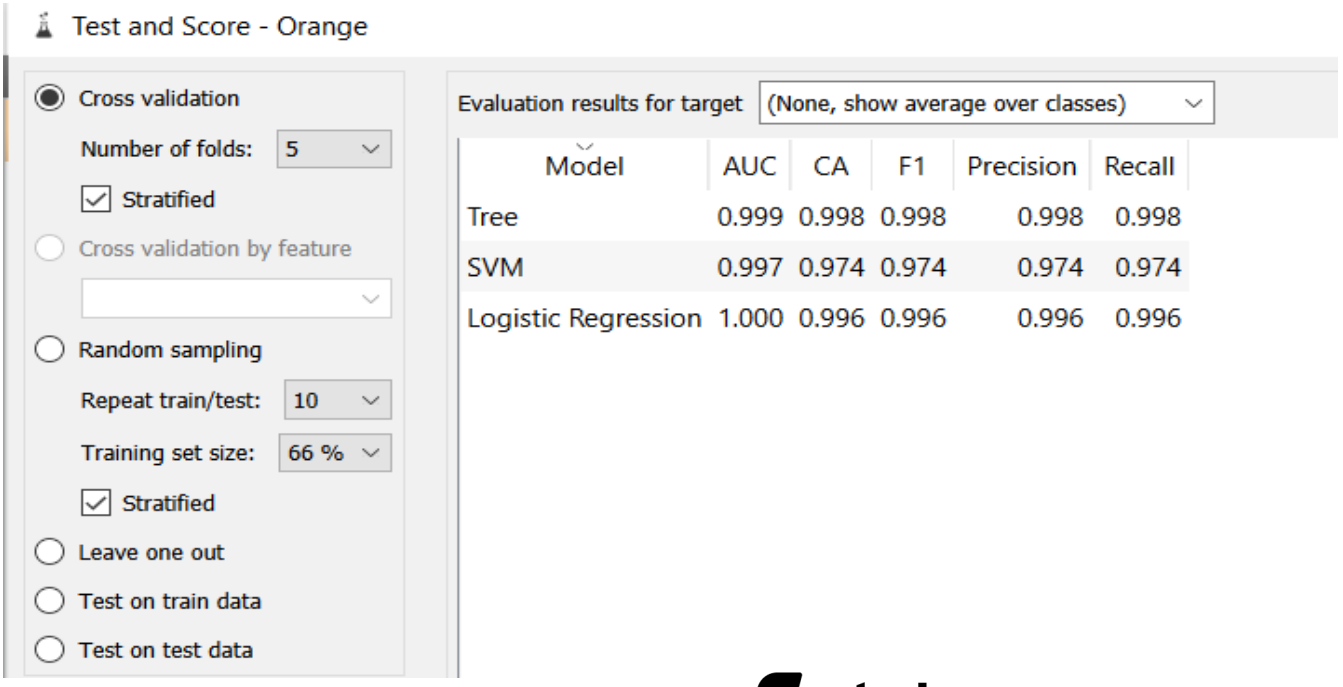

*Hình. Kết quả đánh giá bài toán 3 theo Logistic Regression*




*Hình. Kết quả đánh giá bài toán 3 theo Tree*

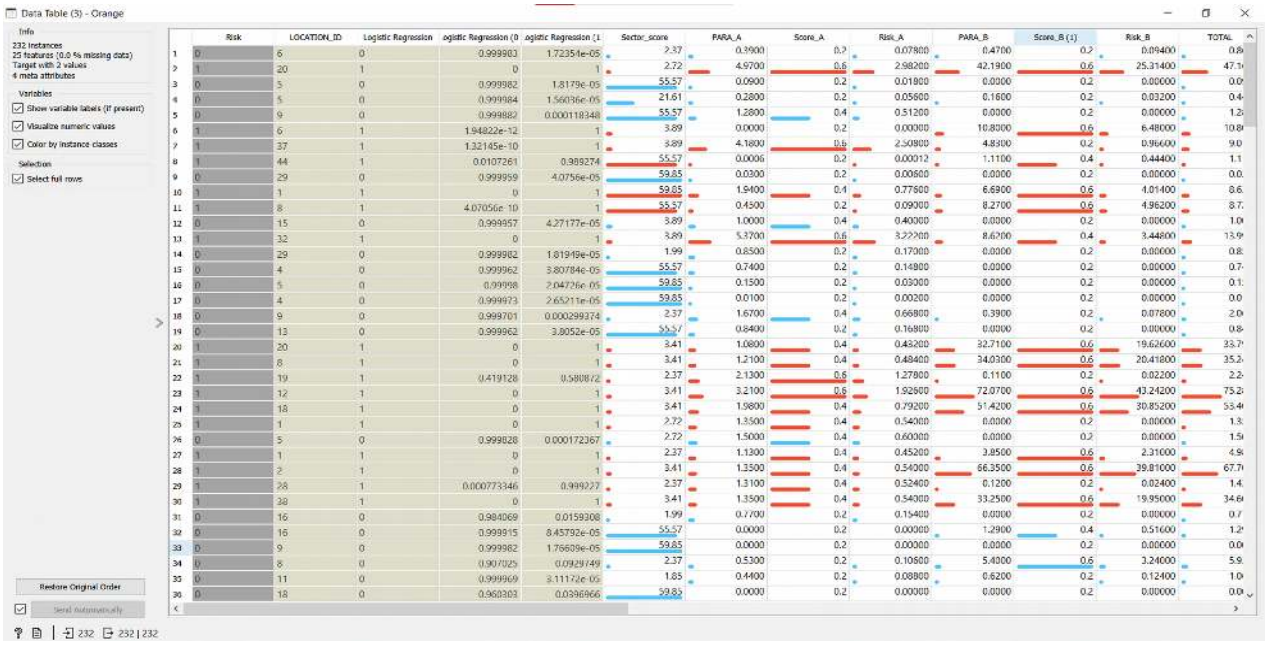
+ Xét sai lầm loại 2 thì SVM=10, LR=2 và Tree=0

+ Theo Test and Score, xét AUC thì LR=1.00 lớn nhất nên ta chọn LR làm mô hình dự báo




*Hình. Kết quả đánh giá bài toán 3 và quyết định lựa chọn phương pháp nghiên cứu*

+ Dự báo kết quả:



# CHƯƠNG III. KẾT QUẢ VÀ KẾT LUẬN

## 1. Đánh giá kết quả

### 1.1. Bài toán 1

Theo bảng đánh giá kết quả, tuy ma trận nhầm lẫn của SVM có sai lầm loại 2 là 60, lớn hơn Tree là 47, nhưng sai lầm loại 1 của SVM là 17 (nhỏ nhất trong ba phương pháp). Thêm vào đó, phương pháp SVM cho kết quả AUC là lớn nhất (0.860). Vì vậy, phương pháp SVM là phương pháp tối ưu.

### 1.2. Bài toán 2

Theo bảng đánh giá kết quả, phương pháp K-Means cho ta thấy nên chia các công ty thành 2 nhóm để so sánh với số điểm Silhouette cao nhất là 0.519 . Khi phân thành 2 nhóm, kiểm toán viên sẽ dựa trên số liệu khác biệt giữa 2 nhóm công ty này nếu có.

### 1.3. Bài toán 3

Theo bảng đánh giá kết quả, tuy ma trận nhầm lẫn của LR có sai lầm loại 2 là 2 lớn hơn Tree là 0, nhưng sai lầm loại 1 của LR là 0 (nhỏ nhấ trong 3 phương pháp). Thêm vào đó, phương pháp LR cho ra kết quả AUC là lớn nhất (=1.00). Vì vậy, LR mới là phương pháp tối ưu hơn cả.

## 2. Kết luận

Gian lận trong báo cáo tài chính càng ngày càng trở nên phổ biến và tinh vi trong cách thực hiện, làm đảo lộn hệ thống tài chính - kinh tế, gây ra các ảnh hưởng nghiêm trọng đến các doanh nghiệp liên quan cũng như các tổ chức kinh tế. Mặc dù các tổ chức, công ty kiểm toán nỗ lực ra sức ngăn chặn các hành vi gian lận, nhưng không thể nào loại bỏ hết những hành vi gian lận này. Bằng chứng là hiện nay vẫn còn rất nhiều công ty thực hiện gian dối trong báo cáo tài chính, và thậm chí trước khi bị phát hiện nó đã gây ra một cuộc đảo lộn lớn trong nền kinh tế.

Điển hình là công ty năng lượng có trụ sở tại bang Texas (Mỹ), trước khi sụp đổ, nếu xét theo doanh thu nó là công ty lớn thứ 7 tại Mỹ. Công ty đã áp dụng các phương pháp kế toán phức tạp có liên quan tới công ty vỏ bọc, Enron đã cắt bỏ được các khoản nợ hàng trăm triệu USD khỏi sổ sách của mình. Nhà đầu tư và các chuyên gia phân tích đã bị lừa vì

tin rằng tình hình tài chính của Enron ổn định hơn nhiều so với thực tế. Mặt khác, các công ty vỏ bọc do giám đốc cấp cao của Enron vận hành đã thổi phồng doanh thu gấp nhiều lần, do vậy, Enron đã tạo ra được con số doanh thu và lợi nhuận ấn tượng. Tuy nhiên, khi mạng lưới gian lận phức tạp bị phanh phui, giá cổ phiếu Enron rơi tự do. Vụ phá sản của Enron cũng kéo theo sự sụp đổ của Arthur Andersen - hãng kiểm toán lớn thứ 5 thế giới.

Nghiên cứu này đã chỉ ra một số cách thức công ty dùng khoản phải thu để gian lận, cách phát hiện gian lận trong báo cáo tài chính và dự đoán khả năng gian lận trên báo cáo tài chính của công ty niêm yết. Khi kiểm toán viên tiến hành kiểm tra, họ sẽ sử dụng mô hình này để hỗ trợ cho công việc của mình, giảm thiểu ít nhất các rủi ro trong quá trình làm việc.

Qua bài nghiên cứu này, nhóm đã hoàn thành mục tiêu đã đề ra là giải quyết các bài toán liên quan vấn đề gian lận nhằm hỗ trợ kiểm toán viên trong quá trình làm việc. Mặc dù không giải quyết được các bài toán lớn và phức tạp nhưng mô hình phát hiện gian lận cũng được xây dựng khá hoàn chỉnh để xác định tính chính xác của vấn đề. Chúng ta có thể xem những phương pháp này là một điều vô cùng quan trọng và cần thiết, mục đích nhằm cải thiện và nâng cao hiệu quả phát hiện các gian lận trong báo cáo tài chính của doanh nghiệp.

## 3. Những hạn chế

Kết quả của bài nghiên cứu chưa có độ chính xác cao trong thực tế vì bộ dữ liệu là có sẵn trong quá khứ, không được cập nhật thêm về các yếu tố khác mà có khả năng ảnh hưởng đến đến biến phụ thuộc. Ví dụ, nếu có thêm một cột là thu nhập bình quân đầu người quốc gia mà có dấu hiệu tăng, thì có thể tác động đến việc làm giảm khả năng không trả tiền của khách hàng.

Đối với phần kết quả xuất ra của các bài toán có thể sẽ bị lệch so với thực tế, bởi vì bộ dữ liệu thô chứa số liệu rất lớn nên nhóm đã chọn mẫu một phần của dữ liệu để làm bài nghiên cứu này.

Vì đề tài liên quan lớn đến thông tin bảo mật các khách hàng của doanh nghiệp nhóm không thể tiến hành khảo sát và lấy số liệu thực tế thêm, chỉ có thể sử dụng bộ dữ liệu có sẵn nên kết quả nghiên cứu còn nhiều hạn chế và mang tính chất tham khảo.

Phạm vi nghiên cứu và ứng dụng bị thu hẹp vì nhóm chỉ đưa ra phương pháp giải quyết bài toán với một chỉ tiêu nhất định là Khoản phải thu đối với khả năng thanh toán của khách hàng mua chịu hàng.

Sinh viên chưa thể áp dụng kiến thức chuyên ngành một cách chuyên sâu nhất để có thể đánh giá bài toán một cách chuyên nghiệp và hoàn chỉnh nhất.