TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN THỊ THÚY VY - 232805404**

**VÕ LUYỆN - 52100911**

**TRƯƠNG THÁI ĐAN HUY - 52100222**

**BÁO CÁO CUỐI KỲ**

**PHÂN TÍCH DỮ LIỆU**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN THỊ THÚY VY - 232805404**

**VÕ LUYỆN - 52100911**

**TRƯƠNG THÁI ĐAN HUY - 52100222**

**BÁO CÁO CUỐI KỲ**

**PHÂN TÍCH DỮ LIỆU**

Người hướng dẫn

**TS. Hoàng Anh**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em xin bày tỏ lòng biết ơn chân thành đến thầy Hoàng Anh vì  
những kiến thức quý báu mà thầy đã truyền đạt và sự tận tâm trong việc giảng dạy  
môn Phân tích dữ liệu. Chúng em cảm nhận được sự chuyên nghiệp và đam mê của thầy trong việc truyền đạt tri thức, và chúng em rất biết ơn vì thầy đã dành thời gian và công sức để hướng dẫn chúng em trong quá trình học tập và tìm hiểu về lĩnh vực này.

Thầy đã truyền đạt những kiến thức sâu sắc và chi tiết về Phân tích dữ liệu, giúp chúng em hiểu rõ hơn về khung phát triển này và cách áp dụng vào thực tế. Nhờ những điều thầy đã truyền dạy, chúng em đã nắm vững cách xử lý dữ liệu một cách hiệu quả, đem lại kết quả tốt trong ứng dụng thực tế.

Chúng em cũng biết ơn vì sự quan tâm và hỗ trợ tận tình của thầy trong quá trình học tập. Thầy đã luôn sẵn sàng trả lời các câu hỏi của chúng em và giúp đỡ chúng em vượt qua những khó khăn trong quá trình nắm bắt kiến thức. Nhờ đó, chúng em đã có thêm niềm tin và động lực để tiếp tục khám phá và phát triển trong lĩnh vực Phân tích dữ liệu.

Chúng em cảm nhận được sự chuyên nghiệp và đam mê của thầy trong việc  
giảng dạy. Sự cống hiến và tâm huyết của thầy đã giúp chúng em có được nền tảng  
vững chắc, đồng thời truyền cảm hứng để chúng em tiếp tục theo đuổi đam mê và  
ước mơ của bản thân.

Với tấm lòng biết ơn sâu sắc, chúng em xin kính chúc thầy Hoàng Anh sức khỏe dồi dào, hạnh phúc và ngày càng thành công trong việc truyền tải tri thức và hỗ trợ sinh viên. Mong rằng những đóng góp của thầy sẽ tiếp tục lan tỏa và mang lại những thành tựu to lớn cho thầy và cả khoa Công nghệ thông tin.

*TP. Hồ Chí Minh, ngày 3 tháng 1 năm 2025*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*VY*

*Nguyễn Thị Thúy Vy*

*LUYỆN*

*Võ Luyện*

*HUY*

*Trương Thái Đan Huy*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Hoàng Anh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 3 tháng 1 năm 2025*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*VY*

*Nguyễn Thị Túy Vy*

*LUYỆN*

*Võ Luyện*

*Huy*

*Trương Thái Đan Huy*

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vi](#_Toc187441934)

[DANH MỤC CÁC CHỮ VIẾT TẮT vii](#_Toc187441935)

[CHƯƠNG 1. BÀI TOÁN 1](#_Toc187441936)

[1.1 Bài toán 1](#_Toc187441937)

[1.2 Cơ sở lý thuyết 1](#_Toc187441938)

[1.2.1 Đặc điểm của giao dịch gian lận: 1](#_Toc187441939)

[1.2.2 Oversampling 1](#_Toc187441940)

[1.2.3 Undersampling 3](#_Toc187441941)

[1.2.4 Hybrid methods: 5](#_Toc187441942)

[CHƯƠNG 2. CÁC PHƯƠNG PHÁP XỬ LÝ BÀI TOÁN 7](#_Toc187441943)

[2.1 Dataset 7](#_Toc187441944)

[2.2 Phân tích dữ liệu: 7](#_Toc187441945)

[CHƯƠNG 3. CÁC MÔ HÌNH ĐỀ XUẤT 12](#_Toc187441946)

[3.1 Bộ dữ liệu huấn luyện 12](#_Toc187441947)

[3.2 Các mô hình học máy cơ bản 12](#_Toc187441948)

[3.2.1 Đánh giá các mô hình 12](#_Toc187441949)

[3.2.2 Đánh giá mô hình: 15](#_Toc187441950)

[3.2.3 Đề xuất cải thiện: 16](#_Toc187441951)

[3.2.4 Kết luận 16](#_Toc187441952)

[3.3 Các mô hình học sâu và ensemble: 17](#_Toc187441953)

[3.3.1 Đánh giá các mô hình 17](#_Toc187441954)

[3.3.2 Đề xuất kinh doanh 19](#_Toc187441955)

[CHƯƠNG 4. KẾT LUẬN 20](#_Toc187441956)

[4.1 Kết luận 20](#_Toc187441957)

[4.2 Hướng phát triển 20](#_Toc187441958)

[Tài liệu tham khảo 21](#_Toc187441959)

# DANH MỤC HÌNH VẼ

[Hình 1‑1 Thuật toán Oversampling 2](#_Toc187441680)

[Hình 1‑2 Công thức nội suy 3](#_Toc187441681)

[Hình 1‑3 SMOTE 3](#_Toc187441682)

[Hình 1‑4 Random Undersampling 4](#_Toc187441683)

[Hình 1‑5 Công thức Smote – Tomek Links 6](#_Toc187441684)

[Hình 2‑1 Phân tích Class 8](#_Toc187441685)

[Hình 2‑2 Phân phối số tiền giao dịch theo Amount 9](#_Toc187441686)

[Hình 2‑3 Tương quan giữa các đặc trưng 10](#_Toc187441687)

[Hình 2‑4 So sánh thuật toán 10](#_Toc187441688)

[Hình 2‑5 Biểu đồ so sánh các phương pháp cân bằng 11](#_Toc187441689)

[Hình 3‑1 Tạo ra bộ dữ liệu huấn luyện 12](#_Toc187441690)

[Hình 3‑2 Xuất bộ dữ liệu huấn luyện và kiểm tra ra .pkl 12](#_Toc187441691)

[Hình 3‑3 Confusion Matrix Naïve Bayes 13](#_Toc187441692)

[Hình 3‑4 Confusion matrix KNN 14](#_Toc187441693)

[Hình 3‑5 Confusion matrix Logictics Regression 14](#_Toc187441694)

[Hình 3‑6 Compairison Roc-accuracy 15](#_Toc187441695)

[Hình 3‑7 Confusion Matrix – Random Forest 17](#_Toc187441696)

[Hình 3‑8 Confusion Matrix – ANN 18](#_Toc187441697)

[Hình 3‑9 Confusion Matrix – LSTM 18](#_Toc187441698)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| KNN | K-Nearest Neighbors |
| ANN | Artificial Neural Network |
| LSTM | Long Short-Term Memory |
| RF | Random Forest |
| PCA | Principal Component Analysis |

# BÀI TOÁN

## Bài toán

Trong lĩnh vực tài chính, giao dịch gian lận gây ra tổn thất nghiêm trọng cho các tổ chức tài chính và khách hàng. Việc phát hiện và ngăn chặn giao dịch gian lận kịp thời là một thách thức quan trọng nhằm bảo vệ tài sản và giảm thiểu rủi ro.

Cụ thể, gian lận trong lĩnh vực **thẻ tín dụng** là một vấn đề nổi bật, nơi các giao dịch bất thường cần được phát hiện nhanh chóng và chính xác. Nhóm nghiên cứu tập trung xây dựng hệ thống tự động phân tích, nhận diện các giao dịch gian lận bằng cách áp dụng các mô hình học máy và học sâu hiện đại, sau đó đánh giá và so sánh kết quả để chọn ra mô hình tối ưu nhất.

## Cơ sở lý thuyết

### Đặc điểm của giao dịch gian lận:

Giao dịch gian lận trong lĩnh vực tài chính, đặc biệt là thẻ tín dụng, thường biểu hiện qua các hành vi bất thường và có khả năng gây thiệt hại nghiêm trọng. Các đặc điểm của giao dịch gian lận bao gồm: số tiền giao dịch lớn vượt mức thông thường, thực hiện giao dịch vào thời điểm không hợp lý (ví dụ: nửa đêm), hoặc từ các địa điểm không phù hợp với hành vi trước đó của người dùng (chẳng hạn giao dịch từ một quốc gia khác mà khách hàng chưa từng đến). Ngoài ra, các hành vi như thực hiện nhiều giao dịch liên tiếp trong thời gian ngắn hoặc sử dụng cùng một tài khoản trên nhiều thiết bị lạ cũng là dấu hiệu phổ biến của gian lận.

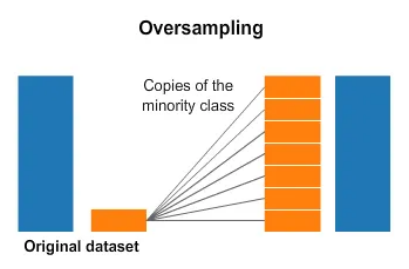
Một thách thức lớn trong phát hiện giao dịch gian lận là vấn đề mất cân bằng dữ liệu, do số lượng giao dịch hợp lệ thường áp đảo giao dịch gian lận. Điều này đòi hỏi các phương pháp xử lý dữ liệu và thuật toán phải đặc biệt nhạy bén để nhận diện chính xác các giao dịch gian lận mà không làm tăng tỷ lệ sai sót (false positives).

### Oversampling

Kỹ thuật oversampling nhằm tăng số lượng mẫu từ lớp thiểu số (giao dịch gian lận) để cân bằng tỷ lệ với lớp đa số (giao dịch hợp lệ). Dưới đây là các kỹ thuật phổ biến:

#### Random Oversampling

Sao chép ngẫu nhiên các mẫu từ lớp thiểu số để đạt được sự cân bằng

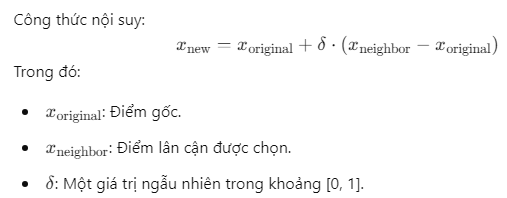


Hình 1‑1 Thuật toán Oversampling

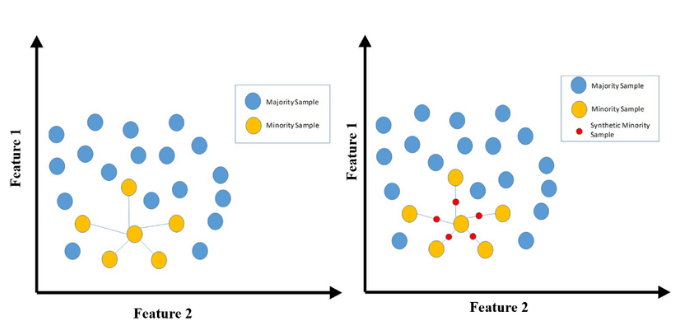
* Ưu điểm:
  + Dễ thực hiện và không yêu cầu nhiều tính toán.
  + Bảo toàn thông tin của lớp thiểu số.
* Hạn chế:
  + Tăng nguy cơ overfitting do lặp lại các mẫu giống nhau.

#### SMOTE (Synthetic Minority Oversampling Technique):

Tạo ra dữ liệu tổng hợp giữa mỗi mẫu của lớp thiểu số và “k” hàng xóm gần nhất của nó. Có nghĩa là, đối với mỗi một trong các mẫu của lớp thiểu số, “k” các láng giềng gần nhất của nó được chọn (theo mặc định là k = 5). Sau đó giữa các cặp điểm được tạo bởi mẫu và từng láng giềng của nó thì ta sẽ có được data mới.



Hình 1‑2 Công thức nội suy



Hình 1‑3 SMOTE

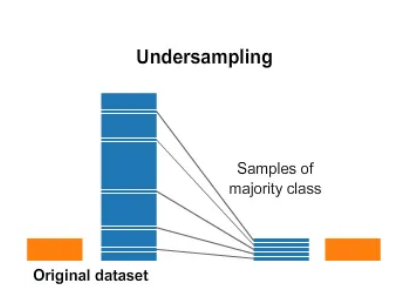
* Ưu điểm:
  + Giảm nguy cơ overfitting so với Random Oversampling.
  + Cải thiện độ phân biệt của mô hình.
* Hạn chế:
  + Không hoạt động tốt nếu dữ liệu có phân bố không đồng nhất.

### Undersampling

Kỹ thuật undersampling giảm số lượng mẫu từ lớp đa số (giao dịch hợp lệ) để cân bằng tỷ lệ với lớp thiểu số. Một số kỹ thuật phổ biến:

#### Random Oversampling

Tương tự Random Oversampling nhưng Random Undersampling là loại bỏ ngẫu nhiên các mẫu từ lớp đa số.



Hình 1‑4 Random Undersampling

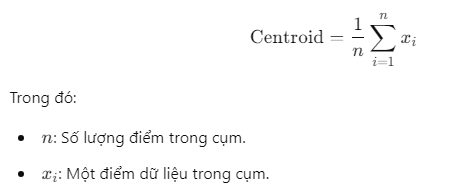
* Ưu điểm:
  + Dễ triển khai, giảm kích thước dữ liệu.
* Hạn chế:
  + Mất thông tin quan trọng từ lớp đa số, có thể làm giảm hiệu suất mô hình.

#### Cluster Centroid

Cluster Centroids là một phương pháp undersampling được sử dụng để giảm số lượng mẫu từ lớp đa số (majority class) trong bài toán dữ liệu mất cân bằng. Phương pháp này sử dụng thuật toán phân cụm (clustering), như k-Means, để xác định các điểm trung tâm (centroids) đại diện cho lớp đa số, thay vì chọn ngẫu nhiên các mẫu từ lớp này.

Nguyên lý hoạt động:

* Phân cụm lớp đa số:
  + Lớp đa số được phân thành 𝑘 cụm (clusters) sử dụng thuật toán như k-Means.
  + Mỗi cụm bao gồm các điểm dữ liệu gần nhau trong không gian đặc trưng (feature space).
* Tính centroid cho mỗi cụm:
  + Centroid là điểm trung tâm của cụm, được tính bằng trung bình các giá trị đặc trưng (feature values) của tất cả các điểm trong cụm



* Giữ lại các centroids
* Tạo tập dữ liệu mới:
  + Thay thế lớp đa số bằng các centroids đã tính toán.
  + Kết hợp lớp thiểu số và các centroids để tạo ra tập dữ liệu cân bằng.

Ưu điểm:

* Giảm số lượng mẫu từ lớp đa số mà vẫn giữ được tính đại diện của dữ liệu
* Giảm kích thước dữ liệu nên giảm thời gian huấn luyện mô hình

Nhược điểm:

* Làm mất chi tiết trong dữ liệu nếu dữ liệu phân bố phức tạp.
* Phụ thuộc vào số cụm k
* Hiệu quả kém với dữ liệu phức tạp

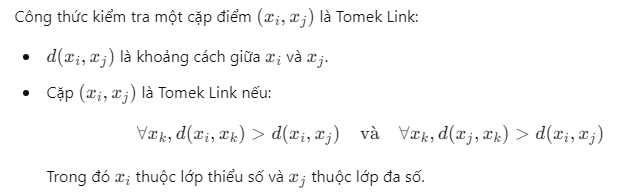
### Hybrid methods:

Phương pháp kết hợp oversampling và undersampling nhằm tận dụng lợi thế của cả hai

#### Smote-Tomek Links:

Phương pháp này kết hợp tăng cường dữ liệu với làm sạch dữ liệu, giúp mô hình học tốt hơn và cải thiện hiệu suất trong các bài toán phân loại với dữ liệu mất cân bằng.

* SMOTE: Tạo mẫu mới từ lớp thiểu số.
* Tomek Links: Loại bỏ các mẫu dư thừa hoặc bị nhiễu từ lớp đa số, giúp làm sạch dữ liệu.



Hình 1‑5 Công thức Smote – Tomek Links

* Nguyên lý hoạt động:
  + Áp dụng SMOTE để tạo ra các mẫu tổng hợp mới dựa trên nội suy giữa các điểm gần kề
  + Phát hiện các cặp Tomek Links giữa hai lớp trong tập dữ liệu mới
  + Loại bỏ các điểm thuộc lớp đa số trong các cặp Tomek Links để làm sạch dữ liệu
* Ưu điểm:
  + Kết hợp khả năng tăng cường dữ liệu và làm sạch nhiễu.
* Hạn chế:
  + Tốn nhiều tài nguyên tính toán.
  + Nếu dữ liệu phức tạp hoặc có quá nhiều nhiễu, Tomek Links có thể không phát hiện hiệu quả.

#### SMOTE-ENN (Edited Nearest Neighbors)

Tương tự SMOTE-Tomek Links, phương pháp kết hợp SMOTE và làm sạch dữ liệu bằng cách loại bỏ các điểm dữ liệu nhiễu hoặc không phù hợp dựa trên nguyên tắc lân cận

Nguyên lý hoạt động:

* Áp dụng SMOTE để tạo ra các mẫu tổng hợp mới dựa trên nội suy giữa các điểm gần kề

Áp dụng ENN:

* Kiểm tra từng điểm dữ liệu: Với mỗi điểm dữ liệu trong tập kết hợp (bao gồm lớp đa số và lớp thiểu số), xác định k-lân cận gần nhất
* Loại bỏ dữ liệu không phù hợp: Một điểm dữ liệu sẽ bị loại bỏ nếu nhãn của nó không khớp với đa số nhãn của k-lân cận gần nhất.

Ưu điểm:

* Cải thiện hiệu suất: Giảm tác động của nhiễu và lỗi dữ liệu từ đó tăng độ chính xác và giảm tỷ lệ lỗi của mô hình.
* Phù hợp với dữ liệu phức tạp

Hạn chế:

* Tốn tài nguyên tính toán
* Loại bỏ nhiều điểm dữ liệu: có thể làm mất thông tin quan trọng nếu không được cấu hình đúng
* Phụ thuộc vào k

# CÁC PHƯƠNG PHÁP XỬ LÝ BÀI TOÁN

## Dataset

Nhóm em sử dụng dữ liệu **Credit Card Fraud Detection Dataset** từ Kaggle để phát hiện gian lận thẻ tín dụng.

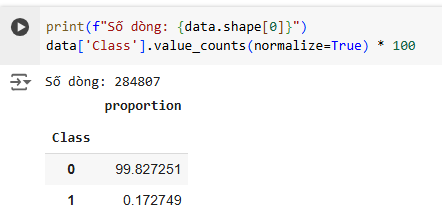
## Phân tích dữ liệu:

Mô phỏng các giao dịch thẻ tín dụng hợp pháp và gian lận gồm 284.807 dòng và có 31 thuộc tính.

Ý nghĩa các thuộc tính:

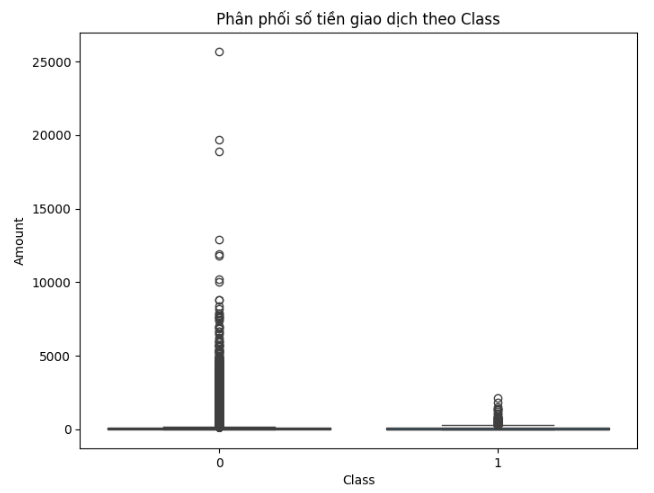
* Time: Thời gian giao dịch (tính bằng giây kể từ giao dịch đầu tiên)
* V1-V28: Các thuộc tính kiểu số thực được ẩn danh đại diện cho các thuộc tính giao dịch khác nhau (ví dụ: thời gian, địa điểm, v.v.) thông qua kỹ thuật PCA.
* Amount: Số tiền giao dịch với kiểu số thực.
* Class: Kiểu nhị phân cho biết giao dịch là gian lận (1) hay không (0)

Phân tích dữ liệu: phân tích các thuộc tính cần xử lý

* Class:
  + Dữ liệu gian lận có 492 dòng với 0.17%
  + Dữ liệu hợp lệ có 284.315 dòng với 99.83%

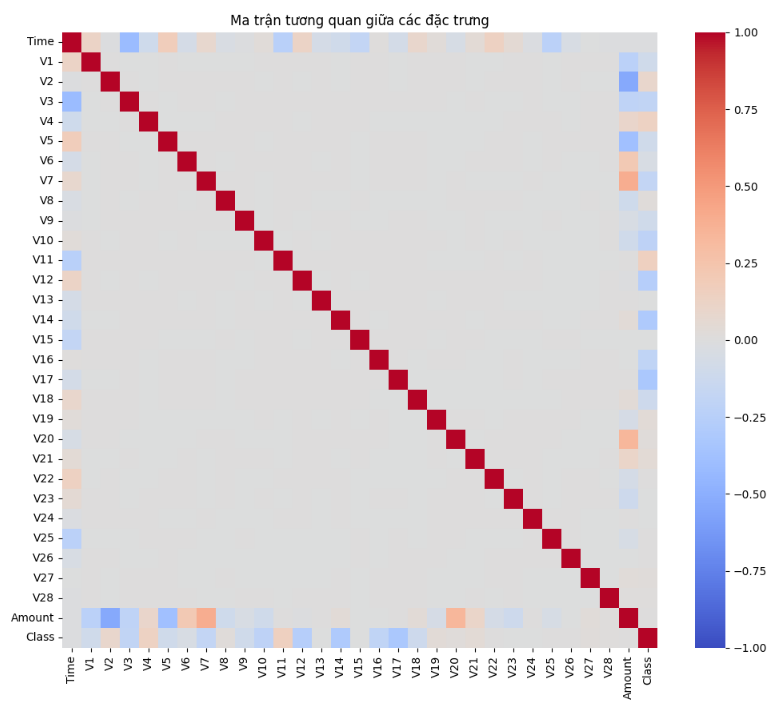
Hình 2‑1 Phân tích Class

* Amount: Phân phối lệch, giao dịch gian lận và số tiền thấp



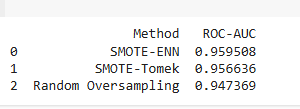
Hình 2‑2 Phân phối số tiền giao dịch theo Amount

* Sử dụng ma trận tương quan để kiểm tra mối quan hệ giữa các đặc trưng => Kết quả các cột V có liên quan nhẹ với Class

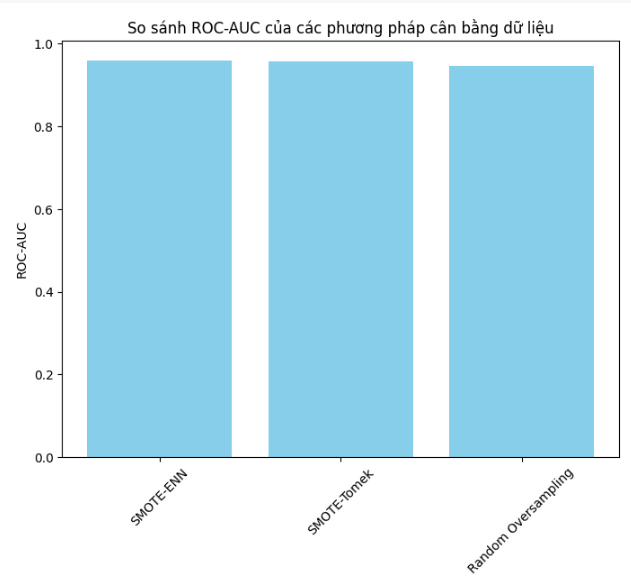
Tổng kết:

Hình 2‑3 Tương quan giữa các đặc trưng

* Thuộc tính phân loại Class: cần cân bằng dữ liệu
* Thuộc tính Amount: là thuộc tính liên tục nhưng phân phối của nó bị lệch: Giao dịch gian lận tập trung ở số tiền thấp nên cần chuẩn hóa dữ liệu

****Kết quả sau khi cân bằng: Phương pháp SMOTE-ENN có kết quả tốt nhất với 96%

Hình 2‑4 So sánh thuật toán

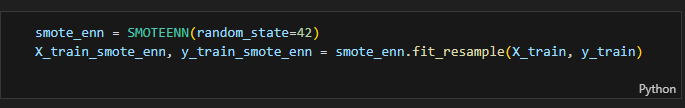
****

Hình 2‑5 Biểu đồ so sánh các phương pháp cân bằng

# CÁC MÔ HÌNH ĐỀ XUẤT

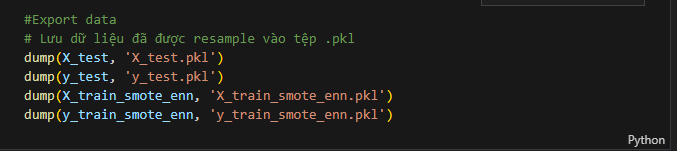
## Bộ dữ liệu huấn luyện

Như đã xử lý dữ liệu bên trên, ta thấy phương pháp SMOTE-ENN đạt được hiệu quả cao nhất với ~96%. Ta quyết định sử dụng bộ dữ liệu sau khi cân bằng của phương pháp này làm bộ dữ liệu huấn luyện cho các thuật toán học máy và học sâu.



Hình 3‑1 Tạo ra bộ dữ liệu huấn luyện

Tạo ra X\_train và y\_train cho các mô hình huấn luyện bằng phương pháp Smote-enn. Tuy nhiên để thuận tiện hơn trong quá trình thực hiện. Lưu các bộ dữ liệu huấn luyện và kiểm tra về dạng “.pkl” để sử dụng lại các giá trị này.



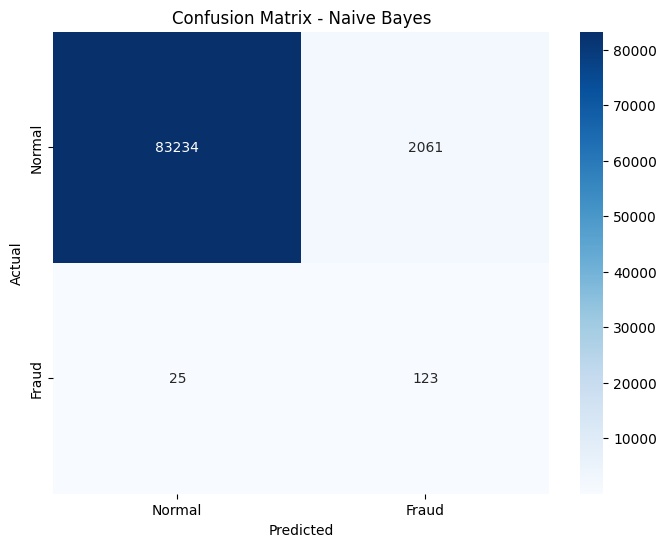
Hình 3‑2 Xuất bộ dữ liệu huấn luyện và kiểm tra ra .pkl

## Các mô hình học máy cơ bản

### Đánh giá các mô hình

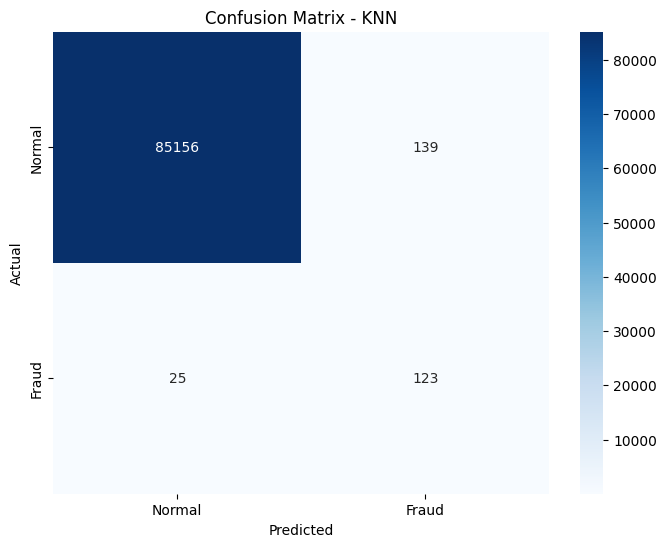
Các mô hình được sử dụng để phát hiện giao dịch gian lận:

1. Naive Bayes (NB): Mô hình phân loại dựa trên định lý Bayes với giả định rằng các đặc trưng là độc lập.
2. K-Nearest Neighbors (KNN): Mô hình phân loại dựa trên khoảng cách, phân loại dựa trên k điểm gần nhất trong không gian đặc trưng.
3. Logistic Regression (LR): Mô hình hồi quy logistic dùng để phân loại nhị phân, mô hình hóa xác suất của một lớp bằng hàm sigmoid.

Naïve Bayes:

Hình 3‑3 Confusion Matrix Naïve Bayes

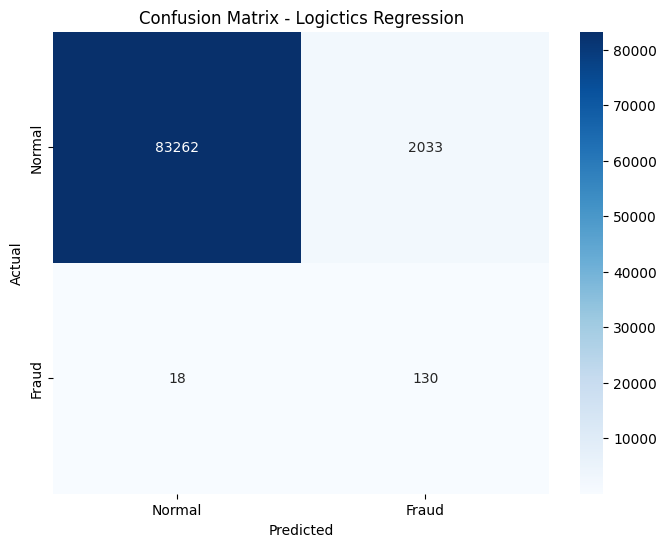
ROC-AUC Score: 0.9547

**KNN:**

Hình 3‑4 Confusion matrix KNN

ROC-AUC Score: 0.9149

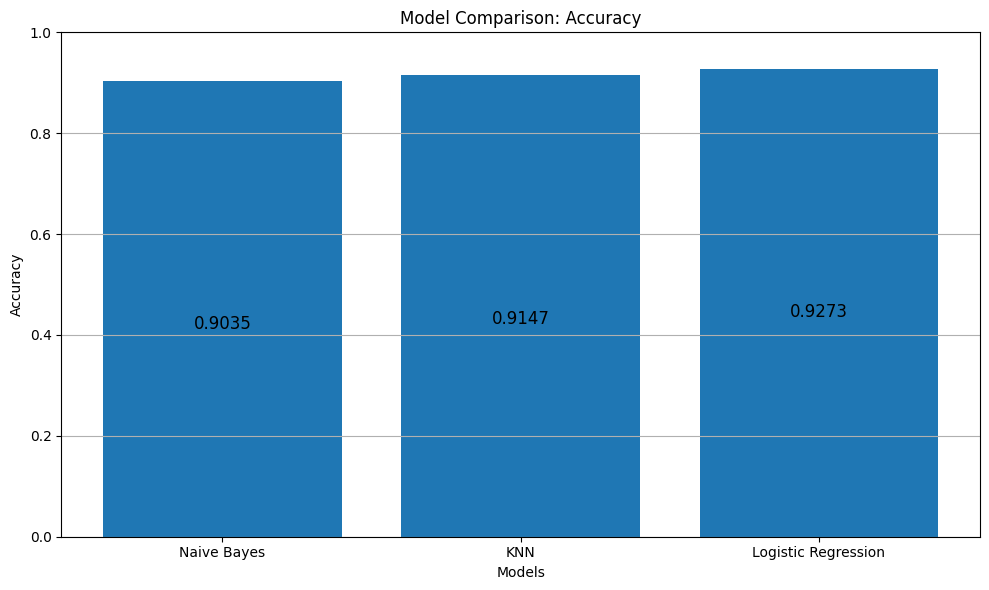
**Logictics Regression:**

****

Hình 3‑5 Confusion matrix Logictics Regression

ROC-AUC Score: 0.9663

So sánh 3 thuật toán:



Hình 3‑6 Compairison Roc-accuracy

### Đánh giá mô hình:

Hiệu Suất Mô Hình:

* Naive Bayes (NB):
  + ROC-AUC Score: 0.9547
  + Nhận xét: Mô hình này thể hiện khả năng phân loại tốt, nhưng phụ thuộc vào giả định về tính độc lập của các đặc trưng. Nó phù hợp cho các bài toán phân loại văn bản và có thể là một lựa chọn tốt trong trường hợp dữ liệu không quá phức tạp.
* K-Nearest Neighbors (KNN):
  + ROC-AUC Score: 0.9149
  + Nhận xét: Mô hình KNN có hiệu suất thấp hơn so với NB và Logistic Regression. KNN có thể không hiệu quả trong các tập dữ liệu lớn và có thể bị ảnh hưởng bởi noise trong dữ liệu. Mô hình này cần điều chỉnh siêu tham số và có thể cải thiện bằng cách sử dụng kỹ thuật giảm chiều.
* Logistic Regression (LR):
  + ROC-AUC Score: 0.9663
  + Nhận xét: Đây là mô hình có hiệu suất tốt nhất, cho thấy khả năng phân loại chính xác cao. Logistic Regression cũng dễ hiểu và có khả năng giải thích các yếu tố ảnh hưởng, điều này rất quan trọng trong lĩnh vực tài chính.

### Đề xuất cải thiện:

Tối Ưu Hóa Mô Hình:

* Cân nhắc việc kết hợp các mô hình (ensemble) để tận dụng sức mạnh của từng mô hình, có thể sử dụng Voting Classifier hoặc Stacking.
* Tiến hành tối ưu hóa siêu tham số cho KNN và Logistic Regression để nâng cao hiệu suất.

Xử Lý Dữ Liệu:

* Đảm bảo dữ liệu đầu vào phong phú và đa dạng. Thêm các đặc trưng mới có thể cải thiện khả năng phân loại.
* Sử dụng các phương pháp xử lý mất cân bằng dữ liệu (như SMOTE) nếu có sự chênh lệch lớn giữa số lượng giao dịch bình thường và gian lận.

Đào Tạo và Giám Sát:

* Định kỳ đào tạo lại mô hình với dữ liệu mới để phản ánh đúng tình hình thực tế.
* Phân tích các trường hợp dự đoán sai để hiểu nguyên nhân và điều chỉnh mô hình khi cần thiết.

Cải Thiện Trực Quan Hóa:

* Sử dụng các biểu đồ trực quan hơn để trình bày kết quả, như biểu đồ cột với giá trị hiển thị rõ ràng, giúp dễ dàng so sánh hiệu suất giữa các mô hình.

### Kết luận

Từ những phân tích và đánh giá trên, Logistic Regression là lựa chọn tốt nhất cho việc phát hiện gian lận trong giao dịch với hiệu suất cao nhất. Tuy nhiên, việc kết hợp nhiều mô hình và cải thiện dữ liệu sẽ giúp tăng cường độ chính xác và hiệu quả của hệ thống phát hiện gian lận trong tương lai.

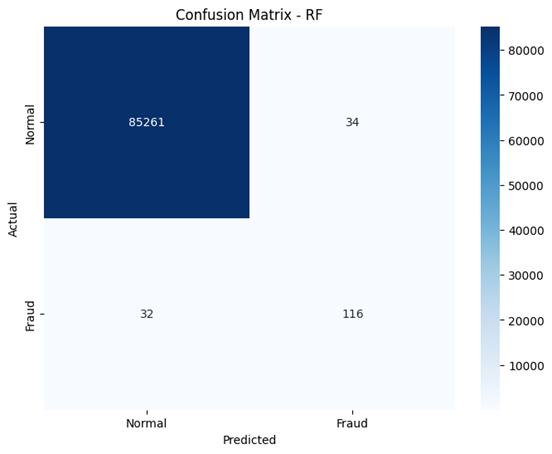
## Các mô hình học sâu và ensemble:

### Đánh giá các mô hình

Các mô hình được sử dụng để phát hiện giao dịch gian lận:

1. **Random Forest (RF):** Mô hình cây quyết định kết hợp.
2. **Artificial Neural Network (ANN):** Mạng nơ-ron nhân tạo với nhiều tầng ẩn.
3. **Long Short-Term Memory (LSTM):** Mạng nơ-ron hồi tiếp phù hợp với dữ liệu tuần tự

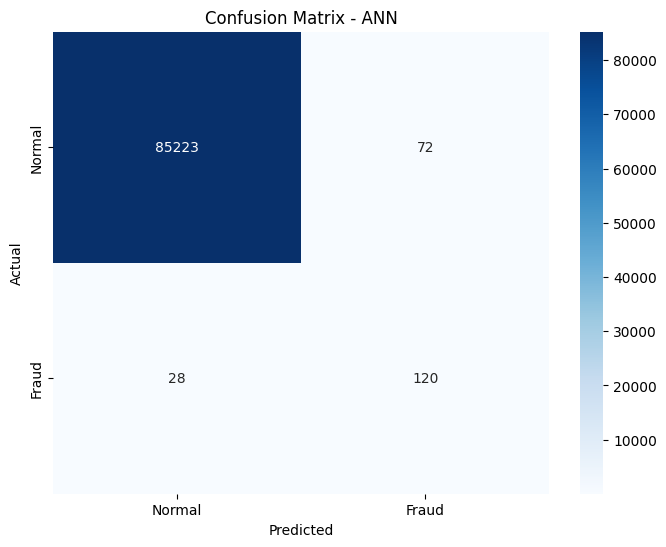
**Random Forest:**

****

Hình 3‑7 Confusion Matrix – Random Forest

ROC - AUC score: 0.9716

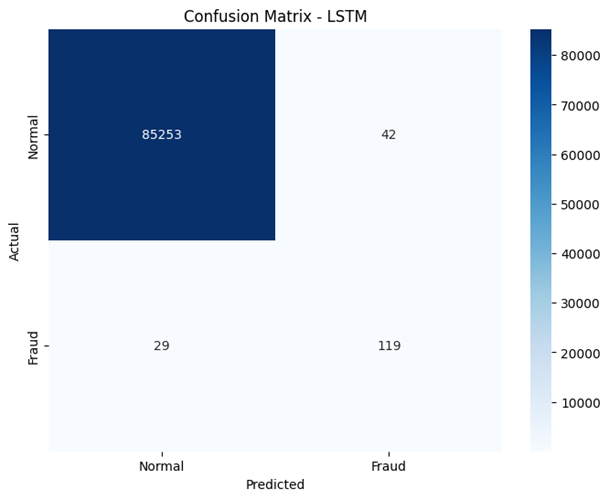
**ANN:**



Hình 3‑8 Confusion Matrix – ANN

ROC – AUC Score : 0.9583

**LSTM:**

****

Hình 3‑9 Confusion Matrix – LSTM

ROC-AUC Score: 0.9657

### Đề xuất kinh doanh

#### Lựa chọn mô hình:

Với Random Forest (ROC-AUC: 0.9716) là mô hình tốt nhất trong các tiêu chí đánh giá tổng thể, chúng tôi đề xuất triển khai mô hình này cho hệ thống phát hiện gian lận.

Mô hình LSTM có thể là lựa chọn bổ sung nhờ khả năng học dữ liệu tuần tự, đặc biệt nếu có thêm dữ liệu thời gian thực.

#### Tối ưu:

**Tăng độ chính xác của mô hình:**

Để cải thiện độ chính xác của mô hình, chúng tôi đề xuất thực hiện các bước sau:

1. Thử nghiệm thêm phương pháp cân bằng dữ liệu khác.
2. Tinh chỉnh tham số mô hình (Hyperparameter Tuning): Sử dụng kỹ thuật Grid Search hoặc Randomized Search để tìm tổ hợp tham số tối ưu cho từng mô hình.
3. Tích hợp thêm đặc trưng (Feature Engineering): Phân tích các đặc trưng hiện tại và thử tạo thêm các đặc trưng mới bằng cách:

Kết hợp các đặc trưng hiện có (ví dụ: tỷ lệ giữa hai giá trị hoặc biến đổi log).

Sử dụng PCA (Principal Component Analysis) hoặc t-SNE để trích xuất thêm các thông tin quan trọng từ dữ liệu.

**Giảm tỷ lệ dự đoán sai (False Positives)**

Tỷ lệ dự đoán sai (False Positives - các giao dịch hợp lệ bị xác định nhầm là gian lận) cần được giảm thiểu vì ảnh hưởng trực tiếp đến trải nghiệm khách hàng và hiệu quả vận hành. Để làm điều này:

Mặc định, mô hình sử dụng ngưỡng 0.5 để quyết định dự đoán lớp nào (gian lận hoặc không gian lận). Thay vì sử dụng ngưỡng này, có thể:

* Chạy phân tích Precision-Recall Curve để xác định điểm cân bằng giữa Precision và Recall.
* Chọn ngưỡng phù hợp giúp giảm False Positives mà vẫn giữ Recall ở mức chấp nhận được.

#### Hỗ trợ quyết định:

Trình bày kết quả dự đoán trên giao diện trực quan hóa (ví dụ: Tableau, Power BI) để dễ dàng phân tích.

Cung cấp danh sách giao dịch đáng nghi nhất cùng xác suất phát hiện để đội ngũ vận hành tập trung kiểm tra.

# KẾT LUẬN

Trong báo cáo này, chúng em đã nêu bật tầm quan trọng của việc phát hiện gian lận trong giao dịch. Chúng em nhận thấy rằng việc áp dụng công nghệ hiện đại không chỉ giúp phát hiện nhanh chóng các hành vi gian lận mà còn bảo vệ doanh nghiệp khỏi tổn thất tài chính.

Ngoài công nghệ, việc xây dựng một văn hóa doanh nghiệp nhạy bén với rủi ro cũng rất cần thiết. Chúng em khuyến khích doanh nghiệp đào tạo nhân viên để nhận biết dấu hiệu gian lận và tạo môi trường làm việc an toàn.

Cuối cùng, sự phối hợp giữa các bộ phận trong doanh nghiệp là yếu tố quan trọng để ngăn chặn gian lận hiệu quả. Đầu tư vào công nghệ mới và cải thiện quy trình quản lý rủi ro sẽ giúp doanh nghiệp không chỉ bảo vệ tài sản mà còn nâng cao uy tín trên thị trường.

Tóm lại, báo cáo này nhấn mạnh sự cần thiết của một hệ thống phòng ngừa gian lận toàn diện và bền vững trong môi trường kinh doanh ngày nay.

# Tài liệu tham khảo

Tiếng việt:

Blog, K. (2020). *Mất cân bằng dữ liệu (imbalanced dataset).* Được truy lục từ https://phamdinhkhanh.github.io/2020/02/17/ImbalancedData.html

Tiếng anh:

Brownlee, J. (2021). *SMOTE for Imbalanced Classification with Python.* Được truy lục từ https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/