**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**
**KHOA CÔNG NGHỆ THÔNG TIN**

**LẠI MINH PHÚ**
**HỒ MINH THANH TÀI**

# PHÁT HIỆN GIAN LẬN TÀI CHÍNH DÙNG CÁC KỸ THUẬT PHÂN TÍCH BẤT THƯỜNG

**KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**
**CHƯƠNG TRÌNH CHẤT LƯỢNG CAO**

**TP. HCM, 2024**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**
**KHOA CÔNG NGHỆ THÔNG TIN**

**LẠI MINH PHÚ – 20127593**
**HỒ MINH THANH TÀI – 20127068**

# PHÁT HIỆN GIAN LẬN TÀI CHÍNH DÙNG CÁC KỸ THUẬT PHÂN TÍCH BẤT THƯỜNG

**KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**
**CHƯƠNG TRÌNH CHẤT LƯỢNG CAO**

**GIÁO VIÊN HƯỚNG DẪN**
**ThS. LÊ PHÚC LỮ**
**ThS. NGUYỄN VĂN QUANG HUY**

**TP. HCM, 2024**

# LỜI CẢM ƠN

Trước hết, chúng em xin gửi lời cảm ơn tới Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh đã tạo nhiều điều kiện về thời gian và tài liệu cho khóa luận này.

Hơn nữa, chúng em xin gửi lời cảm ơn chân thành đến tất cả các thầy cô tại Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh đã trang bị những kiến thức quý báu trong bốn năm qua để chúng em có thể hoàn thành khóa luận này. Chúng em không thể biết ơn hơn về tất cả những bài học mà thầy cô đã truyền trao và những đóng góp mà thầy cô đã cống hiến.

Với toàn bộ tâm huyết và cố gắng trong suốt thời gian qua, cùng với những kiến thức đã học được tại trường và một số kinh nghiệm tích lũy trong quá trình thực tập đã giúp chúng em hoàn thành khóa luận tốt nghiệp này. Tuy nhiên, do kiến thức và kinh nghiệm áp dụng vào thực tế còn hạn chế nên chúng em khó tránh khỏi những sai sót. Vì vậy, chúng em rất mong nhận được sự thông cảm và đóng góp ý kiến của quý thầy cô và các bạn.

Đặc biệt chúng em xin bày tỏ lòng biết ơn sâu sắc tới ThS. Lê Phúc Lữ và ThS. Nguyễn Văn Quang Huy – giảng viên hướng dẫn của chúng em vì đã chia sẻ, giúp đỡ, động viên và truyền cảm hứng cho chúng em.

Chúng em xin chân thành cảm ơn!

# MỤC LỤC

# MỤC LỤC HÌNH ẢNH

# MỤC LỤC BẢNG BIỂU

# DANH MỤC CÁC THUẬT NGỮ

| STT | Thuật ngữ Tiếng Anh | Thuật ngữ Tiếng Việt |
|---|---|---|
| 1 | Financial Fraud | Gian lận tài chính |
| 2 | Anomaly Detection | Phát hiện bất thường |
| 3 | MRI (Magnetic Reasonance Imaging) | Hình ảnh cộng hưởng từ |
| 4 | SMOTE (Synthetic Minority Oversampling Technique) | Phương pháp tăng cường số mẫu lớp thiểu số |
| 5 | ROSE (Random oversampling) | Phương pháp tăng cường ngẫu nhiên số mẫu lớp thiểu số |
| 6 | RUS (Random Undersampling) | Phương pháp giảm thiểu ngẫu nhiên số mẫu lớp đa số |
| 7 | ADASYN (Adaptive Synthetic Sampling) | Phương pháp lấy mẫu thích ứng nhân tạo |
| 8 | ACFE (Association of Certified Fraud Examiners) | Hiệp hội Các nhà điều tra gian lận được công chứng |
| 9 | Point Anomalies | Điểm bất thường |
| 10 | Contextual Anomalies | Bất thường ngữ cảnh |
| 11 | Collective Anomalies | Bất thường tập hợp |
| 12 | Univariate | Đơn biến |
| 13 | Bivariate | Hai biến |
| 14 | Multivariate | Đa biến |
| 15 | Non-graphical analysis | Phân tích phi đồ họa |
| 16 | Graphical analysis | Phân tích đồ họa |
| 17 | Parametric Tests | Kiểm định tham số |
| 18 | Non-Parametric Tests | Kiểm định phi tham số |
| 19 | ANOVA (Analysis of Variance) | Phân tích phương sai |
| 20 | Covariance | Hiệp phương sai |

| | | |
|---|---|---|
| 21 | Machine Learning | Học máy |
| 22 | Supervised Learning | Học có giám sát |
| 23 | Unsupervised Learning | Học không giám sát |
| 24 | Semi-Supervised Learning | Học bán giám sát |
| 25 | Ensemble Learning | Học kết hợp |
| 26 | Classification | Phân loại |
| 27 | Regression | Hồi quy |
| 28 | Overfitting | Quá khớp |
| 29 | Underfitting | Vị khớp |
| 30 | EDA (Exploratory Data Analysis) | Phân tích khám phá dữ liệu |
| 31 | Central tendency | Mức độ tập trung dữ liệu |
| 32 | Mean | Giá trị trung bình |
| 33 | Median | Giá trị trung vị |
| 34 | Mode | Giá trị nhất suất |
| 35 | Spread | Khoảng dữ liệu |
| 36 | Variance | Phương sai |
| 37 | Standard Deviation | Độ lệch chuẩn |
| 38 | Stem-and-leaf plots | Biểu đồ thân và lá |
| 39 | Biểu đồ tần suất | Biểu đồ tần suất |
| 40 | Boxplots | Biểu đồ hộp |
| 41 | Scatter plot | Biểu đồ phân tán |
| 42 | Pair plot | Biểu đồ cặp |
| 43 | Heat map | Bản đồ nhiệt |
| 44 | DBSCAN (Density-based spatial clustering of applications with noise) | Phương pháp phân cụm dựa trên mật độ |
| 45 | LOF (Local Outlier Factor) | Nhân tố ngoại lai cục bộ |
| 46 | ANN (Artificial Neural Network) | Mạng nơ-ron nhân tạo |
| 47 | CNN (Convolution Neural Network) | Mạng nơ-ron tích chập |

| | | |
|---|---|---|
| 48 | RNN (Recurrent Neural Network) | Mạng nơ-ron tuần tự |
| 49 | KNN (K-nearest neighbors) | K láng giềng gần nhất |
| 50 | LR (Logistic Regression) | Mô hình hồi quy Logistic |
| 51 | DT (Decision Trees) and RF (Random Forests) | Cây quyết định và rừng ngẫu nhiên |
| 52 | Boostrap Aggregation / Bagging | Phương pháp bỏ túi |
| 53 | Tăng cường | Phương pháp tăng cường |
| 54 | Stacking (Stacked Generalization) | Phương pháp ngăn xếp |
| 55 | GBM (Tăng cường độ dốc) | Thuật toán tăng cường độ dốc |
| 56 | XGBoost (Extreme Tăng cường độ dốc) | Thuật toán tăng cường độ dốc cực đại |
| 57 | CatBoost (Categorical Tăng cường) | Thuật toán tăng cường phân loại |
| 58 | AdaBoost (Adaptive Tăng cường) | Thuật toán tăng cường thích ứng |
| 59 | LightGBM (Light Tăng cường độ dốc Machine) | Thuật toán học máy tăng cường độ dốc ánh sáng |
| 60 | GOSS (Gradient-based One-Side Sampling) | Lấy mẫu một phía dựa trên độ dốc |
| 61 | GPU (Graphic Processing Unit) | Bộ xử lý đồ họa |
| 62 | Hyperparameter Tuning / Hyperparameter Optimization | Tinh chỉnh/ tối ưu hóa siêu tham số |
| 63 | Class imbalance | Mất cân bằng lớp |
| 64 | Majority Class | Lớp đa số hay lớp “âm” |
| 65 | Minority Class | Lớp thiểu số hay lớp “dương” |
| 66 | Support Vector Machines | Máy véc-tơ hỗ trợ |
| 67 | Neural Networks and Deep Learning | Mạng nơ-ron và học sâu |
| 68 | Outlier Detection | Phát hiện ngoại lệ |
| 69 | PCA (Principal Component Analysis) | Phân tích thành phần chính |
| 70 | Precision | Độ chuẩn xác |
| 71 | Recall | Độ phủ |

| | | |
|---|---|---|
| 72 | F1 Score | Trung bình điều hòa giữa *độ chuẩn xác* và *độ phủ* |
| 73 | Accuracy | Độ chính xác |
| 74 | ROC – AUC / AUC-ROC (Area Under Curve Reciever Operating Charactertics | Diện tích dưới đường cong ROC |
| 75 | PR – AUC / AUC – PR (Area Under Precision – Recall Curve) | Diện tích dưới đường cong Precision-Recall |
| 76 | TPR (True Positive Rate) | Đúng dương tính |
| 77 | FPR (False Positive Rate) | Sai dương tính |
| 78 | Confusion Matrix | Ma trận nhầm lẫn |
| 79 | Correlation Matrix | Ma trận tương quan |
| 80 | TP (True Positive) | Dương tính thực |
| 81 | TN (True Negative) | Âm tính thực |
| 82 | FP (False Positive) | Dương tính giả |
| 83 | FN (False Negative) | Âm tính giả |
| 84 | MCC (Matthews Correlation Coefficient) | Hệ số tương quan Matthews |
| 85 | Gradient descent | suy giảm độ dốc |
| 86 | Loss function | hàm mất mát |
| 87 | Ordered boosting | Tăng cường theo trình tự |
| 88 | Parallel learning | học song song |
| 89 | Distributed learning | phân tán |
| 90 | Histogram-based learning | Học dựa trên biểu đồ tần suất |
| 91 | Learning rate | Tốc độ học |
| 92 | Bin | Khoảng giá trị |
| 93 | Leaf-wise | Theo chiều lá |
| 94 | GA (Genetic Algorithm) | Thuật toán di truyền |
| 95 | Nominal data | Dữ liệu định danh |

| | | |
|---|---|---|
| 96 | Ordinal data | Dữ liệu thứ bậc |
| 97 | Interval data | Dữ liệu khoảng cách |
| 98 | Level-wise | Theo cấp bậc |
| 99 | Isolation forest | Rừng cô lập |
| 100 | Elliptic envelope | Đường cong E-líp |

# TÓM TẮT KHÓA LUẬN

Gian lận tài chính đang trở thành một vấn đề ngày càng nghiêm trọng đối với các doanh nghiệp, chính phủ và các nhà đầu tư trên khắp thế giới. Gian lận tài chính không chỉ đe dọa đến niềm tin của công chúng, mà còn gây hậu quả nặng nề cho các bên liên quan. Vì vậy, việc phát hiện gian lận tài chính đóng vai trò quan trọng trong việc bảo vệ uy tín của doanh nghiệp và tạo ra môi trường kinh doanh minh bạch và công bằng.

Phương pháp phát hiện bất thường là một công cụ quan trọng giúp xác định các biểu hiện bất thường trong dữ liệu tài chính, từ đó giúp nhận diện sớm các hành vi gian lận và đưa ra biện pháp phòng ngừa hiệu quả. Đó là chủ đề chính của khóa luận này.

Trong khóa luận này, chúng tôi thực hiện nghiên cứu và tổng hợp các cơ sở lý thuyết về phát hiện bất thường, phân loại các bất thường. Bên cạnh đó, khóa luận còn đề cập đến các kỹ thuật thống kê, là cơ sở nền tảng trong nhiệm vụ phát hiện bất thường. Sau đó, khóa luận điều hướng đến phần trọng tâm, giới thiệu về học máy trong phát hiện bất thường và đi sâu vào bốn mô hình: Random Forest, XGBoost, CatBoost và LightGBM. Chúng tôi thực hiện việc phát hiện bất thường qua bốn mô hình này trên tập dữ liệu thực tế chứa các giao dịch được thực hiện bằng thẻ tín dụng vào tháng 9 năm 2013 bởi chủ thẻ ở Châu Âu. Sau khi đã có kết quả, bằng các phương pháp đánh giá mô hình, khóa luận đi đến kết luận và nhận định lại những mặt còn hạn chế, từ đó đưa ra những hướng phát triển trong tương lai.

Nghiên cứu nhằm cung cấp hiểu biết toàn diện về ưu điểm và hạn chế của mỗi mô hình trong việc phát hiện bất thường trong bối cảnh phức tạp của dữ liệu tài chính. Mong muốn lớn nhất của chúng tôi đối với khóa luận là có thể đem các kết quả từ nghiên cứu này đóng góp vào việc phát triển các hệ thống phát hiện bất thường, nâng cao khả năng xác định các khuyết điểm và rủi ro tiềm ẩn một cách kịp thời.

# CHƯƠNG 1. GIỚI THIỆU

## 1.1. Mở đầu

### 1.1.1. Lí do chọn đề tài

Trong sự phát triển không ngừng của thị trường tài chính, việc bảo vệ sự trung thực và độ tin cậy của hệ thống tài chính trở nên càng quan trọng hơn bao giờ hết. Chính sách và công nghệ ngày càng tiến bộ, nhưng cùng với đó là sự gia tăng của các hình thức gian lận tài chính tinh vi và phức tạp. Vì vậy, việc lựa chọn đề tài "Phát hiện gian lận tài chính dùng các kỹ thuật phân tích bất thường" của chúng tôi không chỉ là sự quyết định đơn thuần mà còn là một nhiệm vụ nhằm cống hiến vào các nghiên cứu giải quyết vấn đề nêu trên.

Trong việc chống lại gian lận tài chính, các kỹ thuật phát hiện bất thường đóng vai trò quan trọng trong việc bảo vệ hệ thống tài chính. Chúng không chỉ dựa vào các mô hình truyền thống mà còn khám phá những bất thường không ngờ trong dữ liệu, từ đó nâng cao khả năng phát hiện và ngăn chặn các hành vi gian lận. Sự kết hợp giữa sức mạnh của khoa học dữ liệu và sự sáng tạo trong công nghệ là chìa khóa để mở ra cánh cửa chống lại gian lận trong thị trường tài chính.

Đề tài này không chỉ hướng đến việc nghiên cứu một lĩnh vực mới mẻ và hấp dẫn mà còn là sứ mệnh của việc bảo vệ sự công bằng và minh bạch trong môi trường tài chính. Qua đó, chúng tôi mong muốn đóng góp vào sự phát triển của các kỹ thuật phát hiện bất thường nhằm phát hiện gian lận tài chính, từ đó góp phần bảo vệ và xây dựng một thế giới tài chính trong sạch và công bằng hơn.

### 1.1.2. Mục đích

Mục đích của đề tài này là tìm hiểu về phát hiện gian lận tài chính bằng cách sử dụng các kỹ thuật phân tích bất thường. Nghiên cứu này nhằm đưa ra một cái nhìn toàn diện về phát hiện bất thường, bao gồm phân tích các nghiên cứu đã thực hiện trước

đó, cũng như tổng hợp các thuật toán phổ biến được sử dụng trong cộng đồng công nghệ toàn cầu để giải quyết vấn đề này.

### 1.1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng và phạm vi nghiên cứu của khóa luận này sẽ bao gồm:

*1. Xác định và phân loại các phương pháp phát hiện bất thường:* Nghiên cứu sẽ bao gồm việc xem xét các phương pháp phát hiện bất thường hiện có, tổng hợp cơ sở lý thuyết, xác định tính chất và phân loại chúng. Sự so sánh giữa các phương pháp này sẽ được thực hiện để đánh giá tính hiệu quả và tính ứng dụng của chúng trong việc phát hiện bất thường.

*2. Nghiên cứu các kỹ thuật phát hiện bất thường:* Phạm vi nghiên cứu sẽ bao gồm khảo sát và phân tích các kỹ thuật phát hiện bất thường, đồng thời đánh giá sâu hơn về cách các kỹ thuật này hoạt động và cách chúng được áp dụng để phát hiện gian lận tài chính.

*3. Đánh giá và so sánh hiệu suất:* Phạm vi nghiên cứu cũng sẽ bao gồm việc đánh giá và so sánh hiệu suất của các kỹ thuật sau khi đã xây dựng mô hình.

*4. Khảo sát về tính ứng dụng và thực tiễn:* Cuối cùng, nghiên cứu sẽ tập trung vào khảo sát về tính ứng dụng và thực tiễn của các kỹ thuật phát hiện gian lận tài chính trong các môi trường thực tế của ngành tài chính, bao gồm cả việc xem xét các thách thức và cơ hội trong việc triển khai chúng.

Các công cụ và ngôn ngữ lập trình sẽ được sử dụng trong nghiên cứu này bao gồm Python, một ngôn ngữ lập trình phổ biến trong lĩnh vực khoa học dữ liệu, và các thư viện như Pandas, Numpy, và Scikit-learn để xử lý và phân tích dữ liệu. Ngoài ra, chúng tôi cũng sẽ sử dụng các công cụ như Jupyter Notebook để lập trình và thực thi, Matplotlib và Seaborn để trực quan hóa dữ liệu.

## 1.2. Tổng quan

Gian lận tài chính là một vấn đề rộng lớn và hiện hữu trong ngành tài chính, dẫn đến tổn thất đáng kể cho khách hàng và các tổ chức doanh nghiệp. Với sự mở rộng và tiến bộ nhanh chóng của công nghệ hiện đại như Internet, các thiết bị như điện thoại, máy tính xách tay và phương tiện truyền thông xã hội, các hoạt động lừa đảo cũng dần gia tăng. *Gian lận tài chính*, theo định nghĩa của Hiệp hội Các nhà điều tra gian lận được công chứng [1] là bất kỳ hành động cố ý nhằm tước đoạt tài sản hoặc tiền bạc của người khác bằng thủ đoạn gian lận, lừa dối hoặc các biện pháp không công bằng khác. Điều này đã dẫn đến tổn thất hàng tỷ đô la cho các hoạt động kinh doanh, thúc đẩy những nỗ lực sâu rộng hướng tới việc khám phá các kỹ thuật phân tích bất thường để phát hiện gian lận. Vì thế, việc phát triển các hệ thống phát hiện gian lận tài chính là một nhu cầu cấp bách.

Phát hiện bất thường là kỹ thuật được sử dụng để xác định các mô hình không bình thường và không tuân theo hành vi mong đợi, nó có nhiều ứng dụng trong kinh doanh, từ xác định các mô hình bất thường trong lưu lượng mạng báo hiệu về một cuộc tấn công [2] đến giám sát sức khỏe hệ thống (phát hiện khối u ác tính trong cắt MRI) [3], và cuối cùng phát hiện gian lận trong giao dịch thẻ tín dụng [4].

Các tổ chức và cá nhân trong lĩnh vực tài chính luôn phải đối mặt với nguy cơ từ các hành vi gian lận như lạm dụng thông tin, gian lận trong giao dịch hay báo cáo tài chính. Trong bối cảnh này, việc sử dụng các kỹ thuật phân tích bất thường để phát hiện và ngăn chặn các hành vi gian lận ấy là vô cùng quan trọng.

Ngoài ra, việc đánh giá các hướng nghiên cứu đã có của các tác giả trong và ngoài nước về phát hiện gian lận tài chính dùng các kỹ thuật phân tích bất thường cung cấp một cái nhìn tổng quan về tình trạng hiện tại của lĩnh vực này, bao gồm:

*1. Phân tích các phương pháp phát hiện bất thường:* Nhiều nghiên cứu đã tập trung vào phát triển và áp dụng các kỹ thuật bất thường để phát hiện gian lận tài chính. Các phương pháp này thường dựa trên việc xác định sự khác biệt giữa dữ liệu thực tế và

các mô hình dự đoán, nhằm phát hiện các điểm bất thường hoặc các sự kiện không thể giải thích được.

*2. Đánh giá hiệu suất và ứng dụng thực tế:* Một số nghiên cứu đã đánh giá hiệu suất của các phương pháp phát hiện gian lận trên các tập dữ liệu thực tế và trong các môi trường tài chính khác nhau. Tuy nhiên, vẫn còn nhiều thách thức trong việc áp dụng các kỹ thuật này vào các hệ thống thực tế do độ phức tạp và tính đa dạng của các hình thức gian lận tài chính.

*3. Vấn đề còn tồn tại:* Mặc dù đã có sự tiến bộ trong phát triển các kỹ thuật phát hiện bất thường, nhưng vẫn còn nhiều vấn đề cần được giải quyết. Các vấn đề này có thể bao gồm như hiệu suất không đảm bảo, độ phủ và độ chuẩn xác không cao, cũng như khả năng tích hợp vào các hệ thống tài chính hiện đại một cách linh hoạt và hiệu quả. Đồng thời, trong quá trình phát hiện gian lận tài chính, việc tích hợp dữ liệu từ nhiều nguồn khác nhau, ví dụ như giao dịch tài chính hay dữ liệu mạng xã hội, vẫn là một thách thức.

*4. Vấn đề cần tập trung nghiên cứu:* Đề tài cần tập trung vào các vấn đề như nâng cao hiệu suất và độ chính xác của các phương pháp phát hiện gian lận, tăng cường tính ứng dụng và thực tiễn trong các môi trường tài chính cụ thể, phát triển các phương pháp mới hoặc cải thiện các phương pháp hiện có để giải quyết các vấn đề cụ thể mà chúng đang gặp phải, cũng như tích hợp các phương pháp phát hiện gian lận vào hệ thống quản lý rủi ro và gian lận tài chính hiện đại.

## 1.3. Các công trình liên quan

Một số công trình liên quan đến nhiệm vụ phát hiện gian lận bằng các kỹ thuật phân tích bất thường. Cả ba công trình dưới đây đều sử dụng tập dữ liệu chứa các giao dịch được thực hiện bằng thẻ tín dụng vào tháng 9 năm 2013 bởi chủ thẻ ở Châu Âu.

*Công trình 1:* Emmanuel Ileberi và các cộng sự [5] đã thực hiện công trình mang tên *Phát hiện gian lận thẻ tín dụng dựa trên học máy sử dụng thuật toán di truyền để*

*chọn lựa đặc trưng* vào tháng 3 năm 2022. Ưu điểm nổi bật của công trình này là phương pháp lựa chọn đặc trưng dựa trên thuật toán di truyền (GA) kết hợp với các mô hình như Rừng ngẫu nhiên, Cây quyết định, Mạng nơ-ron nhân tạo, Naive Bayes (NB) và Hồi quy tuyến tính. Công trình này sử dụng các độ đo như độ chính xác, độ phủ, độ chuẩn xác và F1-score để đánh giá mô hình.

Sau khi triển khai thuật toán di truyền (GA) trên tập dữ liệu gian lận thẻ tín dụng, công trình đã thu được năm vector đặc trưng tối ưu ($v_1$ đến $v_5$) được hiển thị trong hình 1. Các vector này chứa tên các đặc trưng đại diện cho các thuộc tính tối ưu nhất sẽ được sử dụng để đánh giá hiệu quả của phương pháp đề xuất.

| Attribute vector | Vector length | Attribute list |
|---|---|---|
| $v_1$ | 18 | V1, V5, V7, V8, V11,V13, V14, V15, V16, V17, V18, V19, V20, V21, V22, V23, V24, Amount |
| $v_2$ | 9 | V1, V6, V13, V16, V17, V22, V23, V28, Amount |
| $v_3$ | 13 | V2, V11, V12, V13, V15, V16, V17, V18,V20, V21, V24, V26, Amount |
| $v_4$ | 9 | V2, V7, V10, V13,V15, V17, V19, V28, Amount |
| $v_5$ | 13 | Time, V1, V7, V8, V9, V11, V12, V14, V15, V22, V27, V28, Amount |

*Hình 1. Các véc-tơ đặc trưng tương ứng trong Công trình 1*

*Nguồn:* [5]

Kết quả thực nghiệm sử dụng các thuộc tính được lựa chọn bởi GA cho thấy GA-RF (sử dụng $v_5$) đạt được độ chính xác tối ưu là 99.98%. Ngoài ra, các mô hình khác như GA-DT cũng đạt được độ chính xác đáng kể là 99.92% sử dụng $v_1$.

| | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **ROC-AUC** |
|---|---|---|---|---|---|
| RF | 95.34% | 72.56% | 82.41% | 99.98% | 0.95 |
| DT | 75.22% | 75.22% | 75.22% | 99.92% | 0.88 |

*Bảng 1. Bảng kết quả các mô hình từ Công trình 1*

*Nguồn:* [5]

*Công trình 2:* Vikas Thammanna Gowda [6] và các cộng sự đã thực hiện công trình tên *Phát hiện gian lận thẻ tín dụng sử dụng các mô hình học có giám sát và học không giám sát* vào tháng 7 năm 2021. Phương pháp nổi bật của công trình này là đánh giá một tập dữ liệu mất cân bằng sử dụng các thuật toán học có giám sát và không giám sát, và xác định thuật toán tốt nhất để phát hiện gian lận thẻ tín dụng. Công trình này sử dụng độ chính xác, độ phủ, độ chuẩn xác, F1-score và ROC để đánh giá mô hình. Kết quả công trình cho thấy thuật toán K-means (độ phủ 91.91%) là tốt nhất trong số các thuật toán học không giám sát và máy vector hỗ trợ (độ chính xác 98.49%) hoạt động tốt trong tất cả các thuật toán đã sử dụng.

| Model | F1 score | Accuracy | Recall | Precision |
|---|---|---|---|---|
| **Logistic Regression** | **0.9326** | **0.9703** | **0.909** | **0.9574** |
| **SVM** | **0.9479** | **0.9849** | **0.9191** | **0.9785** |
| **Random Forest** | **0.9435** | **0.9715** | **0.9293** | **0.9583** |
| **Isolation Forest** | **0.67** | **0.9977** | **0.67** | **0.67** |
| **Local Outlier Factor** | **0.25** | **0.9967** | **0.25** | **0.25** |
| **K-means** | **0.926** | **0.9982** | **0.879** | **0.9798** |

*Hình 2. Kết quả của các thuật toán được thực hiện trong Công trình 2*

*Nguồn:* [6]

Tuy nhiên, công trình này vẫn chưa có một đánh giá tổng quan về tập dữ liệu trước khi tiến hành áp dụng các kỹ thuật phát hiện bất thường. Vì sự mất cân bằng của tập dữ liệu rất lớn, việc không xử lý tập dữ liệu và sử dụng các độ đo như độ chính xác vẫn chưa đem lại độ tin cậy về hiệu suất của mô hình.

*Công trình 3:* Aaron Rosenbaum [7] đã thực hiện công trình *Phát hiện gian lận thẻ tín dụng với Học Máy* vào năm 2019. Ưu điểm nổi bật của công trình này là khám phá và đánh giá các phương pháp để giải quyết vấn đề mất cân bằng của dữ liệu. Đặc biệt khi các đặc trưng đã được mã hóa do những mối quan tâm về quyền riêng tư, họ không thể giải thích được mối quan hệ giữa hầu hết các đặc trưng trong tập dữ liệu. Đối với hầu hết các mô hình, công trình đã sử dụng tất cả các biến có sẵn trong tập dữ liệu. Công trình cho rằng lựa chọn đặc trưng thông thường không cải thiện hiệu suất dự đoán do kích thước mẫu lớn.

Để giải quyết vấn đề mất cân bằng lớp, công trình đã sử dụng bốn kỹ thuật lấy mẫu để tăng tỷ lệ các ví dụ thuộc lớp thiểu số trong tập huấn luyện bao gồm phương pháp giảm thiểu số mẫu lớp đa số, phương pháp tăng cường số mẫu lớp thiểu số, phương pháp lai giảm thiểu số mẫu lớp đa số và tăng cường số mẫu lớp thiểu số, và phương pháp tăng cường ngẫu nhiên số mẫu lớp thiểu số. Công trình chứng minh rằng các phương pháp tăng cường số mẫu lớp thiểu số và tạo dữ liệu tổng hợp, khi được điều chỉnh đúng cách, có thể dẫn đến hiệu suất dự đoán vượt trội đối với tập dữ liệu mất cân bằng.

Công trình sử dụng PR_AUC làm độ đo chính để đánh giá hiệu suất do sự mất cân bằng giữa các lớp. Kết quả công trình cho thấy mô hình Rừng ngẫu nhiên dễ triển khai và xử lý mạnh mẽ đối với vấn đề này.

| | AUROC | AUPRC | Accuracy* | Sensitivity* | Specificity* | F1* |
|---|---|---|---|---|---|---|
| Linear logistic (ROSE) | 0.9836798 | 0.8347616 | 0.9995 | 0.797619 | 0.999831 | 0.842673 |
| Quad logistic (Over) | 0.9839819 | 0.8840936 | 0.9993 | 0.734043 | 0.999873 | 0.816568 |
| Random Forest (Over) | 0.9839583 | 0.9095681 | 0.9997 | 0.929577 | 0.999810 | 0.904110 |
| Neural Net (ROSE) | 0.9828006 | 0.7316877 | 0.9995 | 0.831169 | 0.999768 | 0.842105 |

*Hình 3. Kết quả của các thuật toán được thực hiện trong Công trình 3*

*Nguồn:* [7]

Tập dữ liệu gian lận thẻ tín dụng thường mất cân bằng do các giao dịch gian lận xảy ra rất ít. Tình huống mất cân bằng lớp có xu hướng chênh lệch về lớp đa số, dẫn đến hiệu suất dự đoán kém cho lớp thiểu số (số giao dịch gian lận). Ngoài ra, các phương pháp lấy mẫu được thường xuyên sử dụng để đối phó với tình huống này. Chính vì thế, khóa luận sử dụng một trong những phương pháp lấy mẫu phổ biến là SMOTE để xử lý tình trạng này. Có thể nói, phương pháp lấy mẫu đóng vai trò quan trọng đến hiệu suất tổng quan các mô hình được triển khai nhằm cân bằng cả hai lớp đa số và thiểu số. Tiếp theo là các công trình liên quan đến phương pháp lấy mẫu SMOTE được ứng dụng như sau:

*Công trình 4:* Vào năm 2019, Varmedja và Bhropade [8] thực hiện thí nghiệm trên bộ dữ liệu châu Âu và so sánh hiệu suất các mô hình như Hồi quy tuyến tính, Rừng ngẫu nhiên sử dụng phương pháp SMOTE, giúp tạo ra các quan sát tổng hợp cho lớp thiểu

số để giảm mất cân bằng dữ liệu. Kết quả của họ cho thấy mô hình Rừng ngẫu nhiên đạt được hiệu suất tốt nhất với sự cân bằng tỷ lệ giữa độ chuẩn xác chiếm 96,38% và độ phủ chiếm 81,63%. Kết luận công trình cho thấy SMOTE cải thiện hiệu suất của các mô hình được sử dụng.

Tuy nhiên, trong các tập dữ liệu có độ chênh lệch lớn, như bộ dữ liệu giao dịch thẻ tín dụng, phương pháp SMOTE có thể tổng quát hóa quá mức lớp thiểu số, dẫn đến mô hình có tỷ lệ dương tính giả quá cao, đồng nghĩa với việc dễ phân loại sai các giao dịch bình thường là gian lận. Để giải quyết vấn đề này, công trình tiếp theo áp dụng các phương pháp SMOTE với các mô hình học giám sát và học kết hợp.

*Công trình 5:* Vào năm 2019, Taneja, Suri và Kothari [9] tiến hành nghiên cứu sử dụng bộ dữ liệu châu Âu và so sánh các kỹ thuật dựa trên SMOTE kết hợp với cả hai mô hình học giám sát và học kết hợp. Kết quả cho thấy mô hình máy hỗ trợ véc-tơ với Rừng ngẫu nhiên sử dụng phương pháp SMOTE đạt hiệu suất tốt nhất về tỷ lệ độ phủ chiếm 80%, độ chuẩn xác chiếm 91% và F1-Score chiếm 85%. Tuy nhiên, họ cho rằng mô hình này có chi phí tính toán cao.

Không phải tất cả các phương pháp lấy mẫu ảnh hưởng đến các mô hình theo cùng một cách. Chính vì thế, công trình tiếp theo áp dụng cùng tập dữ liệu và so sánh nhiều kỹ thuật lấy mẫu trên nhiều mô hình khác nhau.

*Công trình 6:* Vào năm 2020. Muaz, Jayabalan và Thiruchelvam [10] tiến hành nghiên cứu trên các mô hình khác nhau. Kết quả của họ cho thấy mô hình Rừng ngẫu nhiên kết hợp với phương pháp SMOTE đạt được tỷ lệ độ phủ cao nhất là 81% và độ chuẩn xác là 86%. So sánh với công trình thứ nhất, Muaz, Jayabalan và Thiruchelvam cũng chứng minh rằng phương pháp SMOTE đạt được kết quả tích cực trong việc phát hiện gian lận thẻ tín dụng.

*Công trình 7:* Vào năm 2020. Sahu, GM và Gourisaria [11] phân tích bộ dữ liệu và xây dựng nhiều mô hình cho việc phát hiện gian lận bằng các phương pháp khác nhau để đối phó với sự mất cân bằng lớp. Công trình cũng áp dụng phương pháp tăng

cường số mẫu và kết quả của họ cho thấy mô hình Rừng ngẫu nhiên vượt trội so với các mô hình khác, đạt được tỷ lệ F1-Score là 95,21%.

Sau khi tìm hiểu các công trình liên quan trên về phát hiện gian lận tài chính dùng các kỹ thuật phân tích bất thường, ta có thể thấy mô hình Rừng ngẫu nhiên mang lại tối ưu đáng kể cho vấn đề này. Với những kết quả đầy hứa hẹn, chúng tôi quyết định tìm hiểu các mô hình học kết hợp bên cạnh Rừng ngẫu nhiên như là XGBoost, CatBoost và LightGBM trên tập dữ liệu chứa các giao dịch thẻ tín dụng vào tháng 9 năm 2013 bởi chủ thẻ ở Châu Âu. Ngoài việc đúc kết từ những ưu và khuyết điểm của các công trình trên, chúng tôi sẽ học tập các phương pháp để xử lý tập dữ liệu mất cân bằng và tập trung nghiên cứu cải tiến hiệu suất của các mô hình đã nêu trên.

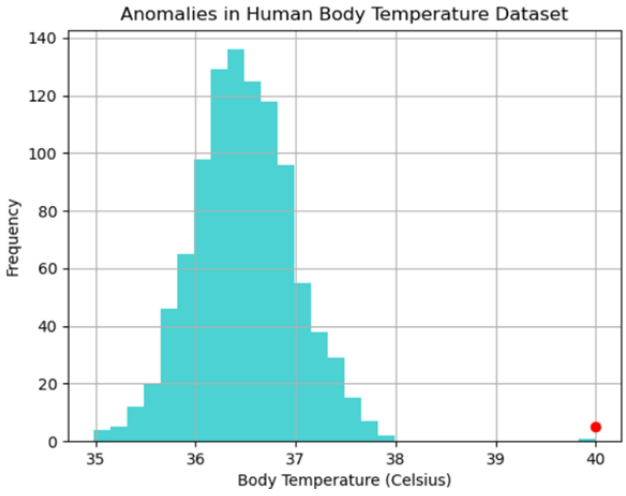
# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Giới thiệu về phát hiện bất thường

### 2.1.1. Định nghĩa về phát hiện bất thường

Theo [12] [13] [14], *phát hiện bất thường*, còn được gọi là *phát hiện ngoại lệ*, thông thường được hiểu là việc xác định các sự kiện hoặc quan sát hiếm, lệch đáng kể và không tuân theo một khái niệm rõ ràng về hành vi bình thường. Những ví dụ như vậy khiến các quan sát này xuất hiện không nhất quán với phần còn lại của tập dữ liệu. Những quan sát này thường ít và xa nhau, đồng thời mang những thông tin quan trọng. Trong ngữ cảnh của một tập dữ liệu rộng lớn, những điểm bất thường này có thể xuất hiện một cách ngẫu nhiên, thường tiết lộ nguyên nhân ẩn sau sự khác biệt này so với phần còn lại của dữ liệu.

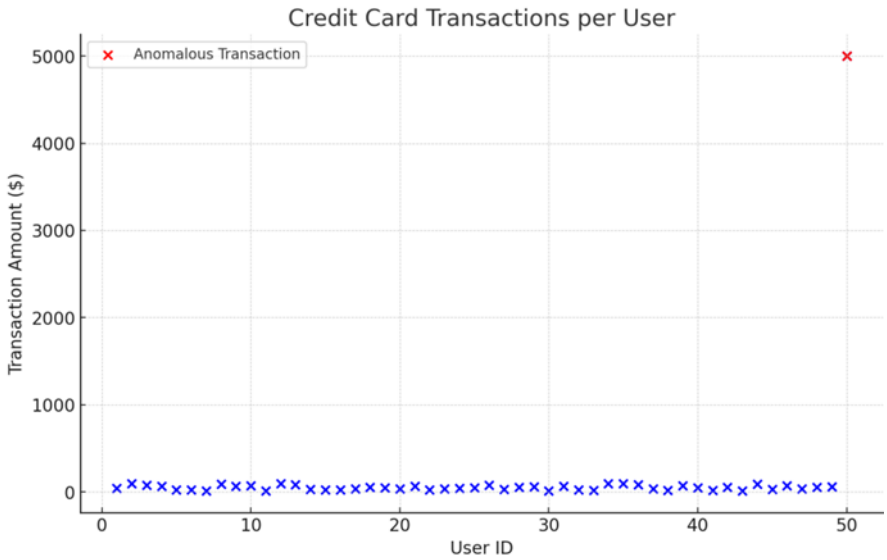
Trong thế giới thực, các bất thường hiển thị dưới nhiều hình thức và phụ thuộc vào một lĩnh vực cụ thể. Ví dụ, trên tập dữ liệu về nhiệt độ cơ thể của con người, một giá trị đọc là 40 độ C sẽ được coi là bất thường, vì nó lệch khá xa so với khoảng nhiệt độ cơ thể bình thường của con người.




*Hình 4. Biểu đồ minh họa về giá trị bất thường trong tập dữ liệu về nhiệt độ cơ thể của con người*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

Tương tự, trong tập dữ liệu giao dịch thẻ tín dụng, một giao dịch lớn đột ngột từ một người dùng thực hiện các giao dịch nhỏ bình thường có thể được coi là một bất thường, vì nó không tuân theo hành vi chi tiêu thông thường của họ.




*Hình 5. Biểu đồ minh họa về sự bất thường khi tồn tại giao dịch lớn đột ngột từ một người dùng thường chỉ thực hiện các giao dịch nhỏ*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

Theo [13], *bất thường* được xác định là một quan sát lệch xa so với các quan sát khác đến mức gây nghi ngờ rằng nó được tạo ra bởi một cơ chế khác. Định nghĩa này làm nổi bật một số khía cạnh quan trọng của bất thường.

- Thứ nhất, bất thường là các điểm dữ liệu khác biệt đáng kể so với các hành vi thông thường.
- Thứ hai, sự khác biệt này không chỉ đến từ nhiễu ngẫu nhiên mà còn gợi ý về một cơ chế hoặc nguyên nhân đằng sau chúng.

Cần lưu ý rằng bất thường không phải lúc nào cũng xấu hoặc không mong muốn. Mặc dù thuật ngữ bất thường thường mang ý nghĩa tiêu cực, ví dụ như trong trường

hợp giao dịch gian lận thẻ tín dụn hoặc xâm nhập mạng [15], nhưng bất thường cũng có thể đại diện cho điều tích cực hoặc có giá trị. Ví dụ, trong lĩnh vực dược phẩm, kết quả bất thường có thể chỉ ra một tác dụng điều trị mới, trước đây chưa biết, của một loại thuốc [16].

### 2.1.2. Phân loại các bất thường

Bất thường, theo bản chất của chúng, lệch khỏi những gì được mong đợi hoặc thông thường trong một tập dữ liệu. Các quan sát này khiến chúng trở nên thú vị và có giá trị. Tuy nhiên, không phải tất cả các bất thường đều được tạo ra như nhau. Tùy thuộc vào đặc điểm của chúng, chúng ta có thể phân loại bất thường thành ba loại riêng biệt [12]:

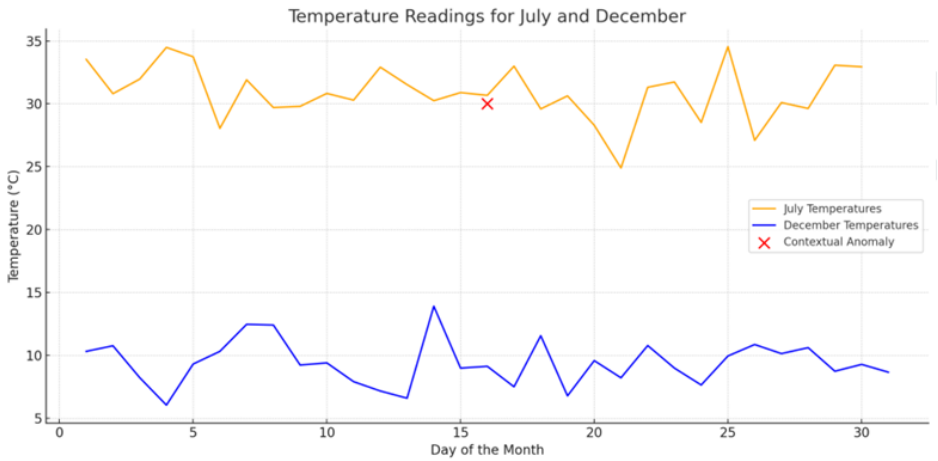
- Điểm bất thường.
- Bất thường ngữ cảnh.
- Bất thường tập hợp.

*Điểm bất thường* đại diện cho hình thức đơn giản nhất của bất thường. Một điểm dữ liệu lệch khỏi phần còn lại của tập dữ liệu một cách đáng kể được coi là một điểm bất thường. Điểm dữ liệu sẽ lệch rõ ràng khi nó nằm ngoài mẫu hoặc phân phối tổng thể của dữ liệu.

Ví dụ, hãy xem xét một tập dữ liệu về chiều cao của con người, trong đó hầu hết các điểm dữ liệu nằm trong khoảng từ 1,5 đến 2 mét. Nếu chúng ta gặp giá trị như 20 mét hoặc 0.2 mét, nó sẽ được coi là một điểm bất thường, vì nó nằm ngoài khỏi khoảng giá trị dự đoán chiều cao của con người.

*Bất thường ngữ cảnh*, hoặc bất thường có điều kiện, là các điểm dữ liệu lệch khỏi mức đáng kể trong một ngữ cảnh cụ thể, nhưng có thể không được coi là bất thường nếu không có ngữ cảnh. Ngữ cảnh đề cập đến tình huống hoặc điều kiện được xác định bằng các thuộc tính dữ liệu khác.

Hãy xem xét các giá trị nhiệt độ từ một trạm thời tiết. Một nhiệt độ 30 độ C vào tháng 7 (mùa hè) có thể không là một bất thường, nhưng cùng một giá trị đọc vào tháng 12 (mùa đông) có thể được xem là một bất thường ngữ cảnh (giả sử trong một khu vực mùa đông thường lạnh).

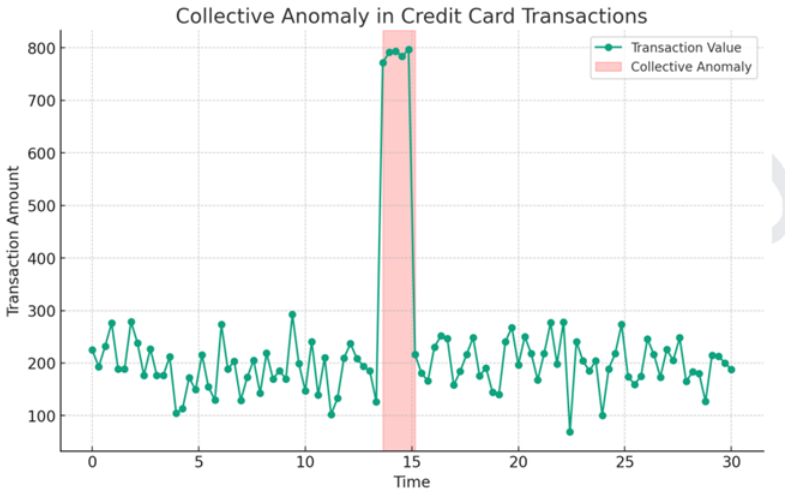



*Hình 6. Biểu đồ thể hiện các giá trị nhiệt độ từ một trạm thời tiết*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

*Bất thường tập hợp* liên quan đến tập hợp các điểm dữ liệu cùng nhau, nhưng có thể không là bất thường khi chúng là những điểm riêng lẻ. Loại bất thường này thú vị trong các trường hợp các điểm dữ liệu có liên quan hoặc kết nối với nhau. Các điểm dữ liệu cá nhân có thể là bất thường khi là những điểm riêng lẻ, nhưng khi xuất hiện cùng nhau thì chúng trở nên bất thường.

Hãy xem xét một tình huống, trong đó một cá nhân đang sử dụng thẻ tín dụng. Một giao dịch có giá trị cao đơn lẻ có thể sẽ không gây nghi ngờ nếu người dùng đã từng có giao dịch tương tự. Tuy nhiên, một loạt các giao dịch có giá trị cao như vậy trong một khoảng thời gian ngắn có thể được coi là một bất thường tập hợp, giúp xác định các gian lận thẻ tín dụng.




*Hình 7. Biểu đồ minh họa một loạt các giao dịch có giá trị cao bất thường trong một khoảng thời gian ngắn*

*Nguồn: Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

Ba loại bất thường này cung cấp cho chúng ta một kiến thức tổng thể để hiểu và dễ xác định bất thường trong các tập dữ liệu khác nhau. Khi chúng ta đi sâu vào các khái niệm và kỹ thuật phát hiện bất thường, các loại bất thường khác nhau có thể yêu cầu các phương pháp phát hiện khác nhau. Hiểu rõ những loại này sẽ giúp chúng ta chọn được kỹ thuật phát hiện bất thường phù hợp nhất cho một tình huống hoặc tập dữ liệu cụ thể [17]. Dựa trên những tính chất của điểm bất thường, bất thường ngữ cảnh và bất thường tập hợp, ta sẽ lập bảng so sánh chúng theo những đặc điểm sau:

| Tính Năng | Điểm Bất Thường | Bất thường ngữ cảnh | Bất thường tập hợp |
|---|---|---|---|
| **Định nghĩa** | Quan sát nổi bật đáng kể so với qui luật ở cách ly. | Sự lệch dữ liệu được đánh giá trong ngữ cảnh của dữ liệu và điều kiện xung quanh. | Những sự bất thường được phát hiện bằng cách xem xét hành vi toàn thể của một nhóm quan sát. |
| **Tập trung** | Các điểm dữ liệu cá nhân. | Dữ liệu trong một ngữ cảnh hoặc cụ thể. | Mẫu và hành vi trong một nhóm quan sát. |
| **Phương pháp phát hiện** | Các độ đo thống kê, như Z-scores hoặc khoảng cách trung bình. | So sánh với môi trường hoặc ngữ cảnh địa phương. | Các kỹ thuật như phân cụm hoặc thuật toán phát hiện bất thường. |
| **Nhạy cảm đối với ngữ cảnh** | Ít nhạy cảm với các yếu tố ngữ cảnh. | Phụ thuộc vào điều kiện và môi trường xung quanh. | Tương đối, vì nó xem xét các mẫu trong một nhóm cụ thể. |
| **Ứng dụng** | Phát hiện gian lận, phát hiện xâm nhập mạng. | Giám sát môi trường, phân tích dữ liệu cảm biến. | Giám sát sức khỏe, phát hiện gian lận |
| **Thách thức** | Có thể bỏ qua những điểm bất thường khi xem xét ngữ cảnh. | Phụ thuộc vào việc định và hiểu ngữ cảnh. | Xác định một nhóm phù hợp cho việc phát hiện bất thường. |
| **Biểu diễn dữ liệu** | Các điểm dữ liệu đơn biến. | Dữ liệu đa biến với các đặc trưng ngữ cảnh. | Dữ liệu đa biến biểu thị hành vi toàn bộ của nhóm. |

*Bảng 2. Bảng so sánh các loại bất thường*

Phát hiện bất thường có nhiều ứng dụng trong các lĩnh vực khác nhau, từ việc phát hiện gian lận thẻ tín dụng, phát hiện xâm nhập mạng và giám sát sức khỏe, đến việc phát hiện hỏng hóc trong xử lý hình ảnh. Việc xác định và hiểu các loại bất thường giúp xác định đúng đắn thông tin quý giá về hành vi, xác định vấn đề và nguy cơ tiềm ẩn, hỗ trợ hiệu quả trong việc quyết định.

Tuy nhiên, phát hiện bất thường vẫn còn đối diện với một số thách thức [12]. Vì bất thường xảy ra rất hiếm, một tập dữ liệu mất cân bằng sẽ có khiến cho quá trình phát hiện bất thường trở nên phức tạp. Ngoài ra, ranh giới giữa hành vi bình thường và bất thường không phải lúc nào cũng rõ ràng. Vì vậy, trong các lĩnh vực khác nhau, quá trình này thường yêu cầu sự kết hợp giữa các kỹ thuật thống kê, thuật toán học máy và kiến thức chuyên môn của lĩnh vực đó.

Trong những chương tiếp theo, khóa luận sẽ đi sâu hơn vào các kỹ thuật khác nhau được sử dụng cho phát hiện gian lận, bắt đầu từ các kỹ thuật thống kê, sau đó chuyển sang các kỹ thuật học máy. Đồng thời, chúng tôi sẽ xem xét kỹ hơn về những phương pháp này, trình bày các nguyên tắc, ứng dụng, ưu điểm và hạn chế của chúng.

Mục tiêu của phát hiện bất thường không chỉ là xác định các điểm ngoại lệ, mà còn là hiểu nguyên nhân đằng sau chúnh. Mỗi bất thường đại diện cho một sự lệch khỏi mô hình thông thường, và khi xác định đúng, ta dễ dàng rút ra những thông tin quý báu và cơ hội cho sự can thiệp và cải thiện.

## 2.2. Các kỹ thuật thống kê trong phát hiện bất thường

Trong chương trước, chúng ta đã hiểu cơ bản về phát hiện bất thường, khám phá điều gì tạo nên một bất thường và các loại bất thường khác nhau: điểm, ngữ cảnh và tập hợp. Chúng ta nhận ra rằng bất thường thực sự có thể tiết lộ thông tin quan trọng trong các lĩnh vực khác nhau, từ tài chính đến y tế v.v... Chúng ta tiếp tục điều hướng đến các kỹ thuật thống kê, là nền tảng của phát hiện bất thường.

Trong chương này, chúng ta sẽ giải mã những kỹ thuật này, bắt đầu với một cái nhìn tổng quan về các kiểm tra thống kê tham số và phi tham số. Chúng ta cũng sẽ khám phá khả năng áp dụng các kiểm tra thống kê cụ thể, chẳng hạn như kiểm định Chi bình phương, kiểm định Grubbs, v.v…

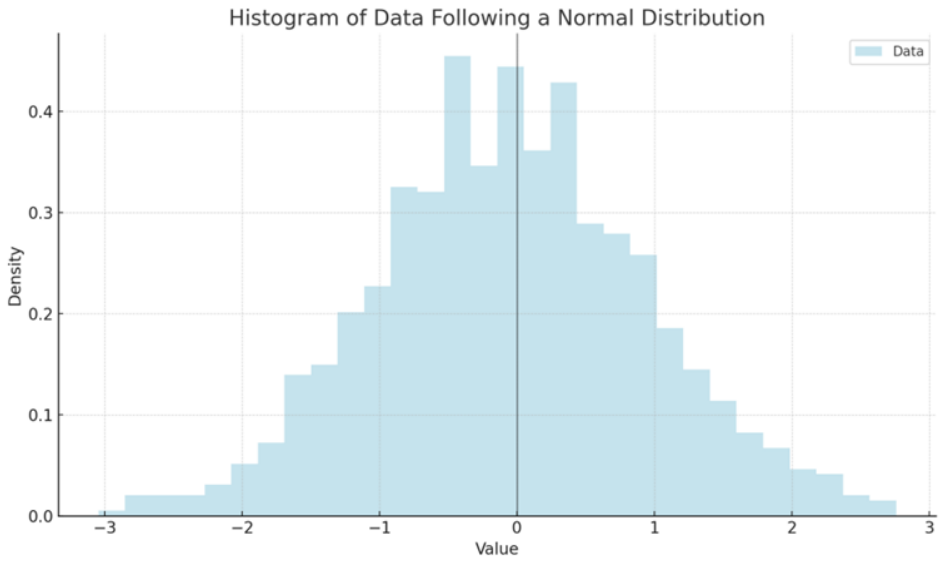
Mỗi kiểm định này sẽ được trình bày nhằm làm sáng tỏ về nền tảng lý thuyết và ứng dụng của chúng trong thế giới thực: Khi nào và tại sao mỗi kiểm tra phù hợp với các loại tập dữ liệu và bất thường cụ thể. Những kiểm định này có thể giúp chúng ta phân biệt giữa một bất thường thực sự và nhiễu thống kê.

Trong lĩnh vực thống kê, các kiểm định đóng vai trò quan trọng trong quá trình ra quyết định và kiểm định các giả thuyết. Các nhà thống kê và nhà khoa học dữ liệu sử dụng những kiểm định này để phân biệt các mẫu và mối quan hệ quan trọng trong dữ liệu, xác minh giả định của họ hoặc bác bỏ một số khẳng định. Các kiểm tra thống kê này có thể được phân loại thành hai loại: tham số và phi tham số.

### 2.2.1. Các phương pháp kiểm định tham số

Khi các kiểm định thống kê đòi hỏi những giả định khá chặt chẽ về phân phối của tổng thể mà từ đó mẫu được chọn ra, các kiểm định ấy được gọi chung là kiểm định tham số vì chúng gắn liền với một một tham số nào đó của tổng thể [18]. Nói một cách khác, các phương pháp kiểm định tham số dựa trên một số giả định về đặc điểm của dữ liệu. Những kiểm tra này giả định rằng dữ liệu tuân theo một phân phối cụ thể, thường là phân phối Gauss hoặc phân phối chuẩn. Họ cũng giả định các điều kiện khác, chẳng hạn như tính đồng nhất của phương sai và sự độc lập của các quan sát.

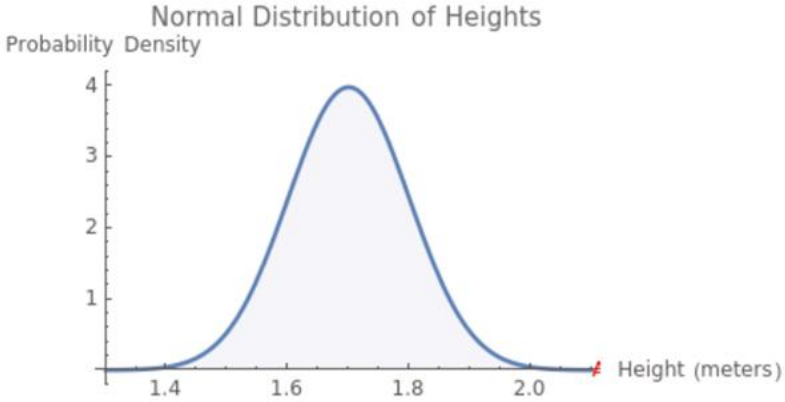
Khi những giả định này được đáp ứng, các kiểm tra tham số là các công cụ mạnh mẽ cho kiểm định giả thuyết và nhạy bén đối với những đặc điểm của dữ liệu.




*Hình 8. Biểu đồ tần suất của dữ liệu tuân theo phân phối chuẩn.*

*Nguồn: https://medium.com/@gabrielpierobon/statistical-techniques-for-anomaly-detection-part-1-parametric-and-non-parametric-tests-1801d07ba3fa*

Đối với phát hiện bất thường, các phương pháp tham số rất hữu ích trong các tình huống mà dữ liệu tuân theo một phân phối đã biết. Ví dụ, trong hình 8, một tập dữ liệu về chiều cao của con người có thể được kỳ vọng tuân theo phân phối chuẩn. Nếu chúng ta gặp một giá trị không bình thường (ví dụ, chiều cao 2,5 mét), chúng ta có thể sử dụng các kiểm tra tham số để đo lường mức độ bất thường của điểm dữ liệu này so với phần còn lại của tập dữ liệu.




*Hình 9. Biểu đồ minh họa tập dữ liệu về chiều cao của con người trong một dân số cụ thể có thể được kỳ vọng tuân theo phân phối chuẩn*

*Nguồn: https://medium.com/@gabrielpierobon/statistical-techniques-for-anomaly-detection-part-1-parametric-and-non-parametric-tests-1801d07ba3fa*

Những kiểm định tham số phổ biến:

- *Z-Test:* Dùng để so sánh trung bình của một nhóm với một giá trị tham chiếu.
- *T-Test:* Dùng để so sánh trung bình của hai nhóm.
- *Phân Tích Phương Sai (ANOVA):* Dùng để so sánh trung bình của ba hoặc nhiều nhóm.
- *Hồi quy tuyến tính:* Là một kỹ thuật dự đoán giả định mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc.

Mỗi kiểm định này phù hợp với các tình huống khác nhau. Ví dụ, Z-Test và T-Test được sử dụng để so sánh trung bình của một hoặc hai nhóm, trong khi ANOVA được sử dụng để so sánh trung bình của ba hoặc nhiều nhóm.

### 2.2.2. Các phương pháp kiểm định phi tham số

Trong phân tích dữ liệu, không phải lúc nào ta cũng gặp được các tình huống thỏa mãn hoàn toàn các giả định cần thiết này, đặc biệt khi chỉ có các mẫu nhỏ. Vì thế, việc nhờ đến những thủ tục đòi hỏi những giả định ít nghiêm ngặt hơn về phân phối

dữ liệu, những thủ tục này được gọi là kiểm định với phân phối bất kì hay còn gọi là kiểm định phi tham số [18]. Những kiểm định này không phụ thuộc vào phân phối và do đó ít mạnh mẽ và linh hoạt hơn so với các phương pháp tham số. Chúng được áp dụng cho nhiều loại dữ liệu khác nhau như dữ liệu định danh, dữ liệu thứ bậc hoặc dữ liệu khoảng cách không có phân phối rõ ràng, do đó khá hiệu quả đối với các giá trị ngoại lệ và không tuân theo phân phối chuẩn.

Trong ngữ cảnh của việc phát hiện bất thường, các phương pháp phi tham số hữu ích khi dữ liệu không tuân theo phân phối đã biết hoặc được xác định rõ, khi bản chất của phân phối không được biết đến.

Một số kiểm định phi tham số phổ biến bao gồm [18]:

- *Kiểm định Chi bình phương:* Thường được sử dụng để kiểm tra mối quan hệ giữa các biến phân loại.
- *Kiểm định U Mann-Whitney:* Dùng để so sánh hai nhóm độc lập.
- *Kiểm định Kruskal-Wallis:* Dùng để so sánh ba hoặc nhiều nhóm độc lập.

Kiểm định tham số và phi tham số phục vụ cho nhiều mục đích khác nhau. Sự lựa chọn giữa hai loại kiểm định này phụ thuộc vào bản chất của dữ liệu, câu hỏi nghiên cứu, giả định. Hơn nữa, ranh giới giữa kiểm định tham số và phi tham số có thể mờ đi. Ví dụ, một số kỹ thuật như KNN có thể được điều chỉnh để hoạt động giống như phương pháp tham số hoặc phi tham số, tùy thuộc vào các tham số được chọn [19].

### 2.2.3. Phân tích khám phá dữ liệu (EDA)

EDA hay còn được gọi là phân tích khám phá dữ liệu là một trong những bước đầu trong quy trình phân tích dữ liệu. Việc hiểu rõ một tập dữ liệu bao gồm khá nhiều công việc, chẳng hạn như nắm rõ các mô tả về dữ liệu, thống kê mô tả của từng đặc trưng để chỉ ra mối tương quan, nhằm phát hiện các xu hướng và tìm ra những bất thường trong dữ liệu. Dựa trên loại tập dữ liệu mà ta xác định những đặc trưng nào là

quan trọng, những kỹ thuật biến đổi hay phân tích cần được sử dụng để xác định mối quan hệ và đồng thời trực quan hóa dữ liệu.

Một quy trình EDA bao gồm các bước như:

*Bước 1:* Thu thập dữ liệu.

*Bước 2:* Xác định các đặc trưng quan trọng và nắm rõ đặc tính của từng đặc trưng.

*Bước 3:* Làm sạch dữ liệu: bao gồm các công việc như loại bỏ dữ liệu có giá trị null, trùng lắp tùy trường hợp, xác định các giá trị ngoại lai, v.v…

*Bước 4:* Xác định các đặc trưng tương quan: sử dụng ma trận tương quan nhằm tìm ra sự tương đồng.

*Bước 5:* Chọn đúng phương pháp thống kê mô tả: tùy vào loại dữ liệu, kích thước dữ liệu, loại đặc trưng cũng như mục đích phân tích.

*Bước 6:* Trực quan hóa và phân tích dữ liệu.

Có ba dạng kỹ thuật khám phá dữ liệu lần lượt là phân tích đơn biến, phân tích hai biến, và phân tích đa biến. Từ đây, các kỹ thuật được phân chia thành hai nhóm là đồ họa và phi đồ họa.

Bởi hai biến vẫn được xem là đa biến nên EDA chỉ có bốn dạng kỹ thuật khám phá dữ liệu nói chung là đơn biến phi đồ hoạ, đồ họa đơn biến, đa biến phi đồ họa và đồ họa đa biến. Các phương pháp phi đồ họa thường bao gồm việc tính toán các số liệu tóm tắt, trong khi các phương pháp đồ họa tóm tắt dữ liệu dưới dạng biểu đồ hoặc hình ảnh.

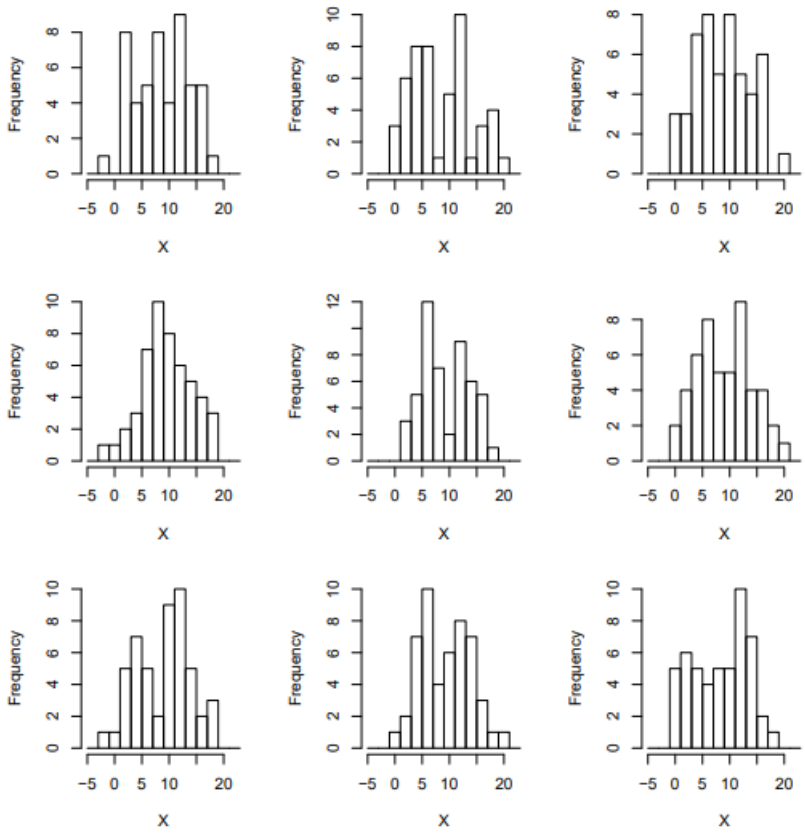
Phân tích đơn biến phi đồ hoạ chủ yếu mô tả, tóm tắt dữ liệu dựa vào các thang đo như mức độ tập trung, khoảng dữ liệu.

Phân tích đồ hoạ đơn biến chủ yếu sử dụng các biểu đồ như biểu đồ thân và lá, biểu đồ tần suất, và biểu đồ hộp. Biểu đồ thân và lá được sử dụng để trực quan hóa dữ liệu ở dạng rút gọn. Biểu đồ tần suất biểu diễn tần suất xuất hiện các giá trị trong tập dữ liệu, ví dụ như giá trị ngoại lai, xu hướng tập trung của từng đặc trưng. Biểu đồ hộp

biểu diễn sự phân tán của các giá trị trong tập dữ liệu, xem các giá trị tập trung trong khoảng nào và từ đó xác định được giá trị ngoại lai của tập dữ liệu.

Phân tích đa biến phi đồ hoạ chủ yếu sử dụng các thông số thống kê như phân tích phương sai, hiệp phương sai và bảng chéo.

Phân tích đồ hoạ đa biến chủ yếu các biểu đồ như biểu đồ hộp, biểu đồ phân tán và phổ biến nhất là biểu đồ cặp. Biểu đồ phân tán thể hiện mức độ tương quan giữa hai biến, tuy nhiên dữ liệu đều phải ở dạng định lượng để biểu diễn. Biểu đổ cặp giúp trực quan hóa mối quan hệ giữa tất cả các biến số trong toàn bộ tập dữ liệu cùng lúc. Ngoài ra, bản đồ nhiệt còn là một biểu đồ được sử dụng khá rộng rãi trong việc xây dựng các mô hình học máy, được trình bày dưới dạng ma trận gồm các cột và hàng, với các trạng thái sắc độ của màu sắc nhằm trực quan hóa mức độ tương quan giữa các đặc trưng.



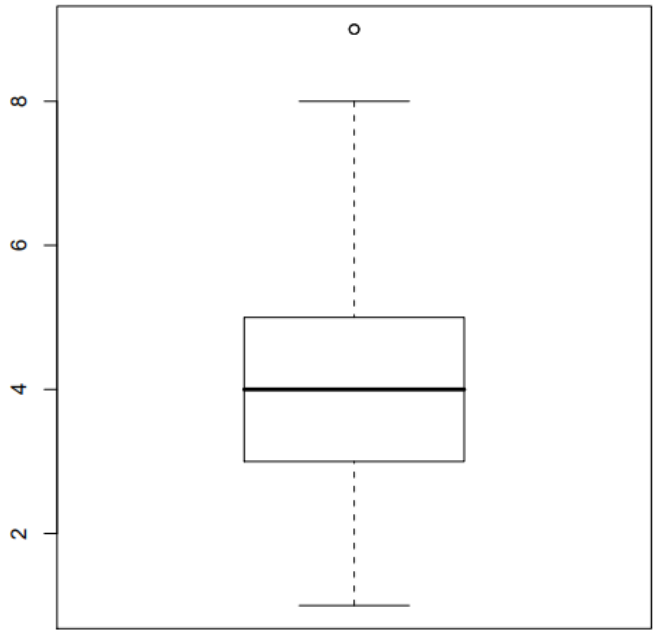

*Hình 10. Biểu đồ tần suất*

*Nguồn:* [20]

```
The decimal place is at the "|".
1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

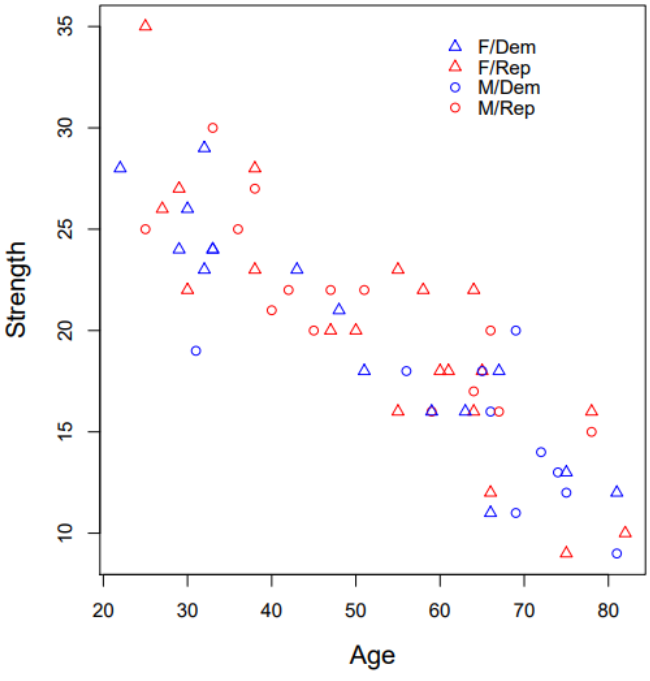
*Hình 11. Biểu đồ thân và lá*

*Nguồn:* [20]



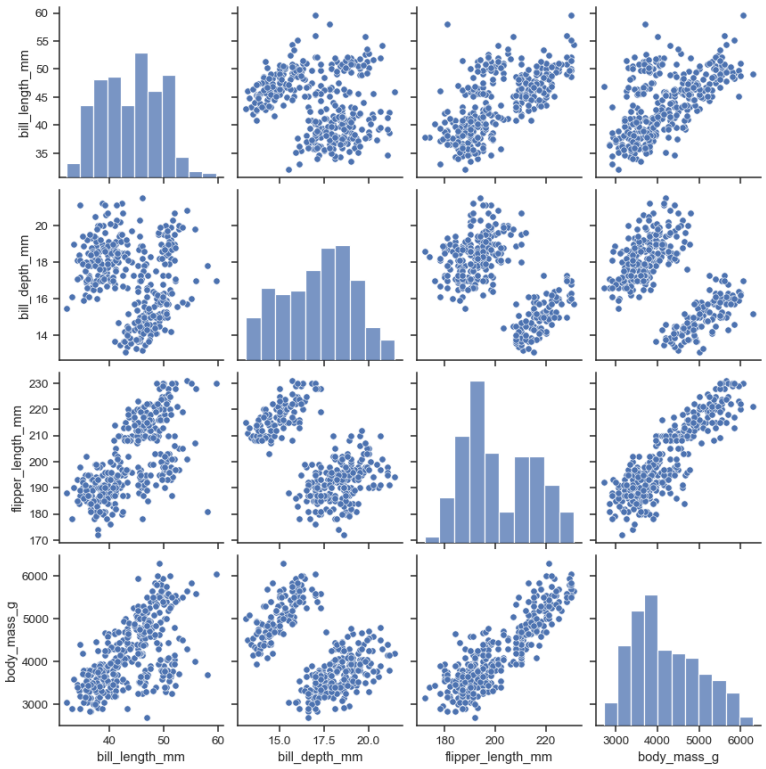

*Hình 12. Biểu đồ hộp*

*Nguồn:* [20]



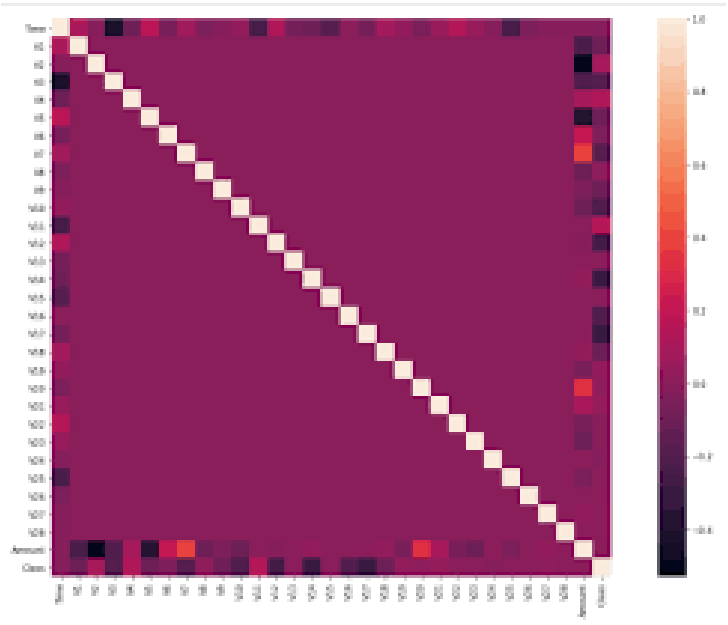

*Hình 13. Biểu đồ phân tán*

*Nguồn:* [20]




*Hình 14. Biểu đồ cặp*

*(Nguồn: seaborn.pairplot — seaborn 0.13.2 documentation (pydata.org))*



*Hình 15. Bản đồ nhiệt*

*(Nguồn: Heat map to visualize correlation matrix | Download Scientific Diagram (researchgate.net))*

## 2.3. Giới thiệu về học máy trong phát hiện bất thường

Trong thế giới dữ liệu, bất thường giống như một cây kim trong đống rơm [7]. Chúng hiếm, khác biệt và ẩn giấu trong lượng lớn dữ liệu bình thường. Tuy nhiên, những bất thường này thường chứa thông tin quý giá. Phát hiện gian lận thẻ tín dụng, phát hiện lỗi hệ thống trong thời gian thực, xác định các bệnh hiếm trong lĩnh vực chăm sóc sức khỏe và nhiều bài toán quan trọng khác phụ thuộc nhiều vào khả năng phát hiện những bất thường này một cách chính xác và hiệu quả.

Trong các chương trước, chúng ta đã khám phá các phương pháp thống kê để phát hiện bất thường. Những phương pháp này, như Z-Score, Kiểm tra Grubbs, Kiểm tra Chi bình phương và Khoảng cách Mahalanobis, đã cung cấp cho chúng ta các công cụ quý giá để xử lý các tình huống đơn giản và các loại phân phối dữ liệu cụ thể. Tuy nhiên, tập dữ liệu thế giới thực thường đi kèm với mức độ phức tạp cao hơn, với các thuộc tính như mô hình phi tuyến, đa chiều, độ nhiễu và mối tương quan phức tạp giữa các biến. Trong những tình huống này, các phương pháp thống kê truyền thống có thể gặp khó khăn, không thể nắm bắt được các mô hình phân biệt dữ liệu bất thường với dữ liệu bình thường.

Để giải quyết các hạn chế của phương pháp thống kê và cung cấp một phương pháp phát hiện bất thường tốt hơn, chúng ta hướng đến học máy. Học máy tập trung vào việc phát triển các thuật toán có thể học từ dữ liệu và đưa ra quyết định dựa trên chúng. Nó có thể khám phá các mô hình ẩn, xác định mối tương quan và dự đoán kết quả tương lai dựa trên dữ liệu lịch sử, làm cho nó trở thành một công cụ hiệu quả cho việc phát hiện bất thường. Các thuật toán học máy có thể lọc qua lượng lớn dữ liệu, mở rộng quá trình phát hiện bất thường để xử lý các tập dữ liệu quy mô lớn thường gặp trong các ứng dụng thực tế.

Các mô hình học máy cho phát hiện bất thường có thể xử lý sự phức tạp và đa dạng của dữ liệu thế giới thực. Chúng xác định bất thường trong không gian dữ liệu đa chiều, nắm bắt mối quan hệ phi tuyến phức tạp giữa các biến và thích nghi với các mô hình dữ liệu thay đổi theo thời gian. Với sức mạnh của học máy, chúng ta có thể

phát hiện bất thường trong nhiều ngữ cảnh khác nhau, từ phát hiện gian lận trong dữ liệu tài chính thời gian thực, phát hiện xâm nhập trong lưu lượng mạng, xác định lỗi hệ thống trong quy trình sản xuất, đến tìm kiếm các mô hình không bình thường trong dữ liệu chăm sóc sức khỏe.

Trước khi tìm hiểu chi tiết về các kỹ thuật học máy cho phát hiện bất thường, việc hiểu rõ về các mô hình học khác nhau mà các kỹ thuật này thuộc về là rất quan trọng. Trong lĩnh vực học máy, có hai mô hình cơ bản là học có giám sát và học không giám sát. Mỗi mô hình có đặc điểm, ứng dụng, ưu điểm và hạn chế riêng.

### 2.3.2. Học không giám sát

Học không giám sát không phụ thuộc vào dữ liệu huấn luyện đã được gán nhãn. Thay vào đó, nó cố gắng học cấu trúc hoặc phân phối cơ bản của dữ liệu để suy luận về dữ liệu mới. Các thuật toán học không giám sát có thể khám phá các mô hình thú vị trong dữ liệu, nhóm các dữ liệu tương tự lại (gom cụm), giảm chiều dữ liệu và nhiều ứng dụng khác.

Học không giám sát đặc biệt phù hợp cho việc phát hiện bất thường vì một số lý do. Thứ nhất, bất thường thường là các sự kiện hiếm, và trong nhiều trường hợp, chúng ta không có các bất thường đã được gán nhãn trong dữ liệu huấn luyện. Các phương pháp học không giám sát xử lý tình huống này bằng cách học cấu trúc của dữ liệu và sau đó xác định các điểm dữ liệu lệch khỏi cấu trúc bình thường này. Thứ hai, bản chất của bất thường có thể thay đổi theo thời gian, làm cho việc dựa vào nhãn bất thường trong quá khứ có thể trở nên khó khăn. Các phương pháp không giám sát, tập trung vào cấu trúc của dữ liệu bình thường, có thể thích nghi với những thay đổi của dữ liệu.

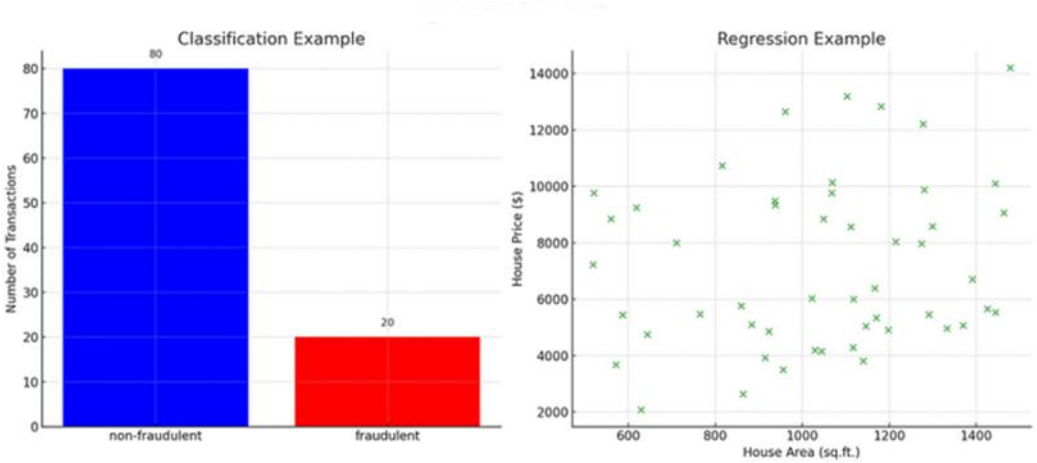
*1. Thuật toán gom cụm:* Các thuật toán gom cụm như K-Means, DBSCAN nhóm các điểm dữ liệu tương tự lại với nhau. Các điểm dữ liệu không phù hợp vào bất kỳ cụm nào hoặc tạo thành các cụm nhỏ riêng biệt có thể được coi là bất thường.

*2. Thuật toán phát hiện ngoại lệ:* Các thuật toán như Nhân tố ngoại lai cục bộ (LOF), Rừng cô lập và đường cong E-líp được thiết kế để phát hiện các điểm ngoại lệ trong dữ liệu. Mỗi thuật toán có một phương pháp khác nhau, nhưng cơ bản, chúng gán một "điểm ngoại lệ" cho mỗi điểm dữ liệu và các điểm dữ liệu có điểm số vượt qua ngưỡng cụ thể được coi là bất thường.

*3. Phân tích thành phần chính (PCA):* PCA là một kỹ thuật giảm chiều dữ liệu có thể được sử dụng cho việc phát hiện bất thường. Nó biểu diễn dữ liệu trong không gian chiều thấp hơn và phát hiện các điểm dữ liệu bất thường khi được biểu diễn bởi các thành phần chính này.

### 2.3.1. Học có giám sát

Trong *học có giám sát*, nhiệm vụ chính là học một ánh xạ từ đầu vào sang đầu ra, hoặc nói cách khác, dự đoán nhãn đầu ra dựa trên dữ liệu đầu vào. Điều này được thực hiện bằng cách huấn luyện một mô hình trên dữ liệu đã được gán nhãn, sau đó mô hình đã được huấn luyện này có thể dự đoán trên dữ liệu chưa thấy trước. Học có giám sát được phân thành hai loại: Phân loại và Hồi quy. Trong phân loại, đầu ra là một danh mục (ví dụ: giao dịch 'gian lận' hoặc 'không gian lận'), trong khi trong hồi quy, đầu ra là một giá trị liên tục (ví dụ: giá của một căn nhà).




*Hình 16. Hình ảnh miêu tả kết quả đầu ra của phân loại và hồi quy*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

Trong ngữ cảnh của việc phát hiện bất thường, học có giám sát rất hiệu quả khi chúng ta có dữ liệu đã được gán nhãn, biết được điểm dữ liệu nào là bất thường và điểm nào không phải. Mô hình học máy có thể học các đặc điểm của các ví dụ bình thường và bất thường trong quá trình huấn luyện, sau đó nó có thể phân loại chính xác các điểm dữ liệu mới là bình thường hoặc bất thường. Tuy nhiên, việc thu thập dữ liệu đã được gán nhãn cho việc phát hiện bất thường là thách thức vì bất thường thường hiếm và thường không biết trước.

Thuật toán học có giám sát đòi hỏi dữ liệu đã được gán nhãn để huấn luyện, tức là chúng ta cần có các điểm dữ liệu đã được phân loại trước là "bình thường" hoặc "bất thường", bao gồm:

*1. Hồi quy Logistic:* Hồi quy Logistic là một kỹ thuật cơ bản được sử dụng cho các bài toán phân loại nhị phân. Trong ngữ cảnh của phát hiện bất thường, chúng ta có thể phân loại dữ liệu là "bình thường" hoặc "bất thường". Mặc dù đơn giản, hồi quy Logistic là một điểm khởi đầu tốt vì sự đơn giản và thời gian huấn luyện nhanh.

*2. Cây quyết định và Rừng ngẫu nhiên:* Cây quyết định phân loại dữ liệu bằng cách đưa ra quyết định dựa trên giá trị đặc trưng. Rừng ngẫu nhiên là một tập hợp các cây quyết định có thể đưa ra dự đoán chính xác và ổn định.

*3. Máy véc-tơ hỗ trợ (SVM):* Trong ngữ cảnh của phát hiện bất thường, SVM có thể được sử dụng trong một phương pháp được gọi là SVM một lớp. Thuật toán này được huấn luyện trên dữ liệu chỉ đại diện cho trạng thái bình thường, phân biệt các điểm dữ liệu mới là tương tự với trạng thái bình thường (không bất thường) hoặc khác biệt (bất thường).

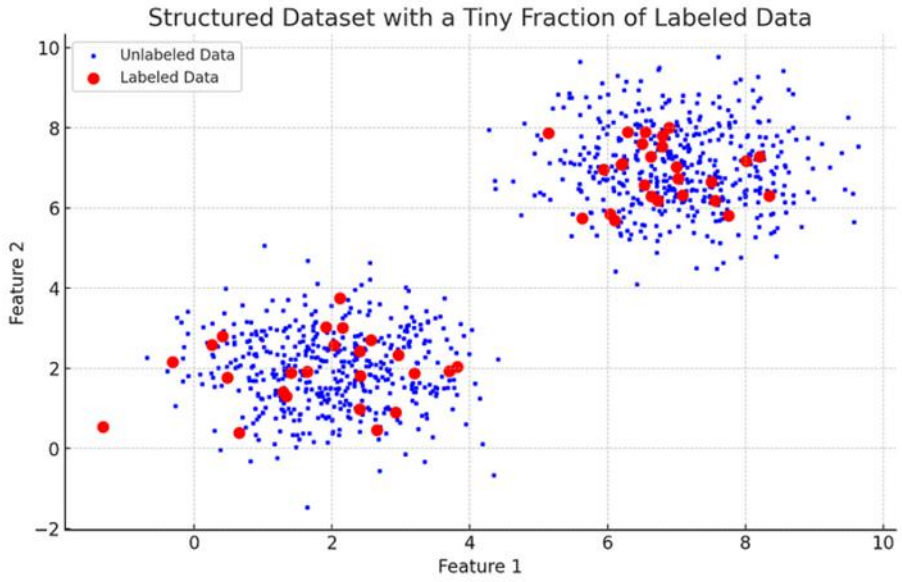
*4. Mạng nơ-ron và học sâu:* Các mô hình phức tạp hơn như mạng nơ-ron có thể nắm bắt các mô hình phức tạp trong dữ liệu. Một loại mạng gọi là Autoencoders đã trở nên phổ biến cho việc phát hiện bất thường. Autoencoders được huấn luyện để tái tạo dữ

liệu đầu vào của chúng, và sai số tái tạo cao có thể chỉ ra một bất thường. Mạng nơ-ron tích chập (CNN) và mạng nơ-ron tuần tự (RNN) cũng có thể được sử dụng cho phát hiện bất thường, đặc biệt là với dữ liệu hình ảnh.

Học có giám sát và học không giám sát đều có vai trò trong phát hiện bất thường. Sự lựa chọn giữa chúng phụ thuộc chủ yếu vào dữ liệu có sẵn đã được gán nhãn và bản chất cụ thể của vấn đề phát hiện bất thường.

### 2.3.3. Học bán giám sát

Mô hình học này tận dụng sức mạnh của học không giám sát để hiểu cấu trúc cơ bản hoặc phân phối của dữ liệu và các mô hình tồn tại trong đó. Đồng thời, nó sử dụng dữ liệu đã được gán nhãn có sẵn để hướng dẫn quá trình học, tạo ra dự đoán hoặc phân loại giống như học có giám sát.



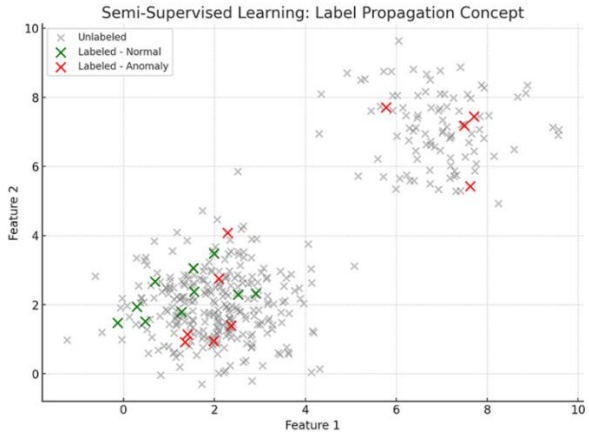

*Hình 17. Biểu đồ minh họa tập dữ liệu với một số ít mẫu được gán nhãn*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

Ý tưởng đằng sau phương pháp này là phân phối dữ liệu giúp xây dựng một mô hình hiệu quả, và số ít các dữ liệu đã được gán nhãn có thể hướng dẫn mô hình này. Bằng cách kết hợp cả dữ liệu đã được gán nhãn và không gán nhãn, học bán giám sát thường có thể đạt được độ chính xác cao hơn so với các phương pháp thuần túy học có giám sát hoặc học không giám sát, đặc biệt là khi không có nhiều dữ liệu đã được gán nhãn.

Mô hình được huấn luyện ban đầu trên dữ liệu bình thường (không gán nhãn) để hiểu rõ hành vi bình thường. Các trường hợp bất thường hạn chế (dữ liệu đã được gán nhãn) sau đó được sử dụng để điều chỉnh mô hình, tạo cơ sở xác định các bất thường.

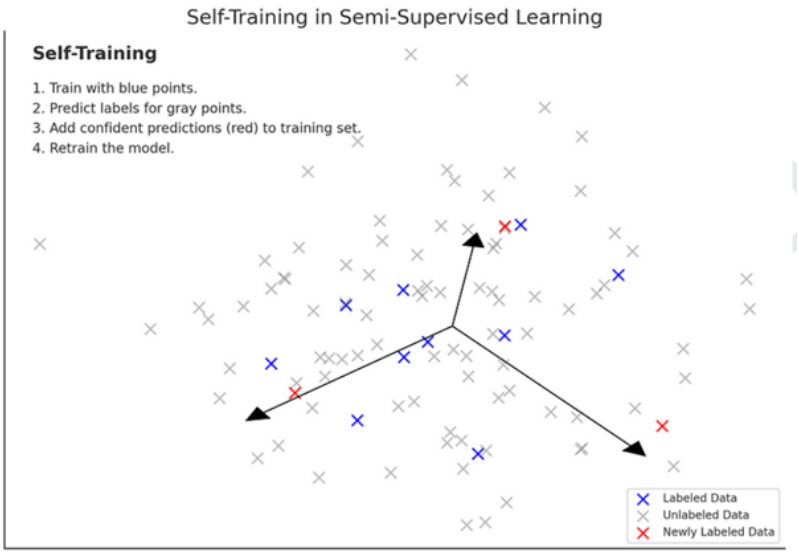
*1. Lan truyền nhãn:* Các phương pháp này bắt đầu với một tập hợp nhỏ các trường hợp đã được gán nhãn và lan truyền nhãn này sang các trường hợp chưa được gán nhãn, sử dụng cấu trúc dữ liệu để hướng dẫn quá trình lan truyền.




*Hình 18. Mô hình học bán giám sát với thuật toán lan truyền nhãn*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

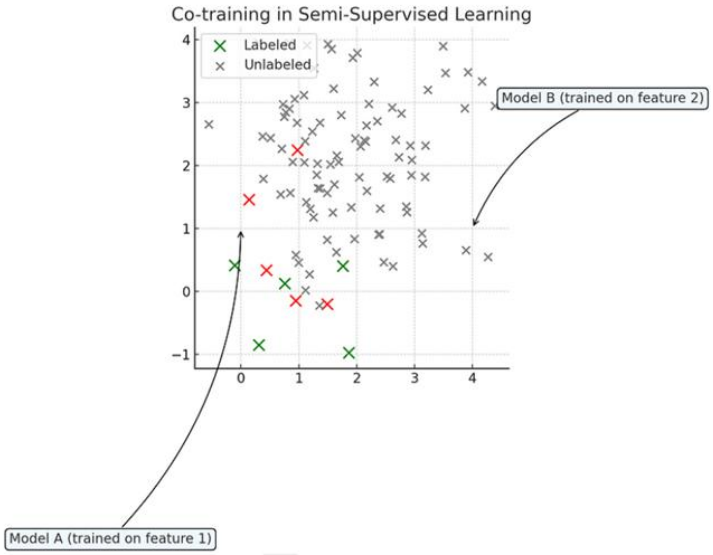
*2. Tự huấn luyện:* Trong tự huấn luyện, mô hình được huấn luyện ban đầu với một tập hợp nhỏ dữ liệu đã được gán nhãn, sau đó mô hình này được sử dụng để dự đoán nhãn cho dữ liệu chưa được gán nhãn. Các nhãn dự đoán có độ tin cậy nhất sau đó được thêm vào tập dữ liệu huấn luyện và mô hình được huấn luyện lại.




*Hình 19. Mô hình học bán giám sát với thuật toán tự huấn luyện*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

*3. Hợp tác huấn luyện:* Trong hợp tác huấn luyện, hai mô hình được huấn luyện độc lập trên dữ liệu (các tập con khác nhau của các đặc trưng), và chúng gán nhãn cho nhau các trường hợp chưa được gán nhãn.




*Hình 20. Mô hình học bán giám sát với thuật toán hợp tác huấn luyện*

*Nguồn: https://medium.com/@gabrielpierobon/introduction-to-machine-learning-for-anomaly-detection-part-1-2c12d7fa236*

## 2.4. Phương pháp đánh giá mô hình

Có nhiều độ đo khác nhau để đánh giá hiệu quả của một thuật toán. Các độ đo này dựa trên số lượng giao dịch phát hiện đúng hoặc sai. Một độ đo phổ biến và hiểu quả là *ma trận nhầm lẫn*. *Ma trận nhầm lẫn* là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp, bao gồm 4 chỉ số sau đối với mỗi lớp phân loại: Dương tính giả (FP), Âm tính giả (FN), Dương tính thực (TP) và Âm tính thực (TN).

*Độ chính xác* là tỷ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

$$\text{Độ chính xác} = \frac{\text{Số lượng dự đoán đúng}}{\text{Tổng số dự đoán}} \quad \text{hoặc} \quad \text{Độ chính xác} = \frac{TP+TN}{TP+TN+FP+FN}$$

*Độ chuẩn xác* là tỷ lệ giao dịch gian lận thật sự trong tổng số các giao dịch được phán đoán là gian lận.

$$\text{Độ chuẩn xác} = \frac{TP}{TP+FP}$$

*Độ phủ* là tỷ lệ những giao dịch được dự đoán đúng là gian lận trong tổng số các gian lận thực tế.

$$TPR = \text{Độ phủ} = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

*F1-score:* Đối với những tập dữ liệu không cân bằng (có sự chênh lệch rất lớn giữa số lượng giao dịch hợp lệ và giao dịch gian lận) thì độ chính xác, độ chuẩn xác hay độ phủ không phản ánh được độ chính xác và hiệu quả của thuật toán. Do vậy, cần sử dụng các độ đo mới, một trong số đó là F1-score.

$$F1 = \left(\frac{2 \times TP}{2 \times TP + FP + FN}\right)$$

*Đường cong ROC:* Để tránh chủ quan khi chỉ lựa chọn một ngưỡng để đánh giá mô hình, có một cách là duyệt qua hết tất cả các ngưỡng có thể được và quan sát ảnh hưởng lên các tỷ lệ dự báo tỷ lệ dương tính thực và tỷ lệ dương tính giả. Khi đó, sẽ dựng được đường cong ROC chứa tất cả các điểm tỷ lệ dương tính thực và điểm tỷ lệ dương tính giả.

Tương tự như đường cong ROC, chúng ta cũng có thể đánh giá mô hình dựa trên việc thay đổi một ngưỡng và quan sát giá trị của độ chuẩn xác và độ phủ. Đường cong độ chuẩn xác và độ phủ, viết tắt là PR-AUC hay AUC-PR, thể hiện sự đánh đổi giữa độ chuẩn xác và độ phủ cho các ngưỡng khác nhau.

# CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

## 3.1. Rừng ngẫu nhiên

*Rừng ngẫu nhiên* là một phương pháp học tập kết hợp cho phân loại hoặc hồi quy Hoạt động bằng cách xây dựng một lượng lớn các cây quyết định tại thời gian đào tạo. Đối với các nhiệm vụ phân loại, kết quả của rừng ngẫu nhiên là lớp được chọn bởi hầu hết các cây. Đối với các nhiệm vụ hồi quy, dự đoán trung bình hoặc trung bình của các cây cá nhân là kết quả được trả về [21] [22].

Thuật toán Rừng ngẫu nhiên được phổ biến rộng rãi vì sự thân thiện đối với người dùng và khả năng thích nghi, cho phép nó giải quyết các vấn đề phân loại và hồi quy một cách hiệu quả. Sức mạnh của thuật toán này nằm ở khả năng xử lý các tập dữ liệu phức tạp và giảm thiểu hiện tượng quá khớp, làm cho nó trở thành công cụ quý giá cho nhiều nhiệm vụ dự đoán trong học máy.

Một trong những đặc điểm quan trọng nhất của Thuật toán Rừng ngẫu nhiên là khả năng xử lý tập dữ liệu chứa biến liên tục trong bài toán hồi quy, và biến phân loại trong bài toán phân loại. Chúng ta sẽ hiểu thuật toán của rừng ngẫu nhiên và các kiến thức liên quan.

### 3.1.1. Kiến thức tổng quan

Trước hết, ta sẽ tìm hiểu về khái niệm cây quyết định. Cây quyết định là một biểu diễn cấu trúc cây, mỗi nút trong cây biểu diễn một quyết định dựa trên các đặc trưng của dữ liệu để phân loại hoặc dự đoán kết quả. Mỗi lá của cây đại diện cho một phân lớp hoặc giá trị dự đoán. Tuy nhiên, cây quyết định đơn lẻ có thể dễ dàng bị quá khớp nếu không được kiểm soát.

Rừng ngẫu nhiên khắc phục vấn đề quá khớp của cây quyết định bằng cách tạo ra nhiều cây quyết định khác nhau và kết hợp chúng lại. Quá trình này bắt đầu bằng việc

chọn ngẫu nhiên một số lượng mẫu từ tập dữ liệu huấn luyện và một số lượng đặc trưng. Sau đó, một cây quyết định được xây dựng dựa trên mẫu và đặc trưng đã chọn. Quá trình này được lặp lại nhiều lần để tạo ra nhiều cây quyết định.

Trước khi hiểu về cách hoạt động của thuật toán rừng ngẫu nhiên trong học máy, chúng ta cần tìm hiểu về kỹ thuật học tập kết hợp. Học tập kết hợp đơn giản là kết hợp nhiều mô hình. Thay vì sử dụng một mô hình riêng lẻ, chúng ta sử dụng một tập hợp các mô hình để đưa ra dự đoán.

Học tập kết hợp sử dụng hai loại phương pháp:

- *Phương pháp bỏ túi:* Phương pháp bỏ túi tạo ra các tập con huấn luyện khác nhau từ dữ liệu huấn luyện mẫu bằng cách thay thế và kết quả cuối cùng dựa trên sự lựa chọn đa số. Ví dụ: Rừng Ngẫu Nhiên.
- *Phương pháp tăng cường:* Phương pháp tăng cường kết hợp các mô hình yếu thành các mô hình mạnh bằng cách tạo ra các mô hình tuần tự sao cho mô hình cuối cùng có độ chính xác cao nhất. Ví dụ: ADABoost, XGBoost.

Như đã đề cập trước đó, Rừng ngẫu nhiên hoạt động dựa trên nguyên tắc Phương pháp bỏ túi. Bây giờ chúng ta hãy đi sâu và hiểu rõ hơn về Phương pháp bỏ túi.

*Phương pháp bỏ túi* là một kỹ thuật kết hợp trong thuật toán Rừng ngẫu nhiên. Dưới đây là các bước thực hiện của Phương pháp bỏ túi:

*1. Lựa chọn Tập con:* Phương pháp bỏ túi bắt đầu bằng việc chọn một mẫu ngẫu nhiên, hoặc tập con, từ toàn bộ tập dữ liệu.

*2. Lấy mẫu Bootstrap:* Sau đó, mỗi mô hình được tạo ra từ các mẫu này, được gọi là mẫu Bootstrap. Quá trình này được gọi là lấy mẫu theo hàng.

*3. Đào tạo Mô hình Độc lập:* Mỗi mô hình được đào tạo độc lập trên tập các mẫu Bootstrap tương ứng của nó. Quá trình đào tạo này tạo ra kết quả cho từng mô hình.

*4. Lựa chọn đa số:* Kết quả cuối cùng được xác định bằng cách kết hợp kết quả của tất cả các mô hình thông qua bỏ phiếu đa số. Kết quả được chọn là kết quả có dự đoán phổ biến nhất giữa các mô hình.

*5. Tổng hợp:* Bước này, liên quan đến việc kết hợp tất cả các kết quả và tạo ra kết quả cuối cùng dựa trên lựa chọn đa số, được gọi là tổng hợp.

*Phương pháp tăng cường* là một trong những kỹ thuật sử dụng khái niệm của học kết hợp. Một thuật toán Phương pháp tăng cường kết hợp nhiều mô hình đơn giản để tạo ra kết quả cuối cùng. Quá trình này được thực hiện bằng cách xây dựng một mô hình bằng cách sử dụng các mô hình yếu theo chuỗi.

Các tham số được sử dụng trong rừng ngẫu nhiên để tăng cường hiệu suất và khả năng dự đoán của mô hình hoặc để làm cho mô hình nhanh hơn.

Ưu điểm của phương pháp Rừng ngẫu nhiên là đa dạng và mạnh mẽ, phù hợp với nhiều loại vấn đề trong phân loại và hồi quy. Phương pháp này giải quyết vấn đề quá khớp bằng cách sử dụng kết quả dựa trên việc bầu chọn đa số hoặc lấy trung bình từ nhiều cây quyết định độc lập. Đặc biệt, nó cũng hoạt động tốt ngay cả khi dữ liệu bị mất mát hoặc thiếu, đảm bảo tính ổn định và đáng tin cậy của kết quả. Mỗi cây quyết định được tạo ra đều độc lập với nhau, duy trì sự đa dạng và giảm không gian đặc trưng. Điều này có ý nghĩa là không cần phải phân chia dữ liệu thành tập huấn luyện và kiểm tra, vì luôn có 30% dữ liệu được sử dụng để tạo ra các cây quyết định từ phương pháp tăng cường.

Tuy nhiên, phương pháp này cũng sở hữu một số nhược điểm cần xem xét. Phương pháp Rừng ngẫu nhiên thường phức tạp hơn nhiều so với cây quyết định đơn lẻ, và do đó, đòi hỏi thời gian đào tạo nhiều hơn. Điều này có thể là một hạn chế đối với những tác vụ yêu cầu sự tinh chỉnh chi tiết hoặc yêu cầu thời gian đào tạo nhanh chóng. Tuy nhiên, với sự đa dạng và tính mạnh mẽ của nó, phương pháp Rừng ngẫu nhiên vẫn là một lựa chọn hấp dẫn cho nhiều ứng dụng học máy.

### 3.1.2. Đặc điểm và các bước xử lý thuật toán

Các đặc điểm quan trọng của rừng ngẫu nhiên:

*1. Đa dạng:* Không phải tất cả các thuộc tính/ biến/ đặc trưng được xem xét khi tạo cây quyết định riêng lẻ; mỗi cây là khác nhau.

*2. Xử lý với tập dữ liệu nhiều chiều:* Vì mỗi cây không xem xét tất cả các đặc trưng, không gian đặc trưng được giảm bớt.

*3. Tích hợp song song:* Mỗi cây được tạo ra độc lập từ dữ liệu và đặc trưng khác nhau. Điều này có nghĩa là chúng ta có thể tận dụng toàn bộ CPU để xây dựng các rừng ngẫu nhiên.

*4. Phân chia Tập huấn luyện - Tập kiểm tra:* Trong một rừng ngẫu nhiên, chúng ta không cần phân tách dữ liệu cho tập huấn luyện và tập kiểm tra vì luôn có 30% dữ liệu không được cây quyết định sử dụng.

*5. Ổn định:* Ổn định xuất phát từ việc kết quả dựa trên bỏ phiếu đa số hoặc trung bình.

Các bước trong thuật toán Rừng ngẫu nhiên

*Bước 1:* Trong mô hình Rừng ngẫu nhiên, một tập con của các điểm dữ liệu và một tập con của các đặc trưng được chọn để xây dựng mỗi cây quyết định.

*Bước 2:* Xây dựng cây quyết định riêng lẻ cho mỗi mẫu.

*Bước 3:* Mỗi cây quyết định sẽ tạo ra một kết quả.

*Bước 4:* Kết quả cuối cùng được xem xét dựa trên bỏ phiếu đa số hoặc trung bình cho phân loại và hồi quy, tương ứng

Đây là một trong những kỹ thuật tốt nhất với hiệu suất cao, được sử dụng rộng rãi trong nhiều ngành công nghiệp vì hiệu quả của nó. Nó có thể xử lý dữ liệu nhị phân, liên tục và phân loại. Nhìn chung, rừng ngẫu nhiên là một mô hình hiệu quả, đơn giản, linh hoạt và mạnh mẽ với một số hạn chế.

## 3.2. XGBoost

XGBoost là một thuật toán học máy rất mạnh mẽ và phổ biến, được phát triển bởi Tianqi Chen tại Đại học Washington. Nó là một biến thể của thuật toán tăng cường độ dốc, tập trung vào việc tối ưu hóa hàm mất mát thông qua việc sử dụng các độ dốc để điều chỉnh các cây quyết định. XGBoost có khả năng xử lý các bài toán có dữ liệu lớn và phức tạp, cũng như cung cấp các kỹ thuật tối ưu hóa để giảm thiểu việc quá khớp và tăng hiệu suất.

Được phát triển dựa trên nguyên lý của thuật toán tăng cường độ dốc, XGBoost không chỉ tối ưu hóa hiệu suất mô hình thông qua việc kết hợp nhiều mô hình yếu lại với nhau, mà còn cung cấp một loạt các tính năng tiên tiến như xử lý hiệu quả dữ liệu thưa thớt, chuẩn hóa để ngăn chặn quá khớp, và khả năng xử lý song song để tăng tốc quá trình đào tạo. XGBoost đã trở thành một công cụ quan trọng trong học máy, được sử dụng rộng rãi trong nhiều lĩnh vực từ phân loại và hồi quy cho đến xếp hạng và dự đoán do người dùng định nghĩa.

Chúng ta sẽ khám phá sâu hơn về cách XGBoost hoạt động, cũng như cách tối ưu hóa và cải tiến nó để đạt được hiệu suất tốt nhất.

### 3.2.1. Kiến thức tổng quan

Tăng cường là một kỹ thuật học tập kết hợp, mục tiêu là tạo ra một mô hình phân loại mạnh từ nhiều mô hình phân loại yếu. Quy trình này được thực hiện thông qua việc xây dựng một chuỗi các mô hình yếu.

Trong quá trình này, mô hình đầu tiên được xây dựng dựa trên dữ liệu đào tạo. Tiếp theo, mô hình thứ hai được xây dựng với mục tiêu cố gắng khắc phục các lỗi mà mô hình đầu tiên đã tạo ra. Quy trình này được lặp lại, với mỗi mô hình tiếp theo cố gắng cải thiện lỗi của mô hình trước đó.

Quy trình này tiếp tục cho đến khi toàn bộ tập dữ liệu đào tạo được dự đoán một cách chính xác hoặc cho đến khi đạt đến số lượng tối đa các mô hình được thêm vào.

Trong thế giới học máy, có ba loại thuật toán thuộc phương pháp tăng cường chính được sử dụng rộng rãi. Mỗi loại thuật toán này có những đặc điểm và ứng dụng riêng biệt, tùy thuộc vào yêu cầu cụ thể của tác vụ học máy.

*1. Thuật toán tăng cường thích ứng:* Adaboost là một kỹ thuật học tập kết hợp, nó kết hợp nhiều mô hình học yếu để tạo ra một mô hình học mạnh. Trong quá trình lặp đi lặp lại, Adaboost gán trọng số bằng nhau cho mỗi mẫu dữ liệu trong lần lặp đầu tiên. Nếu một dự đoán không chính xác xảy ra, Adaboost sẽ tăng trọng số cho quan sát đó. Quy trình này được lặp lại trong các giai đoạn lặp tiếp theo và tiếp tục cho đến khi đạt được mức độ chính xác mong muốn.

*2. Thuật toán tăng cường độ dốc:* GBM là một thuật toán tăng cường hoạt động dựa trên nguyên tắc của thuật toán suy giảm độ dốc. Suy giảm độ dốc là cốt lõi của các quy trình học máy hiện đại, được định nghĩa là một quá trình tối ưu hóa lặp đi lặp lại giúp tìm ra giá trị cực tiểu hoặc cực đại cục bộ của một hàm cho trước. Trong trường hợp này, hàm cho trước là hàm mất mát và suy giảm độ dốc hoạt động theo cách tiến hóa để giảm thiểu hàm mất mát bằng cách tìm các tham số tối ưu cho vấn đề học máy đang xử lý.

*3. Thuật toán tăng cường độ dốc cực đại :* XGBoost, như tên gọi, có thể được coi như GBM nhưng với sự tăng cường hoạt động tối ưu nhất. XGBoost là một thuật toán học máy trong đó các cây quyết định được xây dựng theo thứ tự tuần tự. Trọng số đóng một vai trò quan trọng trong thuật toán này. Cụ thể, trọng số được gán cho tất cả các biến độc lập, sau đó các biến này được đưa vào cây quyết định để dự đoán kết quả. Nếu một biến nào đó được dự đoán sai bởi cây, trọng số của biến đó sẽ được tăng lên. Sau đó, biến này sẽ được đưa vào cây quyết định tiếp theo. Quá trình này được lặp lại, với mỗi cây quyết định tiếp theo cố gắng cải thiện lỗi của cây trước đó. Cuối cùng, tất cả các cây quyết định này được kết hợp lại để tạo ra một mô hình mạnh mẽ và

chính xác hơn. XGBoost có thể được áp dụng cho nhiều loại vấn đề học máy, bao gồm hồi quy, phân loại, xếp hạng và các vấn đề dự đoán do người dùng định nghĩa.

Ưu điểm của XGBoost:

*1. Hiệu suất:* XGBoost tạo ra kết quả vượt trội trong các bài toán học máy khác nhau.

*2. Khả năng mở rộng:* XGBoost được thiết kế để đào tạo hiệu quả và mở rộng các mô hình học máy, làm cho nó phù hợp với các tập dữ liệu lớn.

*3. Tính tùy chỉnh:* XGBoost có một loạt các tham số có thể được điều chỉnh để tối ưu hóa hiệu suất, làm cho nó có thể tùy chỉnh cao.

*4. Xử lý giá trị thiếu:* XGBoost có hỗ trợ tích hợp để xử lý giá trị bị thiếu, làm cho nó dễ dàng làm việc với dữ liệu thế giới thực thường có giá trị bị thiếu.

*5. Khả năng giải thích:* Khác với một số thuật toán học máy có thể khó để giải thích, XGBoost cung cấp tính quan trọng của đặc trưng, cho phép hiểu rõ hơn về các biến quan trọng nhất trong việc đưa ra dự đoán.

Nhược điểm của XGBoost:

*1. Độ phức tạp tính toán:* XGBoost yêu cầu nhiều tài nguyên tính toán, đặc biệt khi đào tạo các mô hình lớn, làm cho nó ít phù hợp với các hệ thống có tài nguyên hạn chế.

*2. Quá khớp:* XGBoost có thể xảy ra quá khớp, đặc biệt khi được đào tạo trên các tập dữ liệu nhỏ hoặc khi sử dụng quá nhiều cây trong mô hình.

*3. Điều chỉnh tham số siêu dữ liệu:* XGBoost có nhiều tham số dùng để điều chỉnh, làm cho việc điều chỉnh tham số một cách chính xác để tối ưu hóa hiệu suất trở nên quan trọng. Tuy nhiên, việc tìm ra bộ tham số tối ưu có thể tốn nhiều thời gian và đòi hỏi chuyên môn.

*4. Yêu cầu về bộ nhớ:* XGBoost có thể yêu cầu nhiều bộ nhớ, đặc biệt khi làm việc với các tập dữ liệu lớn, không phù hợp với các hệ thống có tài nguyên bộ nhớ hạn chế.

### 3.2.2. Đặc điểm và các bước xử lý thuật toán

*Bước 1. Khởi tạo dữ liệu:* Đầu tiên, chúng ta cần chuẩn bị dữ liệu cho mô hình XGBoost. Điều này bao gồm việc chia dữ liệu thành tập huấn luyện và tập kiểm tra.

*Bước 2. Xây dựng mô hình XGBoost đầu tiên:* Sử dụng thư viện XGBoost, chúng ta khởi tạo một mô hình XGBoost với các tham số mặc định.

*Bước 3. Huấn luyện mô hình:* Mô hình được huấn luyện trên tập huấn luyện với một hàm mất mát cụ thể và một tập hợp các tham số.

*Bước 4. Đánh giá mô hình:* Sau khi mô hình đã được huấn luyện, chúng ta sử dụng nó để dự đoán trên tập kiểm tra và đánh giá hiệu suất của nó.

*Bước 5. Tinh chỉnh tham số:* Dựa trên kết quả đánh giá, chúng ta có thể điều chỉnh các tham số của mô hình để cải thiện hiệu suất.

*Bước 6. Trực quan hóa mô hình:* XGBoost cung cấp các công cụ để trực quan hóa mô hình, giúp chúng ta hiểu rõ hơn về cách mô hình đưa ra dự đoán.

*Bước 7. Lưu và sử dụng mô hình:* Cuối cùng, mô hình được lưu lại để sử dụng sau này

## 3.3. CatBoost

CatBoost được phát triển vào năm 2017 bởi các nhà nghiên cứu và kỹ sư học máy tại công ty công nghệ Yandex (Nga). CatBoost phục vụ nhiều mục đích chức năng như: Hệ thống đề xuất, Trợ lý cá nhân, Xe tự lái, Dự báo thời tiết, và nhiều nhiệm vụ khác. Giải thuật CatBoost là một thành viên khác của kỹ thuật tăng cường độ dốc trên cây quyết định. Một trong những tính năng độc đáo mà giải thuật CatBoost cung cấp là

khả năng tích hợp để làm việc với các loại dữ liệu đa dạng để giải quyết một loạt các vấn đề dữ liệu mà nhiều doanh nghiệp đối mặt. Đồng thời, CatBoost cung cấp độ chính xác giống như các giải thuật khác trong các mô hình dùng cây quyết định.

### 3.3.1. Kiến thức tổng quan

CatBoost hỗ trợ các đặc trưng số, danh mục và văn bản nhưng có kỹ thuật xử lý tốt cho dữ liệu danh mục. Giải thuật CatBoost có khá nhiều tham số để điều chỉnh các đặc trưng trong giai đoạn xử lý.

Tăng cường độ dốc là một giải thuật học máy mạnh mẽ hoạt động tốt khi được sử dụng để cung cấp giải pháp cho nhiều vấn đề khác nhau như phát hiện gian lận, hệ thống đề xuất, dự báo. Ngoài ra, nó có thể trả về kết quả vượt trội với số lượng dữ liệu ít. Khác với các giải thuật học máy khác chỉ hoạt động tốt sau khi học từ dữ liệu rộng lớn.

CatBoost có thể cải thiện hiệu suất của mô hình trong khi giảm quá khớp và thời gian dành cho việc tinh chỉnh. CatBoost có một số tham số để tinh chỉnh. Tuy nhiên, nó giảm bớt nhu cầu tinh chỉnh siêu tham số rộng lớn vì các tham số mặc định đã tạo ra kết quả tốt. Quá khớp là một vấn đề phổ biến trong tăng cường độ dốc, đặc biệt khi tập dữ liệu nhỏ hoặc nhiễu. CatBoost có một số đặc điểm giúp giảm quá khớp. Một trong số đó là kỹ thuật điều chuẩn dựa trên độ dốc mới gọi là tăng cường theo trình tự. Một đặc điểm khác là việc sử dụng tỷ lệ học tập theo từng vòng lặp, cho phép mô hình thích ứng với độ phức tạp của vấn đề tại mỗi vòng lặp.

Các giá trị thiếu là một vấn đề phổ biến trong các tập dữ liệu thực tế. Các khung công cụ tăng cường độ dốc truyền thống yêu cầu phải điền giá trị thiếu trước khi huấn luyện mô hình. Tuy nhiên, **CatBoost** có khả năng xử lý giá trị thiếu một cách tự động. Trong quá trình huấn luyện, thuật toán học cách di chuyển theo hướng tối ưu trên độ dốc cho mỗi giá trị thiếu, dựa trên các mẫu trong dữ liệu.

Sau khi xử lý và làm sạch dữ liệu, dữ liệu được chuyển đổi thành các đặc trưng số học để máy tính hiểu và đưa ra dự đoán. Quá trình mã hóa hoặc chuyển đổi này tốn nhiều thời gian và tài nguyên tính toán. CatBoost hỗ trợ làm việc với các đặc trưng không phải số, giúp tiết kiệm thời gian và cải thiện kết quả huấn luyện.

CatBoost cung cấp giao diện thân thiện với người dùng. Đồng thời, thuật toán CatBoost được sử dụng trong Python với scikit-learn, R và các giao diện dòng lệnh. Đồng thời, CatBoost sử dụng GPU phân tán, cho phép CatBoost học nhanh hơn và thực hiện dự đoán nhanh nhiều lần so với các thuật toán khác. Mặt khác, CatBoost có khả năng tìm ra những đặc trưng quan trọng trên tập dữ liệu. Nó cũng hỗ trợ việc trực quan hóa, giúp ta hiểu cấu trúc của mô hình.

CatBoost có nhiều ưu điểm nổi bật:

*1. Thời gian huấn luyện ngắn:* Khác với một số mô hình học kết hợp, CatBoost hoạt động tốt với tập dữ liệu nhỏ. Tuy nhiên, cần lưu ý về việc quá khớp và điều chỉnh các tham số nhằm tạo ra kết quả tốt hơn.

*2. Làm việc với tập dữ liệu nhỏ:* Đây là một trong những điểm mạnh của thuật toán CatBoost. Giả sử tập dữ liệu có các đặc trưng phân loại và việc chuyển đổi chúng sang định dạng số học là một công việc khá lớn. Trong trường hợp này, ta có thể tận dụng sức mạnh của CatBoost để xây dựng mô hình.

*3. Làm việc với tập dữ liệu phân loại:* CatBoost phù hợp khi làm việc với tập dữ liệu phân loại. Quá trình chia nhỏ, cấu trúc cây và quá trình huấn luyện được tối ưu hóa để đem lại kết quả tốt nhất.

Mặc dù CatBoost có nhiều ưu điểm, nhưng cũng có một số khuyết điểm cần xem xét:

*1. Tốn thời gian huấn luyện với dữ liệu lớn:* CatBoost hiệu quả với dữ liệu nhỏ, nhưng với tập dữ liệu lớn, quá trình huấn luyện có thể mất nhiều thời gian.

*2. Khả năng quá khớp:* Như bất kỳ thuật toán nào khác, CatBoost cũng tồn tại nguy cơ quá khớp. Cần kiểm tra và điều chỉnh các tham số để tránh hiện tượng này.

*3. Khó kiểm soát các tham số:* CatBoost tự động xử lý nhiều tham số, nhưng việc điều chỉnh những tham số quan trọng để tối ưu hóa kết quả vẫn cần sự hiểu biết về thuật toán.

### 3.3.2. Đặc điểm và các bước xử lý thuật toán

Dưới đây là các bước hoạt động cơ bản của CatBoost:

*Bước 1. Tính toán sơ bộ các phân chia:* CatBoost xác định các phân chia sơ bộ ban đầu dựa trên các đặc trưng và giá trị nhãn trong tập dữ liệu huấn luyện.

*Bước 2. Xử lý các đặc trưng phân loại:* CatBoost hỗ trợ làm việc với các đặc trưng phân loại mà không cần mã hóa chúng thành đặc trưng số học. Điều này giúp tiết kiệm thời gian và cải thiện kết quả huấn luyện.

*Bước 3. Xây dựng cây quyết định:* CatBoost dựa trên việc xây dựng một tập hợp các cây quyết định liên tiếp. Mỗi cây tiếp theo được xây dựng với tổn thất giảm so với các cây trước đó. Số lượng cây được kiểm soát bằng các tham số khởi đầu. Để ngăn chặn quá khớp, sử dụng bộ phát hiện quá khớp.

*Bước 4. Tính toán giá trị trong các lá:* CatBoost tính toán giá trị dự đoán trong các lá của cây. Các giá trị này được sử dụng để đưa ra dự đoán cho các đối tượng mới.

XGBoost và CatBoost đều là các thuật toán mạnh mẽ trong lĩnh vực học máy, tuy nhiên, mỗi thuật toán có những đặc điểm và ưu điểm riêng. XGBoost tập trung vào việc tối ưu hóa hàm mất mát và xử lý các dữ liệu lớn, trong khi CatBoost chủ yếu tập trung vào việc xử lý dữ liệu phân loại một cách tự động và hiệu quả

## 3.4. LightGBM

LightGBM là một framework sử dụng thuật toán tăng cường độ dốc và cây quyết định. Khác với các phương pháp tăng cường độ dốc truyền thống khác, LightGBM

xây dựng cây quyết định bằng cách sử dụng phương pháp dựa trên biểu đồ tần suất để phân chia các giá trị đặc trưng liên tục thành các bin.

Phương pháp này giúp giảm việc sử dụng bộ nhớ và tăng tốc quá trình huấn luyện, biến nó trở thành một trong những thuật toán tăng cường nhanh nhất hiện có. Hơn nữa, nó có thể dễ dàng xử lý các tập dữ liệu quy mô lớn, là lựa chọn phổ biến cho các ứng dụng như phát hiện bất thường và phân loại hình ảnh.

### 3.4.1. Kiến thức tổng quan

LightGBM là một mô hình học kết hợp được sử dụng rộng rãi nhờ khả năng xử lý các tập dữ liệu lớn với những đặc trưng phức tạp. Mô hình này dựa trên thuật toán tăng cường độ dốc và huấn luyện cây quyết định, hiệu quả trong việc xử lý cả dữ liệu có cấu trúc và phi cấu trúc.

Một trong những điểm mạnh của LightGBM là khả năng xử lý các tập dữ liệu lớn với đặc trưng phức tạp, là lựa chọn phổ biến cho các ứng dụng thực tế. Nó còn có các ưu điểm so với các framework khác, bao gồm tốc độ huấn luyện nhanh hơn, sử dụng bộ nhớ thấp hơn và độ chính xác cao hơn. Mô hình hỗ trợ học song song, phân tán cũng như học trên GPU.

Một trong những điểm mạnh của LightGBM là tốc độ. LightGBM sử dụng phương pháp dựa trên biểu đồ tần suất để tăng tốc quá trình huấn luyện, điều này có thể là một lợi thế lớn khi xử lý các tập dữ liệu lớn. Phương pháp học dựa trên biểu đồ tần suất, nhóm dữ liệu thành các khoảng giá trị, giảm số lượng tính toán cần thiết để xây dựng cây quyết định. Từ đó, mô hình sở hữu thời gian huấn luyện nhanh hơn và sử dụng tài nguyên tính toán hiệu quả hơn. Ngoài ra, LightGBM có tính tùy chỉnh cao, với nhiều siêu tham số khác nhau nhằm cải thiện hiệu suất.

Ưu điểm của LightGBM:

*1. Hiệu suất cao:* LightGBM vượt trội so với các framework khác về tốc độ huấn luyện và kích thước tập dữ liệu, xử lý các tập dữ liệu lớn với hàng triệu đặc trưng.

*2. Tốc độ học nhanh:* LightGBM sử dụng thuật toán Gradient-based One-Side Sampling (GOSS), một biến thể của thuật toán tăng cường độ dốc truyền thống giúp giảm thời gian tìm phân chia tốt nhất cho mỗi nút cây bằng cách thực hiện ít tính toán hơn.

*3. Xử lý đặc biệt:* LightGBM có thể xử lý dữ liệu thưa hoặc chứa giá trị bị thiếu. Đồng thời, triển khai các thuật toán đặc biệt cho nhiệm vụ xếp hạng và phân loại đa lớp.

*4. Hỗ trợ tích hợp:* LightGBM cho phép tích hợp tinh chỉnh siêu tham số nhằm tìm ra bộ tham số tối ưu cho bài toán.

*5. Hỗ trợ tự động:* LightGBM có hệ thống hỗ trợ tự động cho phép người dùng nhanh chóng gửi báo cáo lỗi và yêu cầu tính năng, giúp đội phát triển cải thiện thư viện một cách tốt hơn.

Khuyết điểm của LightGBM:

*1. Giới hạn về lựa chọn đặc trưng:* LightGBM không thể chọn ra các đặc trưng quan trọng cho một tập dữ liệu cụ thể. Thông thường, LightGBM sẽ chọn các đặc trưng có mối tương quan cao nhất với mục tiêu. Tuy nhiên, điều này không nhất thiết có nghĩa rằng những đặc trưng được chọn này là quan trọng nhất hoặc có khả năng dự đoán cao nhất.

*2. Quá khớp:* LightGBM có nguy cơ bị hiện tượng quá khớp khi mô hình tìm ra các mẫu trong dữ liệu huấn luyện mà không tồn tại trong dữ liệu kiểm tra.

*3. Ràng buộc về thời gian và bộ nhớ:* LightGBM có khả năng tốn nhiều thời gian và bộ nhớ khi xây dựng một bộ cây quyết định phức tạp, gây khó khăn cho các tập dữ liệu lớn.

*4. Siêu tham số:* Tìm ra sự kết hợp tốt nhất của các siêu tham số để tối đa hóa độ chính xác có thể gặp khó khăn và tốn thời gian.

### 3.4.2. Đặc điểm và các bước xử lý thuật toán

LightGBM là một thuật toán tăng cường hoạt động bằng cách kết hợp nhiều cây quyết định yếu để tạo ra một mô hình mạnh.

Thuật toán bắt đầu bằng việc tạo một cây quyết định dự đoán biến mục tiêu dựa trên các đặc trưng đầu vào. Sau đó, việc lặp lại nhiều cây quyết định khác vào mô hình, với mỗi cây cố gắng sửa chữa các sai sót của cây trước đó.

Một số đặc điểm quan trọng của LightGBM làm cho nó đặc biệt hiệu quả, bao gồm việc sử dụng học cây quyết định, phương pháp học dựa trên biểu đồ tần suất, tăng trưởng cây theo chiều lá và tinh chỉnh tham số '*regularization'*.

*1. Tăng cường độ dốc:*

LightGBM sử dụng thuật toán tăng cường độ dốc để liên tục cải thiện hiệu suất của các cây quyết định. Phương pháp này thêm một cây mới vào mô hình ở mỗi vòng lặp để sửa chữa những khuyết điểm của các cây trước đó.

Để làm điều này, nó tính toán độ dốc của hàm mất mát đối với dự đoán của mô hình hiện tại và sử dụng thông tin này để cập nhật trọng số của các mẫu huấn luyện. Trọng số được điều chỉnh sau đó được sử dụng để huấn luyện một cây mới tập trung vào các trường hợp mà các cây trước đó đã phân loại sai. Phương pháp này được lặp lại cho đến khi điều kiện dừng được thỏa mãn.

Một lợi thế của thuật toán này là xử lý dữ liệu phức tạp với số lượng đặc trưng lớn, bởi vì thuật toán có thể xác định mối quan hệ phi tuyến giữa các đặc trưng và biến mục tiêu.

*2. Phương pháp học dựa trên biểu đồ tần suất:*

Phương pháp học dựa trên biểu đồ tần suất là một kỹ thuật mà LightGBM sử dụng để tăng tốc quá trình tính toán liên quan đến việc học cây quyết định và thuật toán tăng cường độ dốc.

Phương pháp này bao gồm việc nhóm các giá trị đặc trưng liên tục thành các khoảng giá trị rời rạc, hoặc biểu đồ tần suất, và sử dụng các khoảng giá trị này để tính toán thông tin tại mỗi điểm chia.

Phương pháp dựa trên biểu đồ tần suất mang lại nhiều lợi ích so với các phương pháp trước đó. Vì các khoảng giá trị được tính toán trước và sử dụng lại qua các vòng lặp, giảm việc sử dụng bộ nhớ và chi phí tính toán của thuật toán.

Khi xây dựng cây quyết định, LightGBM không tính toán phân chia tại tất cả các giá trị của đặc trưng. Thay vào đó, nó chọn các điểm phân chia thông minh để xử lý dữ liệu lệch hoặc dữ liệu thưa một cách hiệu quả hơn.

LightGBM sử dụng phương pháp lấy mẫu một phía dựa trên độ dốc (GOSS) để cải thiện hiệu suất của phương pháp dựa trên biểu đồ tần suất. GOSS xác định các mẫu góp phần nhiều nhất vào độ dốc và sử dụng chúng thường xuyên hơn trong quá trình huấn luyện. Điều này giúp giảm tổng số mẫu cần đánh giá trong khi vẫn duy trì độ chính xác của mô hình.

Tổng quan, phương pháp học dựa trên biểu đồ tần suất là một tính năng quan trọng giúp LightGBM trở thành một thuật toán nhanh chóng và hiệu quả.

*3. Tăng trưởng cây theo chiều lá:*

Tăng trưởng cây theo chiều lá là một kỹ thuật mà LightGBM sử dụng để xây dựng cây quyết định một cách hiệu quả hơn. Thay vì phát triển cây theo từng cấp, LightGBM phát triển cây bằng cách mở rộng nút lá có tác động cao nhất.

Phương pháp này có một số ưu điểm so với các phương pháp truyền thống, tạo ra một cấu trúc cây cân bằng và chính xác hơn, vì nó chọn phân chia tại nút lá có tác động lớn nhất đối với hàm mất mát tổng thể. Một lợi ích khác là nó giảm chi phí tính toán của việc xây dựng cây, vì không yêu cầu tính toán tất cả các phân chia có thể tại mỗi cấp.

Tuy nhiên, tăng trưởng cây theo chiều lá có thể dẫn đến hiện tượng quá khớp nếu không được điều chỉnh đúng cách. Để ngăn chặn hiện tượng quá khớp, LightGBM sử dụng một số kỹ thuật điều chuẩn tham số, chẳng hạn như độ sâu tối đa và dữ liệu tối thiểu trong lá, để hạn chế sự phát triển của cây.

*4. Tinh chỉnh tham số 'regularization':*

Điều chỉnh tham số '*regularization'* là một trong những cách để ngăn chặn hiện tượng quá khớp và cải thiện khả năng tổng quát của mô hình. LightGBM sử dụng nhiều phương pháp điều chỉnh để tăng tính linh hoạt và độ chính xác của mô hình.

Tinh chỉnh tham số *'độ sâu tối đa' (max_depth)* là một trong những kỹ thuật điều chuẩn phổ biến được áp dụng trong LightGBM. Lựa chọn này giới hạn độ sâu tối đa của cây quyết định, ngăn mô hình phát triển quá phức tạp và quá khớp dữ liệu huấn luyện.

Một kỹ thuật điều chuẩn khác được sử dụng trong LightGBM là tinh chỉnh tham số *'số dữ liệu tối thiểu trong lá' (min_data_in_leaf)*, xác định số lượng tối thiểu các điểm dữ liệu cần có trong mỗi nút lá của cây quyết định. Điều này giúp giảm nguy cơ quá khớp trong dữ liệu.

Dưới đây là quy trình tổng quan về cách thuật toán LightGBM hoạt động:

*Bước 1. Xử lý dữ liệu:* Bước đầu tiên là chuẩn bị dữ liệu cho việc huấn luyện mô hình. Điều này bao gồm xử lý giá trị bị thiếu, chuyển đổi đặc trưng danh mục thành số và chuẩn hóa nếu cần.

*Bước 2. Chia dữ liệu:* Dữ liệu sau đó được chia thành tập huấn luyện và tập xác thực. Tập huấn luyện được sử dụng để huấn luyện mô hình, và tập xác thực được sử dụng để đánh giá hiệu suất của mô hình trong quá trình huấn luyện.

*Bước 3. Khởi tạo:* LightGBM bắt đầu với một cây quyết định duy nhất được phát triển từ nút gốc.

*Bước 4. Phân chia các nút:* Thuật toán sau đó phân chia các nút bằng cách sử dụng phương pháp dựa trên độ dốc. Nó tính toán độ dốc của hàm mất mát đối với các giá trị dự đoán và tìm phân chia tốt nhất.

*Bước 5. Xây dựng cây:* Thuật toán tiếp tục xây dựng cây quyết định bằng cách đệ quy phân chia các nút cho đến khi thỏa mãn điều kiện dừng. Điều này có thể là độ sâu tối đa của cây, số lượng mẫu tối thiểu cho mỗi lá.

*Bước 6. Tăng cường độ dốc:* Sau khi xây dựng cây quyết định đầu tiên, LightGBM tạo ra một tập hợp các cây bằng cách liên tục thêm cây mới để sửa chữa sai sót của các cây trước đó. Mỗi cây mới được huấn luyện trên các sai số của cây trước đó, tức là sự khác biệt giữa các giá trị dự đoán và giá trị thực.

*Bước 7. Ngăn chặn quá khớp:* LightGBM áp dụng nhiều kỹ thuật để ngăn chặn hiện tượng quá khớp, chẳng hạn như cắt tỉa cây, giới hạn số lá và áp dụng điều chuẩn L1 hoặc L2 cho trọng số lá.

*Bước 8. Dự đoán:* Khi tập hợp các cây đã được huấn luyện, LightGBM sử dụng nó để thực hiện dự đoán trên dữ liệu mới bằng cách lấy trung bình của các dự đoán của tất cả các cây trong tập hợp.

*Bước 9. Đánh giá mô hình:* Hiệu suất của mô hình được đánh giá trên tập xác thực bằng các chỉ số phù hợp như độ chính xác hoặc AUC-ROC.

*Bước 10. Tinh chỉnh tham số:* Cuối cùng, các siêu tham số của mô hình được điều chỉnh bằng cách sử dụng tìm kiếm lưới hoặc tìm kiếm ngẫu nhiên để cải thiện hiệu suất của mô hình.

LightGBM là một thuật toán phổ biến cho nhiệm vụ phát hiện gian lận do khả năng xử lý các tập dữ liệu không cân bằng, tốc độ huấn luyện nhanh và độ chính xác cao. Bằng cách huấn luyện mô hình LightGBM trên một tập dữ liệu chứa cả giao dịch gian lận và không gian lận, chúng ta có thể phát hiện và ngăn chặn các hoạt động gian lận trong thời gian thực.

# CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT QUẢ

## 4.1. Bộ dữ liệu sử dụng

### 4.1.1. Mô tả bộ dữ liệu

Tập dữ liệu chứa các giao dịch được thực hiện bằng thẻ tín dụng vào tháng 9 năm 2013 bởi chủ thẻ ở Châu Âu.

Tập dữ liệu trình bày các giao dịch xảy ra trong hai ngày và có 492 vụ gian lận trong tổng số 284.807 giao dịch. Tập dữ liệu rất mất cân bằng; loại tích cực (lừa đảo) chiếm 0.172% tổng số giao dịch.

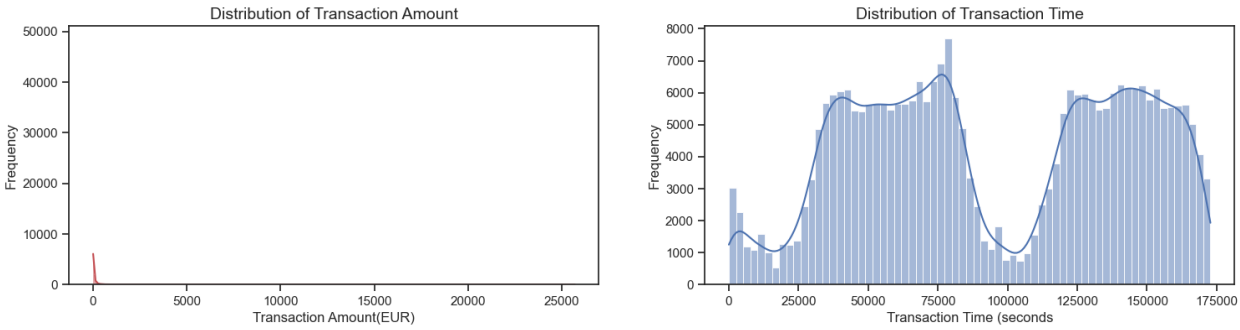
Tập dữ liệu này chứa các biến đầu vào dạng số, là kết quả của phép biến đổi PCA. Do các vấn đề về bảo mật thông tin danh tính người dùng, bộ dữ liệu không cung cấp các tính năng ban đầu hoặc thông tin cơ bản về chúng. Những đặc trưng V1, V2, …, V28 là những thành phần chính có được từ PCA, các đặc trưng duy nhất chưa được chuyển đổi bằng PCA là “Thời gian” và “Số tiền”. Đặc trưng “Thời gian” chứa số giây trôi qua giữa mỗi giao dịch và giao dịch đầu tiên trong tập dữ liệu. Đặc trưng 'Số tiền' là số tiền giao dịch, đặc trưng này có thể được sử dụng cho việc học phụ thuộc vào ví dụ và nhạy cảm với chi phí. Đặc trưng 'Lớp' nhận giá trị 0 trong trường hợp bình thường và 1 trong trường hợp gian lận.

### 4.1.2. Tiền xử lý dữ liệu

Tập dữ liệu bao gồm các giao dịch thẻ tín dụng trong thời gian hai ngày, hầu hết các thuộc tính đều là số thập phân là kết quả của phép biến đổi PCA và được sử dụng để dự đoán liệu một giao dịch có bình thường hay gian lận. Kết quả cuối cùng cho bài toán phân loại nhị phân được mã hóa là '0' và '1', tương ứng với 'bình thường' và 'gian lận'.

EDA được thực hiện trên tập dữ liệu để có cái nhìn tổng quan và cấu trúc cơ bản của dữ liệu. Mặc dù giao dịch gian lận chỉ chiếm 0.172% trong tổng lượng giao dịch, điều này chứng tỏ rằng tuy không thường xuyên xảy ra các giao dịch gian lận, nhưng nếu không phát hiện chúng thì sẽ dẫn đến các thiệt hại lớn về tài chính. Có những giá trị ngoại lệ trong biến 'Amount' đối với giao dịch bình thường, nhưng không có giá trị ngoại lệ nào đối với giao dịch gian lận. Không có giá trị thiếu và không có giá trị null trong bộ dữ liệu.

Hai biểu đồ về phân phối số tiền giao dịch (Distribution of transaction amount) và phân phối thời gian giao dịch (Distribution of transaction time) mô tả rõ hơn hoạt động cũa từng giao dịch. Trong khi các giá trị của lượng giao dịch biến ‘Amount’ khá nhỏ, giá trị biến 'Time' của giao dịch lừa đảo và giao dịch bình thường được xem xét. Nếu so sánh theo số lượng giao dịch trong cả hai trường hợp, thuộc tính này có vẻ giống nhau, nhưng các giao dịch bình thường được phân bố đều, trong khi giao dịch gian lận thì không.




*Hình 21. Biểu đồ về phân phối số tiền giao dịch và thời gian giao dịch*

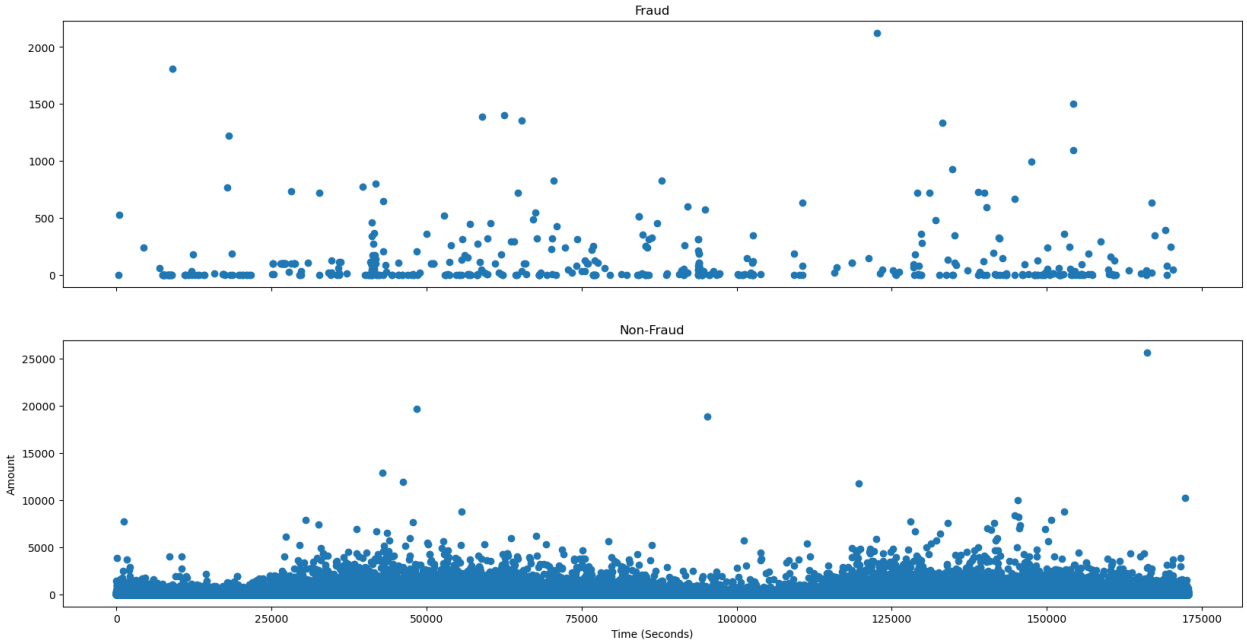
Vì thế, cột chứa các giá trị thời gian trong tập dữ liệu không quá hữu ích và có thể được loại bỏ. Ngoài ra, các số lượng giao dịch bình thường và gian lận được xem khi dữ liệu được scale. Có thể thấy, hầu hết các số tiền giao dịch bình thường khá nhỏ, với giá trị trung bình khoảng 88 và phần trung bình 50% của quan sát nằm giữa 5 và 77.

```
# Xem số lượng giao dịch Not - Fraud trong trường hợp Scale nhỏ
print(df.Amount[df.Class == 0 ].describe())
```

```
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

*Hình 22. Dataframe biểu diễn các thông số thống kê của các giao dịch bình thường*

Giá trị lớn nhất là khoảng 25.691, đang làm cho phân phối nghiêng lên và có thể là một điểm ngoại lệ. Chẳng hạn như ai đó đã mua một chiếc ô tô bằng thẻ tín dụng của họ. Vì thế, biểu đồ phân tán cho đặc trưng thời gian 'Time' và đặc trưng số tiền giao dịch 'Amount' để có cái nhìn tốt hơn về cách dữ liệu được phân bố.



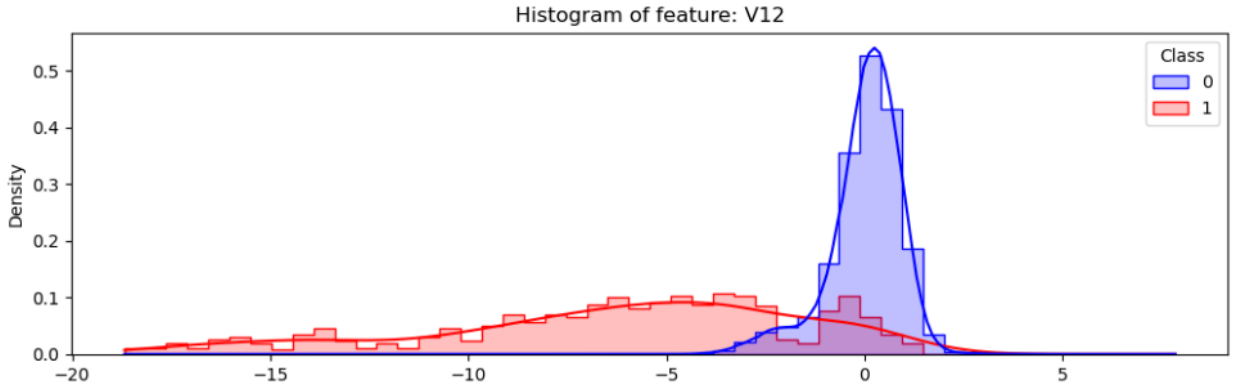

*Hình 23. Biểu đồ phân tán cho các đặc trưng thời gian 'Time' và số tiền giao dịch 'Amount'*

Ngoài ra, phân phối của các đặc trưng trong tập dữ liệu có thể được xem xét sử dụng biểu đồ tần suất (biểu đồ tần suất) cho từng biến. Vì có quá nhiều biến, các biểu đồ cho ta cái nhìn tổng quan về những đặc trưng nào cần thiết cho việc huấn luyện mô hình. Ở đây, ta chia thành ba nhóm như sau:

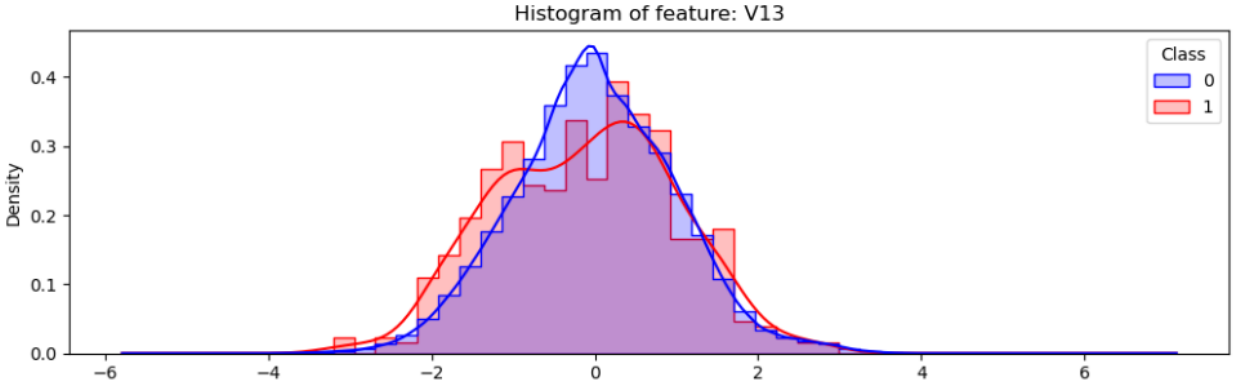
*Nhóm 1:* bao gồm các đặc trưng quan trọng, nên dùng do tần suất giữa giao dịch bình thường và gian lận khác nhau, ví dụ như biểu đồ của đặc trưng V12.

*Nhóm 2:* bao gồm các đặc trưng không quan trọng, không nên dùng do tần suất giữa giao dịch bình thường và gian lận quá giống nhau, ví dụ như biểu đồ của đặc trưng V13.

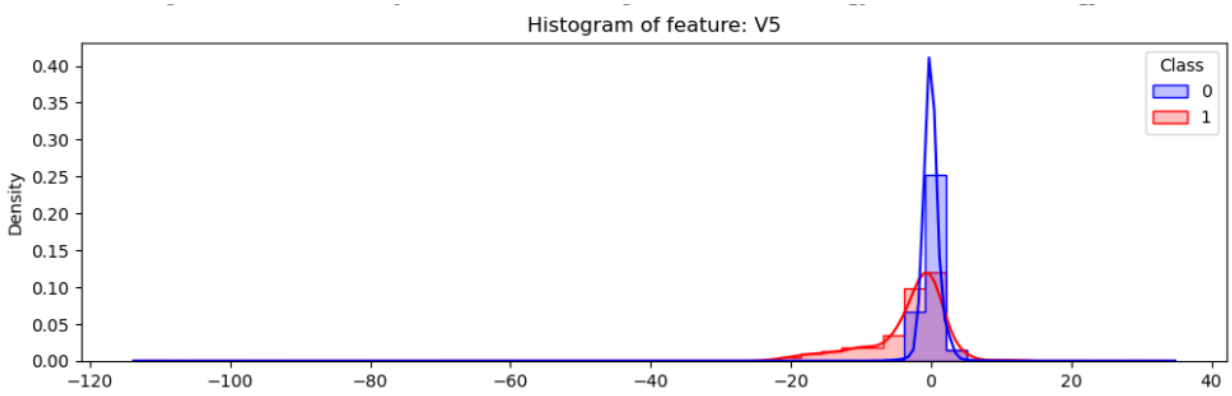
*Nhóm 3:* bao gồm các đặc trưng không chắc chắn, cần được kiểm thử do tần suất giữa giao dịch bình thường và gian lận gần giống nhau, ví dụ như biểu đồ của đặc trưng V5.




*Hình 24. Biểu đồ tần suất  của đặc trưng V12*




*Hình 25. Biểu đồ tần suất  của đặc trưng V13*

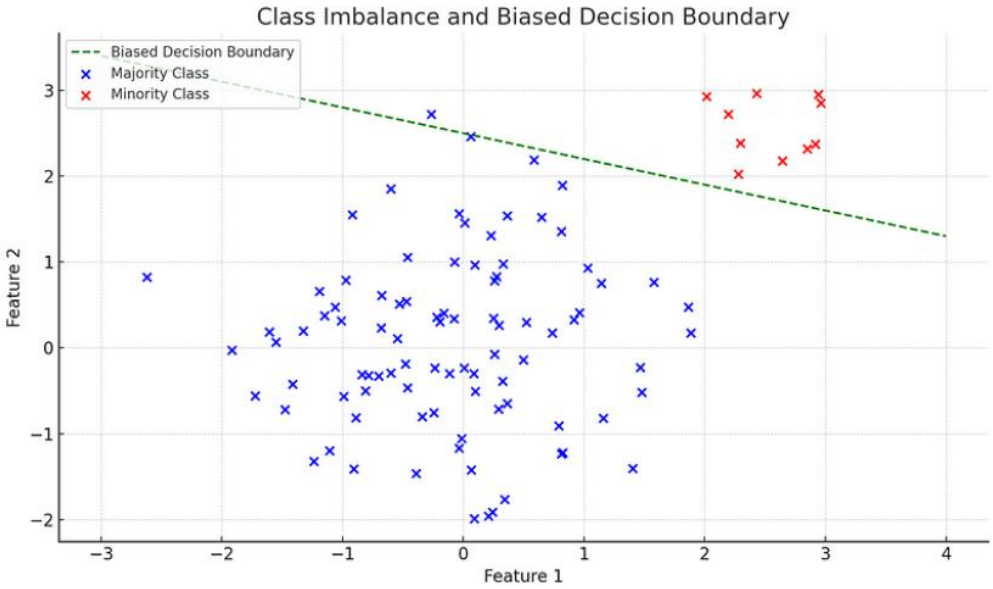



*Hình 26. Biểu đồ tần suất  của đặc trưng V5*

Tuy hầu hết các đặc trưng đã là kết quả của phép biến đổi PCA và việc lựa chọn cho việc huấn luyện mô hình không quá cần thiết nhưng chúng đóng vai trò quan trọng trong việc hiểu sâu thuộc tính của từng đặc trưng khi áp dụng trên các tập dữ liệu tương tự.

Mặt khác, khi gặp những bài toán về phân lớp, trong đó phân phối của dữ liệu thuộc các lớp không cân bằng, một lớp có thể có nhiều dữ liệu đáng kể hơn so với lớp khác, là một tình huống khá phổ biến. Tình huống này được biết đến là mất cân bằng lớp. Lớp được phân loại thành lớp majority (lớp đa số hoặc lớp "âm") và lớp minority (lớp thiểu số hoặc lớp "dương"). Thuật ngữ "dương" và "âm" thường phụ thuộc vào ngữ cảnh của vấn đề.

Chẳng hạn như, trong vấn đề chẩn đoán y tế, lớp "dương" có thể là sự xuất hiện của một căn bệnh. Mất cân bằng lớp có thể xảy ra trong nhiều lĩnh vực, như phát hiện thư rác trong email (nơi hầu hết các email không phải là thư rác), chẩn đoán y tế (nơi hầu hết bệnh nhân có thể không có một căn bệnh hiếm, và phát hiện gian lận tài chính (nơi các giao dịch gian lận thường chỉ là một phần nhỏ của tất cả các giao dịch).




*Hình 27. Đồ thị biểu hiện mất cân bằng lớp*

*Nguồn: https://medium.com/@gabrielpierobon/dealing-with-imbalanced-classes-in-supervised-learning-a9979ba498b4*

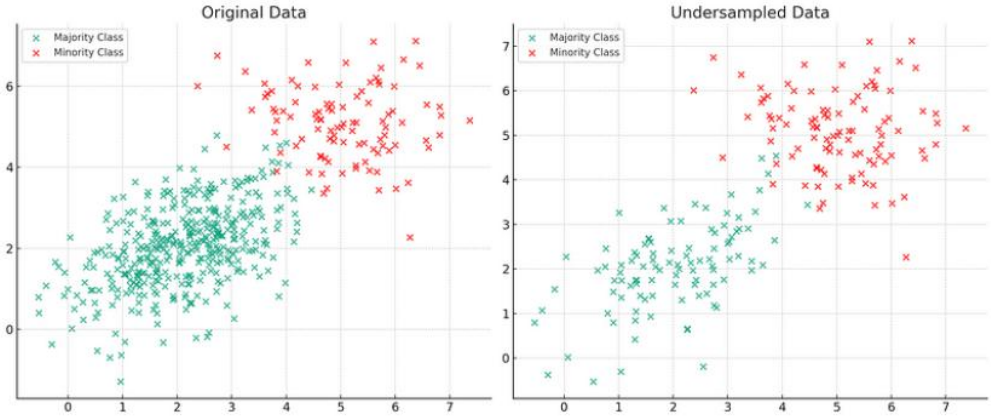
Trong khi nhiều thuật toán học máy hoạt động tốt nhất với phân phối cân bằng giữa các lớp, tình huống mất cân bằng lớp xảy ra khá thường xuyên và dẫn đến hậu quả nghiêm trọng đến hiệu suất mô hình, đặc biệt là đối với lớp thiểu số. Các mô hình truyền thống thường giả định rằng các lớp trong phân loại là cân bằng. Khi xảy ra phân phối các lớp mất cân bằng, những mô hình này có xu hướng chênh lệch đáng kể về lớp đa số, dẫn đến hiệu suất dự đoán kém cho lớp thiểu số.

Điều này dẫn đến việc lớp đa số được dự đoán hầu hết thời gian, độ chính xác rất cao nhưng độ thu hồi hoặc độ chính xác thấp cho lớp thiểu số. Điều này xảy ra vì mô hình "học" rằng nó có thể tối đa hóa độ chính xác tổng thể (thường là chỉ số mặc định để tối ưu hóa mô hình) bằng cách luôn dự đoán cho lớp đa số.

Tuy nhiên, trong nhiều tình huống thực tế, chi phí của việc phân loại sai các trường hợp lớp thiểu số có thể cao đáng kể. Cụ thể trong vấn đề phát hiện gian lận, việc không xác định được một giao dịch gian lận (một trường hợp của lớp thiểu số) có thể mang lại hậu quả nghiêm trọng hơn so với việc phân loại sai một giao dịch không gian lận.

Có nhiều phương pháp để xử lý các lớp mất cân bằng. Mỗi phương pháp đều có ưu điểm và nhược điểm của mình, các kỹ thuật resampling sẽ được sử dụng ở đây để điều chỉnh phân phối lớp trong dữ liệu huấn luyện. Trong đó bao gồm:

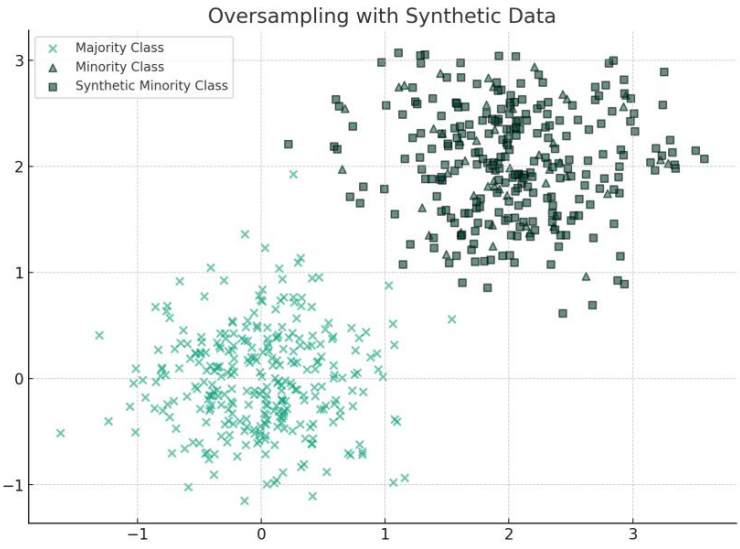
*Phương pháp giảm thiểu số mẫu lớp đa số (Undersampling):* giảm số lượng trường hợp của lớp đa số. Mặc dù điều này có thể giúp cân bằng giữa các lớp, nhưng nó cũng dẫn đến mất mát dữ liệu, điều có thể gây vấn đề nếu tập dữ liệu gốc không đủ lớn.




*Hình 28. Đồ thị biểu diễn undersampling*

*Nguồn: https://medium.com/@gabrielpierobon/dealing-with-imbalanced-classes-in-supervised-learning-a9979ba498b4*

*Phương pháp tăng cường số mẫu lớp thiểu số (Oversampling):* bao gồm việc tăng số lượng trường hợp của lớp thiểu số. Mục tiêu là tạo ra một tập dữ liệu cân bằng hơn bằng cách tăng số lượng mẫu của lớp thiểu số. Một phương pháp phổ biến là SMOTE, tạo ra các trường hợp tổng hợp của lớp thiểu số. Tuy nhiên, có thể dẫn đến hiện tương quá khớp, khi mô hình quá tập trung vào dữ liệu thiểu số.



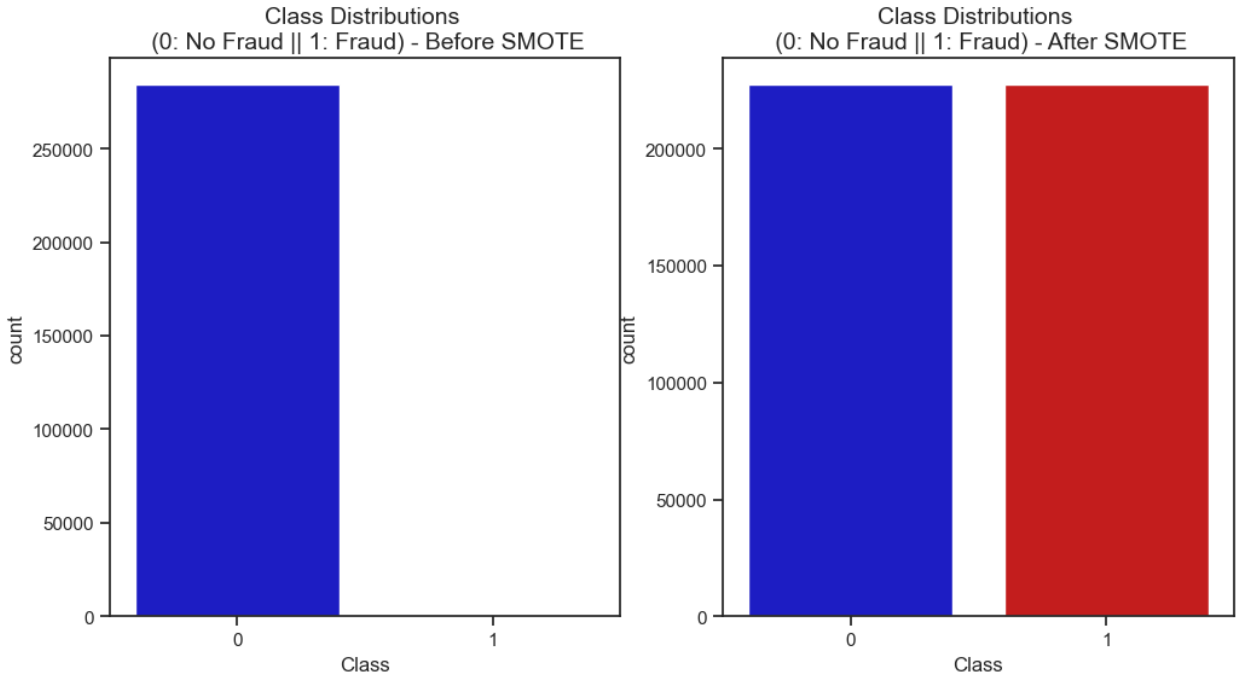

*Hình 29. Đồ thị biểu diễn oversampling*

*Nguồn: https://medium.com/@gabrielpierobon/dealing-with-imbalanced-classes-in-supervised-learning-a9979ba498b4*

Với tập dữ liệu khá mất cân bằng, phương pháp lấy mẫu SMOTE sẽ được sử dụng cho việc huấn luyện các mô hình được đề xuất. Phương pháp này kết hợp việc chọn lựa dữ liệu được sao chép của lớp thiểu số. Một phần của dữ liệu được chọn ngẫu nhiên từ lớp thiểu số, sau đó các mẫu tổng hợp mới cho lớp thiểu số được tạo ngẫu nhiên dựa trên thuật toán k láng giềng gần nhất, sử dụng khoảng cách Euclid giữa các quan sát. Các mẫu này sau đó được thêm vào tập huấn luyện gốc. Điều này dẫn đến việc có nhiều giao dịch gian lận hơn xuất hiện trong dữ liệu huấn luyện, giúp việc nhận diện các mô hình gian lận từ dữ liệu trở nên hiệu quả.

Một tập dữ liệu huấn luyện lớn, khoảng 227.845 quan sát sẽ khiến việc huấn luyện các mô hình trở nên tốn kém tài nguyên tính toán. Tuy nhiên, các tham số lấy mẫu đã được điều chỉnh thủ công sao cho lớp thiểu số được tăng kích thước lên 20 lần so với lớp đa số trong tập dữ liệu huấn luyện ban đầu và tránh trường hợp tồn tại quá ít dữ liệu huấn luyện cho mô hình nếu sử dụng các phương pháp undersampling. Ngoài ra, sử dụng phương pháp SMOTE đòi hỏi cách xử lý dữ liệu tỉ mỉ sao cho các mô hình không bị quá khớp.

Chính vì vậy, nếu chỉ sử dụng độ chính xác làm độ đo chính thức cho tình huống mất cân bằng lớp thì sẽ không đáng tin cậy khi đánh giá hiệu suất các mô hình. Bộ dữ liệu phát hiện gian lận thẻ tín dụng gần như có 99% trường hợp thuộc lớp đa số. Một mô hình chỉ cần dự đoán lớp đa số cho mọi trường hợp sẽ có độ chính xác là 99%, mặc dù hoàn toàn thất bại trong việc nhận diện lớp thiểu số. Do đó, ta sẽ sử dụng các độ đo khác để đánh giá các trường hợp gian lận và không gian lận và xác định mô hình huấn luyện nào là tốt nhất để tối ưu hóa cho việc phát hiện gian lận tài chính.




*Hình 30. Biểu đồ phân phối cân bằng lớp trước và sau khi sử dụng SMOTE*

## 4.2. Kết quả từ các mô hình

Các phương pháp học trên cây quyết định, sẽ dự đoán lớp biến mục tiêu hay lớp đặc trưng phân loại "Lớp" (dùng để phân biệt một giao dịch là bình thường hoặc gian lận) bằng cách học các quy tắc quyết định đơn giản từ tập dữ liệu huấn luyện, và được sử dụng để xây dựng các mô hình học kết hợp. Các mô hình này sử dụng các phương pháp được đề xuất như trên nhằm đưa ra kết quả các độ đo, nhằm cải thiện tính ổn định và hiệu suất dự đoán của chúng.

Các mô hình dưới đây học theo nhóm dựa trên cây quyết định, bao gồm Rừng ngẫu nhiên, XGBoost, CatBoost và LightGBM.

### 4.2.1. Rừng ngẫu nhiên.

Độ chuẩn xác, thu hồi và F1-score cho mỗi lớp như sau:

Đối với lớp "bình thường" (0):

- Độ chuẩn xác: độ chuẩn xác cao, gần như hoàn hảo, ở mức 0.99, thể hiện rằng hầu hết tất cả các giao dịch mà mô hình phân loại là bình thường đều thực sự bình thường.
- Độ phủ: khả năng phủ hoặc khả năng tìm thấy các giao dịch thông thường của mô hình là 0.99, cho thấy 99% giao dịch thông thường thực tế đã được xác định chính xác.
- F1-score: sự kết hợp giữa độ chính xác và độ phủ, là 0.99, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

Đối với lớp "gian lận" (1):

- Độ chuẩn xác: độ chuẩn xác cao, ở mức 0.91, thể hiện rằng trong số tất cả các giao dịch mà mô hình phân loại là gian lận, có tận 91% là lừa đảo.
- Độ phủ: độ phủ cao, ở mức 0.84, cho thấy mô hình có thể xác định chính xác đến 84% giao dịch gian lận thực tế.
- F1-score: cao, ở mức 0.87, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

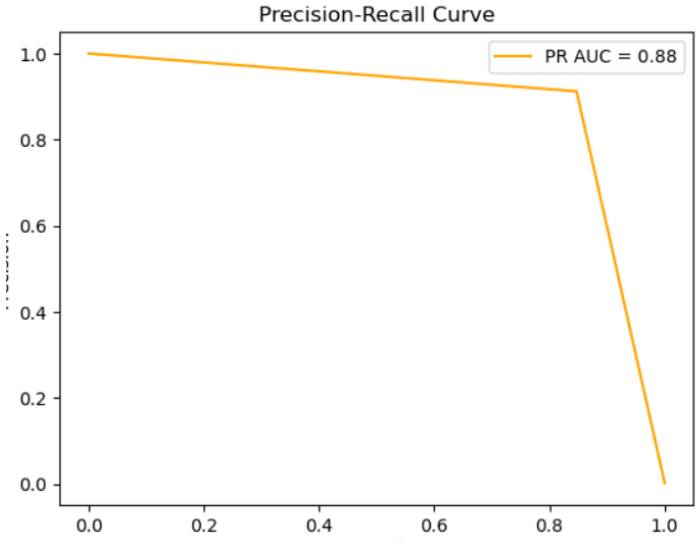
Mô hình Rừng ngẫu nhiên có tỷ lệ phát hiện gian lận cao (84,69%). Ngoài ra, 91% số giao dịch được dự đoán là gian lận được phân loại đúng. Mô hình này có độ chính xác cao nhất (91%) trong số tất cả các bộ phân loại. Độ chính xác cao dẫn đến năng

suất làm việc về tổng thể cao nhất khi độ phủ chiếm 84,69% và F1-Score chiếm 87,8%) so với các bộ phân loại khác.

```
Time taken = 853.4276304244995
              precision    recall  f1-score   support

           0    0.99974   0.99986   0.99980     56864
           1    0.91209   0.84694   0.87831        98

    accuracy                        0.99960     56962
   macro avg    0.95591   0.92340   0.93905     56962
weighted avg    0.99959   0.99960   0.99959     56962
```

*Hình 31. Báo cáo phân loại của mô hình Rừng ngẫu nhiên*

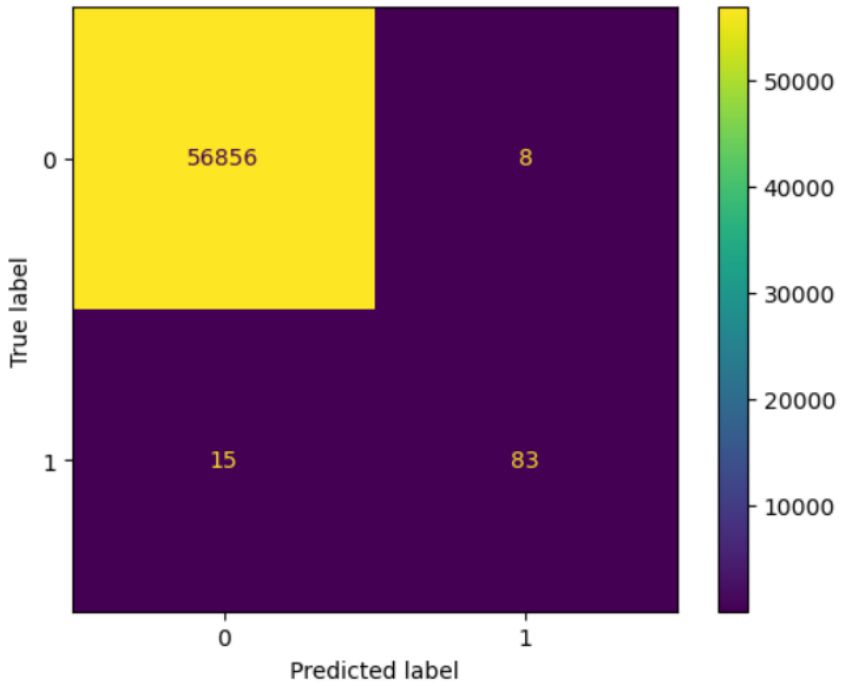
Như các mô hình khác, Rừng ngẫu nhiên có điểm AUC-PR cao, tổng hợp hiệu suất của bộ phân loại qua các ngưỡng khác nhau là 0.87, cho thấy đến 87% khả năng mô hình phân biệt giữa giao dịch gian lận và hợp lệ. Mô hình Rừng ngẫu nhiên cho ra kết quả tốt nhất về các chỉ số đánh giá nhưng tốn khá nhiều thời gian huấn luyện lên tới 853 giây hoặc xấp xỉ 14 phút) nên mô hình này chưa đủ tiêu chí để ta tối ưu hóa và tinh chỉnh.




*Hình 32. Đồ thị biểu diễn AUC-PR của mô hình Rừng ngẫu nhiên*

Ma trận nhầm lẫn dự đoán kết quả như sau:

- Âm tính thực (56856): Mô hình dự đoán chính xác 56.856 trường hợp không có gian lận xảy ra.
- Dương tính giả (8): Mô hình dự đoán sai 8 trường hợp gian lận. Đây là những giao dịch bình thường bị gắn cờ sai là gian lận.
- Âm tính giả (15): Mô hình dự đoán sai 15 trường hợp xảy ra gian lận nhưng không bị gắn cờ. Đây là những giao dịch gian lận mà mô hình không phát hiện được.
- Dương tính thực (83): Mô hình đã gắn cờ chính xác 83 trường hợp gian lận.




*Hình 33. Ma trận nhầm lẫn của mô hình Rừng ngẫu nhiên*

### 4.2.2. XGBoost

Độ chuẩn xác, thu hồi và F1-score cho mỗi lớp như sau:

Đối với lớp "bình thường" (0):

- Độ chuẩn xác: độ chuẩn xác cao, gần như hoàn hảo, ở mức 0.99, thể hiện rằng hầu hết tất cả các giao dịch mà mô hình phân loại là bình thường đều thực sự bình thường.
- Độ phủ: khả năng phủ hoặc khả năng tìm thấy các giao dịch thông thường của mô hình là 0.99, cho thấy 99% giao dịch thông thường thực tế đã được xác định chính xác.
- F1-score: sự kết hợp giữa độ chính xác và độ phủ, là 0.99, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

Đối với lớp "gian lận" (1):

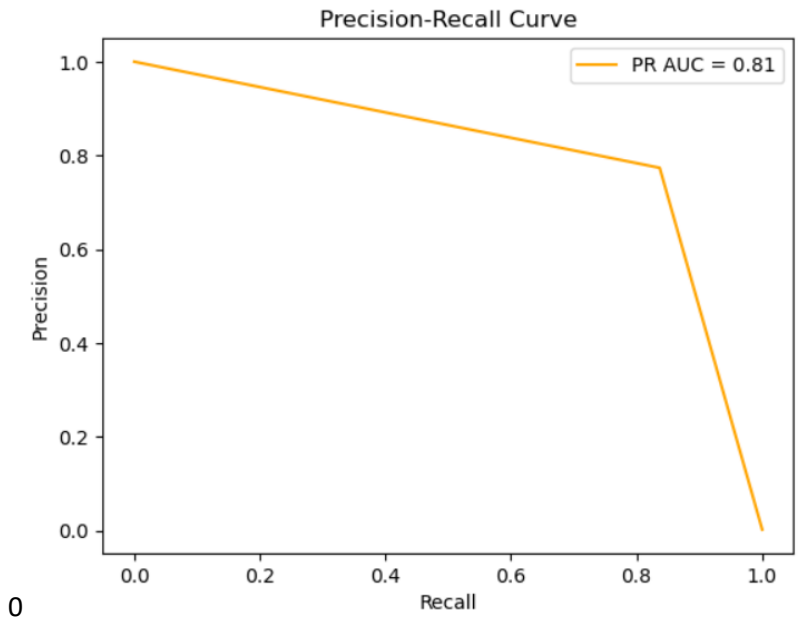
- Độ chuẩn xác: độ chuẩn xác tương đối, ở mức 0.77, thể hiện rằng trong số tất cả các giao dịch mà mô hình phân loại là gian lận, có tận 77% là lừa đảo.
- Độ phủ: độ phủ cao, ở mức 0.83, cho thấy mô hình có thể xác định chính xác đến 83% giao dịch gian lận thực tế.
- F1-score: cao, ở mức 0.8, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

```
Time taken = 2.1530675888061523
              precision    recall  f1-score   support

           0    0.99972   0.99958   0.99965     56864
           1    0.77358   0.83673   0.80392        98

    accuracy                        0.99930     56962
   macro avg    0.88665   0.91816   0.90178     56962
weighted avg    0.99933   0.99930   0.99931     56962
```

*Hình 34. Báo cáo phân loại của mô hình XGBoost*

Mô hình XGBoost có tỷ lệ phát hiện gian lận cao là 83% và độ chính xác khá cao là 77%. Mô hình này phát hiện được 83% tổng số lượng giao dịch gian lận, và 77% trong số các giao dịch được xác định là gian lận được phân loại đúng. Điều này dẫn đến hiệu suất làm việc tổng thể khá cao khi F1-Score chiếm 80%, khá cao trong số các mô hình học kết hợp. XGBoost cũng có AUC-PR cao, là 0.81, cho thấy có 81%

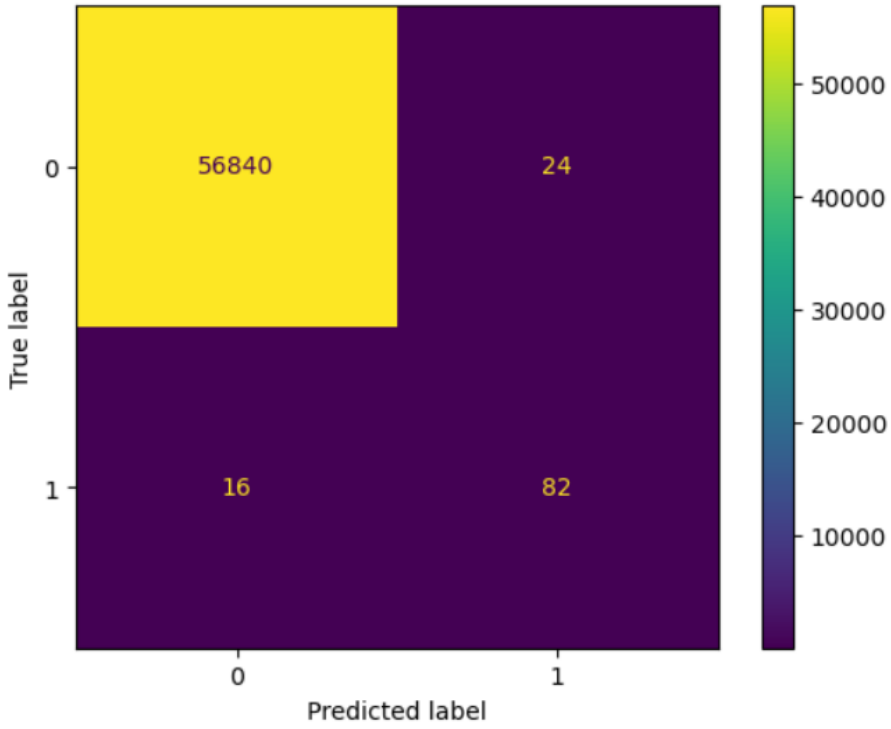
khả năng mô hình phân biệt giữa giao dịch hợp lệ và gian lận. Điều này gợi ý rằng bộ phân loại có khả năng phân biệt tốt. Với thời gian huấn luyện khá lý tưởng chỉ 2 giây, mô hình XGBoost là trong những mô hình có tiêu chí để tối ưu hóa.




*Hình 35. Đồ thị biểu diễn AUC-PR của mô hình XGBoost*

Ma trận nhầm lẫn dự đoán kết quả như sau:

- Âm tính thực (56840): Mô hình dự đoán chính xác 56.840 trường hợp không có gian lận xảy ra.
- Dương tính giả (24): Mô hình dự đoán sai 24 trường hợp gian lận. Đây là những giao dịch bình thường bị gắn cờ sai là gian lận.
- Âm tính giả (16): Mô hình dự đoán sai 16 trường hợp xảy ra gian lận nhưng không bị gắn cờ. Đây là những giao dịch gian lận mà mô hình không phát hiện được.
- Dương tính thực (82): Mô hình đã gắn cờ chính xác 82 trường hợp gian lận.




*Hình 36. Ma trận nhầm lẫn của mô hình XGBoost*

### 4.2.3. CatBoost

Độ chuẩn xác, độ phủ và F1-score cho mỗi lớp như sau:

Đối với lớp "bình thường" (0):

- Độ chuẩn xác: độ chuẩn xác cao, gần như hoàn hảo, ở mức 0.99, thể hiện rằng hầu hết tất cả các giao dịch mà mô hình phân loại là bình thường đều thực sự bình thường.
- Độ phủ: độ phủ hoặc khả năng tìm thấy các giao dịch thông thường của mô hình là 0.99, cho thấy 99% giao dịch thông thường thực tế đã được xác định chính xác.
- F1-score: sự kết hợp giữa độ chính xác và độ phủ, là 0.99, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

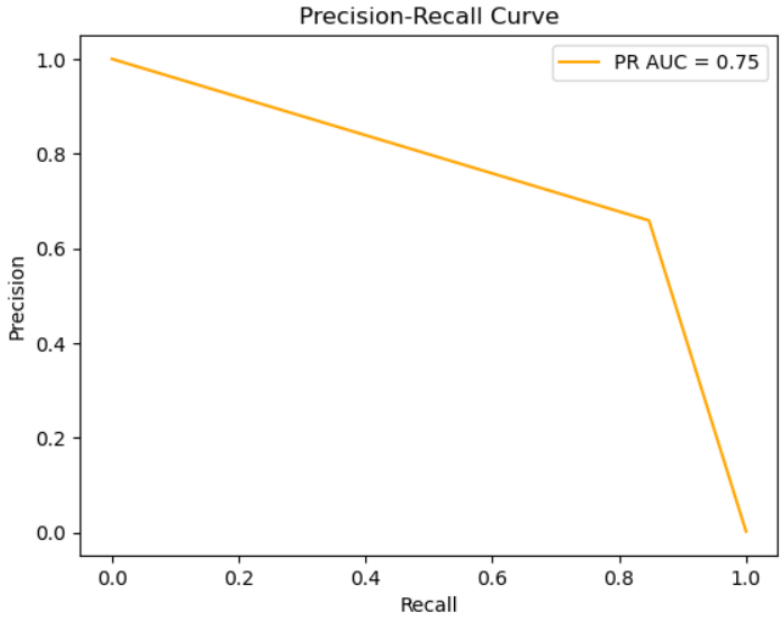
Đối với lớp "gian lận" (1):

- Độ chuẩn xác: độ chuẩn xáctương đối, ở mức 0.65, thể hiện rằng trong số tất cả các giao dịch mà mô hình phân loại là gian lận, có tận 65% là lừa đảo.
- Độ phủ: độ phủ cao, ở mức 0.84, cho thấy mô hình có thể xác định chính xác đến 84% giao dịch gian lận thực tế.
- F1-score: cao, ở mức 0.74, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

```
Time taken = 121.83303165435791
              precision    recall  f1-score   support

           0    0.99974   0.99924   0.99949     56864
           1    0.65873   0.84694   0.74107        98

    accuracy                        0.99898     56962
   macro avg    0.82923   0.92309   0.87028     56962
weighted avg    0.99915   0.99898   0.99905     56962
```

*Hình 37. Báo cáo phân loại của mô hình CatBoost*

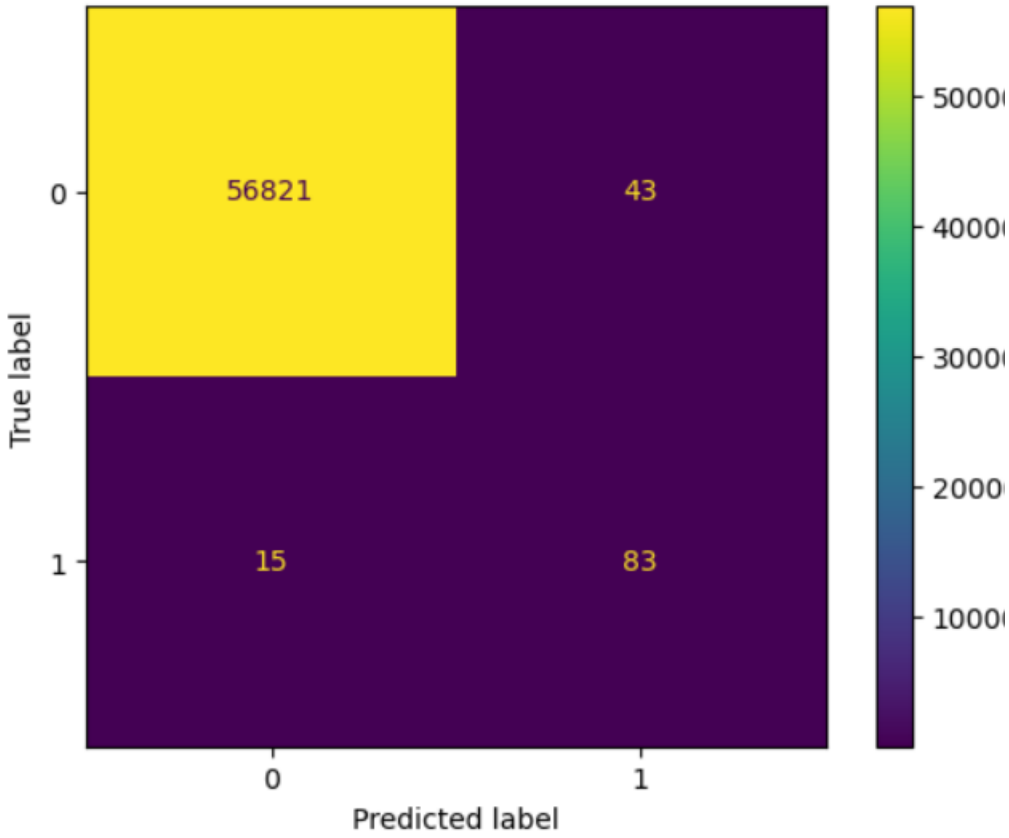
Mô hình CatBoost có tỷ lệ phát hiện gian lận cao là 84% và độ chính xác tương đối 65%. Mô hình này phát hiện được 88% tổng số lượng giao dịch gian lận, và 65% trong số các giao dịch được xác định là gian lận được phân loại đúng. Điều này dẫn đến hiệu suất tổng thể khá cao, ở mức 74% về chỉ số F1-Score, chỉ vỏn vẹn ít hơn so với XGBoost. Tương tự với các mô hình khác, CatBoost có khả năng phân biệt tốt, với AUC-PR ở mức 0.75, cho thấy có 75% khả năng mô hình có khả năng phân biệt được hai lớp. Tuy thời gian huấn luyện mô hình không quá nhanh (121 giây hoặc xấp xỉ 2 phút rưỡi), CatBoost vẫn có khả năng đáp ứng cao về các độ đo đánh giá nói riêng và tối ưu mô hình nói chung.




*Hình 38. Đồ thị biểu diễn AUC-PR của mô hình CatBoost*

Ma trận nhầm lẫn dự đoán kết quả như sau:

- Âm tính thực (56821): Mô hình dự đoán chính xác 56.821 trường hợp không có gian lận xảy ra.
- Dương tính giả (43): Mô hình dự đoán sai 43 trường hợp gian lận. Đây là những giao dịch bình thường bị gắn cờ sai là gian lận.
- Âm tính giả (15): Mô hình dự đoán sai 15 trường hợp xảy ra gian lận nhưng không bị gắn cờ. Đây là những giao dịch gian lận mà mô hình không phát hiện được.
- Dương tính thực (83): Mô hình đã gắn cờ chính xác 83 trường hợp gian lận.




*Hình 39. Ma trận nhầm lẫn của mô hình CatBoost*

### 4.2.4. LightGBM

Độ chuẩn xác, độ phủ và F1-score cho mỗi lớp như sau:

Đối với lớp "bình thường" (0):

- Độ chuẩn xác: độ chuẩn xác cao, gần như hoàn hảo, ở mức 0.99, thể hiện rằng hầu hết tất cả các giao dịch mà mô hình phân loại là bình thường đều thực sự bình thường.
- Độ phủ: độ phủ hoặc khả năng tìm thấy các giao dịch thông thường của mô hình là 0.99, cho thấy 99% giao dịch thông thường thực tế đã được xác định chính xác.

- F1-score: sự kết hợp giữa độ chính xác và độ phủ, là 0.99, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

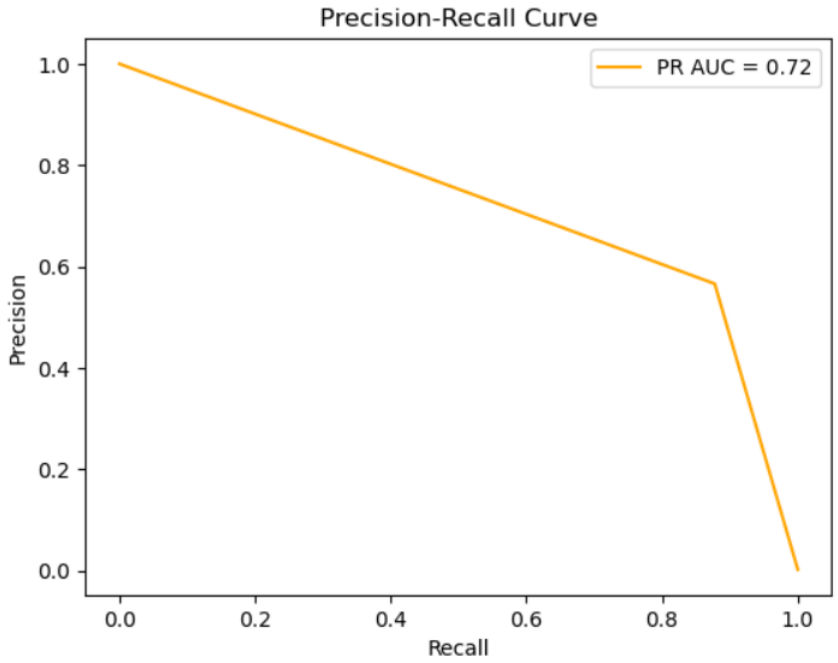
Đối với lớp gian lận (1):

- Độ chuẩn xác: độ chuẩn xáckhông quá cao, ở mức 0.56, thể hiện rằng trong số tất cả các giao dịch mà mô hình phân loại là gian lận, có tận 56% là lừa đảo.
- Độ phủ: độ phủ cao, ở mức 0.87, cho thấy mô hình có thể xác định chính xác đến 87% giao dịch gian lận thực tế.
- F1-score: tương đối, ở mức 0.68, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

```
Time taken = 4.534712553024292
              precision    recall  f1-score   support

           0    0.99979   0.99884   0.99931     56864
           1    0.56579   0.87755   0.68800        98

    accuracy                        0.99863     56962
   macro avg    0.78279   0.93820   0.84366     56962
weighted avg    0.99904   0.99863   0.99878     56962
```

*Hình 40. Báo cáo phân loại của mô hình LightGBM*

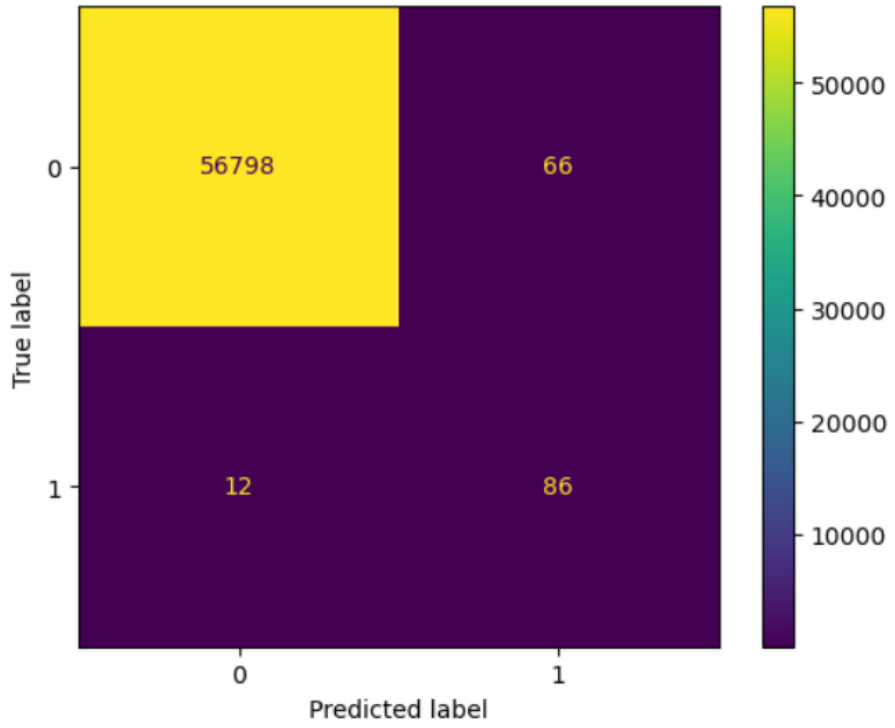
Mô hình LightGBM có tỷ lệ phát hiện gian lận là 87% và độ chính xác là 56%. Mô hình có thể phát hiện được 87% tổng số lượng giao dịch gian lận, và 56% trong số các giao dịch được xác định là gian lận mà thực sự là gian lận. Điều này dẫn đến hiệu suất tổng thể là 69% về F1-Score, mức độ khá cao trong số tất cả các mô hình học kết hợp. So với các mô hình khác, LightGBM đạt được một điểm AUC-PR khá cao là 0.72, cho thấy có 72% khả năng mô hình sẽ có khả năng phân biệt giữa giao dịch hợp lệ và gian lận. Thời gian huấn luyện khá lý tưởng xấp xỉ 5 giây, chỉ sau XGBoost vài giây.




*Hình 41. Đồ thị biểu diễn AUC-PR của mô hình LightGBM*

Ma trận nhầm lẫn dự đoán kết quả như sau:

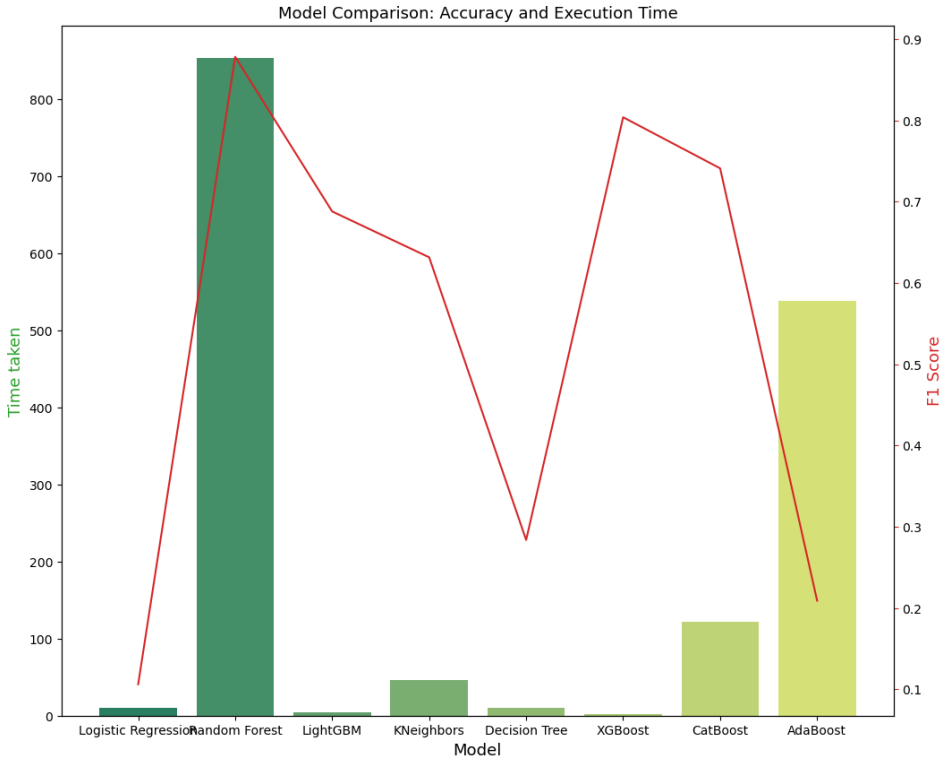
- Âm tính thực (56798): Mô hình dự đoán chính xác 56.798 trường hợp không có gian lận xảy ra.
- Dương tính giả (66): Mô hình dự đoán sai 66 trường hợp gian lận. Đây là những giao dịch bình thường bị gắn cờ sai là gian lận.
- Âm tính giả (12): Mô hình dự đoán sai 12 trường hợp xảy ra gian lận nhưng không bị gắn cờ. Đây là những giao dịch gian lận mà mô hình không phát hiện được.
- Dương tính thực (86): Mô hình đã gắn cờ chính xác 86 trường hợp gian lận.




*Hình 42. Ma trận nhầm lẫn của mô hình LightGBM*

So với các mô hình học kết hợp như Rừng ngẫu nhiên, XGBoost, CatBoost thì LightGBM có tỷ lệ phát hiện gian lận cao nhất là 87 % hay độ phủ cao nhất là 0.87. Tuy mô hình Rừng ngẫu nhiên hiệu suất làm việc tổng thể là tốt nhất nhưng thời gian huấn luyện rất lâu và tốn tài nguyên tính toán. Ngoài ra, mô hình XGBoost cũng khá lý tưởng và cân bằng về hiệu suất làm việc tổng thể nhưng độ phủ không cao bằng các mô hình kia. Chính vì thế, mô hình LightGBM này sẽ được chọn là mô hình để tối ưu hóa và tinh chỉnh.

Dưới đây là biểu đồ cột miêu tả đánh giá các mô hình dựa trên độ đo là F1-score và thời gian huấn luyện, ở đây các mô hình như AdaBoost, Hồi quy Logistic và KNeighborsClassifier cũng được triển khai để có thêm cái nhìn tổng quan về hiệu suất làm việc của các mô hình.




*Hình 43. Biểu đồ hiệu suất và thời gian làm việc của các mô hình theo độ đo F1-Score*

So với các mô hình đã được triển khai, mô hình được đề xuất để đánh giá khả năng phát hiện gian lận chính là mô hình LightGBM kết hợp với phương pháp tinh chỉnh siêu tham số, hay còn được gọi là Tinh chỉnh siêu tham số. Optuna, một framework hỗ trợ tự động điều chỉnh siêu tham số sẽ được triển khai cho mô hình LightGBM.

Với tập dữ liệu đã được cân bằng nhờ phương pháp SMOTE, khi huấn luyện mô hình LightGBM, trước tiên ta chạy một hàm tổng quan, trong đó đưa các tham số cần được tinh chỉnh sử dụng hàm suggest_int (), suggest_float () hoặc suggest_categorical (). Ngoài ra, phương pháp Stratified KFold cross-validation còn được sử dụng để đồng thời đánh giá hiệu suất mô hình nhờ độ đo cross_val_score, trong đó ta truyền vào F1-score của mô hình. Tập dữ liệu lúc này được chia thành năm fold và được huấn luyện năm lần, mỗi lần sử dụng một fold khác nhau làm tập kiểm thử và các fold còn lại làm tập huấn luyện. Lúc này, Optuna sẽ tự động đề xuất các bộ tham số thử nghiệm (trial).

Với các tham số cần được tinh chỉnh như max_depth, learning_rate, num_leaves, … mô hình sẽ được học chạy tổng cộng 100 bộ tham số thử nghiệm và in ra bộ tham số tốt nhất để thế vào. Sau khi chạy xong bộ tham số kiểm nghiệm, kết quả bộ tham số tối ưu được in ra. Cuối cùng, ta thế vào mô hình LightGBM và chạy trên tập dữ liệu kiểm thử.

```
Best value (Accuracy): 0.99955
Best params:
        max_depth: 8
        learning_rate: 0.20156246967856106
        num_leaves: 215
        min_data_in_leaf: 97
        max_bin: 292
        lambda_l1: 5
        lambda_l2: 0
        min_gain_to_split: 1.848207090633376
        bagging_fraction: 0.8
        bagging_freq: 1
        feature_fraction: 0.8
```

*Hình 44. Kết quả bộ tham số tối ưu sau khi chạy Optuna*

Độ chuẩn xác, độ phủ và F1-score cho mỗi lớp như sau:

Đối với lớp “bình thường” (0):

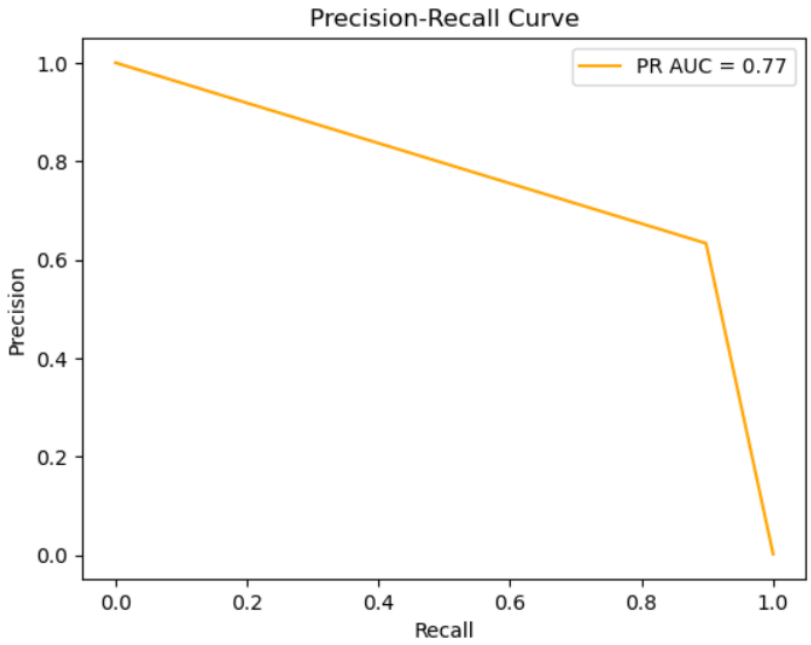
- Độ chuẩn xác: độ chuẩn xác cao, gần như hoàn hảo, ở mức 0.99, thể hiện rằng hầu hết tất cả các giao dịch mà mô hình phân loại là bình thường đều thực sự bình thường.
- Độ phủ: độ phủ hoặc khả năng tìm thấy các giao dịch thông thường của mô hình là 0.99, cho thấy 99% giao dịch thông thường thực tế đã được xác định chính xác.
- F1-score: sự kết hợp giữa độ chính xác và độ phủ, là 0.99, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

Đối với lớp gian lận (1):

- Độ chuẩn xác: độ chuẩn xáctương đối, ở mức 0.63, thể hiện rằng trong số tất cả các giao dịch mà mô hình phân loại là gian lận, có tận 56% là lừa đảo.
- Độ phủ: độ phủ khá cao, ở mức 0.89, cho thấy mô hình có thể xác định chính xác đến 89% giao dịch gian lận thực tế.
- F1-score: tương đối, ở mức 0.74, thể hiện sự cân bằng tốt giữa độ chính xác và độ phủ.

```
Time taken = 4.043440103530884
              precision    recall  f1-score   support

           0    0.99982   0.99910   0.99946     56864
           1    0.63309   0.89796   0.74262        98

    accuracy                        0.99893     56962
   macro avg    0.81646   0.94853   0.87104     56962
weighted avg    0.99919   0.99893   0.99902     56962
```

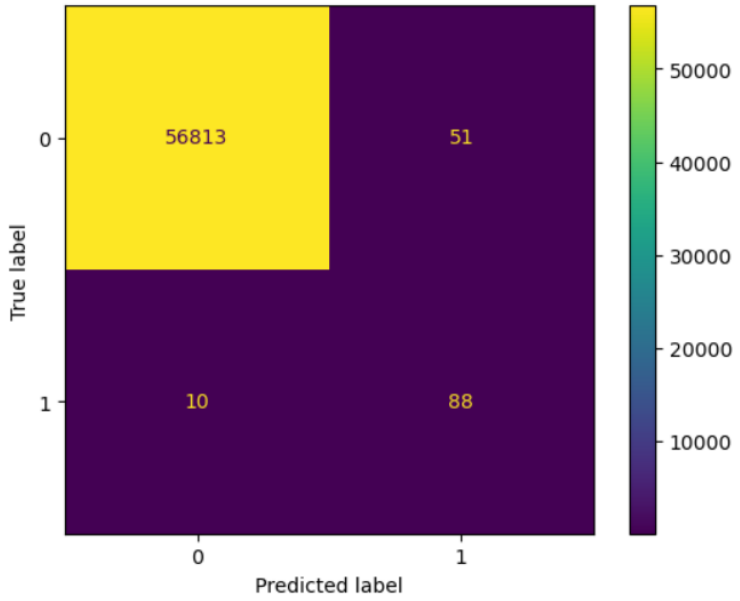
*Hình 45. Báo cáo phân loại của mô hình LightGBM với Tinh chỉnh siêu tham số*




*Hình 46. Đồ thị biểu diễn AUC-PR của mô hình LightGBM với Tinh chỉnh siêu tham số*

Ma trận nhầm lẫn dự đoán kết quả như sau:

- Âm tính thực (56813): Mô hình dự đoán chính xác 56.813 trường hợp không có gian lận xảy ra.
- Dương tính giả (51): Mô hình dự đoán sai 51 trường hợp gian lận. Đây là những giao dịch bình thường bị gắn cờ sai là gian lận.
- Âm tính giả (10): Mô hình dự đoán sai 10 trường hợp xảy ra gian lận nhưng không bị gắn cờ. Đây là những giao dịch gian lận mà mô hình không phát hiện được.
- Dương tính thực (88): Mô hình đã gắn cờ chính xác 88 trường hợp gian lận.




*Hình 47. Ma trận nhầm lẫn của mô hình LightGBM với tinh chỉnh siêu tham số*

| **Mô hình** | **Giá trị siêu tham số** |
|---|---|
| **Rừng ngẫu nhiên** | Giá trị mặc định |
| **XGBoost** | n_estimators = 20. max_depth = 16 |
| **CatBoost** | n_estimators = 20. max_depth = 16 |
| **LightGBM** | Giá trị mặc định |

| | |
|---|---|
| **LightGBM với Tinh chỉnh siêu tham số** | max_depth = 8, learning_rate = 0.20156246967856106, num_leaves = 215, min_data_in_leaf = 97, max_bin = 292, lambda_l1 = 5, lambda_l2 = 0. min_gain_to_split = 1.848207090633376, bagging_fraction =0.8, bagging_freq = 1, feature_fraction = 0.8 |

*Bảng 3. Bảng tóm tắt giá trị siêu tham số của các mô hình*

# CHƯƠNG 5. KẾT LUẬN

## 5.1. Thảo luận về kết quả thực nghiệm

Mục tiêu của khóa luận hướng tới việc cải tiến và tinh chỉnh một trong những mô hình học kết hợp đã triển khai nhằm phát hiện các giao dịch gian lận một cách chính xác nhất. Các mô hình dưới đây học theo nhóm dựa trên cây quyết định, bao gồm Rừng ngẫu nhiên, XGBoost, CatBoost và LightGBM. Ngoài ra, nhằm tạo nên sự đa dạng, các mô hình như Hồi quy logistic, KNN, Cây quyết định và AdaBoost cũng được triển khai thêm nhằm mục đích so sánh và đánh giá chi tiết để tìm ra mô hình nào tốt nhất trong việc phát hiện gian lận.

| Mô hình | Thời gian | Độ chuẩn xác | Độ phủ | F1-Score | AUC-PR |
|---|---|---|---|---|---|
| **LightGBM** | 4.53 | 0.56 | 0.87 | 0.68 | 0.72 |
| **Rừng ngẫu nhiên** | 853.42 | 0.91 | 0.84 | 0.87 | 0.87 |
| **XGBoost** | 2.15 | 0.77 | 0.83 | 0.80 | 0.8 |
| **CatBoost** | 121.83 | 0.65 | 0.84 | 0.74 | 0.75 |
| **AdaBoost** | 538.67 | 0.11 | 0.87 | 0.20 | 0.49 |
| **Cây quyết định** | 10.21 | 0.17 | 0.81 | 0.28 | 0.49 |
| **Hồi quy Logistic** | 10.16 | 0.05 | 0.91 | 0.10 | 0.48 |
| **KNN** | 45.82 | 0.5 | 0.85 | 0.63 | 0.67 |
| **LightGBM với Tinh chỉnh siêu tham số** | **4.04** | 0.63 | **0.89** | 0.74 | **0.76** |

*Bảng 4. Thống kê các độ đo đánh giá hiệu suất tổng quan huấn luyện các mô hình*

Khóa luận nhằm mục đích tối ưu hóa mô hình có khả năng dự đoán gian lận chính xác bằng cách áp dụng các mô hình học kết hợp và một số mô hình học giám sát khác trên tập dữ liệu thẻ tín dụng. Các mô hình sử dụng thuật toán Tăng cường là các mô hình được triển khai chủ yếu trong khóa luận này.

Các mô hình này đã được đánh giá trên tập dữ liệu kiểm thử, bao gồm 284.807 quan sát và có sự mất cân bằng cao khi chỉ có 0.172% các giao dịch là gian lận. Phân phối lớp của dữ liệu kiểm thử không cân bằng do dữ liệu thực về gian lận thẻ tín dụng thường có một phân phối lớp bị lệch cao. Các độ đo dưới đây được sử dụng để đánh giá các mô hình.

Các thuật toán tăng cường được sử dụng để huấn luyện các mô hình. Theo kết quả được hiển thị trong bảng trên, các mô hình sử dụng phương pháp Tăng cường cho ra kết quả vượt trội so với các mô hình giám sát khác trong việc phát hiện giao dịch gian lận. Tổng quan các mô hình trên đều đạt được tỷ lệ phát hiện giao dịch gian lận cao (chiếm 81%) hoặc độ phủ không dưới 0.81. Mặc dù mô hình Rừng ngẫu nhiên có độ chính xác tốt nhất là 0.91 và điểm AUC-PR tốt nhất là 0.87 nhưng thời gian huấn luyện mô hình lâu nhất. Trong khi đó, các mô hình khác như Decision Tree, Logistic Regression và AdaBoost có độ chính xác rất thấp, lần lượt là 0.17, 0.05 và 0.11, cho thấy rất ít giao dịch được dự đoán gian lận là gian lận thực sự.

Độ chuẩn xác là một độ đo về chất lượng dự đoán lớp gian lận và độ phủ là một độ đo về số lượng. Sự đánh đổi giữa precision và recall là một vấn đề thách thức trong tập dữ liệu gian lận mất cân bằng, vì trong phát hiện gian lận thẻ tín dụng, cả Dương tính giả (FP) và Âm tính giả (FN) đều nên được giữ ở mức tối thiểu, nhằm tránh quá trình xác minh không cần thiết và chi phí phản kháng, đồng thời tránh thiệt hại về mặt tài chính do không phát hiện gian lận.

F1-score, kết hợp giữa độ chuẩn xác và độ phủ, nhằm đo lường độ chính xác của mô hình, là một trong những độ đo được sử dụng. Tuy nhiên, F1-score không xem xét số lượng Âm tính thực (TN), mà có giá trị khá cao trong lĩnh vực phát hiện gian lận thẻ tín dụng; dẫn đến kết quả một phần không đáng tin cậy. Do đó, điểm AUC-PR được

sử dụng như một trong những độ đo chính. XGBoost và LightGBM với Tinh chỉnh siêu tham số cũng vượt trội so với các mô hình khác về điểm AUC-PR, lần lượt là 0.8 và 0.76, sau đó tới CatBoost với 0.75. Điều này cho thấy các mô hình sử dụng thuật toán Tăng cường có khả năng phân biệt tốt hơn giữa các giao dịch hợp lệ và gian lận. Để cải thiện hiệu suất dự đoán của các mô hình, siêu tham số của một vài mô hình sẽ được tối ưu hóa. Mô hình Rừng ngẫu nhiên, khác biệt so với các mô hình khác, đạt được kết quả tốt hơn với siêu tham số mặc định. Các mô hình còn lại cho thấy việc tăng số lượng của ước lượng cải thiện hiệu suất của chúng. Tuy nhiên, điều này cũng làm tăng thời gian huấn luyện. Đối với các thuật toán Tăng cường, với tốc độ học từ 0.2 đến 0.3, mô hình LightGBM với tinh chỉnh siêu tham số và XGBoost đạt được kết quả tốt với thời gian huấn luyện hợp lý.

Tổng thể, các mô hình sử dụng thuật toán Tăng cường vượt trội hơn các mô hình khác trong việc phát hiện gian lận thẻ tín dụng. LightGBM với siêu tham số và XGBoost đạt được kết quả tốt nhất về độ phủ, F1-Score và AUC-PR, tiếp theo là CatBoost.

## 5.2. Những mặt còn hạn chế

Bộ dữ liệu giao dịch thẻ tín dụng được sử dụng rất mất cân bằng, chứa đựng cả dữ liệu liên tục và dữ liệu phân loại. Sự thiếu hụt về các bộ dữ liệu tương tự ngoài đời thực và hầu hết các đặc trưng đã được ẩn danh vì vấn đề bảo mật khiến cho việc lựa chọn những đặc trưng quan trọng không tối ưu hóa các mô hình học kết hợp đã được áp dụng.

Ngoài ra, khóa luận áp dụng các mô hình học kết hợp và chỉ sử dụng các thuật toán Tăng cường. Vì thế, các thuật toán khác trong học kết hợp như phương pháp bỏ túi hoặc phương pháp ngăn xếp cũng là một trong những kỹ thuật phổ biến để triển khai.

Hơn thế nữa, phương pháp tăng cường số mẫu dữ liệu như SMOTE được áp dụng dễ xảy ra trường hợp các mô hình bị quá khớp. Các phương pháp tăng cường hay giảm thiểu số mẫu khác hoặc lai giữa cả hai phương pháp cho việc xử lý tình huống mất cân bằng lớp nên được thực thi.

Tuy mô hình LightGBM kết hợp Tinh chỉnh siêu tham số đưa ra độ phủ cao nhất trong các mô hình và điểm AUC-PR tương đối cao, đồng nghĩa với khả năng phát hiện nhiều giao dịch gian lận nhất nhưng độ chuẩn xác và F1-Score còn tương đối thấp (lần lượt là 0.63 và 0.74). Nếu xét về ma trận nhầm lẫn thì mô hình khóa luận đề xuất đạt được 51 Dương tính giả (FP) và 10 Âm tính giả (FN), đồng nghĩa với việc phát hiện nhiều giao dịch gian lận nhất nhưng dễ đoán sai các giao dịch thực chất không là gian lận.

## 5.3. Hướng phát triển

Hướng phát triển trong khóa luận này bao gồm các đề xuất như sau:

*Thứ nhất,* việc tăng giá trị tối thiểu của dữ liệu trong một lá của mô hình LightGBM vớ Tinh chỉnh siêu tham số có thể giảm thiểu trường hợp mô hình quá khớp. Ngoài ra, việc tăng số lượng thử nghiệm cho framework Optuna để tạo ra nhiều giá trị siêu tham số hơn mang lại lợi ích về hiệu suất tổng quan mô hình, cũng như sử dụng các phương pháp điều chỉnh khác nhau như tìm kiếm lưới và tìm kiếm ngẫu nhiên.

*Thứ hai,* sử dụng các phương pháp khác nhau để xử lý sự mất cân bằng lớp. Ngoài SMOTE, các phương pháp như Borderline-SMOTE, ADASYN (Adaptive Synthetic Sampling), RUS (random undersampling) hoặc phương pháp lai giữa oversampling và undersampling thực hiện trên tập dữ liệu này. Ngoài ra, thay vì lấy mẫu dữ liệu, phương pháp nhằm tăng trọng số cho lớp thiểu số (gian lận) cũng là một đề xuất cho hướng phát triển về xử lý tình huống mất cân bằng lớp.

*Thứ ba,* về các độ đo đánh giá, nên sử dụng thêm các độ đo như MCC hoặc điểm G-Mean. MCC (Matthews Correlation Coefficient), hay còn được gọi là hệ số Tương quan Matthews, xem xét tất cả bốn mục của ma trận nhầm lẫn để đánh giá hiệu suất của mô hình, là một độ đo với nhiều thông số, có ích với tập dữ liệu có sự mất cân bằng cao. Cũng như điểm G-Mean, một trong những độ đo hữu ích cho tập dữ liệu mất cân bằng vì nó đánh giá hiệu suất của mô hình phân loại của cả hai lớp đa số và thiểu số.

*Hơn nữa*, các mô hình hợp tác sử dụng các phương pháp bỏ túi thay vì tăng cường cũng nên được điều tra. Các thuật toán học sâu, như ANN, và việc sử dụng bộ tự mã hóa để trích xuất đặc trưng cũng nên được xem xét. Ngoài ra, kết hợp thêm các mô hình học không giám sát, cụ thể như thuật toán phân cụm dữ liệu như gom cụm K-Means và DBSCAN để chỉ ra mối tương quan giữa hai cụm dữ liệu giao dịch bình thường và gian lận cũng là một đề xuất nhằm tối ưu hóa hiệu suất tổng quan của mô hình nhằm phát hiện gian lận tài chính.

# TÀI LIỆU THAM KHẢO


[1] S. Pedneault, "Fraud 101: Techniques and Strategies for Understanding Fraud," *Association of Certified Fraud Examiners,* 2009.

[2] N. H. Dương and H. Đ. Hải, "Phát hiện lưu lượng mạng bất thường trong điều kiện dữ liệu huấn luyện chứa ngoại lai," *Tạp chí Khoa học Công nghệ Thông tin và Truyền thông,* pp. 3-16, 2016.

[3] K. M. v. Hespen, J. J. M. Zwanenburg, J. W. Dankbaar, M. I. Geerlings, J. Hendrikse and H. J. Kuijf, "An anomaly detection approach to identify chronic brain infarcts on MRI," *Scientific Reports,* vol. 11, 2021.

[4] W. Hilal, S. A. Gadsden and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *ScienceDirect,* vol. 193.

[5] E. Ileberi, Y. Sun and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *Journal of Big Data,* vol. 9, 2022.

[6] V. T. Gowda, "Credit Card Fraud Detection using Supervised and Unsupervised Learning," *Wichita State University, USA,* 2021.

[7] A. Rosenbaum, "Detecting Credit Card Fraud with Machine Learning," *Department of Statistics, Stanford University,* 2019.

[8] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, 2019.

[9] S. S. B. K. C. Taneja, "Application of balancing techniques with ensemble approach for credit card fraud detection," in *2019 International Conference on Computing, Power, and Communication Technologies (GUCON)*, New Delhi, India, 2019.

[10] A. Muaz, M. Jayabalan and V. Thiruchelvam, "A Comparison of Data Sampling Techniques for Credit Card Fraud Detection," *International Journal of Advanced Computer Science and Applications(IJACSA),* vol. 11, no. 6, 2020.

[11] A. Sahu, H. GM and M. K. Gourisaria, "A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques," in *2020 IEEE 17th India Council International Conference (INDICON)*, New Delhi, India, 2020 .

[12] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys,* vol. 41, no. 3, 2009.

[13] D. M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.

[14] V. Barnett and T. Lewis, Outliers in Statistical Data, 3rd Edition, Wiley, 1994.

[15] M. Gölyeria, S. Çelika, F. Bozyiğitbc and D. Kılınç, "Fraud Detection on E-commerce Transactions Using," *Artificial Intelligence Theory and Applications,* 2023.

[16] I. Zingman, B. Stierstorfer, C. Lempp and F. Heinemann, "Learning image representations for anomaly detection: application to discovery of histological alterations in drug development," *Medical Image Analysis,* vol. 92, 2024.

[17] E. d. R. d. l. Roy, T. Recht, A. Zemmari, P. Bourreau and L. Mora, "Data analytics for smart buildings: a classification method for anomaly detection for measured data," *Journal of Physics: Conference Series,* vol. 2042, 2021.

[18] H. Trọng and C. N. M. Ngọc, Phân tích dữ liệu nghiên cứu với SPSS tập 1, NXB Hồng Đức, 2008.

[19] D. Lombardi and S. Pant, "A non-parametric k-nearest neighbor entropy estimator," *Physical Reviev E,* vol. 93, no. 1, 2016.

[20] H. J. Seltman, Experimental Design and Analysis, Carnegie Mellon University, 2012.

[21] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995.

[22] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, 1998.