|  |  |
| --- | --- |
|  |  |
| **TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**    **ĐỒ ÁN CUỐI KỲ - HK2 (2022-2023)**  **MÔN HỌC PHÂN TÍCH DỮ LIỆU** | |
| **Đề tài: Phân tích và trực quan hóa dữ liệu khách hàng để đưa ra khuyến nghị cho chiến lược marketing của ngân hàng. Xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng.** | |
|  | |
| **Giảng viên hướng dẫn:**  TS. Nguyễn Thôn Dã | |
|  | |
|  | |
| *TP.HCM, tháng 5 năm 2023* | |

# Danh sách sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| MSSV | Họ và tên | Email |
| K214041239 | Trần Thị Thu Ngân | nganttt21404b@st.uel.edu.vn |
| K214142070 | Nguyễn Thị Linh | linhnt21414@st.uel.edu.vn |
| K214142078 | Đoàn Thị Phương Nhi | nhidtp21414@st.uel.edu.vn |
| K214142079 | Trần Thị Kim Oanh | oanhttk21414@st.uel.edu.vn |
| K214142096 | Nguyễn Minh Tú | tunm21414@st.uel.edu.vn |

# Lời cảm ơn

Nhóm ONNTL xin gửi lời cảm ơn chân thành đến **Tiến sĩ Nguyễn Thôn Dã**, giảng viên giảng dạy môn Phân tích dữ liệu, vì đã đồng hành cùng chúng em trong suốt thời gian học tập vừa qua đặc biệt trong quá trình thực hiện đồ án cuối kỳ. Sự tận tâm giảng dạy, giải đáp mọi thắc mắc cũng như những kiến thức quý báu của thầy đã giúp chúng em hiểu rõ hơn về bộ môn phân tích dữ liệu, từ đó có những kiến thức thức để xây dựng và hoàn thiện đồ án cuối kỳ một cách tốt nhất.

Chúng em cảm nhận được giá trị to lớn của những kiến thức đã được tích lũy trong quá trình học tập, và đó chính là nền tảng giúp chúng em hoàn thành đồ án này. Mặc dù nhóm đã cố gắng hết sức, nhưng chắc chắn rằng bài làm vẫn có thể mắc phải những thiếu sót. Vì vậy, chúng em rất mong nhận được sự đánh giá và góp ý từ thầy để cải thiện bài làm của chúng em và áp dụng kiến thức vào tương lai.

Một lần nữa, nhóm ONNTL xin chân thành cảm ơn thầy vì sự hướng dẫn và những đóng góp quý báu trong suốt thời gian vừa qua.

**Mục lục**

[**Danh sách sinh viên thực hiện:**](#_99nb49c5bwjl) **2**

[**Lời cảm ơn**](#_h8d6ycs8dqj5) **3**

[**1. Giới thiệu**](#_9bqqc2xaugv1) **7**

[**2. Các nghiên cứu liên quan**](#_cvo4rtmbukms) **8**

[2.1. Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques](#_p5h5sfyj37x9) 8

[2.2. Bank Deposit Prediction Using Ensemble Learning](#_v4t7qjoz3l2s) 8

[2.3.A fair classifier chain for multi-label bank marketing strategy classification](#_s3at7cxgie8) 9

[2.4 Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and knearest neighbour.](#_if928ekz0yj9) 9

[2.5.Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining](#_cpa3a6neb6jc) 10

[2.6 Báo cáo “Identifying Long-Term Deposit Customers: A Machine Learning Approach”](#_h5mwy69ef49d) 11

[2.7 Nghiên cứu “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”](#_ve7qjgy5vhcf) 11

[2.8 “Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription”](#_vj9iy9qepcqn) 12

[2.9 “Designing a data mining process for the financial services domain”](#_48235xt932ht) 12

[2.10 “Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms”](#_b3ij7zuvyol0) 12

[**3. Nền tảng lý thuyết:**](#_kegw5gloz1e) **13**

[3.1. Thuật toán Decision Tree:](#_898u3uiirtp2) 13

[3.1.1. Giới thiệu sơ lược và cách hoạt động Decision Tree](#_9bsim3h701jv) 13

[3.1.2. Thuật toán ID3](#_oy1mx9n8g8z7) 14

[3.1.3. Thuật toán Entropy](#_dy6jtpstsa33) 15

[3.1.4. Ưu điểm và nhược điểm của thuật toán Decision Tree](#_qmmguqfu7bg3) 16

[3.2. Thuật toán Gradient Boosting (thuật toán tăng cường tốc độ )](#_qegxz2jnqqzu) 17

[3.2.1. Giới thiệu và cách hoạt động của thuật toán Gradient Boosting](#_85ev2nwqis1w) 17

[3.2.2. LightGBM](#_ss5bkopcrtrr) 18

[3.2.3. XGBoost](#_5lku51rq2dpx) 18

[3.2.4. Ưu và nhược điểm của thuật toán Gradient Boosting](#_oh8qy06pf718) 19

[3.3. Thuật toán Random Forest](#_ifo62q5whjlr) 20

[3.3.1. Giới thiệu và cách hoạt động của thuật toán Random Forest](#_fnwgt4t7b5c3) 20

[Giới thiệu](#_nons26thcs13) 20

[3.3.2. Vấn đề tối ưu tham số](#_5nt3wimniv4l) 23

[3.3.3 Ưu và nhược điểm của thuật toán Random Forest](#_oqxgj01dawsg) 24

[3.4. Các chỉ số đánh giá](#_ekks2ce8tjw8) 25

[3.4.1. Accuracy](#_64l7txaa6b87) 25

[3.4.2. Precision](#_qyen9brtlg) 26

[3.4.3. Recall](#_nqkt5ymajbnp) 26

[3.4.4. F1-Score](#_buti1klkshko) 27

[3.4.5. Kappa Score](#_6209wjxi6a71) 27

[**4. Phương pháp luận nghiên cứu**](#_ir0s6370ugrn) **28**

[4.1 Mô hình tổng quan](#_v4w20lwqghzv) 28

[4.2 Thu thập dữ liệu](#_8fqa9yqq1b1t) 29

[4.3 Mô tả dữ liệu](#_34v1i3p192yy) 29

[4.4 Tiền xử lý dữ liệu và trích xuất đặc trưng](#_8qllfhbyc5r6) 31

[4.4.1 Tiền xử lý dữ liệu](#_ta75gx9oqm1) 31

[**5. Kết quả thử nghiệm và phân tích**](#_evmpfrs9ximg) **38**

[5.1 Đưa ra các kết quả phân tích từ biểu đồ các biến độc lập với tỷ lệ chuyển đổi](#_og4dzqu3qa4h) 38

[5.1.1. Độ tuổi và quyết định đăng ký khoản gửi của khách hàng.](#_xp98zj6ikb6b) 39

[5.1.2. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên nghề nghiệp](#_e6280okefjf5) 40

[5.1.3. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên trình độ giáo dục.](#_a61x16l42b3v) 41

[5.1.4. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tình trạng hôn nhân.](#_3rrcwcu628rv) 41

[5.1.5. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tình trạng nợ vay mua hoặc thuê nhà và nợ vay cá nhân.](#_oafd1kfa7fls) 42

[5.1.6. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua phương thức liên hệ.](#_sdweyz6ksdoi) 43

[5.1.7. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tháng và ngày liên hệ gần nhất.](#_mgpp8eyj7rng) 44

[5.1.8. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua số lần liên hệ được thực hiện.](#_uj12kje6dri3) 45

[5.1.9. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng và kết quả của chiến dịch lần trước.](#_hl2gaqol4mbg) 46

[5.1.10. Kết luận:](#_8nazpcs7bmsy) 47

[5.2 Kết quả mô hình dự đoán](#_cjj3vtvmu7bw) 47

[**6. Kết luận**](#_3gtf8g1fy21) **50**

[6.1 Ưu điểm](#_v85s8vqj5okl) 50

[6.2 Hạn chế](#_96nq19o3anbg) 51

[6.3 Hướng phát triển](#_npergj8uxpu3) 51

[**Tài liệu tham khảo**](#_3qceo9rpuxzo) **51**

**Tóm tắt**

Đồ án "Phân tích và trực quan hóa dữ liệu khách hàng để đưa ra khuyến nghị cho chiến lược marketing của ngân hàng. Xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng." nhằm mục đích giúp ngân hàng đưa ra các chiến dịch phù hợp với từng đối tượng khách hàng, tăng cường hiệu quả chiến dịch tiếp thị và xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng. Với mẫu nghiên cứu được thu thập từ hồ sơ của 41188 khách hàng thuộc một ngân hàng ở Bồ Đào Nha, đồ án tập trung vào việc phân tích và trực quan hóa dữ liệu khai thác dữ liệu khách hàng để đưa ra khuyến nghị cho chiến lược marketing của ngân hàng về nhóm đối tượng khách hàng mục tiêu và từng phương hướng cụ thể tác động đến quyết định đăng ký khoản gửi có kỳ hạn đồng thời xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng, là một giải pháp quan trọng và hữu ích cho các ngân hàng trong việc tối ưu hóa chiến lược marketing và tăng cường sự trung thành của khách hàng.

**Từ khóa: phân tích và trực quan hóa dữ liệu, khai thác dữ liệu, chiến lược marketing, mô hình dự đoán, quyết định đăng ký khoản gửi có kỳ hạn.**

# 1. Giới thiệu

Ngân hàng là một lĩnh vực đóng vai trò quan trọng trong việc hỗ trợ và phát triển các hoạt động kinh tế của một quốc gia. Với sự phát triển của khoa học và công nghệ, các ngân hàng đang phải đối mặt với rất nhiều thách thức mới trong việc thu hút và giữ chân khách hàng của mình. Điều này làm cho việc phân tích thông tin khách hàng để đưa ra khuyến nghị cho chiến lược marketing của ngân hàng và xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng trở nên ngày càng quan trọng.

Vì vậy việc nghiên cứu, phân tích và khai thác dữ liệu khách hàng là một việc làm cần thiết, giúp ngân hàng hiểu rõ hơn về hành vi tiêu dùng, nhu cầu và sở thích của khách hàng, từ đó đưa ra các chiến lược marketing phù hợp và tối ưu hóa chiến lược kinh doanh của mình. Ngoài ra, mô hình dự đoán cũng sẽ giúp ngân hàng dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng và đưa ra các chiến lược phù hợp để tăng cường sự hài lòng và tăng trưởng doanh thu cho ngân hàng.

Xuất phát từ nhu cầu trên, nhóm chọn đề tài “Phân tích và trực quan hóa dữ liệu khách hàng để đưa ra khuyến nghị cho chiến lược marketing của ngân hàng. Xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng.” để nghiên cứu. Kết quả của đồ án này sẽ giúp ngân hàng phát triển các chiến lược marketing đúng đối tượng, đúng phương pháp. Bằng cách khai thác dữ liệu khách hàng và phân tích dữ liệu một cách hiệu quả, ngân hàng có thể đưa ra những khuyến nghị phù hợp với từng khách hàng, nâng cao hiệu quả chiến dịch tiếp thị và tăng cường sự hài lòng của khách hàng. Bên cạnh đó, nhóm cũng sẽ xây dựng mô hình dự đoán quyết định đăng ký khoản gửi có kỳ hạn của khách hàng.

# 2. Các nghiên cứu liên quan

## 2.1. Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques

Thảo luận về lựa chọn tính năng trên bộ dữ liệu tiếp thị ngân hàng gồm 1000 phiên bản bằng cách chọn một tập hợp con các thuộc tính và xây dựng bộ phân loại mạng thần kinh sâu để dự đoán khách hàng nào sẽ đăng ký tiền gửi có kỳ hạn bằng cách sử dụng dữ liệu được cung cấp. Một cách tiếp cận khác để xác định khách hàng đăng ký dịch vụ ngân hàng được thảo luận trong [7] bằng cách đạt được hai mục tiêu: mục tiêu đầu tiên là dự đoán phản hồi của từng khách hàng và mục tiêu thứ hai là xác định các tính năng có liên quan bằng cách sử dụng phân cụm. Trong [8], nó đã được đề xuất để cải thiện kết quả của các chiến dịch tiếp thị bằng cách phân nhóm khách hàng dựa trên sự tương đồng và sau đó tạo ra một bộ phân loại. Các thuật toán mới như SMOTE [9] và Rotation Forest (PCA)–J48 [10] cũng đã được đề xuất để cải thiện kết quả của các chiến dịch tiếp thị. Hiệu suất của thuật toán được so sánh với các bộ phân loại hiện có [11-12] trên một số tham số, cụ thể là độ chính xác, lỗi bình phương trung bình và kết quả được phân tích trong các nghiên cứu khác nhau [13-14]. Trong [15], phân tích thành phần chính trên tập dữ liệu tiếp thị ngân hàng chứa các thuộc tính hỗn hợp đã được đề xuất để thực hiện trích xuất tính năng nhằm xây dựng bộ phân loại mạng thần kinh sâu. Tập dữ liệu đầu vào có 1000 phiên bản và phương sai tích lũy của việc giảm thứ nguyên từ 16 xuống 3 là 99,99740%. Dự đoán giá đóng cửa cổ phiếu bằng cách sử dụng 2D PCA được thảo luận trong [16], nơi các mạng lưới niềm tin sâu sắc được sử dụng để thực hiện phân loại. Một nghiên cứu khác về dự đoán giá cổ phiếu liên quan đến mạng thần kinh Elman [17] được trình bày. Trong [18-20], SVM dựa trên PCA và SVM mờ được thảo luận để đánh giá rủi ro tín dụng ngân hàng. Tầm quan trọng của PCA trong việc giảm quy mô cho vay và chẩn đoán bệnh phổi giai đoạn đầu đã được phân tích trong [21].

## 2.2. Bank Deposit Prediction Using Ensemble Learning

Nhiều loại thuật toán dự đoán hồ chứa đã được đề xuất. Các phương pháp được sử dụng phổ biến nhất là Naive Bayes, MLPNN, TAN, Hồi quy logistic, Cây quyết định (C5.0, J48), LADT, RBFN, SVM và Mạng miễn dịch nhân tạo. Elsalamony et al. [6] đã tiến hành một nghiên cứu so sánh về các mạng thần kinh tri giác nhiều lớp (MLPNN) với các phân loại khác nhau, chẳng hạn như B. Tree Augmented NAIVE BAYES (TAN), hồi quy logistic (LR) và C5.0 để phân tích độ chính xác, độ nhạy và kết quả giới tính cụ thể. MLPNN cho thấy độ chính xác tốt nhất là 90,49%, LR cho thấy tỷ lệ độ nhạy tốt nhất là 65,53% và C5.0 cho thấy tỷ lệ độ đặc hiệu tốt nhất là 93,23% để dự đoán tiền gửi ngân hàng. Wisaeng et al. [7] đã phân tích giá trị trọn đời của khách hàng và mạng lưới thần kinh để cải thiện dự đoán về đăng ký tiền gửi ngân hàng trong các chiến dịch tiếp thị qua điện thoại. Moro et al. Ba lần lặp lại của phương pháp Quy trình tiêu chuẩn xuyên ngành để khai thác dữ liệu (CRISP-DM) đã được thực hiện. [9] thu được kết quả chính xác từ mô hình DM bằng cách sử dụng bộ dữ liệu chiến dịch tiếp thị trực tiếp từ Ngân hàng Bồ Đào Nha. Trong công việc này, thời lượng cuộc gọi được coi là yếu tố dự đoán quan trọng nhất cho vấn đề phân loại này. Bài báo chỉ ra rằng các mô hình tốt hơn do mạng nơ-ron (NN) triển khai có thể đạt được hiệu suất dự đoán cao hơn.

## 2.3.A fair classifier chain for multi-label bank marketing strategy classification

Phân loại nhiều nhãn là một biến thể của bài toán phân loại trong đó nhiều nhãn có thể được gán cho mỗi trường hợp (Herrera et al., 2016). Ngoài ra, phân loại theo nhiều nhãn có thể được mô tả như một dạng khái quát hóa của phân loại theo nhiều lớp trong đó không có ràng buộc nào về số lượng nhãn có thể được gán cho mỗi trường hợp (Wu và cộng sự, 2020). Ai cũng biết rằng phân loại đa nhãn được sử dụng trong nhiều ứng dụng trong thế giới thực như phân loại hình ảnh (Allaouzi và Ahmed, 2019), phân đoạn hình ảnh (Zhang và cộng sự, 2020), chú thích hình ảnh (Loukas và Sgouros, 2020), ghi nhãn văn bản y sinh (Du và cộng sự, 2019), theo dõi huyết áp (Douibi và cộng sự, 2019), dự đoán chấn thương khi trượt tuyết (Radovanovic và cộng sự, 2020a), xử lý tự động và khai thác kiến thức ´ cho mạng xã hội ( Hashemi và Hall, 2020) và lọc cộng tác (Zhou et al., 2019)

## 2.4 Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and knearest neighbour.

Một nghiên cứu của Tekouabou, Cherif và Silkan (2019) cho thấy hiệu quả của năm kỹ thuật máy học phụ thuộc vào việc áp dụng tiếp thị qua điện thoại hiệu quả của ngân hàng. Nghiên cứu tập trung vào năm kỹ thuật máy học, bao gồm hồi quy logistic, máy vectơ hỗ trợ, mạng trung lập nhân tạo, cây quyết định và Bayes ngây thơ. Sử dụng các phân tích về độ chính xác và thước đo F, nghiên cứu cho thấy rằng các kỹ thuật cây quyết định và mạng thần kinh nhân tạo đạt được độ chính xác cao hơn 93%. Kết quả cho thấy ANN và DT là những yếu tố quan trọng quyết định khách hàng đăng ký sử dụng sản phẩm dịch vụ ngân hàng. Việc áp dụng các kỹ thuật máy học bị ảnh hưởng bởi khả năng diễn giải kết quả của ngân hàng và tính sẵn có của dữ liệu. Tekouabou và cộng sự. (2019) đã phát triển một khung đại diện cho phương pháp mô hình hóa dữ liệu cho các thách thức phân loại. Các phương pháp tính toán dựa trên cấu trúc hỗ trợ các chiến lược khai phá dữ liệu. Công nghệ đã cho phép các ngân hàng thu thập thông tin liên quan từ khách hàng. Việc sử dụng các chiến lược máy học phụ thuộc vào sự sẵn có của thông tin và phương pháp tính toán được sử dụng bởi ngân hàng. Khuôn khổ cũng chỉ ra rằng mô hình học tập và phương pháp học tập có giám sát ảnh hưởng đến việc sử dụng tiếp thị trực tiếp. Một nghiên cứu được lựa chọn bởi Daniel, Grayzna và Barbara (2016) sử dụng chiến dịch phân loại có giám sát đã phát hiện ra rằng sự phức tạp của dữ liệu mà các ngân hàng sử dụng có thể dẫn đến việc sử dụng các kỹ thuật khai thác dữ liệu. Hình thức mà các ngân hàng sử dụng trong tiếp thị trực tiếp ảnh hưởng đến khả năng nhận được thêm thông tin từ khách hàng. Nghiên cứu cho thấy hơn 50% ngân hàng đã sử dụng chiến lược tiếp thị trực tiếp trong các chiến dịch của họ (Rubtcova & Pavenkov, 2019). Khả năng khách hàng trả lời tiền gửi có kỳ hạn phụ thuộc vào việc thông tin do ngân hàng cung cấp như một phần của chương trình khuyến mãi có đầy đủ hay không.

## 2.5.Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining

Bài báo thảo luận về tầm quan trọng của tiếp thị qua điện thoại, bằng cách xử lý hiệu quả dữ liệu khách hàng luôn là nền tảng để xây dựng các mối quan hệ tích cực và kết nối mạng lưới trong việc quảng bá và cung cấp sản phẩm mới [19]. Bài viết trên sử dụng phương pháp khai thác dữ liệu tích hợp và học máy đối với dữ liệu ngân hàng. Mục tiêu cũng giống như đã thảo luận trong nghiên cứu này, đó là tìm ra giải pháp tối ưu cho ngân hàng trong việc thực hiện chiến dịch tiếp thị qua điện thoại. Mô hình đề xuất sử dụng hai mô hình khác nhau; Hồi quy logistic và perceptron đa lớp. Mục đích là để phân biệt hiệu suất giữa phân loại và mô hình đa lớp của nơ-ron. Trong quá trình xác thực chéo, các kết hợp khác nhau của 72 Fawaz J. Alsolami et al. đã sử dụng. được áp dụng, cung cấp độ chính xác đo lường hỗn hợp. Các công việc liên quan được thảo luận trong phần này đã giải quyết các vấn đề chính được nêu ra trong các nghiên cứu này. Công việc trước đây đã chỉ ra tầm quan trọng của ý tưởng nghiên cứu khi một số nghiên cứu đã được công bố trong hai năm qua. Do đó, nghiên cứu này có thể có ý nghĩa quan trọng đối với lĩnh vực ngân hàng, nơi các nhà hoạch định chính sách có thể áp dụng phương pháp CNTT tích hợp để phân tích dữ liệu và thu thập bằng chứng thống kê trước khi phát triển một chiến lược tiếp thị qua điện thoại mới. Dựa trên tổng quan tài liệu, phương pháp được trích xuất và sử dụng phổ biến nhất trong các phương pháp tiếp cận công việc trước đây là hồi quy logistic, cây quyết định và perceptron đa lớp. Các mục liên quan phổ biến nhất và kết quả lập mô hình liên quan được trình bày trong Bảng 1, trong đó cùng một bộ dữ liệu được sử dụng trong nghiên cứu này đã được chọn. Bảng cho thấy hiệu suất thấp nhất được đo bằng cách sử dụng cây quyết định, trong khi hiệu suất cao nhất đạt được bằng cách sử dụng perceptron đa lớp.

|  |  |  |  |
| --- | --- | --- | --- |
| Techniques / Previous Work | UCI Data, Using R Tool, [17] | UCI Data using Weka, ,2019 [19] | UCI Data Using R Tool, 2014[18] |
| Decision Tree | 86.80% | X | 92.90% |
| Logistic Regression | X | 90.18% | 90. 0% |
| Multilayer Perceptron | X | 89.40% | 92.90% |

## 2.6 Báo cáo “Identifying Long-Term Deposit Customers: A Machine Learning Approach”

Theo nghiên cứu của Mohammad Abu Tareq Rony, Md. Mehedi Hassan, Eshtiak Ahmed, Asif Karim, Sami Azam và D. S. A. Aashiqur Reza, báo cáo “Identifying Long-Term Deposit Customers: A Machine Learning Approach” nghiên cứu về việc sử dụng các kỹ thuật phân tích dữ liệu, lựa chọn đặc trưng và học máy để phân tích các đặc trưng của khách hàng cũng như các biến có thể ảnh hưởng đến quyết định của khách hàng một cách đáng kể. Các phương pháp này có thể được sử dụng để xác định người tiêu dùng trong các danh mục khác nhau để dự đoán liệu một khách hàng có đăng ký gửi tiền gửi dài hạn hay không, cho phép chiến lược tiếp thị trở nên thành công hơn. Trong nghiên cứu này, tác giả đã sử dụng phân tích dữ liệu tương tác (EDA), phân tích thành phần chính (PCA), phân tích yếu tố và tương quan trong tập dữ liệu được cung cấp. Ngoài ra, mục tiêu của nghiên cứu là sử dụng ít nhất ba thuật toán phân loại điển hình giữa Logistic Regression, Random Forest, Support Vector Machine và K-nearest neighbors để xây dựng các mô hình dự đoán về khách hàng đăng ký gửi tiền gửi dài hạn. Tác giả đã sử dụng Logistic Regression và đạt được độ chính xác tốt nhất là 90,64% và độ nhạy là 99,05%. Kết quả được phân tích bằng điểm chính xác, độ nhạy và độ đặc hiệu của các thuật toán này.

## 2.7 Nghiên cứu “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”

Nghiên cứu “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology” mô tả một dự án khai thác dữ liệu (Data Mining) dựa trên phương pháp CRISP-DM. Dữ liệu thực tế được thu thập từ một chiến dịch tiếp thị của Bồ Đào Nha liên quan đến đăng ký tiền gửi ngân hàng. Mục tiêu kinh doanh là tìm một mô hình có thể giải thích thành công của một liên hệ, tức là nếu khách hàng đăng ký tiền gửi. Mô hình này có thể tăng hiệu quả chiến dịch bằng cách xác định các đặc điểm chính ảnh hưởng đến thành công, giúp quản lý tốt hơn các nguồn lực có sẵn (ví dụ: nỗ lực nhân sự, cuộc gọi điện thoại, thời gian) và lựa chọn một tập khách hàng tiềm năng chất lượng cao và giá cả phải chăng.

## 2.8 “Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription”

Bài báo “Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription” nghiên cứu về việc sử dụng các mô hình học máy để dự đoán khách hàng đăng ký gửi tiền gửi ngân hàng. Bài báo này được viết bởi Sipu Hou, Zongzhen Cai, Jiming Wu và Peng Xie và được xuất bản trên tạp chí International Journal of Information Technology & Decision Making vào tháng 1 năm 2022. Bài báo này giới thiệu một phương pháp dự đoán khách hàng đăng ký gửi tiền gửi ngân hàng bằng cách sử dụng mô hình học máy logistic regression. Kết quả cho thấy phương pháp này có hiệu quả và có thể được áp dụng để tìm khách hàng mục tiêu và cải thiện độ chính xác của marketing.

## 2.9 “Designing a data mining process for the financial services domain”

Bài báo “Designing a data mining process for the financial services domain” được đăng trên tạp chí Journal of Business Analytics giới thiệu một quy trình khai thác dữ liệu cho lĩnh vực dịch vụ tài chính được thiết kế để đáp ứng yêu cầu về quản lý rủi ro, tuân thủ quy định và đảm bảo chất lượng trong các dự án khai thác dữ liệu. Quy trình này được gọi là Financial Industry Process for Data Mining (FIN-DM) và được thiết kế để tích hợp tuân thủ quy định và quản lý rủi ro vào vòng đời của dự án khai thác dữ liệu. Bài báo cũng giới thiệu các khái niệm cơ bản về khai thác dữ liệu và các quy trình khai thác dữ liệu tiêu chuẩn như CRISP-DM.

## 2.10 “Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms”

Bài báo “Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms” nghiên cứu về việc dự đoán nhu cầu sử dụng các phương pháp telemarketing để quảng bá các khoản tiền gửi ngân hàng dài hạn đến khách hàng tiềm năng. Nghiên cứu này khám phá nhu cầu về khoản tiền gửi ngân hàng dài hạn bằng cách sử dụng các thuật toán học máy khác nhau như Random Forest (RF), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Decision Tree (DT) và Logistic Regression (LR). Tập dữ liệu liên quan đến các chiến dịch tiếp thị trực tiếp (cuộc gọi điện thoại) của một tổ chức ngân hàng Bồ Đào Nha được xem xét để phân tích. Kết quả cho thấy rằng mô hình LR cung cấp độ chính xác 92,48%, đây là mô hình tốt nhất để dự đoán khách hàng tiềm năng có quan tâm đến khoản tiền gửi dài hạn thông qua telemarketing. Kết quả của nghiên cứu cũng cung cấp thông tin chi tiết cho các ngân hàng để đưa ra quyết định chính sách telemarketing trong việc thành công của khoản tiền gửi ngân hàng cho khách hàng hiện tại và tiềm năng của họ.

# 3. Nền tảng lý thuyết:

## 3.1. Thuật toán Decision Tree:

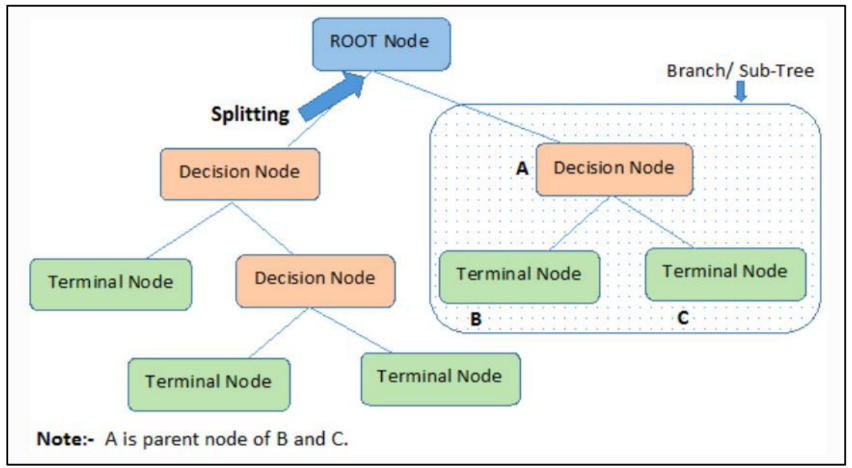
### 3.1.1. Giới thiệu sơ lược và cách hoạt động Decision Tree

Bạn có biết rằng trong cuộc sống hàng ngày, bạn vẫn sử dụng phương pháp cây quyết định. Ví dụ, bạn đi siêu thị mua sữa cho cả gia đình. Những từ đầu tiên trong đầu bạn sẽ là: bạn cần mua bao nhiêu sữa? Bạn sẽ biết chắc: nếu là ngày thường gia đình bạn sẽ dùng hết 1 lít sữa, đến cuối tuần sẽ là 1,5 lít. Vì vậy, tùy vào ngày mà bạn sẽ quyết định mua bao nhiêu thực phẩm cho gia đình mình. Đó là một loại cây quyết định nhị phân

.

Khái niệm Decision Tree

Cây quyết định là một thuật toán học có giám sát phi tham số, được sử dụng cho cả nhiệm vụ phân loại và hồi quy. Gồm cấu trúc cây, phân cấp, bao gồm cả nút gốc, các nhánh, nút bên trong và nút lá. Dựa vào dãy các luật, cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân double), Định danh(Nominal), Thứ tự( Ordinal), Số lượng(Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là double hoặc Ordinal. Có thể nói, dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) , thuật toán sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết. Trong việc tính toán các xác suất có điều kiện, cây quyết định cũng là một phương tiện có tính mô tả . Có mô tả thuật toán trên như là sự kết hợp kỹ thuật toán học để hỗ trợ việc mô tả, phân loại và tổng quát hóa tập dữ liệu đã cho.

* Hình (1) Mô hình phân loại các nút trong Decision Tree*

Cách hoạt động thuật toán Decision Tree

Xây dựng cây: Thuật toán bắt đầu với tập dữ liệu huấn luyện ban đầu. Đầu tiên, thuật toán chọn một thuộc tính để phân chia tập dữ liệu ban đầu thành các tập con. Quá trình chọn thuộc tính phân chia được thực hiện bằng cách sử dụng các độ đo như Information Gain, Gini Index hoặc Chi-square để đánh giá mức độ phân chia của thuộc tính.

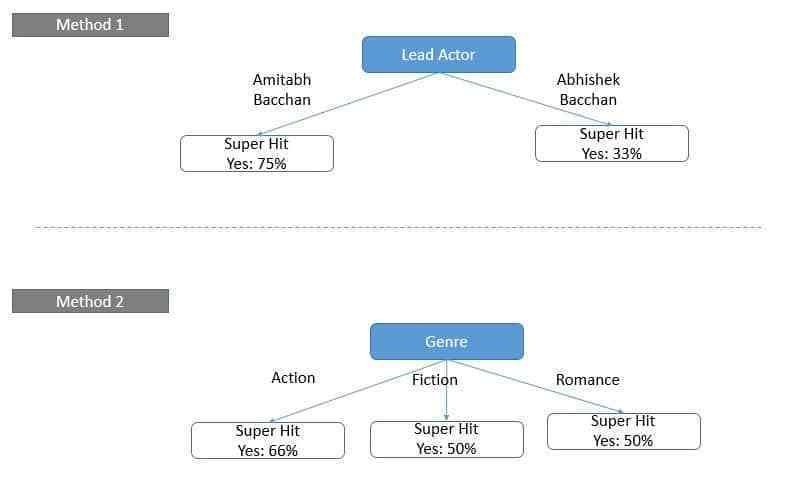
Tách tập dữ liệu: Sau khi chọn thuộc tính phân chia, cây quyết định được tách thành các nhánh con tương ứng với các giá trị riêng biệt của thuộc tính phân chia. Quá trình này tiếp tục đệ quy cho đến khi đạt được một điều kiện dừng, chẳng hạn như đạt đến một độ sâu tối đa hoặc không còn thuộc tính nào có thể phân chia hiệu quả.

Dự đoán: Khi cây quyết định đã được xây dựng, nó có thể được sử dụng để dự đoán nhãn cho các mẫu mới dựa trên quy tắc phân loại trong các nút lá của cây.

### 3.1.2. Thuật toán ID3

ID3(J.R. Quinlan 1993) sử dụng phương pháp tham lam tìm kiếm từ trên xuống thông qua không gian của các nhánh có thể không có backtracking. ID3 sử dụng Entropy và Information Gain để xây dựng một cây quyết định.

Giả sử, bạn muốn xác định độ thành công của bộ phim chỉ trên 1 yếu tố, bạn sẽ có hai cách thực hiện sau qua diễn viên chính của phim và qua thể loại phim.

**** *Hình (2) Sơ đồ hai phương pháp*

Qua sơ đồ, ta có thể thấy rõ ràng ràng, với phương pháp thứ nhất, ta phân loại được rõ ràng, trong khi phương pháp thứ hai, ta có một kết quả lộn xộn hơn. Và tương tự, cây quyết định sẽ thực hiện như trên khi thực hiện việc chọn các biến.

Có rất nhiều hệ số khác nhau mà phương pháp cây quyết định sử dụng để phân chia. Dưới đây, tôi sẽ đưa ra hai hệ số phổ biến là Information Gain và Gain rate( ngoài ra còn hệ số Gini).

### 3.1.3. Thuật toán Entropy

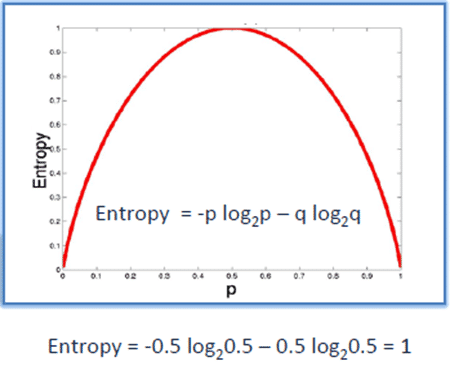
Entropy là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau

Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1, x2,, xn. pi =p ( x = xi).

p = ( p1, p2,, pn). Entropy của phân phối này được định nghĩa là H( p) = – ∑ nn = 1 pi log( pi)

Giả sử bạn tung một đồng xu, entropy sẽ được tính như sau

H = -(0.5 ln(0.5)0.5 ln(0.5))

****

Ta có thể thấy rằng, entropy đạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

P tinh khiết pi = 0 hoặc pi = 1

P vẩn đục pi = 0.5, khi đó hàm Entropy đạt đỉnh cao nhất

### 3.1.4. Ưu điểm và nhược điểm của thuật toán Decision Tree

### 

Ưu điểm của thuật toán Decision Tree:

* Dễ hiểu và diễn giải: Cây quyết định tạo ra một mô hình rõ ràng và dễ hiểu, giúp người dùng hiểu cách quyết định được đưa ra. Cây quyết định có thể được biểu diễn dưới dạng đồ thị hoặc cây, giúp thể hiện các quy tắc phân loại một cách trực quan.
* Xử lý dữ liệu hỗn loạn: Decision Tree có khả năng xử lý dữ liệu có nhiễu và thiếu giá trị. Nó có khả năng chịu được các dạng dữ liệu không đồng nhất và không cần tiền xử lý dữ liệu phức tạp.
* Phân loại đa lớp: Decision Tree có khả năng phân loại đa lớp, tức là nó có thể xử lý các bài toán phân loại có nhiều hơn hai nhãn.
* Tính tương tác: Decision Tree có khả năng xác định tương tác giữa các thuộc tính và quyết định phân loại. Nó có thể phát hiện quy tắc phức tạp và tương tác phức tạp giữa các biến độc lập.

Nhược điểm của thuật toán Decision Tree:

* Dễ bị overfitting: Decision Tree có xu hướng overfitting, đặc biệt khi cây quá phức tạp và không được kiểm soát. Overfitting xảy ra khi cây quyết định quá khớp với dữ liệu huấn luyện, làm giảm khả năng tổng quát hóa cho dữ liệu mới.
* Không ổn định với các biến nhỏ: Decision Tree có thể rất nhạy cảm với sự thay đổi nhỏ trong dữ liệu huấn luyện. Một biến nhỏ có thể làm thay đổi quyết định phân loại của cây.
* Tính chất đơn điểm: Mỗi cây quyết định xây dựng dựa trên một tập dữ liệu cụ thể. Điều này có nghĩa là sự thay đổi nhỏ trong tập dữ liệu huấn luyện có thể dẫn đến các cây quyết định khác nhau. Điều này làm cho tính ổn định của cây quyết định không cao.
* Không xử lý tốt các quy tắc phức tạp: Decision Tree có khả năng xây dựng các quy tắc phân loại đơn giản.

## 3.2. Thuật toán Gradient Boosting (thuật toán tăng cường tốc độ )

### 3.2.1. Giới thiệu và cách hoạt động của thuật toán Gradient Boosting

Gradient Boosting là một thuật toán học máy mạnh mẽ và phổ biến, nó được sử dụng để xây dựng các mô hình dự đoán có hiệu suất cao. Đây là một thuật toán thuộc họ ensemble learning, trong đó nhiều mô hình yếu được kết hợp lại để tạo ra một mô hình mạnh hơn.

Tăng cường độ dốc là một thuật toán bắt nguồn từ thuật toán cây quyết định, thuật toán này tuần tự xây dựng và học nhiều cây quyết định

Cách hoạt động của Gradient Boosting:

* Xây dựng mô hình ban đầu: Ban đầu, một mô hình yếu được xây dựng, thường là một cây quyết định đơn giản. Mô hình yếu này được gọi là "base learner" hoặc "weak learner".
* Đánh giá sai số: Sai số giữa kết quả dự đoán của mô hình ban đầu và kết quả thực tế được tính toán. Đây là sai số cần được giảm thiểu trong quá trình huấn luyện.
* Xây dựng mô hình tiếp theo: Một mô hình yếu mới được xây dựng với mục tiêu giảm sai số. Để làm điều này, Gradient Boosting sẽ tạo ra một phiên bản "cải tiến" của mô hình yếu ban đầu bằng cách tìm kiếm hướng đi tốt nhất để giảm sai số.
* Cập nhật dự đoán: Dự đoán của mô hình ban đầu và mô hình mới được kết hợp lại để tạo ra dự đoán cuối cùng. Quá trình này giúp cải thiện dự đoán ban đầu bằng cách tập trung vào các điểm dữ liệu mà mô hình ban đầu không dự đoán đúng.
* Lặp lại quá trình: Các bước 2-4 được lặp lại để xây dựng các mô hình yếu tiếp theo và cải thiện dự đoán cho đến khi đạt được số lượng mô hình yếu đã xác định trước đó hoặc đạt đến điều kiện dừng khác nhau.

### 3.2.2. LightGBM

LightGBM là một framework mã nguồn mở được phát triển bởi Microsoft và được sử dụng để xử lý thuật toán tăng cường độ dốc (Gradient Boosting). Nó là một phiên bản tối ưu và nhanh chóng của thuật toán Gradient Boosting.

Các đặc điểm và ưu điểm của LightGBM:

* Tốc độ cao: LightGBM sử dụng một số kỹ thuật tối ưu để cải thiện tốc độ huấn luyện và dự đoán. Nó sử dụng cách tiếp cận "leaf-wise" (tìm kiếm theo lá) để xây dựng cây quyết định, giúp tăng tốc quá trình huấn luyện. LightGBM cũng hỗ trợ việc xử lý song song và có thể chạy trên nhiều CPU.
* Tiêu thụ bộ nhớ thấp: LightGBM sử dụng kỹ thuật "Histogram-based Gradient Boosting" để giảm bộ nhớ cần thiết cho việc xây dựng cây quyết định. Thay vì lưu trữ tất cả các giá trị của các thuộc tính, nó xây dựng histogram của dữ liệu và chỉ lưu trữ các giá trị thống kê. Điều này giúp giảm bộ nhớ được sử dụng và tăng tốc độ huấn luyện.
* Hỗ trợ các tính năng phong phú: LightGBM cung cấp nhiều tính năng hữu ích như xử lý dữ liệu thiếu, xử lý các biến định tính và hạng mục, tạo tập dữ liệu phân loại tỷ lệ cân bằng và kiểm tra quá khớp. Nó cũng hỗ trợ tùy chỉnh các hàm mất mát và đánh giá để phù hợp với các vấn đề cụ thể.
* Khả năng mở rộng và tích hợp: LightGBM có khả năng mở rộng để xử lý các tập dữ liệu lớn với hàng triệu mẫu và hàng nghìn thuộc tính. Nó cũng hỗ trợ tích hợp với các công cụ và thư viện phổ biến khác như scikit-learn và pandas.
* Hiệu suất dự đoán tốt: LightGBM tạo ra các mô hình có khả năng dự đoán cao và độ chính xác tốt trên các tập dữ liệu mới.

### 3.2.3. XGBoost

XGBoost (Extreme Gradient Boosting) là một công cụ mạnh mẽ khác sử dụng thuật toán Gradient Boosting và cũng được rất ưa chuộng và sử dụng phổ biến trong cộng đồng học máy, đặc biệt là trong các cuộc thi trên Kaggle.

Các đặc điểm và ưu điểm của XGBoost:

* Hiệu suất cao: XGBoost được tối ưu hóa để đạt hiệu suất cao, đồng thời giảm thiểu thiếu hụt thông qua việc sử dụng các kỹ thuật như tối ưu hóa hàm mất mát, xử lý song song và quản lý bộ nhớ hiệu quả.
* Xử lý tốt dữ liệu lớn và nhiều thuộc tính: XGBoost có khả năng xử lý các tập dữ liệu lớn có hàng triệu mẫu và hàng nghìn thuộc tính. Nó cũng hỗ trợ nhiều định dạng dữ liệu đầu vào và có khả năng làm việc với các biến độc lập không đồng nhất.
* Quản lý quá khớp: XGBoost có nhiều cơ chế để giảm overfitting, bao gồm giới hạn độ sâu của cây, đánh giá cắt tỉa, và điều chuẩn L1 và L2.
* Hỗ trợ các hàm mất mát tùy chỉnh: XGBoost cho phép người dùng xác định và tùy chỉnh các hàm mất mát phù hợp với bài toán cụ thể, giúp tối ưu hóa mô hình cho mục tiêu cụ thể.
* Tích hợp dễ dàng: XGBoost có khả năng tích hợp với các thư viện và framework phổ biến như scikit-learn, pandas và numpy, giúp dễ dàng tích hợp vào quy trình làm việc hiện có.
* Khả năng diễn giải: XGBoost cung cấp thông tin về độ quan trọng của các thuộc tính trong quá trình huấn luyện, giúp người dùng hiểu rõ hơn về mô hình và tạo ra giải thích dễ hiểu.

### 3.2.4. Ưu và nhược điểm của thuật toán Gradient Boosting

Ưu điểm của thuật toán Gradient Boosting:

* Hiệu suất cao: Gradient Boosting là một thuật toán mạnh mẽ với hiệu suất cao, đạt được độ chính xác tốt trên các bài toán dự đoán.
* Xử lý tốt các mô hình phức tạp: Gradient Boosting có khả năng xây dựng các mô hình phức tạp và mô hình hóa mối quan hệ phi tuyến giữa các biến.
* Khả năng xử lý các loại dữ liệu: Gradient Boosting có khả năng xử lý cả dữ liệu số và dữ liệu hạng mục. Nó có thể tự động mã hóa các biến hạng mục để sử dụng trong quá trình huấn luyện.
* Xử lý tốt dữ liệu không cân bằng: Gradient Boosting có khả năng xử lý tốt các tập dữ liệu không cân bằng, trong đó số lượng mẫu của một lớp ít hơn so với các lớp khác.
* Tích hợp các kỹ thuật chống quá khớp: Gradient Boosting cung cấp các kỹ thuật chống quá khớp như cắt tỉa cây, giới hạn độ sâu cây, và điều chuẩn để giảm quá khớp trong quá trình huấn luyện.

Nhược điểm của thuật toán Gradient Boosting:

* Thời gian huấn luyện lâu: Gradient Boosting có thể mất nhiều thời gian để huấn luyện, đặc biệt là khi số lượng cây và kích thước dữ liệu lớn.
* Dễ bị quá khớp: Gradient Boosting có nguy cơ bị quá khớp trên tập huấn luyện, đặc biệt khi số lượng cây lớn. Việc sử dụng các kỹ thuật chống quá khớp và kiểm soát siêu tham số là cần thiết để giảm quá khớp.
* Phụ thuộc vào siêu tham số: Gradient Boosting có một số siêu tham số cần được tinh chỉnh phù hợp. Việc chọn sai siêu tham số có thể ảnh hưởng đến hiệu suất của mô hình.
* Không thể xử lý dữ liệu có đặc trưng không rõ ràng: Gradient Boosting có khả năng tốt trong việc xử lý dữ liệu có đặc trưng rõ ràng và quan hệ tuyến tính. Tuy nhiên, nếu dữ liệu có các quan hệ phi tuyến phức tạp hoặc không có đặc trưng rõ ràng, Gradient Boosting có thể không đạt được hiệu suất tốt.
* Không ổn định với dữ liệu nhạy cảm: Gradient Boosting có thể nhạy cảm với các thay đổi nhỏ trong dữ liệu đầu vào. Một sự thay đổi nhỏ có thể dẫn đến sự thay đổi lớn trong mô hình đầu ra, điều này có thể làm mất tính ổn định của mô hình.

## 3.3. Thuật toán Random Forest

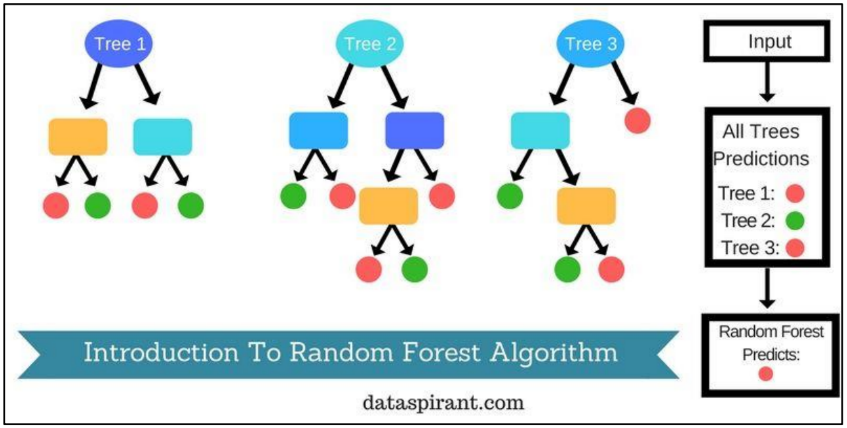
### 3.3.1. Giới thiệu và cách hoạt động của thuật toán Random Forest

### Giới thiệu

Thuật ngữ Random Forest xuất hiện lần đầu tiên vào năm 1995. Năm 1996, nó được kết hợp với công nghệ chọn thuộc tính ngẫu nhiên của Leo Breiman Năm 2001, Leo Breiman thiết lập thuật toán Random Forest bổ sung thêm thuộc tính ngẫu nhiên.

Ngẫu nhiên hóa các danh mục. Ngoài việc sử dụng các mẫu dữ liệu khác nhau cho mỗi cây, các khu rừng ngẫu nhiên đã được thay đổi để xây dựng các cây hồi quy và phân loại khác nhau. Thư viện triển khai rừng ngẫu nhiên đầu tiên được xây dựng ở Fortran bởi Leo Breiman và Cutler. Một khu rừng ngẫu nhiên được mô hình hóa như

cây phân loại Tuy nhiên, rừng ngẫu nhiên sử dụng các mẫu ngẫu nhiên cho cây và lựa chọn thuộc tính ngẫu nhiên khi phân chia cây.Thuật toán rừng ngẫu nhiên đã được chứng minh là chính xác hơn và nhanh hơn khi được huấn luyện trên không gian dữ liệu lớn với nhiều thuộc tính và việc sử dụng dự đoán của tất cả các cây trong 15 khu rừng để phân loại hoặc hồi quy giúp cho kết quả thuật toán chính xác hơn.

****

*Hình (3) Thuật toán Random Forest*

Rừng ngẫu nhiên có thể giải quyết cả vấn đề hồi quy và phân loại. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Như tên gợi ý, Random Forest (RF) dựa trên:

Random: Tính ngẫu nhiên. Rừng: Nhiều cây quyết định.

Rừng ngẫu nhiên tạo cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên,

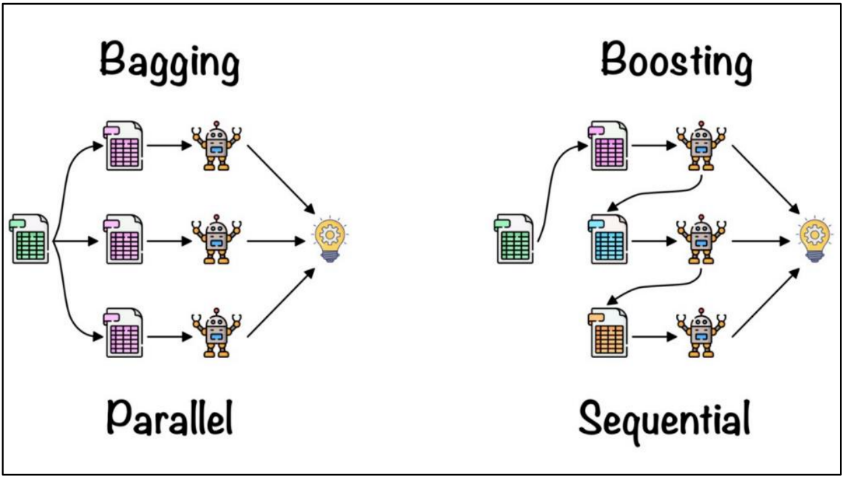
Từ mỗi cây, giải pháp tốt nhất được dự đoán và chọn bằng cách bỏ phiếu.Cũng cung cấp một chỉ số tốt về tầm quan trọng của tính năng. rừng ngẫu nhiên có nhiều ứng dụng như công cụ đề xuất, phân loại và lựa chọn hình ảnh tính năng. Nó có thể được sử dụng để phân loại những người xin vay trung thành, xác định hoạt động gian lận và dự đoán dịch bệnh. Trước khi tìm hiểu hoạt động của RF, chúng ta phải xem xét về phương pháp kết hợp

( phương pháp Ensemble). Có thể sử dụng để tăng độ chính xác trên tập dữ liệu là kết

hợp một số mô hình với nhau. Phương pháp này gọi là esemble. Phương pháp

Ensemble Learning được chia thành những loại sau đây

* Bagging( kỹ thuật đóng gói) xây dựng một lượng lớn các mô hình( thường là cùng loại) trên những mẫu con( subsamples) khác nhau từ tập huấn luyện dữ liệu một cách song song nhằm đưa ra dự đoán tốt hơn.
* Boosting( kỹ thuật tăng cường) xây dựng một lượng lớn các mô hình( thường là cùng loại). Tuy nhiên quá trình huấn luyện trong phương pháp này diễn ra tuần tự theo chuỗi( sequence). Trong chuỗi này mỗi mô hình sau sẽ học cách sửa những lỗi của dữ liệu trong các mô hình trước. Cụ thể là mục tiêu của kỹ thuật tăng cường này sẽ khắc phục và giải quyết các dữ liệu mà mô hình trước dự đoán sai nhằm đưa ra dự đoán tốt hơn trong thuật toán.

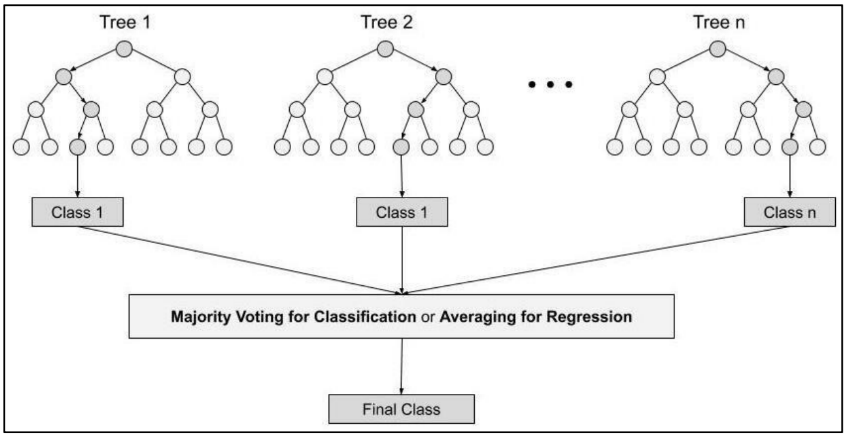


*Hình (4) Hai kỹ thuật đặc trưng trong phương pháp Ensemble*

**Cách hoạt động thuật toán Random Forest**

**-Kỹ thuật đóng gói (Bagging) trong RF hoạt động theo các bước sau:**

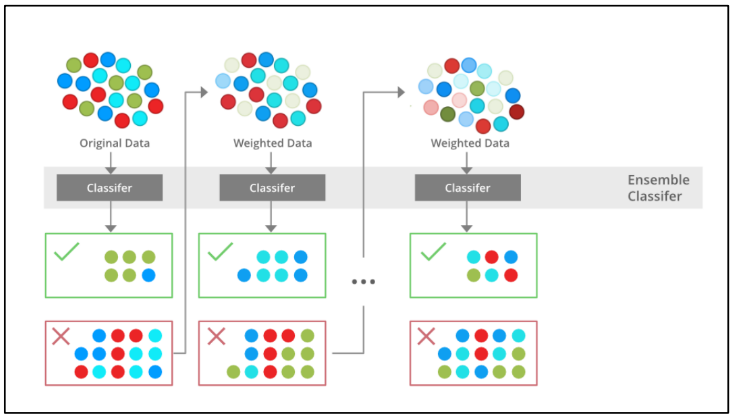
* Bước 1: Chọn một mẫu ngẫu nhiên từ tập dữ liệu đã cho (chọn điểm dữ liệu "k" ngẫu nhiên từ tập huấn luyện "m", lưu ý rằng k < m). Từ tập hợp các thuộc tính "k", Nút tính toán "d" là lựa chọn tốt nhất cho nút phân loại. Chia các nút con cho. Nút tốt nhất vừa được tìm thấy. Lặp lại cho đến khi đạt được k nút Tạo "n" cây quyết định.
* Bước 2: Xây dựng cây quyết định cho từng mẫu và lấy kết quả dự đoán từ cây quyết định mỗi cây quyết định.
* Bước 3: Bình chọn cho từng dự đoán. kiểm phiếu Toàn bộ khu rừng cho mỗi kết quả.
* Bước 4: Chọn kết quả xét theo đa số phiếu (dự đoán) nhiều nhất là dự đoán cuối cùng) hoặc trung bình của phân loại và hồi quy tương ứng.



*Hình (5) Các bước hoạt động của Bagging trong thuật toán Random Forest*

**-Kỹ thuật tăng cường (Boosting) trong RF hoạt động theo các bước sau:**

* Bước 1 - Khởi tạo tập dữ liệu và gán trọng số bằng nhau cho các điểm dữ liệu.
* Bước 2 Cung cấp dữ liệu này làm đầu vào cho mô hình và xác định điểm dữ liệu bị phân loại sai.
* Bước 3 Tăng trọng số của các điểm dữ liệu được phân loại sai và giảm trọng số của số điểm dữ liệu được phân loại đúng. Sau đó chuẩn hóa trọng số của tất cả các điểm dữ liệu.
* Bước 4: Đưa ra kết quả dự đoán sau khi chuẩn hóa dữ liệu để thực hiện phân loại và hồi quy tương ứng

****

*Hình (6) Minh họa về quá trình huấn luyện trong kỹ thuật tăng cường diễn ra tuần tự theo chuỗi*

### 3.3.2. Vấn đề tối ưu tham số

Đối với các thuật toán nói chung và rừng ngẫu nhiên nói riêng, việc tối ưu tham số dãy là rất quan trọng và cần thiết. Bằng cách thay đổi giá trị hoặc phép tính của tham số chữ số trong thuật toán, bạn có thể thay đổi độ chính xác, thời gian thực hiện và tài nguyên hệ thống tổng thể của thuật toán. Random Forest cũng có một số tham số đào tạo. Do đó, cả trong đào tạo và với trong một số ứng dụng nhất định, điều quan trọng là chọn các tham số theo cách sao cho thuật toán có hiệu quả cao và cho kết quả tốt.

Siêu tham số tối ưu hóa quan trọng trong thuật toán: Siêu tham số rừng ngẫu nhiên được sử dụng để tăng khả năng dự đoán của mô hình hoặc để làm cho mô hình

chữ số nhanh hơn:

– Tăng khả năng dự đoán của ước tính:

* n\_estimators siêu tham số, chỉ định số lượng cây sau đó thuật toán

được tạo trước nếu số phiếu bầu tối đa hoặc giá trị trung bình của các dự đoán được chấp nhận.Nói chung, nhiều cây hơn sẽ tăng hiệu suất và đưa ra dự đoán ổn định hơn, nhưng chúng cũng làm chậm quá trình tính toán.

* Một siêu tham số quan trọng khác là max\_features. Đây là số lượng tối đa

thuộc tính mà RF được kiểm tra trong một cây. Việc tăng max\_features thường cải thiện hiệu suất của mô hình vì có nhiều tùy chọn hơn để xem xét cho mỗi nút hiện có. Tuy nhiên, điều này không nhất thiết phải đúng vì nó làm giảm tính đa dạng của các cây riêng lẻ tạo nên đề xuất bán hàng độc nhất (USP) của khu rừng ngẫu nhiên.Do đó, cần phải đạt được số dư

và chọn các thuộc tính tối ưu cho phù hợp.

* Siêu tham số quan trọng cuối cùng là min\_sample\_leaf, chỉ định số lượng lá tối thiểu cần thiết để chia một nút bên trong.

– Model Boost:

* Siêu tham số n\_jobs cho động cơ biết nó có thể sử dụng bao nhiêu bộ xử lý. Nếu giá trị là một, bạn chỉ có thể sử dụng một bộ xử lý.

Giá trị "-1" có nghĩa là không giới hạn.

* Siêu tham số Random\_state làm cho đầu ra của mô hình

có thể tái tạo được. Mô hình luôn trả về cùng một kết quả với giá trị trạng thái ngẫu nhiên, sử dụng cùng một siêu tham số và cung cấp cùng một dữ liệu đào tạo.

* Cuối cùng, có oob\_score (còn được gọi là lấy mẫu oob), một phương pháp để xác thực chéo các cấu trúc ngẫu nhiên. Trong ví dụ này, khoảng một phần ba dữ liệu không được sử dụng để đào tạo mô hình và có thể được sử dụng để đánh giá hiệu suất của nó. Mẫu này được gọi là mẫu bỏ túi.Nó rất giống với việc để lại một người nhận hợp lệ. Phương pháp này chỉ đơn giản là dán nhãn khác nhau cho từng quan sát trong số quan sát được sử dụng trong cây. Và sau đó tìm số phiếu bầu tối đa trong quan sát.

### 3.3.3 Ưu và nhược điểm của thuật toán Random Forest

Ưu điểm của thuật toán Random Forest:

* Chống quá khớp (Overfitting): Random Forest có khả năng chống quá khớp tốt hơn so với các thuật toán khác như Decision Tree. Bằng cách kết hợp nhiều cây quyết định độc lập, nó giúp giảm thiểu overfitting và tăng tính tổng quát hóa của mô hình.
* Xử lý dữ liệu lớn: Random Forest có khả năng xử lý tập dữ liệu lớn mà không yêu cầu quá nhiều thời gian huấn luyện. Thuật toán này có thể được song song hóa để tận dụng tối đa tài nguyên tính toán.
* Độ tin cậy cao: Random Forest tích hợp nhiều cây quyết định độc lập và tổng hợp kết quả từ tất cả các cây. Điều này giúp cung cấp dự đoán ổn định và tin cậy hơn so với một cây quyết định đơn lẻ.
* Xử lý cả dữ liệu số và hạng mục: Random Forest có khả năng xử lý cả dữ liệu số và dữ liệu hạng mục. Nó có thể tự động xử lý biến hạng mục và không yêu cầu mã hóa thủ công.
* Dễ dàng xác định tầm quan trọng của đặc trưng: Random Forest cung cấp một đánh giá tầm quan trọng của từng đặc trưng trong quá trình huấn luyện. Điều này giúp hiểu rõ hơn về tác động của từng đặc trưng lên mô hình và cung cấp thông tin hữu ích cho việc lựa chọn đặc trưng.

Nhược điểm của thuật toán Random Forest:

* Kích thước mô hình lớn: Random Forest tạo ra nhiều cây quyết định độc lập và kết hợp chúng lại, điều này dẫn đến kích thước mô hình lớn hơn so với một cây quyết định đơn lẻ. Điều này có thể tốn kém về tài nguyên lưu trữ và tốc độ dự đoán.
* Không thể xử lý tốt với dữ liệu có giá trị bị khuyết: Random Forest không xử lý tốt với dữ liệu có giá trị bị khuyết (missing values). Một số phương pháp đơn giản như điền giá trị trung bình hoặc giá trị phổ biến có thể được sử dụng để xử lý giá trị bị khuyết trước khi huấn luyện Random Forest.
* Yêu cầu điều chỉnh siêu tham số: Random Forest có một số siêu tham số cần được tinh chỉnh phù hợp. Các siêu tham số như số lượng cây, độ sâu cây, số lượng đặc trưng được chọn mỗi lần chia, và mức độ ngẫu nhiên trong việc xây dựng các cây đều có thể ảnh hưởng đến hiệu suất của mô hình. Việc chọn sai siêu tham số có thể dẫn đến overfitting hoặc underfitting.
* Không phù hợp cho dữ liệu dạng chuỗi (sequential data): Random Forest không phù hợp cho việc xử lý dữ liệu dạng chuỗi, như dữ liệu thời gian hoặc dữ liệu có mối quan hệ thứ tự. Trong trường hợp này, các thuật toán khác như Recurrent Neural Networks (RNNs) hoặc Long Short-Term Memory (LSTM) có thể là lựa chọn tốt hơn.

## 3.4. Các chỉ số đánh giá

### 3.4.1. Accuracy

Accuracy (Độ chính xác) là một thước đo phổ biến để đánh giá hiệu suất của mô hình phân loại. Nó đo lường tỷ lệ phần trăm các dự đoán đúng trên tổng số điểm dữ liệu kiểm định.Công thức tính Accuracy như sau:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Trong đó:

* TP (True Positive) là số lượng điểm dữ liệu dương tính (positive) được dự đoán đúng.
* TN (True Negative) là số lượng điểm dữ liệu âm tính (negative) được dự đoán đúng.
* FP (False Positive) là số lượng điểm dữ liệu âm tính (negative) được dự đoán sai thành dương tính (false positive).
* FN (False Negative) là số lượng điểm dữ liệu dương tính (positive) được dự đoán sai thành âm tính (false negative).

Accuracy có giá trị từ 0 đến 1, với 1 cho biết mô hình dự đoán đúng tất cả các điểm dữ liệu và 0 cho biết mô hình không dự đoán đúng bất kỳ điểm dữ liệu nào.

Độ chính xác (Accuracy) là một thước đo đơn giản và dễ hiểu. Tuy nhiên, nó có thể không phản ánh được một cách chính xác hiệu suất của mô hình khi dữ liệu không cân bằng, tức là tỷ lệ giữa các lớp dương và âm không đồng đều. Trong trường hợp này, mô hình có thể có độ chính xác cao nhưng không thực sự tốt trong việc phân loại các điểm dữ liệu thuộc lớp thiểu số. Do đó, khi đánh giá mô hình, nên xem xét các thước đo khác như Precision, Recall, F1-score để có cái nhìn toàn diện hơn về hiệu suất của mô hình phân loại.

### 3.4.2. Precision

Precision (Độ chính xác) là một thước đo đánh giá hiệu suất của mô hình phân loại, tập trung vào tỷ lệ các dự đoán dương tính (positive) đúng so với tổng số dự đoán dương tính. Công thức tính Precision như sau:

Precision = TP / (TP + FP)

Trong đó:

* TP (True Positive) là số lượng điểm dữ liệu dương tính (positive) được dự đoán đúng.
* FP (False Positive) là số lượng điểm dữ liệu âm tính (negative) được dự đoán sai thành dương tính (false positive).

Precision có giá trị từ 0 đến 1, với 1 cho biết mô hình dự đoán đúng tất cả các điểm dữ liệu dương tính và 0 cho biết mô hình không dự đoán đúng bất kỳ điểm dữ liệu dương tính nào.

Precision đo lường khả năng của mô hình phân loại để xác định chính xác các điểm dữ liệu thuộc lớp dương tính. Nó quan tâm đến việc tránh sai sót trong việc đánh giá một điểm dữ liệu là dương tính khi nó thực sự không phải. Precision là một thước đo quan trọng trong các bài toán mà việc tránh sai sót dương tính (false positive) là quan trọng, chẳng hạn như trong phát hiện bệnh, phát hiện gian lận, hay phát hiện email rác.

### 3.4.3. Recall

Recall (còn được gọi là Sensitivity, True Positive Rate) là một thước đo đánh giá khả năng của mô hình phân loại trong việc tìm ra tất cả các điểm dữ liệu thuộc lớp dương tính (positive). Nó tập trung vào tỷ lệ các điểm dữ liệu thuộc lớp dương tính được dự đoán đúng so với tổng số điểm dữ liệu thực sự thuộc lớp dương tính. Công thức tính Recall như sau:

Recall = TP / (TP + FN)

Trong đó:

* TP (True Positive) là số lượng điểm dữ liệu dương tính (positive) được dự đoán đúng.
* FN (False Negative) là số lượng điểm dữ liệu dương tính (positive) được dự đoán sai thành âm tính (false negative).
* Recall có giá trị từ 0 đến 1, với 1 cho biết mô hình dự đoán đúng tất cả các điểm dữ liệu dương tính và không có false negative (sai sót trong việc dự đoán các điểm dữ liệu dương tính). Giá trị 0 cho biết mô hình không dự đoán đúng bất kỳ điểm dữ liệu dương tính nào.

Recall đo lường khả năng của mô hình phân loại trong việc tìm ra tất cả các điểm dữ liệu dương tính. Nó quan trọng trong các bài toán mà việc tránh sót các điểm dữ liệu dương tính (false negative) là quan trọng, chẳng hạn như trong phát hiện bệnh, phát hiện gian lận, hay phát hiện spam email

### 3.4.4. F1-Score

F1-Score là một thước đo kết hợp giữa Precision và Recall để đánh giá hiệu suất tổng thể của mô hình phân loại. Nó cung cấp một con số tổng hợp đánh giá cân bằng giữa khả năng phân loại chính xác các điểm dữ liệu dương tính và khả năng tìm ra tất cả các điểm dữ liệu dương tính.Công thức tính F1-Score như sau:

F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)

F1-Score có giá trị từ 0 đến 1, với 1 cho biết mô hình hoàn hảo trong việc phân loại cả Precision và Recall và 0 cho biết mô hình không phân loại chính xác bất kỳ điểm dữ liệu dương tính nào hoặc không tìm ra bất kỳ điểm dữ liệu dương tính nào.

F1-Score được sử dụng khi cần đánh giá một mô hình phân loại một cách tổng thể, không tập trung quá nhiều vào Precision hoặc Recall mà tạo ra sự cân bằng giữa cả hai. Nó thường được sử dụng trong các bài toán mà việc cân nhắc cả Precision và Recall là quan trọng, ví dụ như phân loại email rác, phát hiện bệnh, hay phân loại hình ảnh. F1-Score là một thước đo hữu ích để so sánh hiệu suất giữa các mô hình phân loại khác nhau và giúp chọn ra mô hình tốt nhất dựa trên cân nhắc giữa Precision và Recall.

### 3.4.5. Kappa Score

Kappa Score, còn được gọi là Cohen's Kappa, là một thước đo đánh giá mức đồng nhất (agreement) giữa các dự đoán của mô hình phân loại và nhãn thực tế. Nó đánh giá khả năng của mô hình phân loại trong việc phân loại đúng dựa trên sự đồng nhất giữa dự đoán và nhãn thực tế, và điều chỉnh cho mức đồng nhất ngẫu nhiên.Công thức tính Kappa Score như sau:

Kappa = (Accuracy - Expected\_Accuracy) / (1 - Expected\_Accuracy)

Trong đó:

* Accuracy là tỷ lệ các điểm dữ liệu được phân loại đúng, được tính bằng số lượng dự đoán đúng chia cho tổng số điểm dữ liệu.
* Expected\_Accuracy là tỷ lệ đồng nhất ngẫu nhiên dựa trên phân phối ngẫu nhiên của các nhãn dự đoán và nhãn thực tế. Nó được tính bằng cách tích chập của tỷ lệ của các nhãn dự đoán và tỷ lệ của các nhãn thực tế.

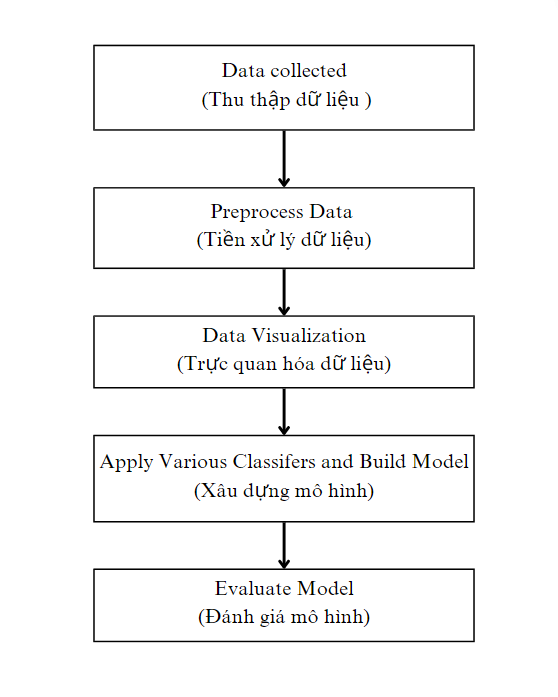
Kappa Score có giá trị từ -1 đến 1, với 1 cho biết mô hình phân loại hoàn toàn đồng nhất với nhãn thực tế, 0 cho biết mô hình không tốt hơn một phân loại ngẫu nhiên, và giá trị âm cho biết mô hình phân loại tệ hơn ngẫu nhiên.

Kappa Score hữu ích khi đánh giá mô hình phân loại trong các tình huống mà mức đồng nhất ngẫu nhiên giữa dự đoán và nhãn thực tế có thể xảy ra. Nó được sử dụng trong các bài toán phân loại đa nhãn hoặc khi mất cân bằng giữa các nhãn.

Kappa Score giúp đánh giá hiệu suất của mô hình phân loại một cách toàn diện hơn so với chỉ sử dụng Accuracy, đồng thời điều chỉnh cho mức đồng nhất ngẫu nhiên giữa các nhãn dự đoán và nhãn thực tế.

# 4. Phương pháp luận nghiên cứu

## 4.1 Mô hình tổng quan



## 4.2 Thu thập dữ liệu

Bộ dữ liệu được tìm kiếm trên Kaggle bao gồm các thông tin liên quan đến chiến dịch tiếp thị của một tổ chức ngân hàng Bồ Đào Nha về khoản gửi có kỳ hạn nhằm dự đoán kết quả của chiến dịch tiếp thị. Chiến dịch thành công nghĩa là khách sạn đồng ý đăng ký khoản gửi có kỳ hạn, ngược lại chiến dịch này thất bại nghĩa là khách hàng không đăng ký khoản gửi có kỳ hạn.

Nguồn của bộ dữ liệu được tìm thấy tại *“[Moro và cộng sự, 2014] S. Moro, P. Cortez và P. Rita. Phương pháp tiếp cận dựa trên dữ liệu để dự đoán sự thành công của tiếp thị qua điện thoại của ngân hàng. Hệ thống hỗ trợ quyết định, Elsevier, 62:22-31, tháng 6 năm 2014”*

## 4.3 Mô tả dữ liệu

Bộ dữ liệu bao gồm 41.188 quan sát với 21 biến, trong đó 20 biến đầu vào liên quan đến chiến dịch tiếp thị trực tiếp (gọi điện thoại) của một tổ chức ngân hàng Bồ Đào Nha nhằm để dự đoán biến mục tiêu là khách hàng có đăng ký một khoản tiền gửi có kỳ hạn hay không (biến y). Nếu khách hàng đăng ký khoản gửi này nghĩa là chiến dịch tiếp thị của ngân hàng có kết quả thành công và ngược lại. Trong tập dữ liệu này sẽ bao gồm các biến phân loại và định lượng được mô tả chi tiết trong bảng dưới đây.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên biến (Attributes ) | Kiểu dữ liệu  (Type) | Mô tả chi tiết (Description) |
| 11 | Age | Định lượng | Độ tuổi của khách hàng |
| 22 | Job | Phân loại | Nghề nghiệp của khách hàng được phân loại theo các nhóm chính sau: *“admin”, “blue-collar”, “entrepreneur”, “housemaid”, “management”, “retired”, “self-employed”, “services”, “student”, “technician”, “unemployed”, “unknown”.* |
| 3  43 | Marital | Phân loại | Tình trạng hôn nhân bao gồm: *“divorced”* (ly hôn), *“married”* ( đã kết hôn), *“single”* (độc thân) và *“unknown”*( chưa xác định được) . |
| 44 | Education | Phân loại | Trình độ được phân loại theo các nhóm sau: *“basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”* và *“unknown”* |
| 55 | Default | Phân loại | Tình trạng sử dụng tín dụng, được phân loại: *“yes”* đang sử dụng tín dụng, *“no”* không sử dụng và *“unknow”* là không xác định. |
| 66 | Housing | Phân loại | Có khoản vay để mua hoặc thuê nhà hay không, được phân loại: *“yes”* là có khoản vay mua nhà, *“no”* là không và *“unknown”* chưa xác định được. |
| 77 | Loan | Phân loại | Khoản vay nợ cá nhân, được phân loại: *“yes”* đang có khoản vay nợ, *“no”* là không và *“unknown”* là chưa xác định được. |
| 88 | Contact | Phân loại | Phương thức liên hệ, được phân loại làm 2 nhóm *“cellular”* (di động) và *“telephone”* ( điện thoại dây). |
| 99 | Month | Phân loại | Tháng liên hệ cuối cùng với khách hàng trong năm, được tính bắt đầu từ tháng 3 đến tháng 12: *“mar”, “apr”, “may”, “jun”, “jul”, “aug”, “sep”, “oct”, “nov”, “dec”.* |
| 110 | Day of week | Phân loại | Ngày liên hệ cuối cùng với khách hàng trong tuần, được phân loại: *“mon”, “tue”, “wed”, “thu”, “fri”.* |
| 11 | Duration | Định lượng | Thời lượng của lần liên lạc cuối cùng với khách hàng được tính bằng giây. |
| 112 | Campaign | Định lượng | Số lần liên hệ cho 1 khách hàng được thực hiện trong chiến dịch hiện tại. |
| 113 | Pdays | Định lượng | Số ngày đã trôi qua kể từ lần liên hệ cuối cùng với khách hàng trong chiến dịch trước. (Số 999 trong dữ liệu có nghĩa là khách hàng chưa từng được liên hệ trước đó). |
| 114 | Previous | Định lượng | Số lần liên hệ cho khách hàng trong chiến dịch trước đó. |
| 115 | Poutcome | Phân loại | Kết quả của chiến dịch tiếp thị trước đó, được phân loại: *“failure”* (thất bại), *“success”* (thành công), *“nonexistent”* (không tồn tại-nghĩa là người này không phải là khách hàng trong chiến dịch lần trước). |
| 116 | Emp.var.rate | Định lượng | Tỷ lệ thay đổi việc làm - chỉ số hàng quý. |
| 117 | Cons.price.idx | Định lượng | Chỉ số giá tiêu dùng - chỉ số hàng tháng. |
| 118 | Cons.conf.idx | Định lượng | Chỉ số niềm tin của người tiêu dùng - chỉ số hàng tháng. |
| 119 | Euribor3m | Định lượng | Tỷ lệ lãi suất trung bình mà các ngân hàng có thể cho vay trong khoảng thời gian 3 tháng - chỉ số hàng ngày. |
| 220 | Nr.employed | Định lượng | số lượng nhân viên - chỉ số hàng quý (số) |
| 221 | y | Phân loại | Biến mục tiêu (biến đầu ra), khách hàng có đăng ký khoản gửi có kỳ hạn không? Được phân loại: *“yes”* là có đăng ký (chiến dịch tiếp thị của ngân hàng thành công), *“no”* là không đăng ký (chiến dịch tiếp thị của ngân hàng thất bại). |

## 4.4 Tiền xử lý dữ liệu và trích xuất đặc trưng

### 4.4.1 Tiền xử lý dữ liệu

Để bộ dữ liệu trở nên phù hợp và đảm bảo được hiệu suất tối ưu của các thuật toán nhóm đã tiến hành tiền xử lý dữ liệu thông qua các bước: xử lý dữ liệu bị thiếu, xóa dữ liệu nhiễu, xử lý giá trị ngoại lai, chọn đặc trưng và mã hóa dữ liệu.

**Xử lý dữ liệu bị thiếu**

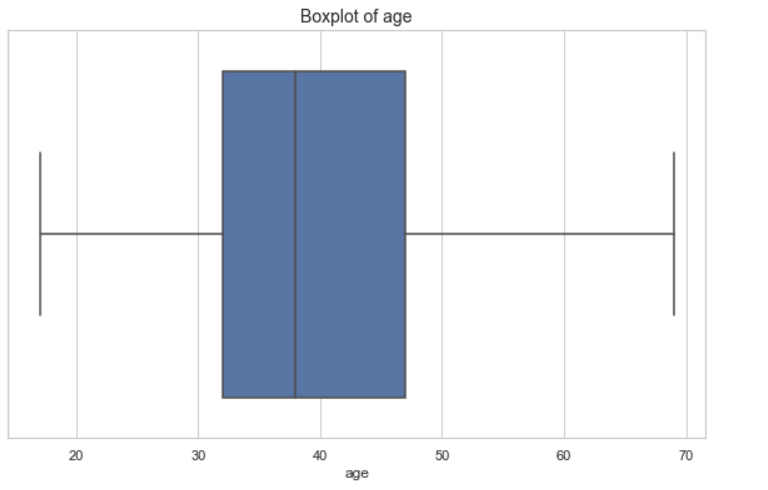
Trong tập dữ liệu sau khi kiểm tra không có biến nào bị thiếu giá trị, tuy nhiên có nhiều biến tồn tại giá trị unknown điều này làm cho mô hình gặp khó khăn trong việc tạo ra dự đoán chính xác cho các quan sát chứa giá trị này. Điều này có thể làm giảm độ chính xác tổng thể của mô hình.

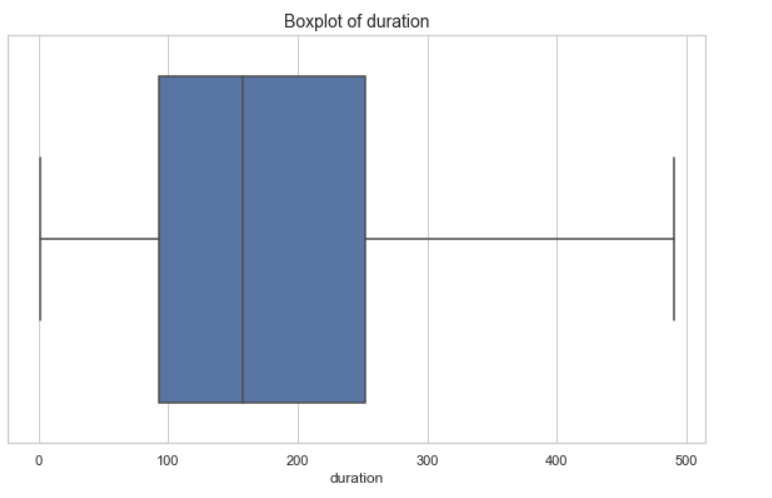
Sau quá trình nghiên cứu nhóm đã quyết định thay thế tất cả giá trị unknown có trong các thuộc tính bằng mode (giá trị mode) là giá trị có tần suất xuất hiện cao nhất để đảm bảo tính toàn vẹn của dữ liệu và độ chính xác của kết quả.

Sau đó tiếp tục tiến hành các bước tiền xử lý dữ liệu như xóa dữ liệu nhiễu: xóa các dòng bị trùng lặp, trong bộ dữ liệu có thuộc tính *“default”* có sự mất cân bằng lớn khi số lượng “no” là 41185 còn “yes” chỉ có 3. Với thuộc tính “education” có một số giá trị tương tự như basic.9y, basic.6y và basic.4y nên gộp lại thành 1 giá trị chung là mid.school. Cuối cùng là thuộc tính “pdays” số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ chiến dịch trước đó với giá trị 999 có nghĩa là khách hàng chưa được liên hệ trước đó nên được thay thế thành 0 để tránh trường hợp giá trị ngoại lai tác động mạnh vào độ chính xác của mô hình.

**Khám phá dữ liệu bằng cách sử dụng phân tích trực quan và xử lý ngoại lai**

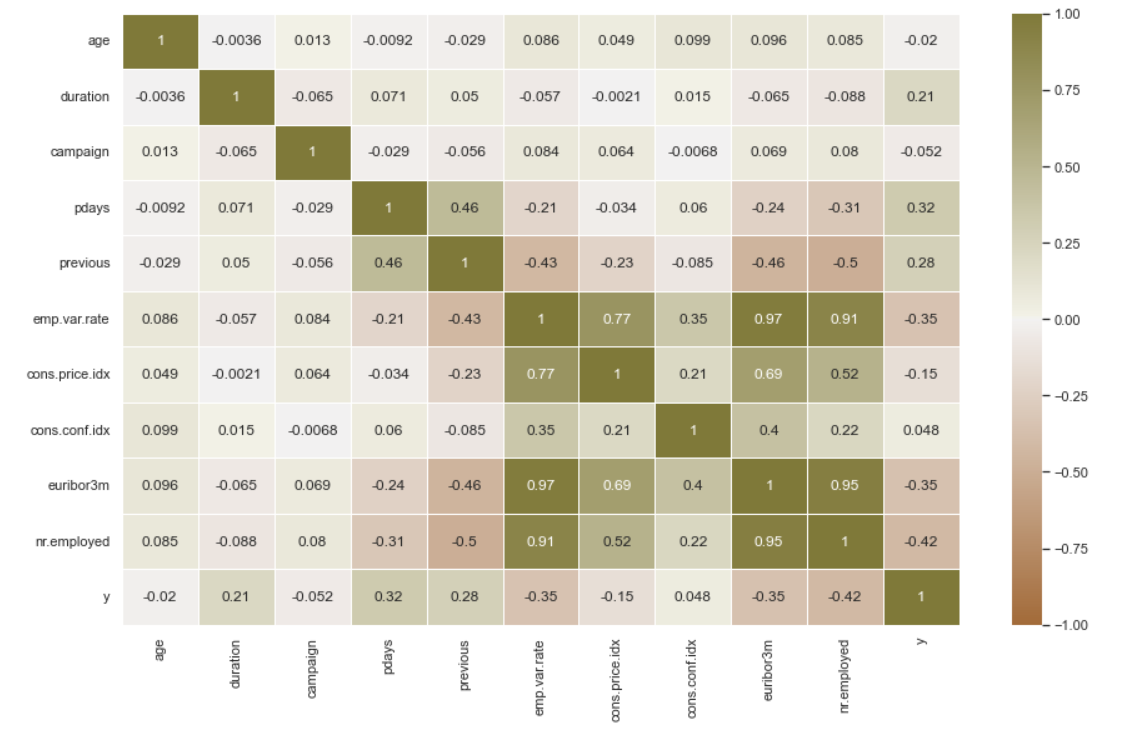
Khi xem xét thống kê mô tả của bộ dữ liệu, nhóm phát hiện tồn tại rất nhiều biến có giá trị ngoại lai cần được xử lý. Để cụ thể hơn tiến hành khám phá dữ liệu tìm hiểu sự phân bố và tính chất của các biến nhóm đã tiến hành vẽ biểu đồ hộp (box plot) để phát hiện các giá trị ngoại lai dựa trên sự khác biệt rõ rệt so với phần còn lại của dữ liệu. Sau đó tiến hành xử lý ngoại lai bằng tính toán z-score cho mỗi quan sát: Z-score được tính toán bằng cách lấy giá trị của mỗi quan sát trừ đi giá trị trung bình của biến và chia cho độ lệch chuẩn của biến. Xác định ngưỡng cho z-score: một giá trị z-score tuyệt đối lớn hơn 3 được coi là ngoại lai và tiến hành loại bỏ những ngoại lai vượt ngưỡng giá trị xác định. Nhóm còn xử lý ngoại lai bằng phương pháp tìm IQR (Interquartile Range): Tính toán IQR cho biến: IQR được tính bằng cách lấy phân vị thứ 75 (Q3) trừ đi phân vị thứ 25 (Q1). Xác định ngưỡng cho IQR: Ngưỡng được chọn dựa trên IQR là một ngưỡng cố định bằng 1.5. Ngưỡng này sẽ được sử dụng để xác định giá trị ngoại lai.Xác định các quan sát có giá trị nằm ngoài phạm vi Q1 - ngưỡng \* IQR và Q3 + ngưỡng \* IQR. Các quan sát vượt quá phạm vi này được xem là giá trị ngoại lai và được xử lý bằng cách loại bỏ.





*Biểu đồ Boxplot của các biến sau khi đã xử lý ngoại lai*

Thêm vào đó nhóm đã tiến hành vẽ biểu đồ thể hiện ma trận tương quan giữa các biến để hiển thị mối quan hệ tương quan giữa các cặp biến trong tập dữ liệu phát hiện sự tồn tại của tương quan cao tiến hành xử lý để giảm đa cộng tuyến trong mô hình. Cụ thể giữa thuộc tính “pdays” và previous có mối tương quan dương lớn nên được xử lý bằng cách biến đổi dữ liệu trong cột "previous" thành một biến nhị phân, trong đó giá trị 1 đại diện cho các mục tiêu trước đó có giá trị lớn hơn 0 và giá trị 0 đại diện cho các mục tiêu trước đó bằng 0 hoặc âm.



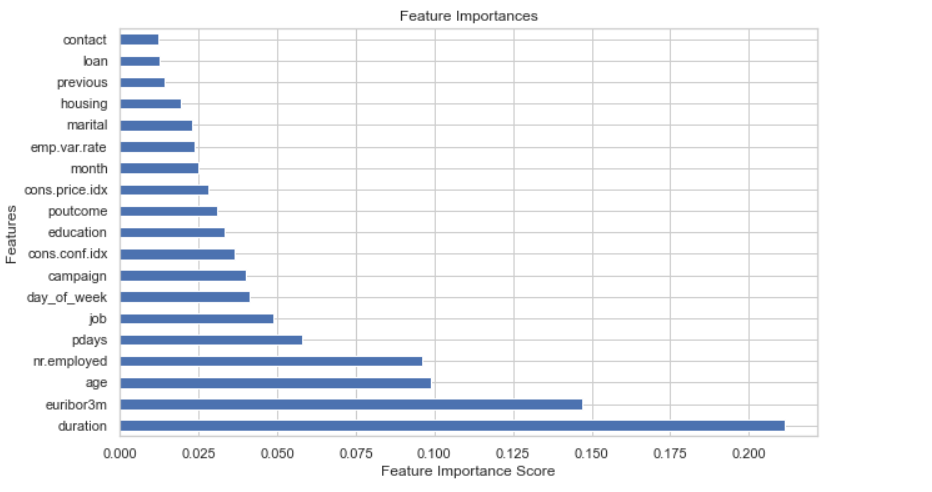
*Biểu đồ thể hiện ma trận tương quan giữa các biến*

**Mã hóa dữ liệu**

Do hầu hết các mô hình học máy yêu cầu dữ liệu đầu vào là dạng số để mô hình có thể xử lý và học từ dữ liệu nhóm đã tiến hành chuyển đổi các biến không số hoặc biến hạng mục thành dạng số để mô hình có thể xử lý. Bằng phương pháp mã hóa nhãn ( (Label encoding) mã hóa các biến thành các số nguyên duy nhất. Và mã hóa one-hot (One-hot encoding) mã hóa các biến mục tiêu thành các biến giả, trong đó mỗi biến giả đại diện cho một giá trị duy nhất của biến ban đầu và chỉ có thể có giá trị 0 hoặc 1.

**Chọn đặc trưng**

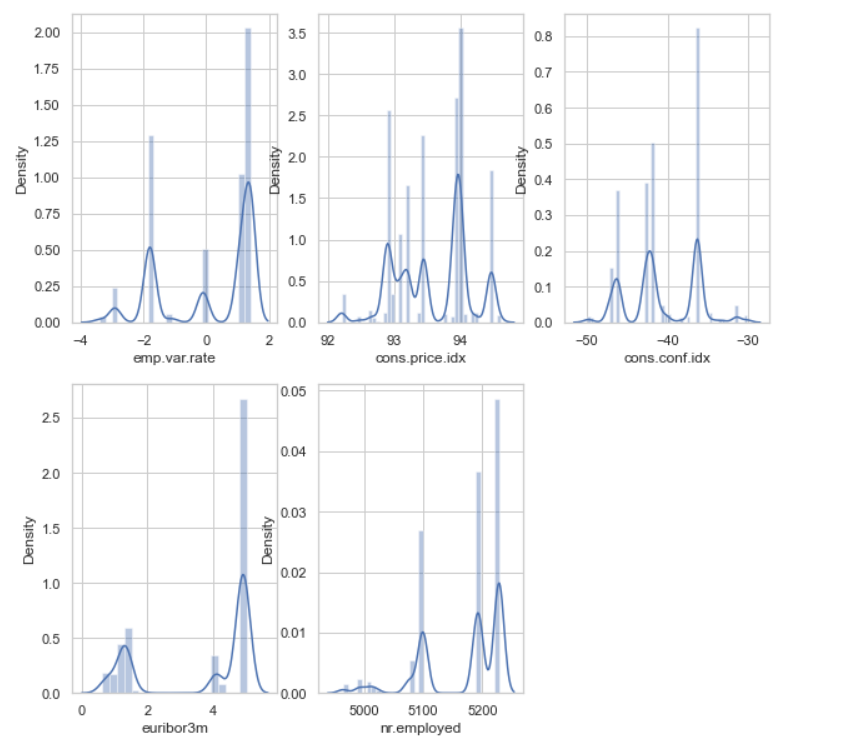
Để tăng hiệu suất cho mô hình dự đoán, chúng tôi đã tiến hành chọn các biến quan trọng và có ý nghĩa nhất từ bộ dữ liệu ban đầu bằng phương pháp sử dụng mô hình dự đoán Random Forest để ước lượng độ quan trọng của các biến.



*Biểu đồ thể hiện mức độ quan trọng của các biến*

Khi quan sát biểu đồ chúng tôi nhận ra các biến *“contact”, “loan”, “previous” và “housing”* chiếm tỷ lệ nhỏ nên có mức độ quan trọng thấp sẽ được loại bỏ ra khỏi bộ dữ liệu để tránh làm giảm hiệu suất của mô hình.

**Trực quan hóa dữ liệu và phân tích ban đầu**



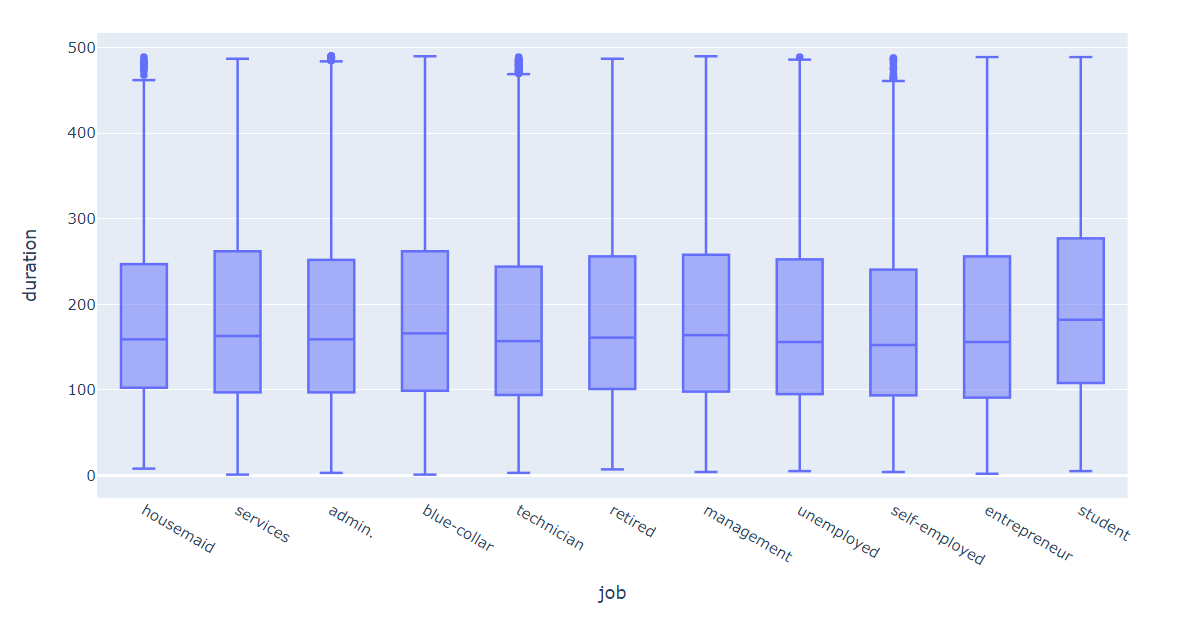
Ta thấy tỷ lệ thay đổi việc làm của xã hội cao, nghĩa là ngân hàng đã thực hiện chiến dịch khi có biến động kinh tế.

Chỉ số giá tiêu dùng cũng tốt cho thấy người dân đang có một mức giá hợp lý để thanh toán các khoản chi tiêu hằng ngày. Vì vậy, đây sẽ là thời điểm tốt để ngân hàng thực hiện các chiến dịch, bởi lẽ, mọi người đang có tiền dư và dễ nảy sinh ý tưởng gửi tiền tiết kiệm.

Chỉ số niềm tin thấp cho thấy người dân đang không có niềm tin vào nền kinh tế đầy biến động. Vì vậy, lối thoát duy nhất sẽ là gửi tiền ngân hàng với lãi suất thấp nhưng an toàn. Ngân hàng cũng nên thực hiện chiến dịch quảng cáo tại những thời điểm này.

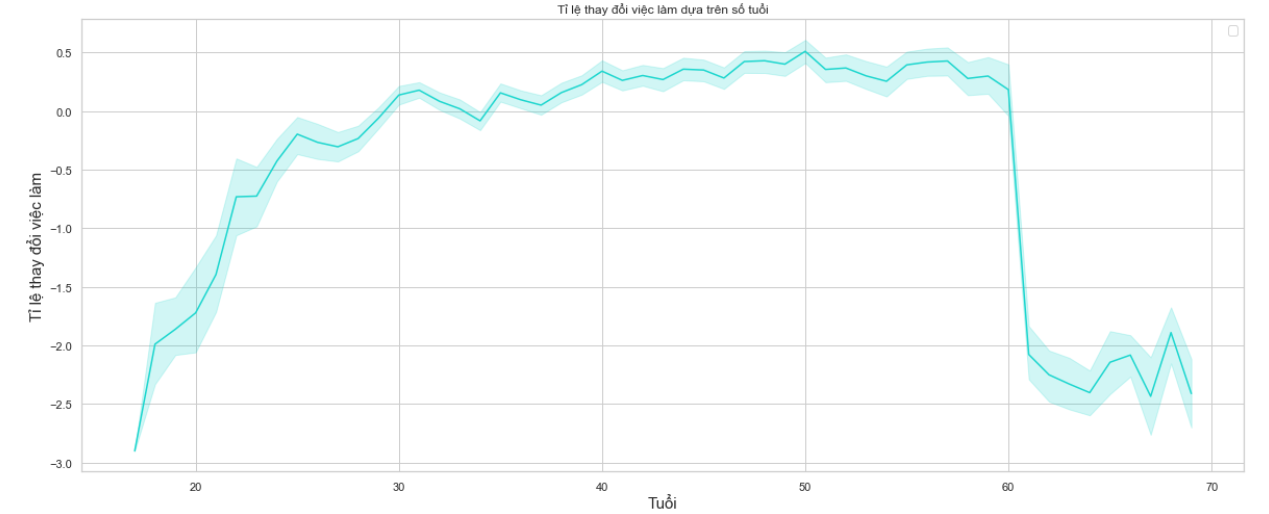
Lãi suất Euribor ba tháng là lãi suất mà một số ngân hàng Châu Âu cho vay lẫn nhau trong thời hạn ba tháng. Nhìn vào biểu đồ ta thấy lãi suất khá cao, vì vậy lượng tiền của hệ thống ngân hàng đang bị cô lập và lưu thông kém.

Số lượng nhân viên ngân hàng đang ở mức cao nhất sẽ tạo ra một nguồn thu nhập cao cho những nhân viên đó. Vì vậy, nhóm khuyến nghị các chiến dịch marketing nhắm vào các nhân viên này.



*Biểu đồ thể hiện thời lượng cuộc gọi của chiến dịch marketing ngân hàng đã thực hiện dựa trên nghề nghiệp tương ứng*

Nhìn chung ở chiến dịch marketing trước ngân hàng tập trung nâng cao thời lượng cuộc gọi ở các nhóm đối tượng là sinh viên và người nghỉ hưu và quản lý bởi vì có cơ sở để xác định đây là nhóm khách hàng mục tiêu của các khoản gửi có kỳ hạn, khi mà sinh viên nhập học thì cần các khoản gửi để chi tiêu cho sinh hoạt và học tập nên gia đình có xu hướng mở các khoản gửi có kỳ hạn cho con cái. Với đối tượng là người nghỉ hưu sau khi nhận bảo hiểm xã hội hay chế độ phúc lợi xã hội khi nghỉ hưu họ mở khoản gửi tiết kiệm để có thể nhận tiền lãi mỗi tháng như là một khoản tiền lương phù hợp cho tiêu dùng hàng tháng. Với nhóm quản lý thì có thể mức lương cao có nhiều dư giả nên họ quyết định mở khoản gửi tiết kiệm để có thể đáp ứng với những kế hoạch dài hạn sau này.



*Biểu đồ thể hiện tỷ lệ thay đổi việc làm dựa trên số tuổi*

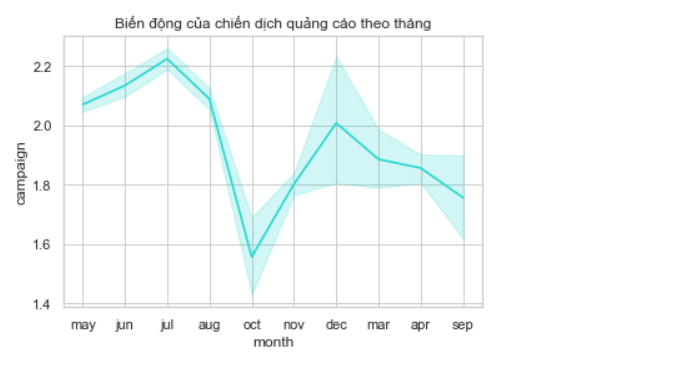
Nhìn vào biểu đồ có thể thấy tỉ lệ biến động việc làm xảy ra đột ngột ở lứa tuổi từ 55 đến 60 tuổi. Vì đây là độ tuổi nghỉ hưu khi nhận được trợ cấp xã hội người ta có xu hướng mở khoản gửi ngân hàng cần phải chú ý đẩy mạnh chiến dịch marketing vào nhóm đối tượng này.

Ngoài ra tỉ lệ người có việc làm tăng dần mạnh từ 20 đến 30 tuổi, cần phải có những chính sách khuyến khích đẩy mạnh marketing vào nhóm đối tượng này vì họ có khả năng trở thành nhóm khách hàng tiềm năng của các khoản gửi có kỳ hạn.



*Biểu đồ thể hiện số lượng người quyết định đăng ký khoản gửi*

Nhìn chung, số lượng người không đăng ký chiếm phần lớn, gấp gần 15 lần so với người đăng ký khoản gửi có kỳ hạn. Như vậy, những quảng cáo của ngân hàng đang không hiệu quả, cần được đổi mới cách tiếp cận để thu hút thêm nhiều đối tượng hơn.



*Biểu đồ thể hiện biến động các chiến dịch quảng cáo theo tháng*

Có thể thấy chiến dịch chủ yếu tập trung vào đầu kỳ ngân hàng (tháng 5, 6, 7). Thông thường giai đoạn giáo dục bắt đầu trong thời gian đó nên có khả năng cha mẹ mở khoản gửi có kỳ hạn dưới tên của con cái họ.

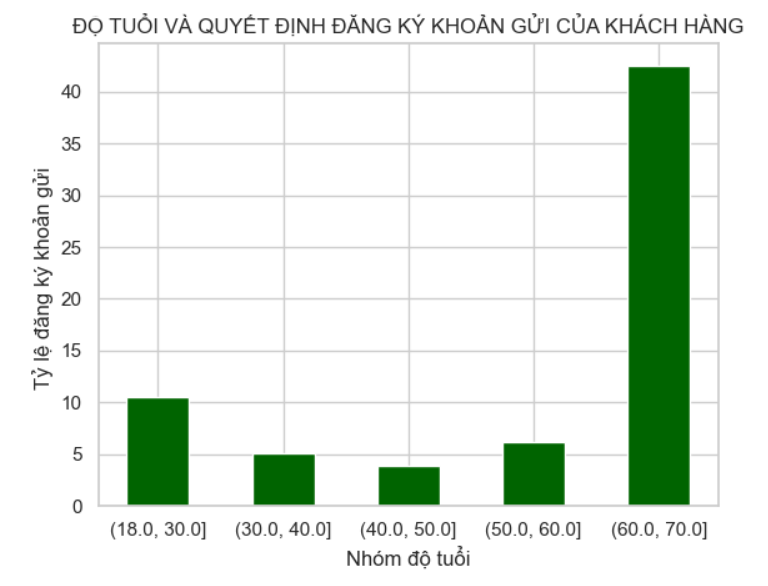
Nhìn vào biểu đồ, ta cũng thấy được ngân hàng cũng tăng mạnh các chiến dịch của mình vào cuối năm (tháng 11 và tháng 12).

# 5. Kết quả thử nghiệm và phân tích

## 5.1 Đưa ra các kết quả phân tích từ biểu đồ các biến độc lập với tỷ lệ chuyển đổi

Sau khi tiến hành trực quan hóa dữ liệu nhằm biểu diễn các dữ liệu dưới dạng biểu đồ, đồ thị,... chúng tôi tiến hành phân tích, tìm ra mối liên quan giữa các biến cũng như xu hướng biến động mà không thể dễ dàng nhận thấy trong biểu đồ gốc. Thông qua đó đưa ra những nhận xét phù hợp với nền tảng lý thuyết và cả thực tế, nhằm áp dụng vào chiến dịch tiếp thị trong tương lai của ngân hàng nói riêng cũng như lĩnh vực tài chính-ngân hàng nói chung.

### 5.1.1. Độ tuổi và quyết định đăng ký khoản gửi của khách hàng.

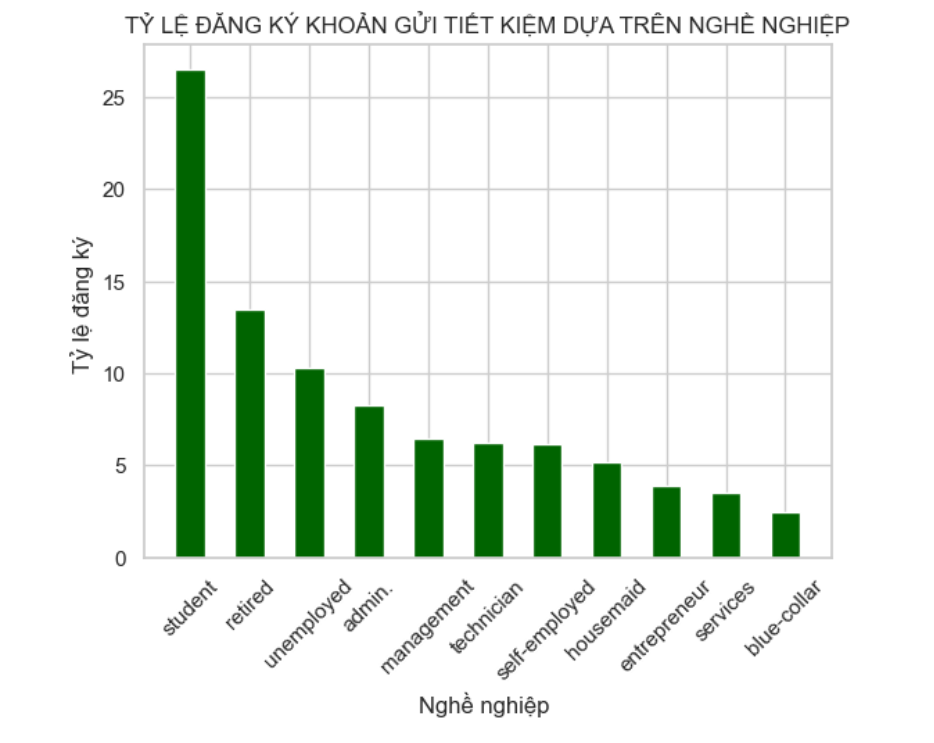


*(Độ tuổi và quyết định đăng ký khoản gửi của khách hàng.)*

Kết quả chiến dịch tiếp thị của ngân hàng cho thấy tỷ lệ đăng ký khoản gửi tiết kiệm của những người trên 60 tuổi đã đạt tới hơn 40% trong số những người cùng nhóm tuổi được tiếp thị. Điều này cho thấy rằng nhóm người trên 60 tuổi phản ứng tốt với chiến dịch tiếp thị của ngân hàng hơn so với các nhóm tuổi khác. Riêng nhóm khách hàng có độ tuổi 30-40 tuổi hay 40-50 tuổi tỷ lệ đăng ký khoản gửi tiết kiệm thấp dưới 5% . Lý do vì họ đa số là những người đã lập gia đình, phải trang trải các chi phí nuôi dạy chăm sóc con cái thậm chí có những người đang chịu các khoản vay nợ hoặc thuê nhà ở. Có thể nói độ tuổi là một trong số các yếu tố đáng cân nhắc trong chiến dịch tiếp thị để ngân hàng có thể đưa ra các kế hoạch tiếp thị phù hợp nhất.

### 

### 5.1.2. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên nghề nghiệp

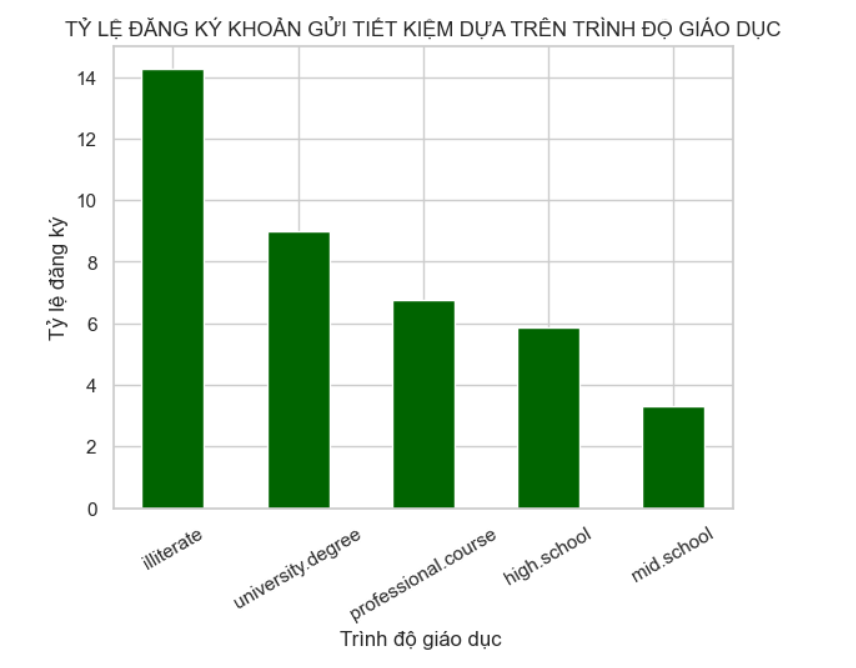


*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên nghề nghiệp)*

Có hai nhóm chính là "student" (sinh viên) và "retired" (người nghỉ hưu) trong phân loại nghề nghiệp của khách hàng có tỷ lệ đăng ký khoản gửi tiết kiệm ngân hàng cao nhất. Riêng với sinh viên có thể họ được phụ huynh mở tài khoản tiết kiệm trước khi bước vào trường đại học, giúp họ tích lũy tiền để đảm bảo chi trả học phí và chi phí sinh hoạt hàng ngày. Hay có những người xem đó là khoản tiết kiệm để làm vốn cho việc kinh doanh sau khi tốt nghiệp. Nhóm thứ 2 có tỷ lệ khoản gửi khá cao là người nghỉ hưu, thông thương sau khi nghỉ hưu thường có một khoản tiền tích lũy, việc gửi tiết kiệm trong ngân hàng giúp họ duy trì và tăng lên số tiền này, tạo ra nguồn thu nhập bổ sung để hỗ trợ cho cuộc sống hưu trí. Qua đây, ngân hàng nên có các chính sách ưu đãi đặc biệt cho 2 đối tượng này như lãi suất ưu đãi hoặc các dịch vụ và sản phẩm tài chính phù hợp với nhu cầu của họ.

### 5.1.3. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên trình độ giáo dục.

Biểu đồ cho thấy khi trình độ giáo dục càng cao thì tỷ lệ đăng ký khoản gửi tiết kiệm ngân hàng càng cao, đứng thứ nhất là University.degree (Người tốt nghiệp Đại học), tiếp theo là Professional.course (Người tốt nghiệp các khóa học chuyên nghiệp), kế tiếp là High.school (Trung học phổ thông) và Mid. school (Trung học cơ sở). Nhìn chung,trình độ học vấn cao thường đi đôi với mục tiêu tài chính rõ ràng và khả năng lập kế hoạch tài chính. Những người này thường có ý thức về việc tích lũy tài sản, tiết kiệm cho tương lai và tạo ra nguồn thu nhập bền vững. Với sự hiểu biết về tài chính cùng mức thu nhập ổn định nên đó là lý do họ có tỷ lệ đăng ký khoản gửi tiết kiệm có kỳ hạn cao hơn so với nhóm còn lại. (Riêng với nhóm đối tượng “illiterate” chúng ta sẽ không xét đến, lý do là vì trong bộ dữ liệu chỉ có 18 quan sát thuộc vào đối tượng này, nên tỷ lệ này cao hoàn toàn không có ý nghĩa)



*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên trình độ giáo dục)*

### 5.1.4. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tình trạng hôn nhân.

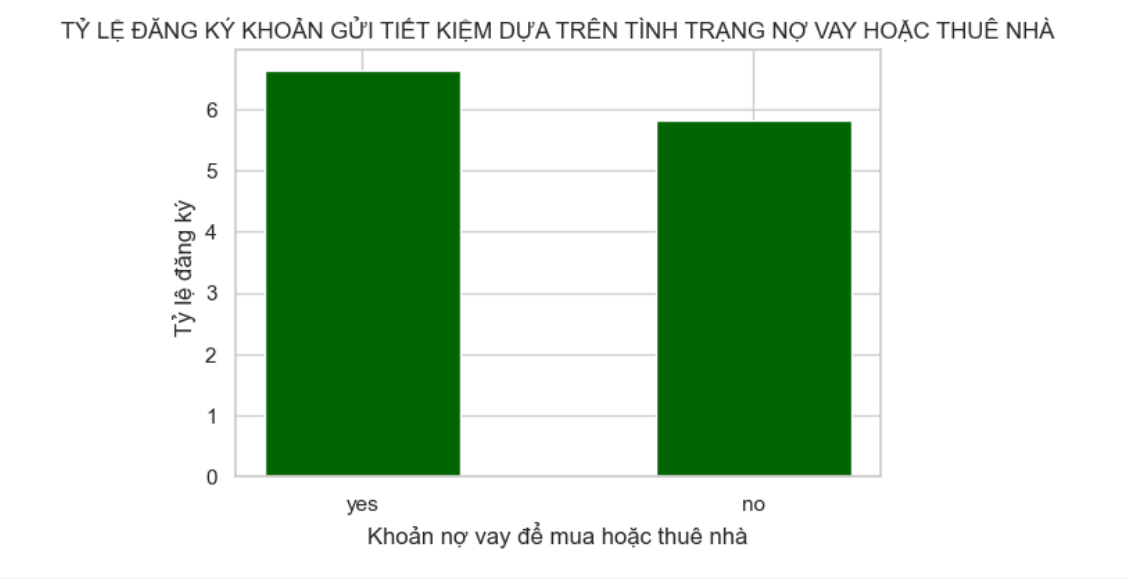
Dựa vào biểu đồ có thể thấy tỷ lệ đăng ký khoản gửi tiết kiệm ở nhóm đối tượng độc thân cao so với những người đã kết hôn và ly hôn. Nguyên nhân chính là vì người độc thân tự do tài chính, có quyền tự quyết đối với việc quản lý tài sản và quyết định đầu tư. Họ có thể dễ dàng đưa ra các quyết định tài chính mà không cần phải tham khảo ý kiến hoặc đồng thuận của người khác. Đồng thời, người độc thân thường có mục tiêu tài chính riêng và quan tâm đến việc tích lũy tài sản cho tương lai cá nhân. Họ có thể tập trung vào việc tiết kiệm và đầu tư vào những mục tiêu cá nhân như mua nhà, du lịch hoặc hỗ trợ giáo dục. Riêng với 2 nhóm còn lại là người đã kết hôn hoặc đang trong quá trình ly hôn thường phải đối mặt với trách nhiệm gia đình và chi phí hàng ngày lớn hơn. Họ có thể phải chia sẻ tài chính và chịu trách nhiệm đối với chi phí gia đình, trẻ em, hoặc các khoản nợ sau ly hôn. Điều này có thể giới hạn khả năng tiết kiệm và đầu tư của họ.

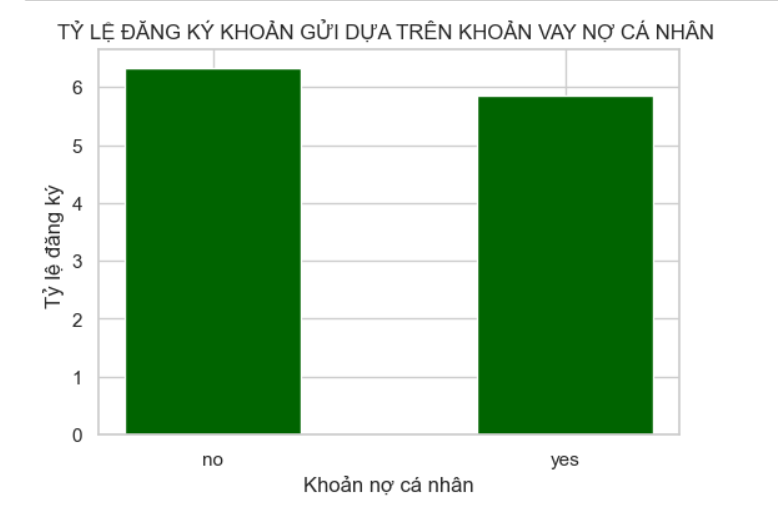


*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tình trạng hôn nhân. )*

### 5.1.5. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tình trạng nợ vay mua hoặc thuê nhà và nợ vay cá nhân.

Nhìn chung việc một người đang có khoản nợ vay để mua hoặc thuê nhà hay các khoản nợ vay cá nhân cũng không ảnh hưởng quá nhiều đến quyết định đăng ký gửi tiền tiết kiệm có kỳ hạn, Vì biểu đồ cho thấy không có sự chênh lệch quá lớn về tỷ lệ giữa các nhóm đối tượng này, cho nên ngân hàng không cần quan tâm quá nhiều vào các yếu tố này trong quá trình thực hiện kế hoạch tiếp thị.





### 5.1.6. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua phương thức liên hệ.

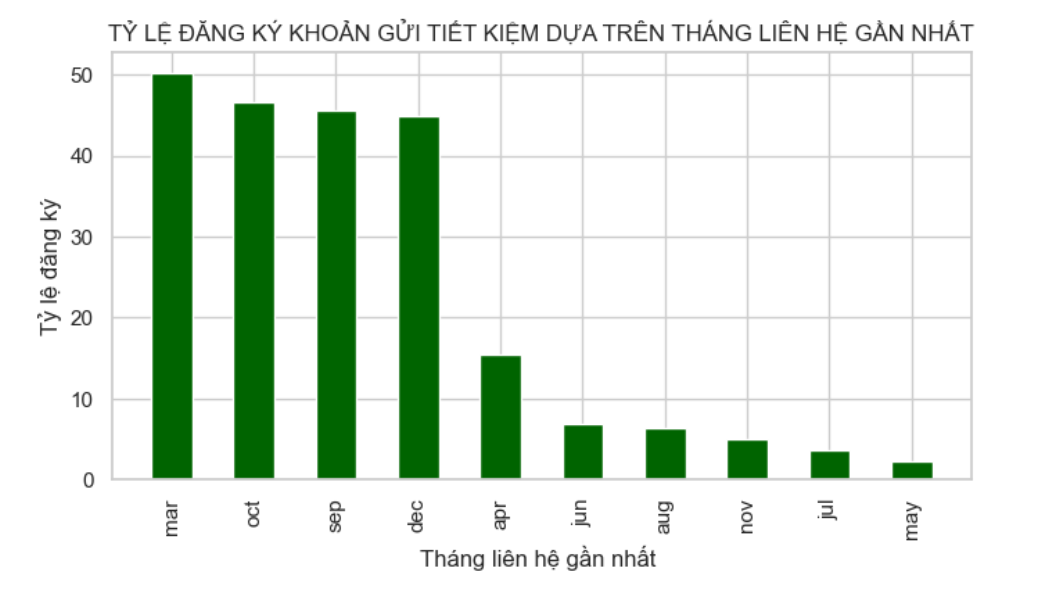
Sự chênh lệch về kết quả chiến dịch marketing thông qua các phương thức liên hệ có thể phụ thuộc vào nhiều yếu tố khác nhau. Tuy nhiên, có thể nhận thấy rằng liên hệ với khách hàng thông qua điện thoại di động thường mang lại hiệu quả tốt hơn so với việc sử dụng điện thoại bàn. Do tính tiện lợi và linh hoạt, điện thoại di động thường có tỉ lệ phản hồi cao hơn so với điện thoại bàn. Khách hàng dễ dàng truy cập và phản hồi nhanh chóng đến các cuộc gọi, tin nhắn hoặc thông điệp tiếp thị của ngân hàng, giúp tạo ra tương tác tích cực và tăng tỷ lệ đăng ký.

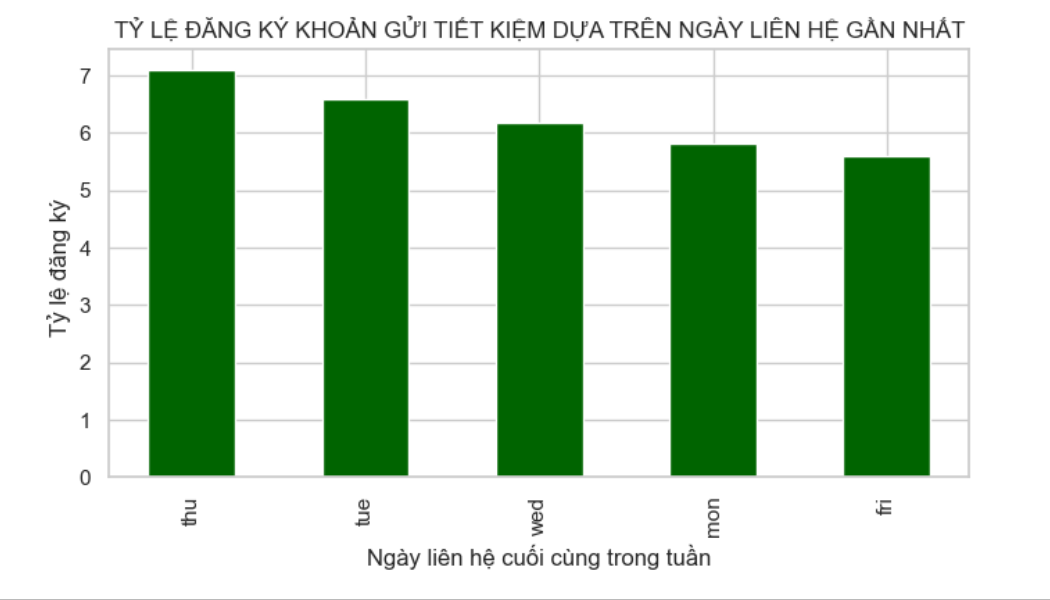


*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua phương thức liên hệ.)*

### 5.1.7. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tháng và ngày liên hệ gần nhất.

Thông quan 2 biểu đồ dưới cho thấy khi liên hệ lần cuối vào tháng Ba, tháng Mười hai, tháng Chín, tháng Mười và các thức như thứ Năm, thứ Ba, thứ Tư thường có tỷ lệ đăng ký cao hơn so với những mốc thời gian còn lại. Lý giải cho điều này có thể là vì Tháng Ba là thời điểm nhiều người tiến hành xem xét lại kế hoạch tài chính hàng năm, người ta thường xem xét và đánh giá các tài sản, đầu tư và các khoản tiết kiệm đã có để điều chỉnh kế hoạch tài chính cho năm tiếp theo. Riêng tháng 9 và 10 cha mẹ thương gửi các khoản tiền tiết kiệm cho con cái vì đây là thời điểm nhập học học của các trường đại học.

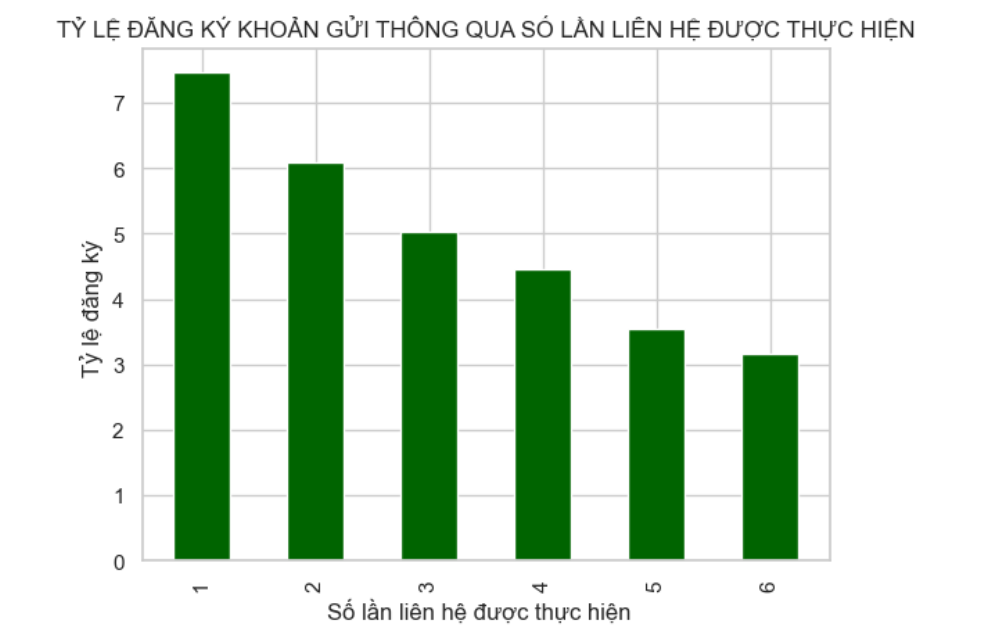




*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng dựa trên tháng và ngày liên hệ gần nhất. )*

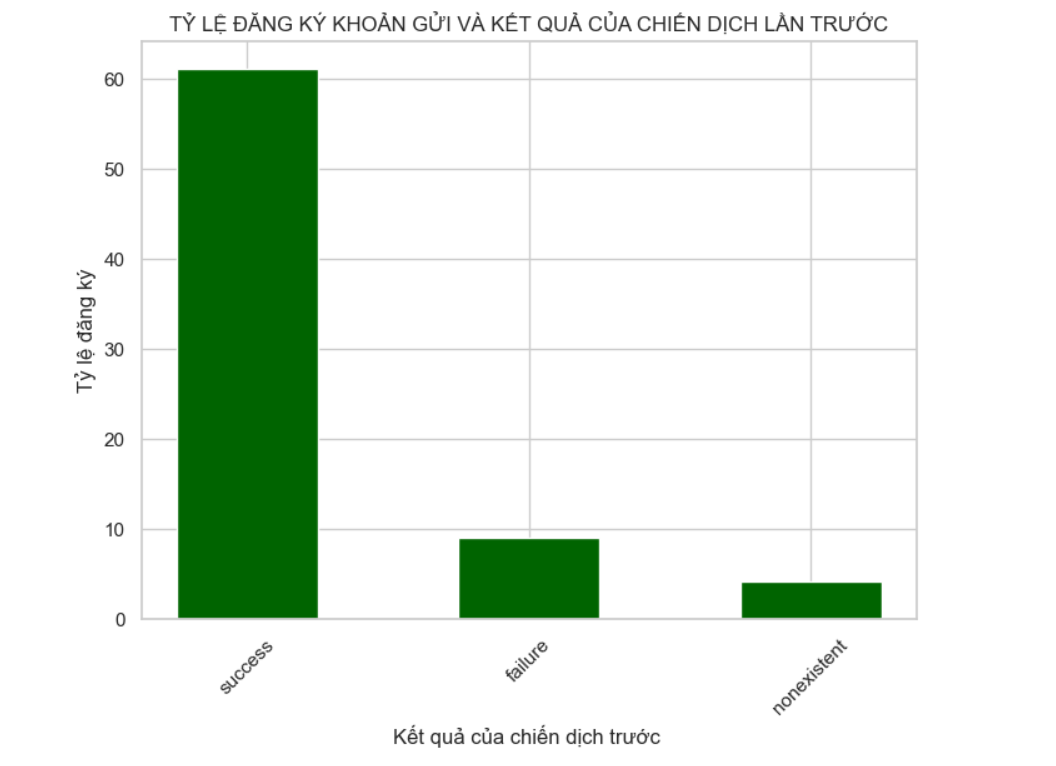
### 5.1.8. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua số lần liên hệ được thực hiện.

Thông tin trên biểu đồ được thể hiện rất rõ ràng, số lần liên hệ nhiều không đồng đồng nghĩa với việc chiến dịch tiếp thị thành công. Điều này được lý giải 1 cách đơn giản thông qua thực tế, ví dụ sau 1 lần liên hệ khách hàng đồng ý mở khoản gửi tiết kiệm thì chiến dịch đã thành công và không cần phải liên hệ thêm. Ngược lại nếu khách hàng chưa có câu trả lời, thì nhân viên ngân hàng sẽ tiếp tục thực hiện các cuộc gọi tiếp thị nhằm mục đích thuyết phục khách hàng.



*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng thông qua số lần liên hệ được thực hiện.)*

### 5.1.9. Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng và kết quả của chiến dịch lần trước.

****

*(Tỷ lệ đăng ký khoản gửi tiết kiệm của khách hàng và kết quả của chiến dịch lần trước.)*

Nhìn chung, nếu kết quả của chiến dịch trước đó thành công , thì ở chiến dịch này tỷ lệ đăng ký cũng khả quan hơn. Lý do có thể là vì những khách hàng cũ đã trở thành khách hàng trung thành, tiếp tục tin tưởng và lựa chọn đăng ký khoản gửi tiền tiết kiệm có kỳ hạn của ngân hàng.

### 5.1.10. Kết luận:

Sau khi phân tích bộ dữ liệu Chiến dịch Tiếp thị của Ngân hàng, chúng tôi rút ra những kết luận sau đây với mong muốn có thể được sử dụng để cải thiện một chiến dịch tiếp thị tương tự, triển khai các chiến dịch mới và khắc phục những hạn chế của chiến dịch trong tương lai.

Đầu tiên, dựa trên tỷ lệ đăng ký mở tài khoản tiết kiệm của các nhóm khác nhau, chúng tôi thấy rằng những người trẻ tuổi từ 20 đến 30 tuổi là sinh viên và những người độc thân, hay những người có trình độ học vấn cao cũng như những người đã nghỉ hưu trên 60 tuổi, có nhiều khả năng trở thành khách hàng so với các đối tượng còn lại. Vì vậy, ngân hàng nên tập trung vào các nhóm đối tượng này và tạo các chương trình tài chính phù hợp và thông điệp tiếp thị khác nhau trong quảng cáo cho từng nhóm.

Tiếp theo là ưu tiên liên hệ với khách hàng qua điện thoại di động khi có thể để nâng cao tỷ lệ phản hồi cũng như cải thiện tỷ lệ đăng ký của khách hàng.

Bên cạnh đó, thời gian phù hợp để thực hiện chiến dịch cũng như liên hệ với khách hàng là các ngày Thứ Năm, Thứ Ba, Thứ Tư; đặc biệt là các Tháng 3, Tháng 12, Tháng 9 và Tháng 10

Cuối cùng là chú trọng và đầu tư phát triển chương trình khách hàng thân thiết cho khách hàng hiện tại bằng cách tặng họ một số phần thưởng và ưu đãi độc đáo. Bởi vì dữ liệu cho thấy những khách hàng trung thành, những người đã từng đăng ký khoản gửi tiết kiệm trong chiến dịch cũ vẫn có tỷ lệ cao trở thành khách hàng mới trong chiến dịch tiếp thị kế tiếp.

## 5.2 Kết quả mô hình dự đoán

Chúng tôi đã thực hiện dự đoán bằng bốn mô hình bao gồm Decision Tree, LightGBM, Random Forest và XGBoost để so sánh và rút ra kết luận

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1- Score** | **Kappa** |
| --- | --- | --- | --- | --- | --- |
| *XGBoost* | 0.955278 | 0.639113 | 0.547496 | 0.589767 | 0.566267 |
| *LightGBM* | 0.953250 | 0.624473 | 0.511226 | 0.562203 | 0.537769 |
| *Random Forest* | 0.954264 | 0.646119 | 0.488774 | 0.556539 | 0.532916 |
| *Decision Tree* | 0.937836 | 0.473016 | 0.514680 | 0.492969 | 0.459920 |

Nhìn vào bảng kết quả, ta thấy mô hình XGBoost, LightGBM, Random Forest, dự đoán tốt hơn các mô hình khác.

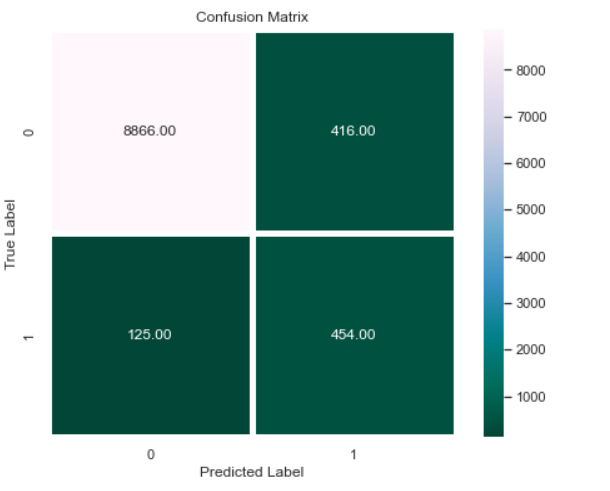
Tiếp theo, bởi vì dữ liệu bị mất cân bằng, chúng tôi sẽ dùng thuật toán oversampling để tăng số lượng mẫu lớp thiểu số để cân bằng với lớp đa số bằng phương pháp SMOTE. Phương pháp SMOTE cho phép nội suy để tạo ra các mẫu nhân tạo mới trong lớp thiểu số, dựa trên việc kết hợp các mẫu gần nhau.

Chúng tôi sẽ áp dụng ba mô hình vào bộ dữ liệu đã được cân bằng:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **LightGBM** | **XGBoost** | **Random Forest** |
| *Accuracy* | 0.945 | 0.949 | 0.95 |
| *Precision* | 0.522 | 0.552 | 0.56 |
| *Recall* | 0.784 | 0.665 | 0.674 |
| *F1-Score* | 0.627 | 0.603 | 0.611 |
| *Kappa Score* | 0.598 | 0.576 | 0.585 |

Nhìn vào bảng kết quả, ta thấy mô hình LightGBM dự đoán với F1-Score cao hơn cả nên chúng tôi lựa chọn mô hình LightGBM

Vẽ biểu đồ confusion\_matrix:

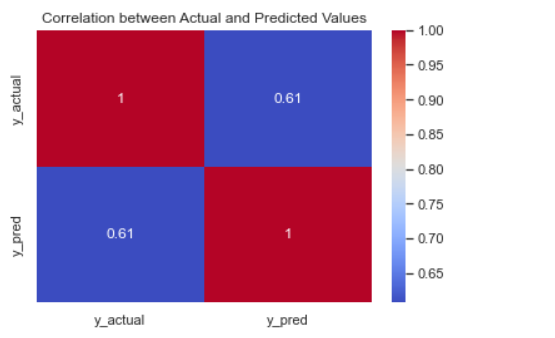


Bảng kết quả dự đoán:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.99 | 0.96 | 0.97 | 9282 |
| 1 | 0.52 | 0.78 | 0.63 | 579 |
| accuracy |  |  | 0.95 | 9861 |
| macro avg | 0.75 | 0.87 | 0.80 | 9861 |
| weighted avg | 0.96 | 0.95 | 0.95 | 9861 |

Nhìn vào kết quả, ta có thể thấy được mô hình dự đoán tốt về khả năng khách hàng không đăng ký dịch vụ (với mã hoá 0) nhưng lại dự đoán không tốt về khả năng khách hàng đăng ký dịch vụ (với mã hoá 1)

Tiếp theo, chúng ta sẽ huấn luyện mô hình và so sánh kết quả mô hình dự đoán với tập dữ liệu bằng biểu đồ heatmap ma trận tương quan



Nhìn chung, kết quả dự đoán tương đối chính xác. Độ chênh lệch giữa biến y\_actual và y\_pred có mối tương quan với nhau

Tính toán Cohen’s Kappa:

Cohen’s Kappa: 0.5983172063158185

# 6. Kết luận

## 6.1 Ưu điểm

* Từ bộ dữ liệu ban đầu về kết quả của các chiến dịch tiếp thị ban đầu của một ngân hàng ở Bồ Đào Nha, chúng tôi đã xác định được các yếu tố thực sự có ảnh hưởng tới quyết định đăng ký khoản gửi có kỳ hạn của khách hàng.
* Đưa ra được nhóm khách hàng mục tiêu mà ngân hàng cần nhắm tới để tăng tỉ lệ thành công trong việc thực hiện các chiến dịch marketing.
* Từ phân tích và trực quan hóa dữ liệu đưa ra những gợi ý về phương thức liên lạc, những tháng cần tập trung thực hiện chiến dịch, những ngày cần gọi điện cho khách hàng và thời lượng cuộc gọi cần thiết trong giao dịch trao đổi để ngân hàng chú trọng hơn trong việc tiếp thị với khách hàng của mình.
* Xây dựng được mô hình dự báo quyết định đồng ý hay không đồng ý đăng ký khoản gửi của khách hàng với các chỉ số như sau

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tên mô hình | Accuracy | Precision | Recall | F1-Score | Kappa Score |
| LightGBM | 0.945 | 0.522 | 0.784 | 0.627 | 0.598 |

## 6.2 Hạn chế

* Bộ dữ liệu bị mất cân bằng khá lớn dù đã thực hiện kỹ thuật oversampling nhưng khó có thể đưa ra một mô hình dự đoán chính xác.
* Trong quá trình thu thập dữ liệu có thể thiếu hoặc thừa những biến quan trọng làm thông tin bị mất dẫn đến hiểu biết không đầy đủ về dữ liệu làm cho độ chính xác của mô hình giảm mất cơ hội khai thác sâu hơn trong dữ liệu
* Chia dữ liệu train-test chỉ 1 lần, có khả năng bỏ qua các điểm dữ liệu quan trọng ảnh hưởng đến kết quả (một số điểm dữ liệu quan trọng chỉ xuất hiện trong một tập dữ liệu (train hoặc test) mà không xuất hiện trong tập còn lại, kết quả dự đoán có thể không chính xác và không đáng tin cậy)

## 6.3 Hướng phát triển

* Tiếp tục thực hiện các thuật toán hiện có để cải thiện kết quả dự đoán cho lớp dữ liệu thiểu số (mã hoá 1) để tăng độ chính xác cho mô hình dự báo.
* Thu thập thêm nhiều dữ liệu để tăng thêm tính khách quan cho bộ dữ liệu, bổ sung thêm dữ liệu có thuộc tính của lớp thiểu số để mô hình không bị lệch dù đã áp dụng oversampling.
* Để đánh giá mô hình một cách hiệu quả và lựa chọn mô hình tốt nhất, tìm hiểu và áp dụng K-Fold validation thay vì chỉ sử dụng một tập huấn luyện và một tập kiểm tra để đánh giá mô hình
* Tiếp tục tìm tòi khám phá dữ liệu, luyện tập khả năng trực quan hoá dữ liệu tốt hơn nhằm đưa ra các kết luận phù hợp với thực tiễn.

# Tài liệu tham khảo

Chittem Leela Krishna, Poli Venkata Subba Reddy, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Patwary MJA, Akter S, Bin Alam MS, Rezaul Karim ANM. Bank Deposit Prediction Using Ensemble Learning. Artificial Intelligence Evolution [Internet]. 2021 Aug. 30 [cited 2023 May28];2(2):42-51. Available from: <https://ojs.wiserpub.com/index.php/AIE/article/view/880>

Hashemi, M., Hall, M., 2020. Multi-label classification and knowledge extraction from oncology-related content on online social networks. Artificial Intelligence Review, 53, 8, 1–38.

Jamiu Olelaken, O. (n.d.). Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and knearest neighbour. . Retrieved from <https://norma.ncirl.ie/4574/1/jamiuolalekanoni.pdf>

Moro, S., Laureano, R. and Cortez, P., “Using data mining for bank direct marketing: An application of the crisp-dm methodology.,” in Proceedings of European Simulation and Modelling Conference, 2011, pp. 117–121.

Mauritsius, T., Braza, A. S., and Fransisca, “Bank Marketing Data Mining using CRISP-DM Approach,” Int. J. Adv. Trends Comput. Sci. Eng., 8(5) 2019.

Nguyễn Thị Thùy Linh, “Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định”, 2005

Adam Shafi**,** Data scientist at Deliveroo, “Random Forest Classification with Scikit-Learn”, Updated Feb 2023

[Anshul Saini](https://www.analyticsvidhya.com/blog/author/anshul508/) — Published On September 15, 2021 and Last Modified On April 26th, 2023, “Master the AdaBoost Algorithm: Guide to Implementing & Understanding AdaBoost”

Rony, M. A. T., Hassan, M. M., Ahmed, E., Karim, A., Azam, S., & Reza, D. A. (2021, December). Identifying Long-Term Deposit Customers: A Machine Learning Approach. In 2021 2nd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). IEEE.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Hou, S., Cai, Z., Wu, J., Du, H., & Xie, P. (2022). Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. International Journal of Business Analytics (IJBAN), 9(1), 1-14.

Plotnikova, V., Dumas, M., Nolte, A., & Milani, F. (2023). Designing a data mining process for the financial services domain. Journal of Business Analytics, 6(2), 140-166.

Borugadda, P., Nandru, P., & Madhavaiah, C. (2021). Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. St. Theresa Journal of Humanities and Social Sciences, 7(1), 91-108.