**NGUYỄN ĐĂNG TRƯỜNG**

TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**TÊN HỌC PHẦN**

**HỌC MÁY NÂNG CAO**

**ĐỀ TÀI:**

**DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ**

**DỊCH VỤ**

**Sinh viên thực hiện**

**:**

**VŨ HẢI PHONG**

**VŨ DUY HIỆP**

**:**

**VŨ VĂN ĐỊNH**

**Giảng viên hướng dẫn**

**:**

**CÔNG NGHỆ THÔNG TIN**

**Ngành**

**:**

**CÔNG NGHỆ PHẦN MỀM**

**Chuyên ngành**

**:**

**CNPM6**

**Lớp**

**Khóa**

**:**

**D**

**15**

***Hà Nội, tháng 10 năm 2023***

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ Và Tên** | **Nội Dung Thực Hiện** | **Điểm** | **Chữ Ký** |
| **1** | **VŨ HẢI PHONG**  (20810310505) |  |  |  |
| **2** | **VŨ DUY HIỆP**  (20810310510) |  |  |  |
| **3** | **NGUYỄN ĐĂNG TRƯỜNG**  (20810310526) |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ Và Tên Giảng Viên** | **Chữ ký** | **Ghi chú** |
| **Giảng Viên Chấm 1:** |  |  |
| **Giảng Viên Chấm 2:** |  |  |

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc152593483)

[CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY 2](#_Toc152593484)

[1.1 Tổng Quan Về Học Máy 2](#_Toc152593485)

[1.1.1 Khái Niệm Về Học Máy 2](#_Toc152593486)

[1.1.2 Các phương pháp về học máy 2](#_Toc152593487)

[1.1.3 Một số giải pháp của Machine learning 3](#_Toc152593488)

[*1.1.3.1. Machine learning được giám sát* 3](#_Toc152593489)

[*1.1.3.2. Machine learning không được giám sát* 3](#_Toc152593490)

[*1.1.3.3. Machine learning bán giám sát* 4](#_Toc152593491)

[*1.1.3.4. Machine learning tăng cường* 4](#_Toc152593492)

[1.2. Lịch sử hình thành Machine learning 4](#_Toc152593493)

[1.3. Những câu hỏi thường gặp về machine learning 5](#_Toc152593494)

[1.3.1. Machine learning được dùng để làm gì? 5](#_Toc152593495)

[1.3.2. Các bước hoạt động của machine learning là gì? 5](#_Toc152593496)

[1.3.3. Machine learning phù hợp với ngôn ngữ nào? 5](#_Toc152593497)

[CHƯƠNG 2: BÀI TOÁN DỰ ĐOÁN KHÁCH HÀNG DỜI BỎ DỊCH VỤ 6](#_Toc152593498)

[2.1 Giới thiệu 6](#_Toc152593499)

[2.2. Phương pháp phân lớp dữ liệu mất cân bằng 6](#_Toc152593500)

[2.2.1. Phân lớp dữ liệu 6](#_Toc152593501)

[2.2.2. Phân lớp dữ liệu mất cân bằng 7](#_Toc152593502)

[2.3. Phương pháp đề xuất nâng cao hiệu quả dự đoán khách hàng rời bỏ dịch vụ 9](#_Toc152593503)

[CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM 13](#_Toc152593504)

[Kết Luận 17](#_Toc152593505)

[TÀI LIỆU THAM KHẢO 18](#_Toc152593506)

# LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Trường Đại học Điện Lực đã đưa môn học Học Máy Nâng Cao vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – thầy Vũ Văn Định đã dạy dỗ, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học Học Máy Nâng Cao của thầy, chúng em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này.

Những năm gần đây với sự phát triển vượt bậc của công nghệ thông tin. “Machine Learning” đang là một cụm từ được mọi người chú ý đến với khả năng tự học hỏi một cách nhanh chóng khi có một luồng dữ liệu nhất định cùng với thuật toán phù hợp.

Hiện nay, càng nhiều người chú ý và quan tâm đến Machine Learning bởi những thành tích đáng kể mà nó mang lại cho con người. Ngoài ra Machine Learning có thể thay thể con người ở một trong công việc như xử lý tính toán ,… Chính sự hiệu quả đó nên nhóm chúng em quyết định chọn đề tài “Dự Đoán Khách Hàng Rời Bỏ Dịch Vụ”. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong cô xem xét và góp ý để bài báo cáo của chúng em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

**CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY**

## 1.1 Tổng Quan Về Học Máy

### 1.1.1 Khái Niệm Về Học Máy

* Học máy hay máy học có tên tiếng anh đầy đủ là Machine Learning, viết tắt là ML. Thuật toán Machine Learning là các chương trình máy tính có khả năng học hỏi và hoàn thành các nhiệm vụ, đồng thời là cách để cải thiện hiệu suất theo thời gian vô cùng hiệu quả.
* Học máy là công nghệ phát triển từ trí tuệ nhân tạo
* Ngoài ra, Machine Learning còn được biết đến là công nghệ phát triển từ lĩnh vực trí tuệ nhân tạo tiên tiến nhất hiện nay. Để đảm bảo không sai lệch và không xuất hiện dữ liệu giả, Machine Leaarning vẫn cần quá trình tìm hiểu và lựa chọn kỹ thuật phân tích dữ liệu từ con người.
* Học máy ngày càng mang tính phổ biết trên toàn thế giới. Sự tăng trưởng vượt bậc của dữ liệu lớn (Big Data) và các thuật toán Machine Learning đã cải thiện độ chính xác của những mô hinhg và dự đoán tương lai.
* Nhiệm vụ của học máy là tập trung phát triển chương trình máy tính truy cậpdữ liệu và sử dụng chúng để tự học. Mục tiêu chính của hoạt động này là giúp máy tính tự hoạt động mà không cần sự hỗ trợ của con người.
* Machine learning được xem là một phần quan trọng thuộc lĩnh vực khoa học dữ liệu. Chúng sử dụng các phương pháp thống kê, thuật toán để phân loại, dự đoán và khám phá những thông tin quan trọng của dữ liệu. Nhờ những chuỗi thôngtin này, người dùng nhanh chóng đưa ra các quyết định trong các hoạt động kinh doanh của mình. Vì vậy, Machine learning chính là giải pháp lý tưởng giúp doanh nghiệp tác động đến chỉ số tăng trưởng doanh thu.

### 1.1.2 Các phương pháp về học máy

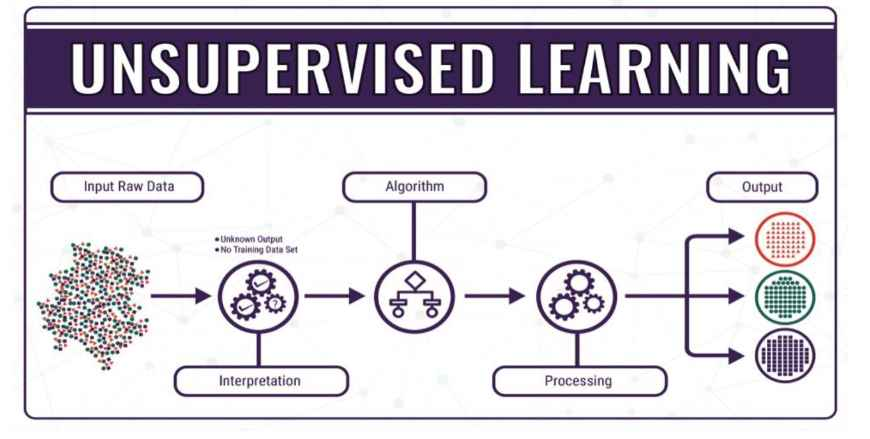
* Thu thập và Chuẩn bị Dữ liệu:
* Thu thập dữ liệu: Bước đầu tiên là thu thập dữ liệu liên quan đến vấn đề mà bạn muốn giải quyết.
* Tiền xử lý dữ liệu: Loại bỏ dữ liệu nhiễu, điền giá trị còn thiếu, chuẩn hóa và biến đổi dữ liệu để làm cho nó phù hợp cho mô hình học máy.
* Chọn Mô hình:
* Chọn một loại mô hình học máy phù hợp với vấn đề cụ thể của bạn, ví dụ: hồi quy tuyến tính, cây quyết định, mạng nơ-ron, học sâu (deep learning), v.v.
* Chia Dữ liệu:
* Chia dữ liệu thành tập huấn luyện, tập kiểm tra và tập kiểm định để đánh giá hiệu suất của mô hình.
* Huấn luyện Mô hình:
* Sử dụng tập huấn luyện để huấn luyện mô hình của bạn. Mô hình sẽ học từ dữ liệu thông qua việc tối ưu hóa hàm mất mát (loss function) thông qua các phương pháp như gradient descent.
* Đánh giá Mô hình:
* Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình. Các độ đo phổ biến bao gồm độ chính xác (accuracy), độ mất mát (loss), và các độ đo đặc biệt tùy thuộc vào bài toán cụ thể.
* Tinh chỉnh Mô hình:
* Nếu mô hình của bạn không đạt được hiệu suất mong muốn, bạn có thể điều chỉnh các siêu tham số (hyperparameters) hoặc chọn mô hình khác.
* Triển khai và Sử dụng:
* Sau khi bạn đã hài lòng với mô hình của mình, bạn có thể triển khai nó vào ứng dụng hoặc hệ thống thực tế để thực hiện dự đoán hoặc phân loại dữ liệu.
* Liên tục Cải thiện:
* Học máy là một quá trình liên tục. Bạn nên theo dõi và cải thiện mô hình khi có dữ liệu mới hoặc khi yêu cầu thay đổi.

### 1.1.3 Một số giải pháp của Machine learning

#### *1.1.3.1. Machine learning được giám sát*

Machine learning được giám sát sử dụng tập dữ liệu được gắn nhãn để phân loạidữ liệu hoặc dự đoán kết quả chính xác nhất. Phương pháp này có thể điều chỉnh trọng lượng khi dữ liệu mới được đưa vào mô hình cho đến khi chúng được lắp một cách thích hợp.Giải pháp Machine learning có giám sát giúp doanh nghiệp xử lý các vấn đề thực trên quy mô lớn, như việc phân loại thư rác trong thư mục riêng. Những thuật toán được sử dụng trong học máy được giám sát là: mạng nơ-ron, hồi quy tuyến tính, hồi quy logistic, máy vectơ hỗ trợ,…

#### *1.1.3.2. Machine learning không được giám sát*

Phương pháp này sử dụng các thuật toán Machine learning để phân tích các tập dữ liệu không được gắn nhãn. Những thuật toán này có khả năng phát hiện nhóm dữ liệu ẩn mà không cần sự hỗ trợ của con người. Đây là giải pháp lý tưởng để: phân tích dữ liệu, phân khúc khách hàng, nhận dạng hình ảnh, chiến lược bán chéo,…

Bên cạnh đó, Machine learning không giám sát còn giảm số lượng tính năng trong một mô hình dựa trên việc giảm kích thước. Chúng có hai cách tiếp cận phổ biến là: phân tích thành phần chính và phân tích giá trị đơn lẻ. Một số thuật toán khác trong phương pháp nay là: phân cụm k-mean, phân cụm xác suất, mạng nơ-ron,…

#### *1.1.3.3. Machine learning bán giám sát*

Đây là sự kết hợp hài hòa giữa học máy được giám sát và không giám sát. Machine learning bán giám sát sử dụng các dữ liệu gắn nhãn và không gắn nhãn đểđào tạo. Phương pháp này giúp giải quyết các vấn đề không có đủ dữ liệu được gắnnhãn để đào tạo thuật toán học máy được giám sát.

#### *1.1.3.4. Machine learning tăng cường*

Machine learning tăng cường là phương pháp học tập dựa trên sự tương tác với môi trường. Chúng có khả năng thực hiện một mục tiêu nhất định như lái xe hoặc chơi trò chơi với đối thủ. Phương pháp này được sử dụng để đào tạo máy tính hoànthành một quy trình gồm nhiều bước.

## 1.2. Lịch sử hình thành Machine learning

Machine learning là thuật ngữ được đặt bởi Arthur Samuel vào năm 1959. Samuel là một IBMer người Mỹ kiêm nhà tiên phong trong lĩnh vực trí tuệ nhân tạo và máy tính chơi game. Năm 1960, thuật ngữ học máy phổ biến hơn thông qua cuốn sách của Nilsson, nội dung đề cập đến việc phân loại máy học.

Machine learning hiện đại bao gồm hai mục tiêu chính: phân loại dữ liệu thông qua mô hình đã được phát triển và đưa ra dự đoán về kết quả trong tương lai dựa trên mô hình này.

## 1.3. Những câu hỏi thường gặp về machine learning

### 1.3.1. Machine learning được dùng để làm gì?

Machine learning góp mặt trong cuộc sống hằng ngày của con người. Một số lĩnh vực đã ứng dụng học máy như:

* + Trợ lý ảo
  + Đưa ra những dự đoán về lưu lượng truy cập
  + Phát hiện gian lận trực tuyến
  + Lọc thư rác
  + Đề xuất sản phẩm
  + Phương tiện không người lái.

### 1.3.2. Các bước hoạt động của machine learning là gì?

Thông thường, công nghệ học máy bao gồm ba bước hoạt động: đào tạo, xác thực và kiểm tra. Trong đó, giai đoạn kiểm tra liên quan đến việc quản lý tiếng ồn và kiểm tra các thông số. Đây chính là ba bước hoạt động cơ bản của phương pháp học máy.

### 1.3.3. Machine learning phù hợp với ngôn ngữ nào?

Ngôn ngữ lập trình tốt nhất đối với công nghệ hiện đại machine learning là: JavaScript, Python, Julia, R, Java. Trong đó, Python chính là ngôn ngữ lập trình được sử

dụng phổ biến nhất khi thiết lập giải pháp học máy. Lượng người dùng sử dụng ngôn ngữ này ngày càng tăng.

# CHƯƠNG 2: BÀI TOÁN DỰ ĐOÁN KHÁCH HÀNG DỜI BỎ DỊCH VỤ

**2.1 Giới thiệu**

* Bài toán dự đoán khách hàng rời bỏ giao dịch vụ là dự đoán liệu khách hàng sẽ không còn mua sản phẩm hoặc dịch vụ của mình trong một khoảng thời gian nhất định nữa hay không.
* Trong thực tế thì bài toán này thường gặp trong các tổ chức kinh doanh dịch vụ như ngân hàng, bảo hiểm, viễn thông... Đây là một bài toán quan trọng đối với bất kỳ một tổ chức nào vì nếu có thể dự đoán được sớm việc khách hàng không tiếp tục sử dụng dịch vụ, tổ chức có thể đưa ra được các phương án để giữ chân khách hàng. Việc chú trọng đến tập khách hàng (có khả năng) rời bỏ dịch vụ luôn được Ban lãnh đạo của các Tổ chức quan tâm vì nhiều lý do. Bởi giữ chân khách hàng sẽ giúp tăng uy tín thương hiệu, tăng doanh thu. Bên cạnh đó, chi phí đầu tư mỗi khách hàng mới nhiều gấp nhiều lần khách hàng cũ và việc tìm kiếm khách hàng mới cũng sẽ bị ảnh hưởng bởi việc khách cũ rời bỏ dịch vụ. Nhận thức được những điều đó, các tổ chức luôn cố gắng níu kéo từng khách hàng một, tìm biện pháp để kịp thời giữ chân khách hàng có nguy cơ rời bỏ dịch vụ của mình.
* Dự đoán khách hàng rời bỏ dịch vụ nói riêng và các bài toán dự báo trong kinh tế, tài chính nói chung đều đã có những nghiên cứu phân tích dữ liệu để giải quyết từ lâu nhưng đến nay vẫn luôn được quan tâm đặc biệt bởi tầm quan trọng của nó. Với sự phát triển của khoa học công nghệ và sự bùng nổ dữ liệu hiện nay xuất hiện các kho dữ liệu khổng lồ (Big Data) thì các phương pháp phân tích dữ liệu truyền thống đòi hỏi những yêu cầu điều tra phức tạp và tốn kém về mặt thời gian. Do đó, xu thế hiện nay để giải quyết hiệu quả hơn các bài toán này là sử dụng các kỹ thuật của khai phá dữ liệu và các thuật toán học máy (Nguyễn Ngọc Tuân, 2016), (H. Ali, 2019). Bài toán dự đoán khách hàng rời bỏ dịch vụ là thuộc dạng phân lớp trong khai phá dữ liệu và có đặc thù dữ liệu thường là mất cân bằng khi số lượng đa số (không rời dịch vụ) có thể sẽ lớn hơn rất nhiều so với số lượng lớp thiểu số (có rời dịch vụ), điều này làm cho các thuật toán phân lớp gặp rất nhiều khó khăn,do đó cần có hướng tiếp cận riêng để giải quyết. (Yanmin , 2009). Mặc dù hiện nay đã có một số phương pháp, thuật toán được đề xuất cho mô hình phân lớp dự đoán khách hàng rời bỏ dịch vụ và đã thu được những kết quả nhất định trong một số trường hợp riêng, tuy nhiên vấn đề này có thể được làm tốt hơn nữa để nâng cao hiệu quả dự đoán.

## 2.2. Phương pháp phân lớp dữ liệu mất cân bằng

### 2.2.1. Phân lớp dữ liệu

Phân lớp (classification) là một kỹ thuật quan trọng trong khai phá dữ liệu, mục đích là gán (dự đoán) nhãn của một phần tử dữ liệu mới (chưa biết nhãn) từ những thuộc tính của phần tử dữ liệu đó. Tập các giá trị nhãn lớp ở đây là hữu hạn, và nếu chỉ có 2 giá trị thì được gọi là phân lớp nhị phân. Ví dụ điển hình của phân lớp dữ liệu như việc phân loại email mới gửi đến là thư rác hay không, nếu là thư rác thì email sẽ được gán nhãn Spam và chuyển vào thư mục spam, còn nếu không thì sẽ được gán nhãn Non-spam và chuyển vào thư mục inbox.

Quá trình phân lớp gồm hai giai đoạn: xây dựng mô hình (learning) và sử dụng mô hình (classification).Giai đoạn xây dựng mô hình là việc học dữ liệu từ một tập dữ liệu huấn luyện (training set) đã biết trước nhãn bằng các thuật toán học máy (machine learning) để tạo ra một mô hình (model) có khả năng dự đoán nhãn lớp cho dữ liệu mới. Tuỳ theo thuật toán học máy được sử dụng thì có những mô hình phân lớp khác nhau, chẳng hạn như: cây quyết định (Decision Tree), k - láng giềng gần nhất (k - Nearest Neighboor), máy véc tơ hỗ trợ (Support Vector Machine), Naïve Bayes, rừng ngẫu nhiên (Random Forest)…

Sau khi xây dựng được mô hình phân lớp ở giai đoạn huấn luyện thì sẽ sử dụng mô hình để phân lớp dữ liệu mới nếu hiệu quả phân lớp chấp nhận được. Để đánh giá mô hình phân lớp là chấp nhận được hay không, ta sử dụng một bộ dữ liệu kiểm tra độc lập với bộ dữ liệu huấn luyện rồi từ đó xác định xem có bao nhiêu phần tử dữ liệu được phân lớp đúng và bao nhiêu phân tử dữ liệu bị phân lớp sai. Một số độ đo đánh giá hiệu quả phân lớp phổ biến là: Accuracy, F-score, Sensitivity (Recall), Specificity, G-mean… (Yanmin Sun, 2009).

### 2.2.2. Phân lớp dữ liệu mất cân bằng

Dữ liệu mất cân bằng là dữ liệu có sự chênh lệch lớn về số lượng phần tử giữa các lớp dự đoán (H. Ali, 2019), nghĩa là số lượng các phần tử đại diện cho một lớp lớn hơn rất nhiều so với các lớp khác, chẳng hạn như tỷ lệ khách hàng không rời bỏ dịch vụ thường là cao hơn rất nhiều so với khách hàng rời bỏ dịch vụ. Hoặc trong việc phát hiện bệnh nhân ung thư thì tỷ lệ bệnh nhân không bị ung thư là cao hơn rất nhiều so với các bệnh nhân bị ung thư... Đối với trường hợp hai lớp (chỉ có 2 trường hợp của lớp dự đoán, chẳng hạn như việc xác định giới tính là Nam/Nữ, hay xác định khả năng khách hàng là có rời bỏ dịch vụ/ không rời bỏ dịch vụ...) thì tỷ lệ này có thể là 1:5, 1:10, 1:100… Lớp chiếm số đông phần tử gọi là lớp đa số (negative), ngược lại lớp có ít phần tử gọi là lớp thiểu số (positive). Khi tiến hành khai phá dữ liệu trên các dữ liệu mất cân bằng thì các thuật toán thường đạt độ chính xác cao với lớp đa số nhưng với lớp thiểu số thì ngược lại.

Trong thực tế, vấn đề mất cân bằng dữ liệu đối với các bộ dữ liệu của bài toán dự đoán khách hàng rời bỏ dịch vụ là phổ biến. Để giải quyết hiệu quả cho bài toán này thì có thể sử dụng các phương pháp phân lớp của khai phá dữ liệu, trong đó có hai hướng tiếp cận chính được tập trung nghiên cứu cho việc nâng cao hiệu quả dự đoán là: hướng tiếp cận ở mức độ thuật toán và hướng tiếp cận ở mức độ dữ liệu.

Hướng tiếp cận ở mức độ thuật toán: Tập trung vào việc điều chỉnh, cải tiến các thuật toán phân lớp chuẩn (như cây quyết định, Naïve Bayes, máy véctơ hỗ trợ SVM, k láng giềng gần nhất KNN...) sao cho phù hợp với dữ liệu mất cân bằng, chẳng hạn như tăng cường học cho lớp thiểu số. Hướng tiếp cận này thường là phức tạp hơn so với hướng tiếp cận ở mức độ dữ liệu và yêu cầu cần phải hiểu rõ về thuật toán phân lớp cần cải tiến.

Hướng tiếp cận ở mức độ dữ liệu: Bao gồm các phương pháp điều chỉnh để giảm sự mất cân bằng dữ liệu bằng cách tăng số lượng phần tử lớp thiểu số (sinh thêm các phần tử thuộc lớp thiểu số một cách ngẫu nhiên, hoặc có chọn lọc, hoặc sinh thêm phần tử nhân tạo), giảm số lượng phần tử lớp đa số hoặc kết hợp cả hai phương pháp. Cả ba phương pháp trên đều hướng đến mục tiêu cân bằng phân bố dữ liệu. Ưu điểm của các phương pháp tiếp cận mức độ dữ liệu là sự linh hoạt, dữ liệu có thể sử dụng để huấn luyện các bộ phân loại khác nhau. Hướng tiếp cận này được tập trung nghiên cứu nhiều hơn và một số phương pháp tiêu biểu có thể kể tới là:

• ROS (Random Over-sampling)

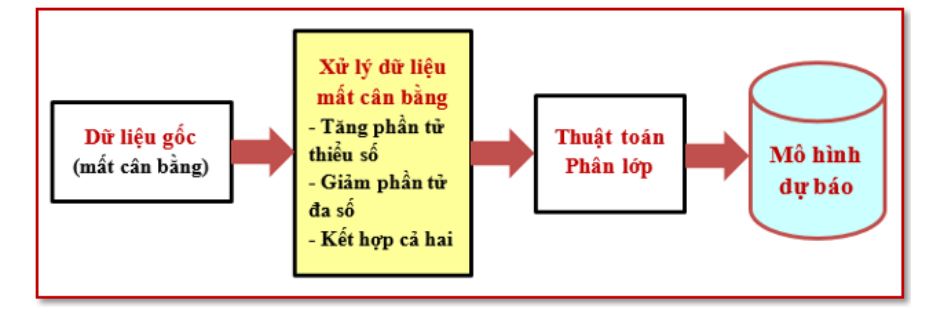
• RUS (Random Under-sampling)

• SMOTE (Synthetic Minority Over-sampling Technique)

• BOS (Boderline SMOTE): sinh thêm phần tử nhân tạo dựa trên đường biên

• SLS (Safe-level SMOTE): sinh thêm phần tử nhân tạo dựa trên mức an toàn

• Một số phương pháp khác: Tomek Link, ADASYN…



**Hình 1.** Phân lớp dữ liệu mất cân bằng theo hướng tiếp cận mức độ dữ liệu

Một số tác giả cũng đã sử dụng phương pháp phân lớp dữ liệu mất cân bằng để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ theo các cách tiếp cận khác nhau. Nhóm tác giả trong (Aamer, 2017) đã kết hợp phương pháp lựa chọn thuộc tính và xử lý dữ liệu mất cân bằng, trong khi đó, Annisa Aditsania và các cộng sự trong (Annisa, 2017) sử dụng phương pháp lấy mẫu tổng hợp thích ứng (ADASYN - một biến thể của SMOTE) và thuật toán lan truyền ngược để xử lý mất cân bằng dữ liệu trong bài toán dự đoán khách hàng rời bỏ dịch vụ. Ngoài ra, các tác giả trong (Uma, 2018) đã đề xuất phương pháp mới SOS-BUS kết hợp tăng số phần tử nhân tạo SMOTE với kỹ thuật giảm số phần tử của họ đề xuất… Về cơ bản là các nghiên cứu tập trung chủ yếu vào việc tiếp cận xử lý mất cân bằng ở cấp độ dữ liệu, đề xuất một số cải tiến kết hợp với thuật toán SMOTE để dự báo khách hàng rời bỏ dịch vụ, tuy nhiên dữ liệu sử dụng mang tính đặc thù và không công khai.

Tại Việt Nam, một số tác giả như (Nguyễn Ngọc Tuân, 2016), (Kien Vu, 2018) đã nghiên cứu việc áp dụng kỹ thuật khai phá dữ liệu để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ trong lĩnh vực kinh doanh viễn thông. Đặc biệt, nhóm tác giả của FTP Telecom (Duyen, 2017) và các đồng nghiệp cũng đã có ứng dụng phân lớp dữ liệu mất cân bằng và phương pháp lựa chọn thuộc tính để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ Internet tại tổ chức của mình.

## 2.3. Phương pháp đề xuất nâng cao hiệu quả dự đoán khách hàng rời bỏ dịch vụ

Mặc dù các phương pháp phân lớp dữ liệu mất cân bằng phổ biến hiện nay như SMOTE và các biến thể của nó đã có những cải tiến mô hình phân lớp bằng cách sinh thêm các phần tử nhân tạo theo những cách khác nhau, tuy nhiên qua khảo sát dữ liệu của một số bài toán dự đoán khách hàng rời bỏ dịch vụ và thực nghiệm thì chúng tôi nhận thấy rằng các phương pháp này còn tồn tại một số hạn chế như:

• Phần tử thiểu số nhân tạo được sinh quá xa với phần tử thiểu số thực sự, và có thể sẽ nằm trong vùng có nhiều phần tử đa số nên sẽ gây nhiễu (giảm hiệu quả phân lớp).

• Những phần tử đa số nằm sâu trong vùng có nhiều phần tử thiểu số là những phần tử có ảnh hưởng nhiễu lớn nhưng các phương pháp hiện tại không xử lý nhiễu với những phần tử này.

• Phương pháp Borderline SMOTE thì chỉ áp dụng cho các trường hợp dữ liệu được phân bố bởi đường biên rõ ràng, hay phương pháp Safelevel SMOTE thì định nghĩa tỷ lệ an toàn dựa trên số phần tử láng giềng có thể sẽ bị nhiễu nếu phân bố dữ liệu không đồng đều về khoảng cách nên khi áp dụng với bài toán khách hàng rời bỏ dịch vụ thì hiệu quả không cao…

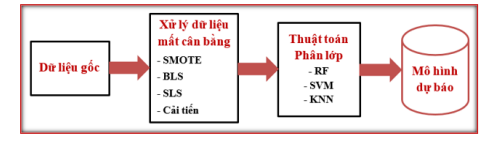
Do đó, chúng tôi đề xuất một phương pháp mới để nâng cao hiệu quả cho mô hình dự đoán khách hàng rời bỏ dịch vụ. Phương pháp này cũng giống như SMOTE là xử lý dữ liệu mất cân bằng trước khi huấn luyện mô hình, các bước thực hiện như sau:

• Lấy giá trị khoảng cách R1 và R2 đủ nhỏ và tham số T ≥ 1.

• Với mỗi phần tử đa số X, nếu T\*n > m, trong đó lần lượt là số phần tử đa số, số phần tử thiểu số có khoảng cách tới X nhỏ hơn R1 thì đổi nhãn của X thành nhãn của lớp thiểu số.

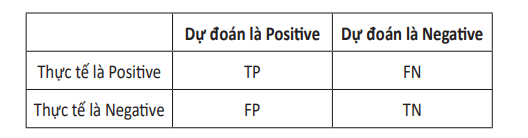
• Với mỗi phần tử thiểu số thực sự P, sinh thêm ngẫu nhiên k phần tử thiểu số nhân tạo có khoảng cách tới P nhỏ hơn R2 .

Để minh chứng tính hiệu quả của phương pháp đề xuất, chúng tôi tiến hành thực nghiệm trên bộ dữ liệu Customer Churn của Kaggle và so sánh kết quả với các phương pháp xử lý dữ liệu mất cân bằng khác, bao gồm: SMOTE, Boderline SMOTE (BLS), Safe-level SMOTE (SLS). Bộ dữ liệu này có kích thước gồm 4250 bản ghi, trong đó tỷ lệ mất cân bằng dữ liệu là 1 : 6,1. Các thuật toán phân lớp sử dụng trong thực nghiệm là: rừng ngẫu nhiên (RF), máy véc tơ hỗ trợ (SVM), k láng giềng gần nhất (KNN).



**Hình 2.** Các bước tiến hành thực nghiệm

Chúng tôi sử dụng các độ đo Sensitivity (Recall), Specificity, G-mean được tính từ ma trận nhầm lẫn như sau:



**Bảng 1.** Ma trận nhầm lẫn

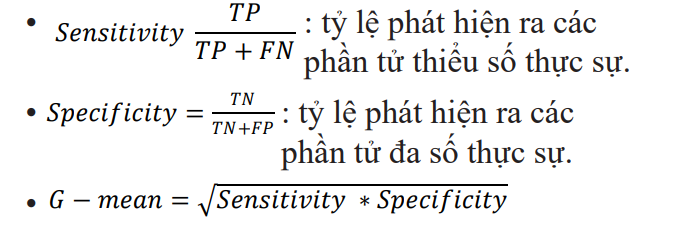
Trong đó các hàng của ma trận là nhãn lớp thực tế, các cột là nhãn lớp dự đoán và:

• TN: số lượng phần tử lớp đa số được phân loại chính xác.

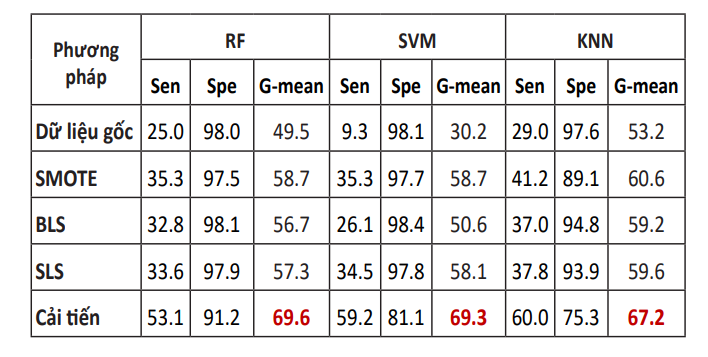
• FN: số lượng phần tử lớp thiểu số bị phân loại nhầm là phần tử lớp đa số.

• TP: số lượng phần tử lớp thiểu số được phân loại chính xác.

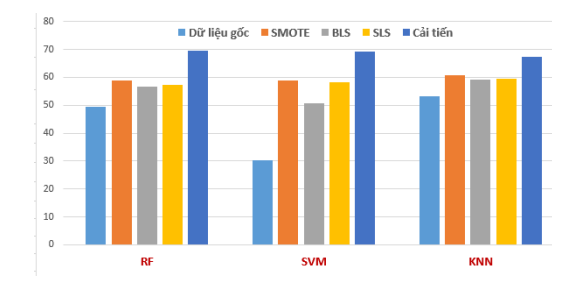
• FP: số lượng phần tử lớp đa số bị phân loại nhầm là phần tử lớp thiểu số.



Kết quả thực nghiệm được thống kê như sau:



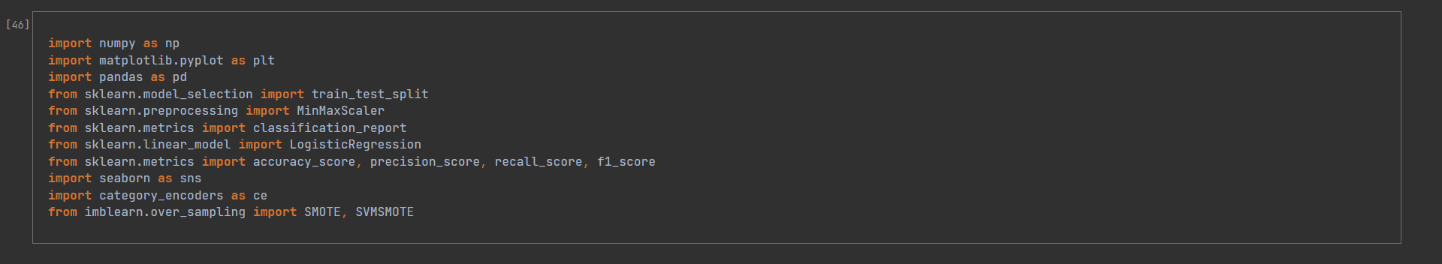
**Bảng 2.** Tổng hợp kết quả thực nghiệm



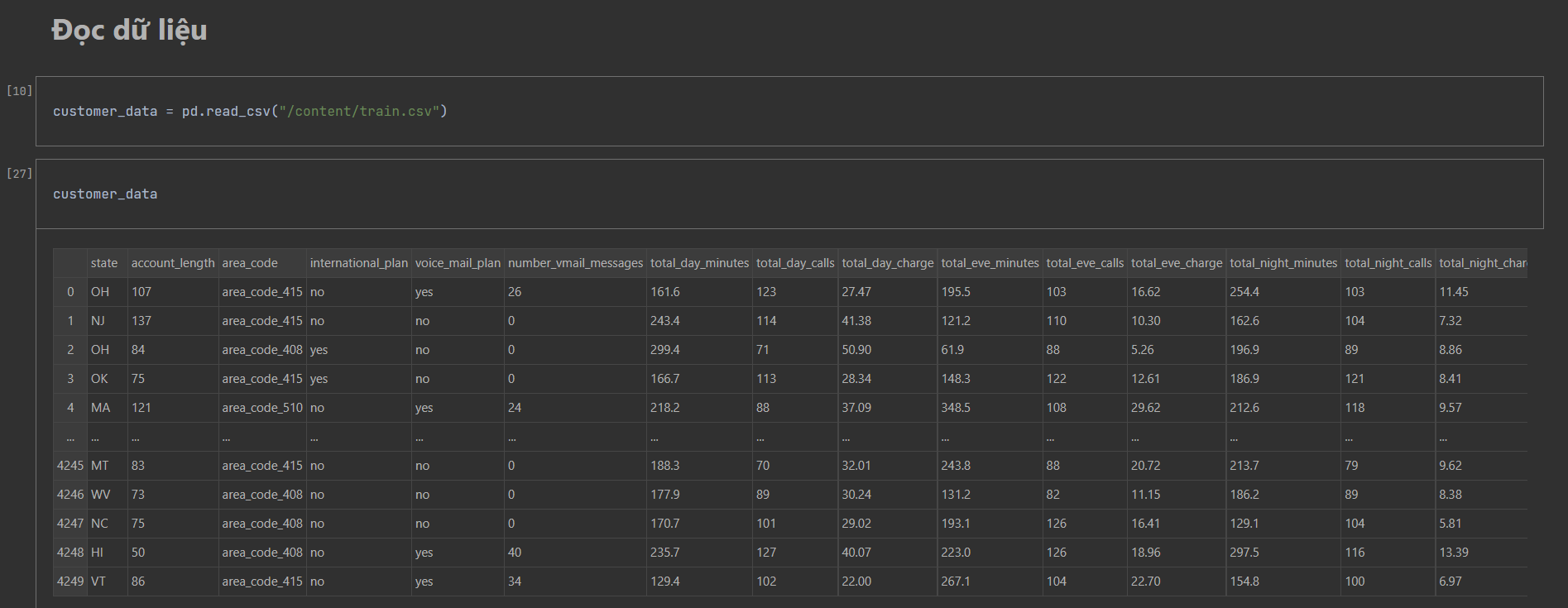
**Hình 3.** Biểu đồ so sánh giá trị G-mean

Từ Bảng 2 và Biểu đồ tại Hình 3 nhận thấy rằng khi chúng ta áp dụng các phương pháp xử lý dữ liệu mất cân bằng thì hiệu quả phân lớp đều tốt hơn so với thực hiện trên dữ liệu gốc ban đầu, mặc dù độ đo Specificity có giảm nhẹ nhưng hai độ đo quan trọng là Sensitivity và G-mean đều tăng, điều này là rất có ý nghĩa với mô hình phân lớp dữ liệu mất cân bằng. Đặc biệt là phương pháp cải tiến của chúng tôi là có hiệu quả cao hơn với 3 phương pháp SMOTE, BLS, SLS trong cả 3 lần thực nghiệm với các thuật toán RF, SVM, KNN. Ngoài ra, theo Biểu đồ so sánh giá trị G-mean thì chúng ta nhận thấy rằng mặc dù phương pháp BLS và SLS là những biến thể của SMOTE nhưng hiệu quả phân lớp trong bài toán dự đoán khách hàng rời bỏ dịch vụ lại không cao hơn so với SMOTE.

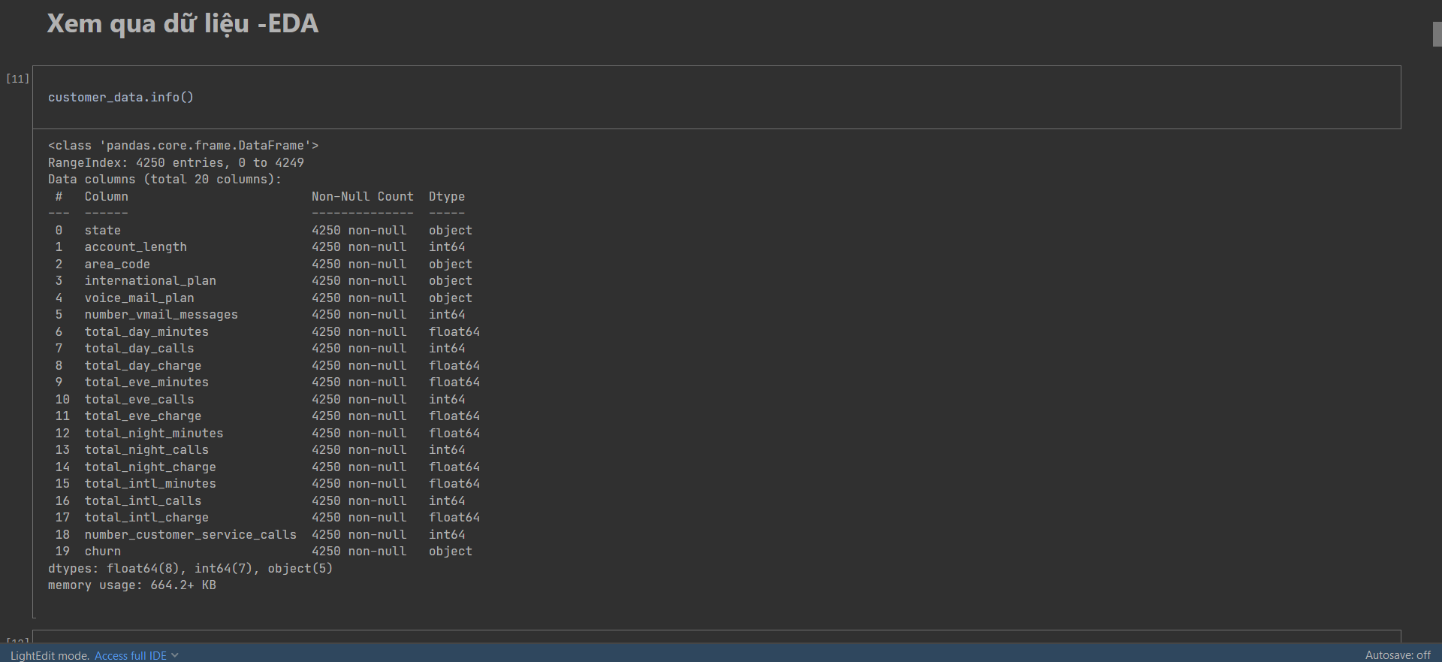
**CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM**

****

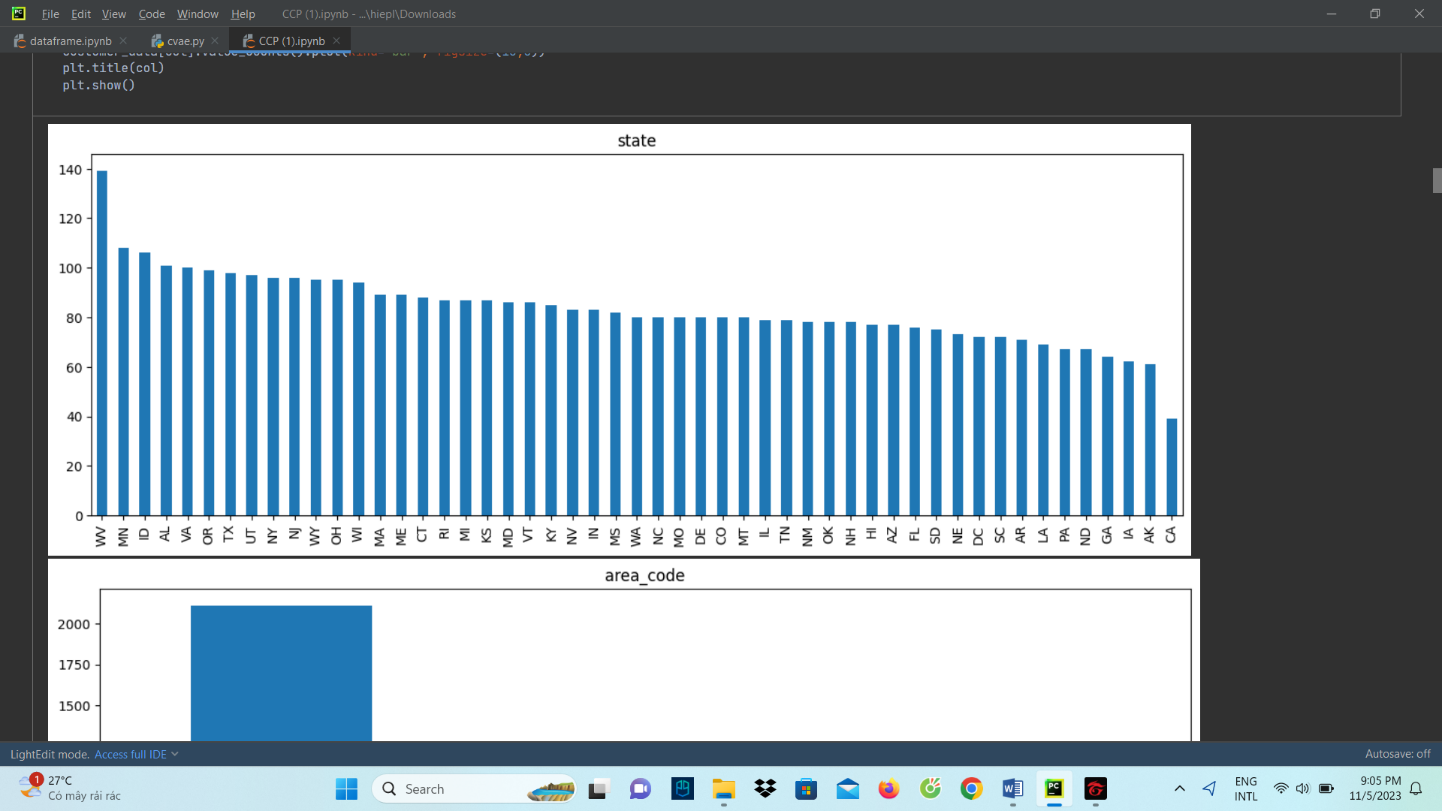
*Hình 3.1: Cài đặt*

****

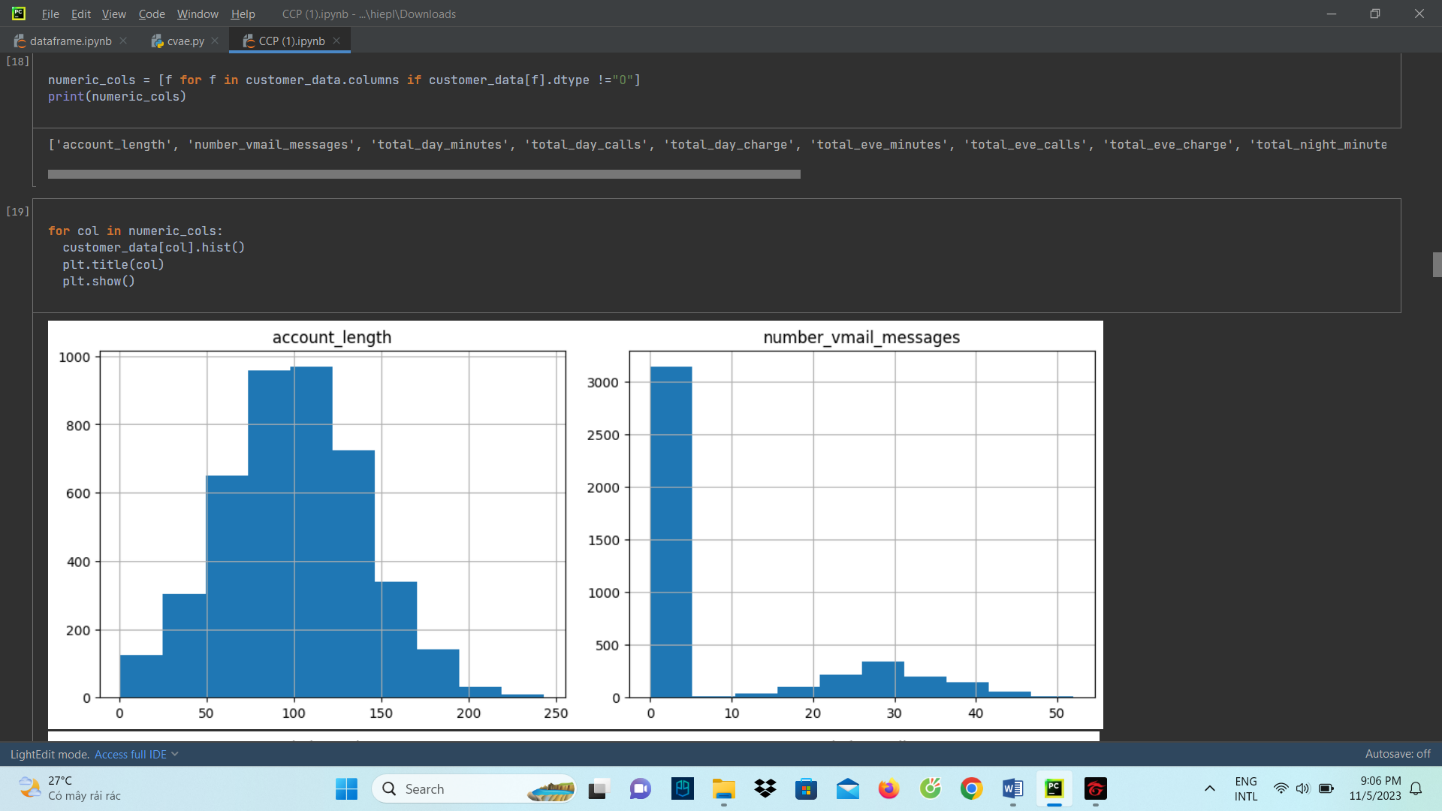
*Hình 3.2: Đọc dữ liệu*

****

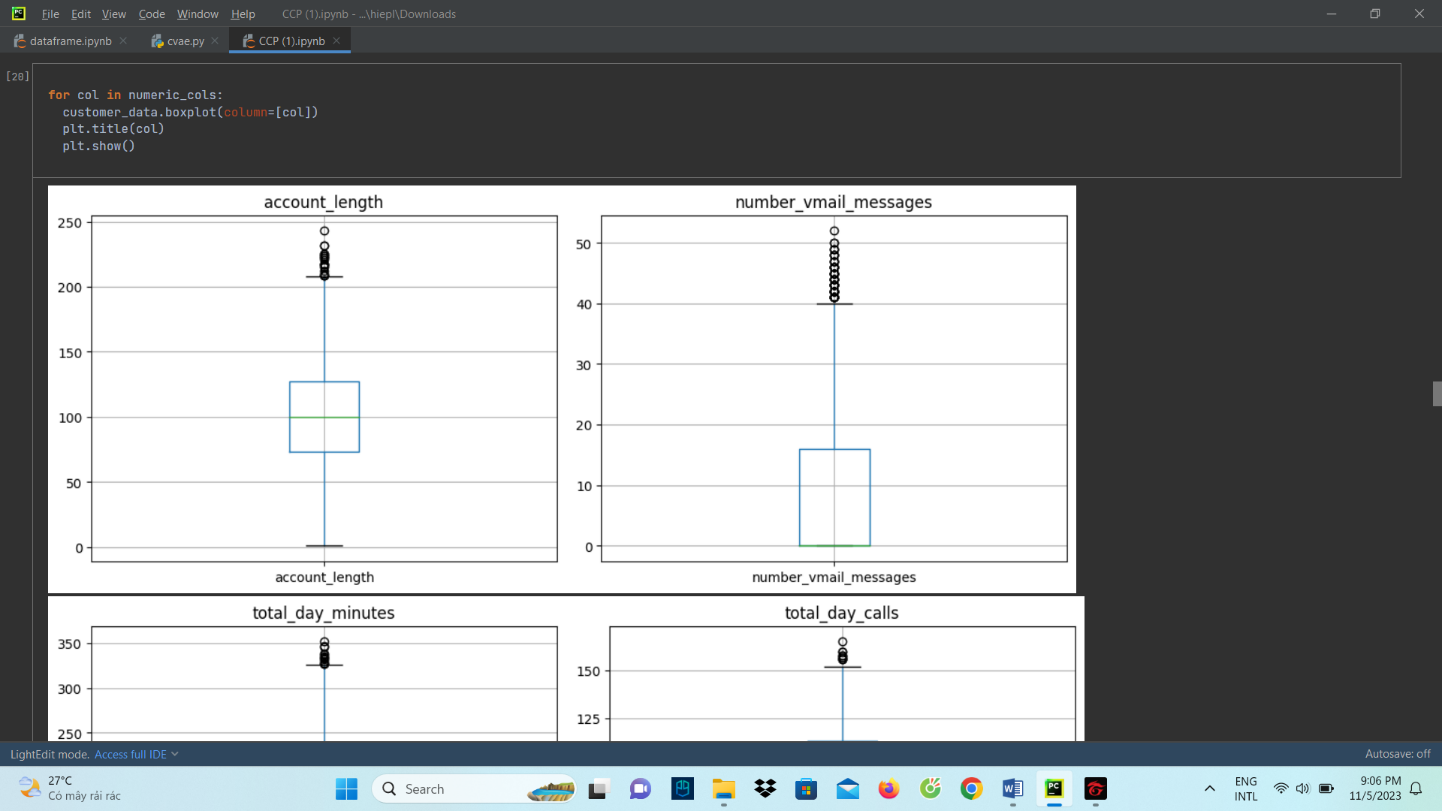
*Hình 3.3: Xem qua dữ liệu*

****

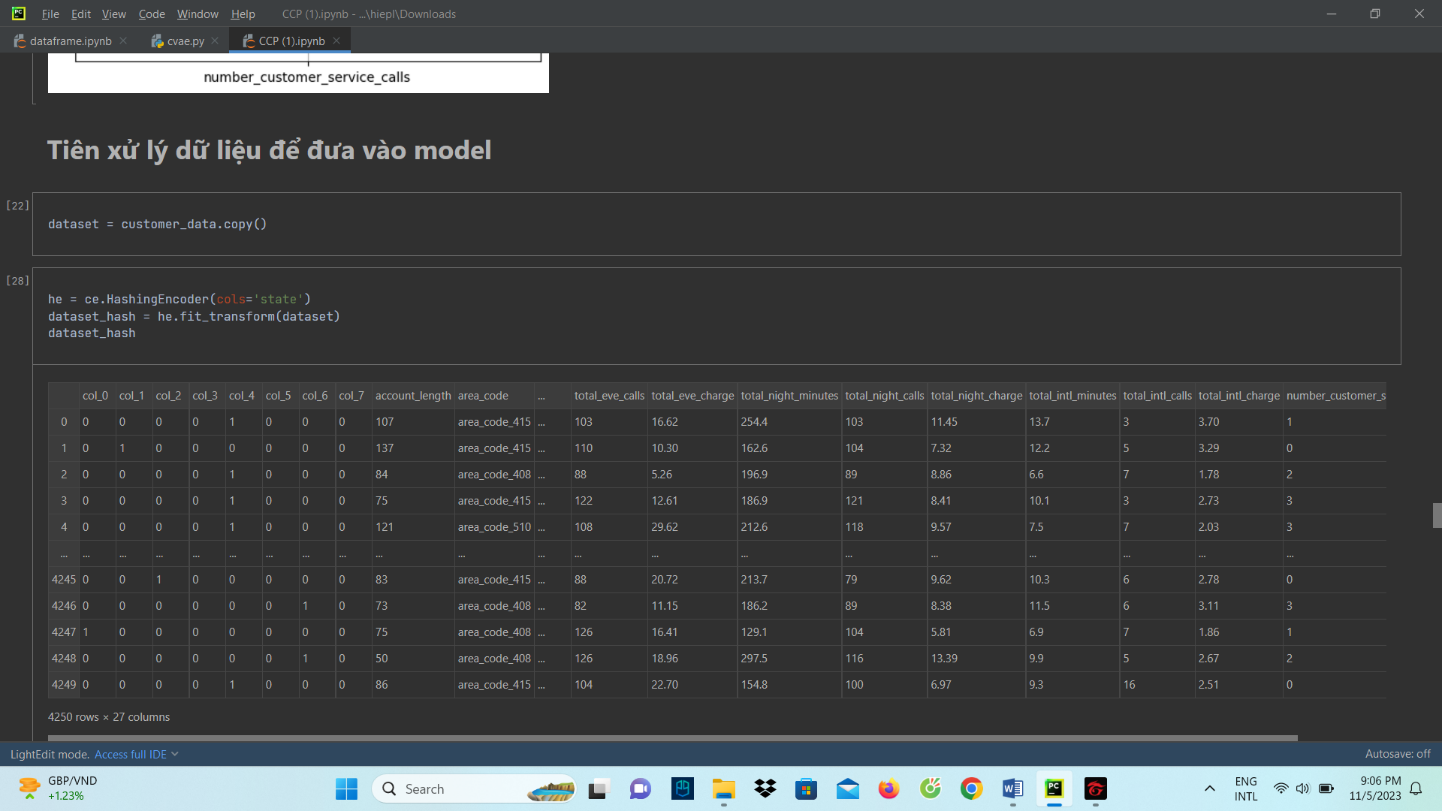
*Hình 3.4: Xem qua dữ liệu*

****

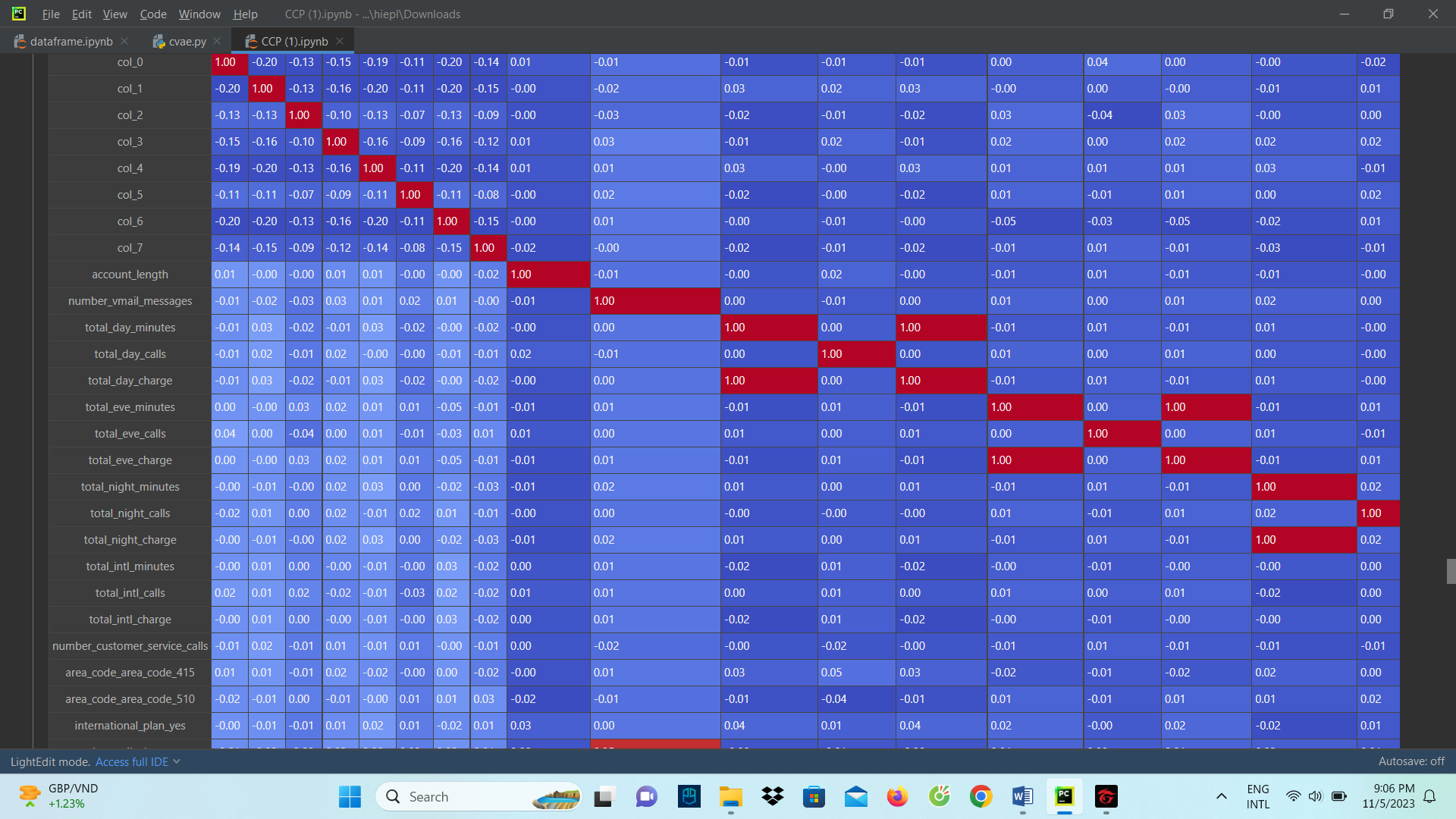
*Hình 3.5: Xem qua dữ liệu*

****

*Hình 3.6: Xem qua dữ liệu*

****

*Hình 3.7: Tiền xử lí dữ liệu*

****

*Hình 3.8: Tiền xử lí dữ liệu*

# 

# KẾT LUẬN

Trong bối cảnh ngày nay, khi sự cạnh tranh giữa các dịch vụ môn học máy nâng cao ngày càng gay gắt, việc dự đoán và ngăn chặn khách hàng rời bỏ trở nên quan trọng hơn bao giờ hết. Bài báo cáo này đã tập trung vào việc phân tích và dự đoán hành vi của khách hàng để đưa ra những giải pháp hiệu quả trong việc giữ chân họ. Sau một thời gian tích cực làm việc, nghiên cứu của nhóm em với sự hướng dẫn, chỉ bảo nhiệt tình của thầy giáo giảng dạy bộ môn – Vũ Văn Định, nhóm em đã hoàn thành bài báo cáo của mình với đề tài: “Dự đoán khách hàng rời bỏ dịch vụ”. Trong quá trình nghiên cứu và xây dựng lên chương trình này, nhóm em đã rất cố gắng nhưng đồ án này vẫn còn nhiều thiếu sót.

**Ở bước đầu kết quả đạt được:**

* Giảm tỷ lệ khách hàng rời bỏ
* Tăng cường trải nghiệm của khách hàng
* Tối ưu hóa duy trì khách hàng
* Tăng cường uy tín thương hiệu

**Hạn chế**

* Phụ thuộc vào dữ liệu
* Thiếu yếu tố con người
* Sự biến đổi của thị trường

**Hướng phát triển**

* Tích hợp Trí tuệ nhân tạo (AI) và Học máy Tiến triển
* Tăng cường tính tương tác
* Quản lý độ đa dạng của dữ liệu
* Áp dụng công nghệ mới

Trong tương lai, chúng ta cần liên tục theo dõi và cập nhật mô hình dự đoán để đảm bảo tính chính xác và khả năng áp dụng. Đồng thời, việc nghiên cứu và áp dụng các công nghệ mới như học máy và trí tuệ nhân tạo sẽ giúp chúng ta tiếp tục nâng cao khả năng dự đoán và phòng tránh rời bỏ khách hàng.

Nhìn chung, thông qua nỗ lực nghiên cứu và ứng dụng các phương pháp dự đoán tiên tiến, chúng ta có thể xây dựng một hệ thống quản lý khách hàng mạnh mẽ và linh hoạt, giúp tối ưu hóa trải nghiệm của khách hàng và duy trì sự hài lòng, đồng thời giảm thiểu nguy cơ mất mát khách hàng.

# TÀI LIỆU THAM KHẢO

[1] Tài liệu về học máy https://vti-solutions.vn/machine-learning-chia-khoa-cho-san-xuat-4-0

[2] Tài liệu về machine learning https://topdev.vn/blog/machine-learning-la-gi

[3] https://cuongquach.com/ebook-machine-learning-co-ban-pdf.html